UNIVERSIDAD DE VALENCIA

PhD Program

“La Traducción y la Sociedad del Conocimiento”

Departamento de Teoría de los Lenguajes y Ciencias de la Comunicación

FACULTAD DE FILOLOGÍA, TRADUCCIÓN Y COMUNICACIÓN

Use and Evaluation of Controlled Languages in Industrial Environments and Feasibility Study for the Implementation of Machine Translation

In Candidacy for the Degree of Doctor of Philosophy,

With the Title of "Doctor Internacional"

PhD Thesis

Submitted by: Laura Ramírez Polo

Supervised by: Dr. Manuel Pruñonosa Tomás

Valencia, June 2012

Table of Contents

Index of Tables

Index of Figures

Abbreviations

ACE	Attempto Controlled English /Avaya: Controlled English
AECMA	European Association of Aerospace Manufacturers
BTE	Boeing Technical English
CASE	Case's Clear and Simple English
CASL	General Motor's Controlled Automotive Service Language
CELT	Controlled English to Logic Translation
CFE	Caterpillar Fundamental English
CL	Controlled Language
CLCE	Common Logic Controlled English
CLIP	Controlled Language for Inference Purposes
CSDG	Controlled Siemens Documentary German
CTE	Caterpillar Technical English
CTL	Computation Tree Logic
DCE	Diebold Controlled English
DLTIL	Distributed Language Translation Intermediate Language
DOCL	Dual Oriented Controlled Language
DRT	Discourse Representation Theory
HAMT	Human-Aided Machine Translation
HELP	Hyster's Easy Language Program

HOCL	Human Oriented Controlled Language
INTERCOM	International Council for Technical Communication
ITM	Interactive Machine Translation
KISL	Kodak International Service Language
MAHT	Machine-Aided Human Translation
MARTIF	Machine-Readable Terminology Interchange Format
MCE	Multinational Customized English
MOCL	Machine Oriented Controlled Language
MT	Machine Translation
PACE	Perkins Approved Clear English
PEP	Plain English Program
PNL	Pseudo Natural Language
ROI	Return on Investment
R-Rules	Readability-Rules
RA	Reparaturanleitung (repair instructions)
SBT	Service Bulletin Technique
SDD	Siemens Dokumentationsdeutsch
SGML	Standard Generalized Markup Language
SI	Service Information
SOS	Sidney OWL Syntax

STE	Simplified Technical English
TM	Translation Memory / Terminology Management
TMS	Translation Memory System
T-Rules	Translatability rules
XML	eXtended Markup Language

RESUMEN*

El presente trabajo de investigación se enmarca en los estudios de doctorado en traducción y la sociedad del conocimiento de la Universidad de Valencia y, en concreto, en la línea de investigación en tecnologías de la traducción, terminología y localización. En este sentido, esta disertación surge por la necesidad de establecer una metodología de investigación y ofrecer resultados empíricos sobre el desarrollo, implementación y evaluación de lenguajes controlados en la documentación técnica y su efecto tanto en los textos originales como en las traducciones de estos documentos.

Así pues, el objetivo ha sido desarrollar una metodología para evaluar el impacto de los lenguajes controlados en la producción de documentación técnica dentro de contextos industriales y, más en concreto, en la elaboración de documentación técnica para el vehículo. El impacto se ha concretado en la mejora de la traducibilidad automática, un concepto que hemos discutido ampliamente en el capítulo 4, así como de la calidad de los textos meta.

Este objetivo general se deriva de tres hipótesis que planteamos desde el principio en este estudio:

- En primer lugar, que los textos escritos de conformidad con las reglas de un lenguaje controlado y la ayuda de una herramienta para su aplicación mejoran su inteligilibilidad, comprensión y traducibilidad.

* De conformidad con: Universitat de València. Consell de Govern. ACGUV 252/2008. Reglament dels Premis Extraordinaris de Doctorat. [pdf]. [en línea]. València: Universitat de València. 5 p. Disponible en: <http://www.uv.es/~sgeneral/Reglamentacio/Doc/Doctorat/E1.pdf> [consulta: 29/02/2012].

- En segundo lugar, que la traducción automática (TA) es una tecnología que puede representar un evaluador "objetivo" respecto a la traducibilidad, ya que no contamos con las diferentes variantes que resultan de la traducción humana.

- Por último y, como efecto colateral, que la TA es una tecnología que permite afrontar la traducción del creciente volumen de documentación técnica. Con procesos bien definidos, esta tecnología puede representar un ahorro considerable en el tiempo y los costes de traducción, sin tener que renunciar necesariamente a la calidad.

Así pues, para alcanzar este objetivo general y corroborar o desechar las hipótesis planteadas, definimos una serie de objetivos específicos que se concretan en los siguientes puntos:

- La elaboración de un marco teórico en el que definir, describir y analizar el concepto de lenguaje controlado, delimitándolo de otros conceptos anejos como el de lenguaje natural o sublenguaje, para después estudiar la aplicación de estos lenguajes en contextos industriales y las herramientas que sirven para automatizar su aplicación, con especial hincapié en MULTILINT/CLAT, con la cual realizaremos la parte empírica de este trabajo. Asimismo, este marco teórico comprenderá un estudio descriptivo de los problemas y particularidades de la traducción de documentación técnica y un análisis de los diferentes métodos de evaluación de tecnologías lingüísticas. Para ello nos centraremos en la evaluación de herramientas y reglas de lenguaje controlado, así como en la evaluación de la traducción automática.

- El diseño de una propuesta metodológica de sólida base teórica para discernir si los textos escritos y editados conforme a las reglas de un lenguaje controlado son más traducibles (automáticamente) que otros. Este aspecto es novedoso ya que hasta la fecha la mayoría de estudios utilizan traductores humanos para las evaluaciones, sin establecer claras diferencias entre las reglas que pueden mejorar

la traducibilidad humana y la automática. Asimismo, no existen estudios como este, en el que se utilizan textos reales del campo de la automoción, y pocos estudios en otras áreas de la industria.

Nuestra metodología se divide en tres fases, a saber:

Fase 1. En esta fase se lleva a cabo una microevaluación para determinar qué recursos son los más idóneos para llevar a cabo la evaluación la fase 2. Se trata de seleccionar por una parte una tipología textual más adecuada para la implementación de la traducción automática y, por otra parte, el sistema de traducción automática más adecuado para nuestros propósitos. Para ello se aplicarán métodos de evaluación humana y automática que darán como resultado la selección de un sistema.

Fase 2. En esta segunda fase se lleva a cabo una macroevaluación con un corpus de textos en lengua origen (en este caso alemán) escritos sin seguir las directrices de un lenguaje controlado y esos mismos textos corregidos tras aplicar las reglas del MULTILINT/CLAT. Una vez elaborado este corpus se traducen los textos de forma automática con el sistema seleccionado en la fase 1. La evaluación de la calidad de ambos corpus nos permitirá extraer conclusiones sobre el impacto de la aplicación de un lenguaje controlado tanto en el texto origen como en el texto meta.

Fase 3. En una última fase llevamos a cabo un estudio económico y de viabilidad para analizar el retorno de la inversión de la implementación de un proceso de traducción con lenguajes controlados y traducción automática en un contexto industrial, teniendo en cuenta la adaptación de los procesos y la idiosincrasia de estas tecnologías.

Tras haber aplicado esta metodología en tres fases los resultados nos han desvelado por una parte qué recursos son los más adecuados para llevar a cabo nuestra investigación y,

por otra, cómo influyen las reglas del lenguaje controlado implementado por la herramienta MULTILINT/CLAT. En concreto, nos ha sido posible dilucidar qué tipo de reglas tienen un mayor efecto en el texto meta, si bien los resultados no son del todo concluyentes debido a la subjetividad de la evaluación y las diferencias entre los evaluadores, aspecto que deberá ser considerado en futuros estudios. Asimismo, el análisis económico y de procesos desvela que para aplicar este tipo de tecnologías es necesario un estudio pormenorizado de todos los factores y la definición de un proceso óptimo. Aún así, esto no implica necesariamente una reducción de costes y, en cualquier caso, no a corto plazo, ya que la implementación de la tecnología viene aparejada con numerosos gastos y la restructuración de procesos. No obstante, sí pueden extraerse otro tipo de ventajas como la reducción temporal de los procesos de traducción o una mejora de la consistencia de los documentos gracias a un mayor control de la terminología y las estructuras lingüísticas.

En general podemos concluir que la implementación de lenguajes controlados sí es percibida como positiva, especialmente para la lengua origen, tal como demuestran los datos presentados en el capítulo 6. No obstante, no es infalible, ya que en algunos casos las reglas pueden no incurrir en mejoras e incluso pueden derivar en un empeoramiento de la calidad del texto. Esto se hace todavía más patente en el texto meta, en este caso el texto en inglés, aunque esto no puede achacarse únicamente al efecto del lenguaje controlado, ya que la misma aplicación de la traducción automática, una tecnología todavía imperfecta, redunda en una merma de la calidad. Una alternativa para intentar solucionar este escollo sería implementar un motor de traducción automática estadística entrenado y adaptado exclusivamente a los textos sobre automoción con los que hemos trabajado. Por otra parte, se podría también aplicar la evaluación con traductores humanos, aunque en ese caso el factor subjetivo de la evaluación aumentaría y además no sería una evaluación óptima según las recomendaciones de White & Taylor (1998), que afirman que un método de evaluación ideal para la traducción automática

[1] "should be readily reusable, with a minimum of preparation and participation of raters

or subjects".

Entre las reglas que tienen un mejor impacto tanto en la lengua origen como en la lengua meta encontramos por una parte normas concernientes a la ortografía y palabras desconocidas y, por otra, al uso de la terminología aprobada así como a oraciones de estructura compleja, que se han de evitar. Esto confirma en parte los postulados teóricos de Reuther (2003) y otros autores que han estudiado los diversos aspectos de la traducibilidad (ver Anexo 2). Lamentablemente no hemos podido extraer resultados concluyentes respecto a las reglas que producen un empeoramiento en la calidad, ya que los resultados están muy sesgados y es posible asignar a una sola regla o grupo de reglas una merma de forma consistente en la calidad.

A continuación ofrecemos una visión global de todo el trabajo resumiendo y exponiendo las conclusiones de cada capítulo. Finalizaremos este resumen con las líneas de investigación futura que nos sugiere este estudio.

En el capítulo 1 hemos analizado con detalle el concepto de lenguaje controlado, acotándolo con respecto a otros términos parejos como lenguaje natural o sublenguajes. Para ello hemos analizado, en primer lugar, los intentos de estudio y sistematización del uso del lenguaje en contextos de especialización, haciendo especial hincapié en la teoría de los sublenguajes postulada por Z.S. Harris. A continuación hemos repasado los diferentes postulados teóricos sobre los lenguajes controlados, concluyendo en la siguiente definición que hemos elaborado a partir de la bibliografía examinada en este capítulo:

> Un lenguaje controlado es un subsistema que contiene tanto elementos propios de un sublenguaje (de especialidad) como del lenguaje estándar, siendo las propiedades de estos elementos una gramática restringida y un léxico controlado.

En el resto del capítulo se han desgranado las ventajas y desventajas argüidas por diferentes autores sobre el uso de estos subsistemas para pasar a continuación a las diferentes clasificaciones, siendo la más habitual y la más apropiada para nuestro estudio la de HOCL (Human-Oriented Controlled Languages o lenguajes controlados

dirigidos a humanos) y MOCL (Machine-Oriented Controlled Languages o lenguajes controlados dirigidos a máquinas). Por último, se han analizado los diferentes niveles de control, que en todos los lenguajes estudiados suelen confluir en tres: léxico, gramatical y de estilo.

El capítulo 2 adopta una perspectiva histórica repasando los intentos por controlar la producción lingüística en la elaboración de textos, remontándonos hasta uno de los primeros intentos por restringir el léxico y las construcciones permitidas, el BASIC English de 1930, y continúa con un repaso a las diversas iniciativas por controlar el lenguaje. En concreto, nos hemos centrado en el contexto industrial y en la lengua inglesa, ya que es donde más ejemplos encontramos de este fenómeno, si bien hemos hecho referencia a iniciativas en otros contextos y en otros idiomas, como por ejemplo el francés, el sueco, el alemán o el español. En el resto del capítulo se analiza la automatización de la aplicación de lenguajes controlados mediante el diseño e implementación de herramientas para la revisión automática. En este sentido, es esencial distinguir entre la especificación y la herramienta: en el primer caso, se trata de un listado de normas y léxico restrictivo. En el segundo caso, hablamos de un instrumento informático que ayuda al autor a aplicar estas normas sin necesidad de consultarlas cada vez.

En el diseño de las herramientas hay dos planteamientos posibles: el prescriptivo, en el que se describen las estructuras permitidas y se detectan aquellas que no pueden ser analizadas, y el proscriptivo, donde se formalizan todas las estructuras que no deben ser utilizadas. Si bien ambos enfoques requieren una descripción y formalización de estructuras, el segundo suele requerir menos trabajo, ya que los desarrolladores pueden concentrarse en aquellos patrones que no son aceptados, que son menos que los que sí lo son. No obstante, en ambos casos es muy complicado cubrir todos los posibles casos, ya que es difícil predecir qué escribirá el autor y cómo se desviará de la regla.

En este capítulo también se han tratado cuestiones relacionadas con la implementación de lenguajes controlados en los procesos de redacción de textos, haciendo un repaso de los diferentes aspectos que se han de tener en cuenta para integrarlos de manera

eficiente en un proceso: el mantenimiento, la formación de los autores y la compatibilidad e integración con otras tecnologías como la traducción automática, son aspectos fundamentales. El capítulo termina con información sobre la herramienta MULTILINT y sus posteriores versiones CLAT y Congree, que es la que se ha empleado en la parte empírica de este trabajo.

En el capítulo 3 hemos abordado aspectos teóricos y prácticos relacionados con la traducción técnica y sus particularidades. Se parte de una introducción histórica para explicar su nacimiento tal como la conocemos hoy en día, se remonta a mediados del s. XIX y tiene su máximo exponente de crecimiento con el estallido de la Primera Guerra Mundial. Mucho ha cambiado desde entonces. En la actualidad el trabajo del redactor técnico no puede concebirse sin ordenador y herramientas electrónicas como plantillas, bases de datos terminológicas y de contenido y sofisticados editores. Sin duda, otro de los cambios ha sido la institucionalización de la profesión del redactor técnico, sobre todo en países industrializados como EE.UU. y Alemania, que cuentan con instituciones académicas y asociaciones profesionales que los respaldan. Tras acotar el término documentación técnica, que entendemos como cualquier documento producido a lo largo del ciclo de vida de un producto, desde su concepción hasta la producción, mantenimiento, uso, disposición y posible reciclaje, el resto del capítulo repasa algunas de las clasificaciones de la documentación técnica hechas por autoras como Reiss, Göpferich, Gamero Pérez y Lehrndorfer y repasa a continuación las distintas tipologías textuales que nos podemos encontrar en el sector de la automoción.

Concluimos el capítulo analizando las particularidades de la traducción técnica, haciendo especial hincapié en las diferencias entre el inglés y el alemán, las lenguas de nuestro estudio, y con algunas reflexiones sobre el uso de la traducción automática para este tipo de textos.

En el capítulo 4 hemos abordado el complejo tema de la evaluación de tecnologías lingüísticas, con el objetivo de ofrecer un marco teórico sólido a la parte empírica de nuestro trabajo. Así pues, nos hemos centrado en la evaluación de lenguajes controlados y traducción automática, las dos tecnologías que hemos abordado en este trabajo de

investigación. En concreto, analizamos los tipos de evaluación, los actores de la misma y la evolución histórica de la evaluación de tecnologías lingüísticas. Hemos revisado asimismo el diseño metodológico de una evaluación, que se desarrolla habitualmente en tres fases:

1. La selección de las herramientas y materiales que han de ser evaluados;

2. La selección de los evaluadores y sus características.

3. La selección de laos parámetros y métricas.

Todo ello dependerá de qué ha de ser evaluado y en qué condiciones, por lo que en primer lugar será fundamental establecer cuál es el contexto de la evaluación, según los postulados teóricos de la evaluación basada en el contexto, para después determinar cómo se evalúa. Esta metodología es válida tanto para los lenguajes controlados como para la traducción automática y es la que también aplica el FEMTI Framework, en el cual nos hemos basado para evaluar las herramientas de traducción automática en la primera fase de nuestro estudio empírico.

Una vez presentado el estado de la cuestión relativo a evaluación de tecnologías lingüísticas, en el capítulo 5 hemos establecido la metodología en la que se basa nuestro estudio empírico. En concreto, esta metodología se fundamenta en las tres fases que hemos expuesto al principio de este resumen:

1. Selección de recursos, que incluye los tres pasos del diseño metodológico de una evaluación.

2. Evaluación de la calidad de textos antes y después de la implementación del lenguaje controlado y la traducción automática.

3. Análisis económico y de viabilidad. La primera fase y parte de la tercera están basadas en el FEMTI Framework. La fase 2 y la otra parte de la tercera fase son de nuevo diseño.

Los resultados de las tres fases de la evaluación se presentan en los capítulos 6 y 7, donde podemos concluir lo siguiente:

a) Respecto a la mejora de la calidad de los textos escritos en lenguaje controlado: se percibe una mejora de la misma en los textos de la lengua origen, que no es tan obvia en textos en inglés, si bien estos resultados pueden achacarse también al bajo nivel de calidad intrínseco de la traducción automática; en algunos casos los evaluadores han percibido un empeoramiento de la calidad de las frases tanto del texto origen como del texto meta, aunque mayoritariamente en este último. Podríamos concluir por tanto que, dependiendo del contexto, algunas reglas pueden tener efectos no deseables para la aplicación de la traducción automática.

b) Respecto al análisis económico y de viabilidad, cuyo desarrollo se hace en el capítulo 7, hemos podido observar que para aplicar este tipo de tecnologías de manera efectiva es necesario que se den dos factores, a saber:

- El diseño de procesos óptimos para garantizar la calidad al mismo tiempo que se busca el ahorro de tiempo o costes;

- Grandes volúmenes de traducción y, si es posible, a varias lenguas.

Para definir a cabo procesos óptimos con garantía de calidad será necesario contar con posteditores o revisores especializados en corregir el resultado de la traducción automática, a los que hay que formar primero y ofrecer tarifas dignas para que realicen un trabajo de calidad, por lo que el retorno de la inversión solo se producirá a medio o largo plazo. En nuestro caso se produce un retorno del 20,76% después de poco más de cinco años. Se trata de un margen bastante estrecho debido al alto nivel de calidad

exigido. Un proceso que no incluyera post-edición conseguiría un retorno en menos de un año, aunque no era este el objetivo de nuestro estudio. Hay que tener en cuenta, no obstante, que la implementación de este tipo de tecnologías puede tener un factor positivo añadido que es el del ahorro de tiempo o la mejora de la comunicación en el seno de la empresa si se implementara un servicio de traducción automática sin post-edición únicamente para este fin.

La evaluación de tecnologías del lenguaje es un asunto complejo que requiere de un análisis profundo del contexto para aplicar la metodología más apropiada, que considere también las restricciones temporales y económicas que condicionan un proyecto. Por ello, cuando nos planteamos un escenario de evaluación, es nefario definir bien cuáles serán los objetivos y el contexto para adaptarla lo mejor posible a nuestras necesidades y hacerla lo más óptima y reutilizable posible en cuanto a los recursos empleados y los resultados obtenidos.

Nuestro estudio ha seguido estas directrices estableciendo los límites de la evaluación y definiendo detalladamente el contexto en el que esta se enmarca. Con el fin de obtener resultados más esclarecedores, un análisis más amplio podría incluir más idiomas meta y más tipos de textos, así como sistemas de traducción automática estadísticos entrenados específicamente para un campo de estudio, que en este caso ha sido el de la automoción. De esta forma podríamos obtener mejores resultados del retorno de la inversión y saber si las reglas de lenguaje controlado que aplicamos en los textos origen tienen los mismos efectos en diferentes lenguas meta o pueden variar. Además, la inclusión de nuevas tipologías textuales nos permitiría saber si las reglas de un lenguaje controlado tienen los mismos efectos en otros textos diferentes a la documentación técnica.

Asimismo es necesario seguir investigando en nuevos estándares y métricas que permitan realizar evaluaciones lo más objetivas y eficientes posibles, optimizando los recursos y permitiendo la correlación con otros parámetros y métricas, para establecer relaciones entre los diferentes aspectos de una evaluación, como por ejemplo, en nuestro caso, entre la comprensibilidad y la traducibilidad de los textos o entre la calidad del

texto origen y el texto meta. En este sentido, sería beneficioso y necesario el desarrollo de aplicaciones informáticas para la evaluación que permitieran una fácil inclusión de los corpus, la selección de métricas y la obtención y análisis de resultados, siguiendo la estela de las que presentan Nießen et al. (2000) o Language Studio, una herramienta de traducción desarrollada por la empresa Asia Online[2] que incluye métricas para la evaluación de la traducción automática. Gracias a estas tecnologías la evaluación se convierte en un proceso accesible a un mayor número de potenciales usuarios y permite una mejora en el proceso evaluativo. No existen, no obstante, herramientas de este tipo para la evaluación de lenguajes controlados, debido en gran parte a las particularidades de cada uno de ellos y a la falta de métricas estándares. Un objetivo para el futuro sería pues el desarrollo de herramientas que facilitaran dicha evaluación, permitiendo por ejemplo la compilación de corpus o la creación de tests para evaluar los efectos y supuestas mejoras que este tipo de tecnología aporta.

0 INTRODUCTION

Damit das Mögliche entsteht, muß immer wieder das Unmögliche versucht werden.
Hermann Hesse, Letter to Wilhelm Gundert , Sept. 1960.

0.1 Motivation

My interest in natural language processing and machine translation goes back to my student time at the Faculty of Human and Social Sciences, while reading for a degree in Translation and Interpreting. My first contact with computational linguistics took place during a extra-curricular course with professor Juan Carlos Ruiz Antón at the Universitat Jaume I. I was impressed by the idea of formalizing human natural language and creating useful applications to improve human-machine interaction.

I decided to study Computational Linguistics in more depth, and I enrolled in the postgraduate program at the University of Munich. Due to my background as a translator, I had a continuing interest in analyzing and creating tools to help the translator. This led me to complete a number of work placements at companies that developed language technologies or worked to implement such technologies in their workflows.

The present project was born seven years ago, while I was working as a student trainee at BMW AG in Munich, Germany. Authors at BMW had been using the tool MULTILINT to adapt their texts according to a set of rules with the aim of improving translation and terminology processes. At that time there was a project in progress to switch to a new version of MULTILINT, named CLAT, and which today is marketed as Congree[4], even though CLAT is still used in some settings. However, there was a

need to empirically demonstrate the associated improvements so that the investment in material and human resources could be justified from a managerial point of view.

This was the trigger that marked the beginning of my research work, and which has taken up my time, with many interruptions, for these last seven years.

During this period there have been moments of despair and disillusion, as well as sparks of motivation and fascination. Especially the last two years have embodied a "drive" due to the increasing relevance and topicality enjoyed by controlled languages and, especially Machine Translation. Projects such as EuroMatrix[5] and its follow-up project, EuroMatrixPlus or Panacea[6], as well as the continuous growth of Statistical Machine Translation and the need to optimize translation processes in an increasing globalised world have helped to award even more significance to this work.

0.2 Breeding Ground

The automotive industry in Germany is characterized by expanding model series coupled with shorter product development cycles and a growing complexity of vehicles. This results in a sharp rise in demand for technical information at the wholesale and retail level. Not only does this imply an increase in source language texts (usually German or English), but also an exploding number of documents in all the different languages into which technical documentation needs to be translated.

It is a fact that the amount of documentation increases annually due to the reasons mentioned above. The need to maintain a high language quality both in the source and in the target texts, without increasing documentation and translation costs, is therefore real and pressing. Companies have long recognized all these hurdles and have been working in the past years on the creation and maintenance of linguistic resources such as terminology databases and translation memories. Though these efforts are valuable and contribute to gaining quality and reducing costs in content creation processes, further options have to be considered and evaluated in order to face the imminent

increase in content and costs. Therefore, it has become necessary to adjust the information flow within the companies.

One of the approaches used by multinational companies such as General Motors, Volkswagen, or BMW to optimize language management and the information flow is that of controlled language (Bernardi, Bocsak, & Porsiel, 2005; Haller & Fottner-Top, 2001; Means & Godden, 1996). As Feely & Harzing (2003) point out, "a controlled language imposes limits on vocabulary and syntax rules so as to make the text produced more easily comprehended by the non-native speaker/reader and equally more amenable to machine translation". I will deal more in depth with the definition of this concept in Chapter 1. Since it is difficult to implement these languages coherently, special tools are used to automate their application in industrial contexts. One example of these tools is MULTILINT, which is a prototype solution that automatically checks the texts with respect to orthography, grammar, style, terminology, abbreviations and consistency. Some companies such as Volkswagen, Daimler, BMW or Siemens have implemented this tool in their authoring processes and are still using it as CLAT or Congree, as is claimed in the informational brochure of the new company marketing the product.

0.2.1 Authoring Processes in the Automotive Industry

The literature produced within the automotive industry is usually varied and heterogeneous. Without taking into account other kinds of documents produced in Engineering or Marketing departments, in the Support and Service areas alone there is a myriad of different document types. These include service information, repairing instructions, tightening torques for tiring changes, inspection sheets, technical data, training documents, diagnosis, technical campaigns, programming data and owner's manuals.

To solve this problem, some companies have worked on implementing single authoring systems to integrate all these document types and profit from the potential synergy between them. In this way, a single information platform supplies all relevant data on the basis of uniform criteria – from adoption, development and production to

compilation, translation and distribution, to be used within the car dealer organization. However, there are still some other companies that work with heterogeneous single solutions. Besides, the information contained in these documents is not always accessed by the same systems in the workshops. As a consequence, every system offers different information, resulting in inconsistency or redundancy. Furthermore, the information is not always available at every workstation and it is necessary to change the location of the vehicle to access the right information.

0.2.2 MULTILINT, CLAT and Congree

One possible solution to tackle the heterogeneity of documents and to attain more consistency among them, is to support authors by an authoring tool when checking and proofreading texts. Automotive companies such as Bertone, BMW, Jaguar, Renault, Rolls-Royce Motor Cars, Rover and Volvo participated in the development of MULTIDOC, an initiative to establish a common basis for collaborative efforts of the European automotive industry in the production, management and translation of technical after-sales information (Haller, 2001). BMW was involved in a previous German National project, MULTILINT, implementing this tool in its processes. It continues to used MULTILINT in the form of a newer version with added functionality, called CLAT. Other automotive companies that use CLAT/Congree in their authoring processes are Volkswagen and Daimler.

MULTILINT was a prototype solution designed by the Institut der Gesellschaft zur Förderung der Angewandten Informationsforschung e.V. (IAI), located in Saarbrücken, Germany. It was designed to automatically check texts with respect to orthography, grammar, style, terminology, abbreviations and consistency. Furthermore, the tool detected potential term candidates and proposed them for integration in a terminology database.

MULTILINT was developed in the frame of the project TETRIS, the goal of which was to develop a prototype system to support technical writers when writing their documents. BMW has been using this tool for the creation of technical documentation

since 1996 (Haller, 1996), contributing to the quality and consistency of documents. This tool was first a prototype solution and was voluntarily used during a period of over 10 years. Today it is formally embedded in the authoring process in the form of CLAT.

The aim of this tool is to create more understandable and translatable texts by checking the terms used in the text, contributing to the creation of short and intelligible sentences, and to apply abbreviations correctly. However, even after its complete deployment, it has been difficult to prove that this tool brings significant improvements to the linguistic quality of the documents, especially with respect to translation. The TETRIS Project[7], in the bosom of which MULTILINT was born, deals with the evaluation of MULTILINT and the evaluation of TERMLINT, a tool developed also in this project for the extraction and administration of terminological resources. The evaluation of MULTILINT was divided in two parts: "Proof-Reading" and "Hit Rate in Translation Memory Systems".

The aim of the first evaluation scenario was to determine the average saving potential obtained by using MULTILINT in contrast to human proofreading. The tests included a statistical macro evaluation, where factors such as different scenarios for the creation of content, usability of the system and general program behaviour were tested. A dynamic micro evaluation was also carried out, focusing on texts checked with MULTILINT. In this case, the results had to be evaluated with regards to the information retrieval measurements "precision" and "recall", that is, how many mistakes were recalled and, out of these, which of them were correctly recalled or not (precision). The conclusion of this first evaluation scenario was that MULTILINT was not sophisticated enough to replace an experienced and specialized human proof-reading.

The second evaluation scenario, "Hit Rate in Translation Memory Systems", intended to prove that the use of MULTILINT could increase the hit rate in translation memory systems by ensuring more consistency in the source texts. Though this scenario was repeated twice, the results were not meaningful enough, due to subjective factors such as the learn effect on MULTILINT and the differences in the writing skills of the different authors.

All in all, it was not possible to assess and prove the quality of MULTILINT in a meaningful way, concluding that determining a significant improvement in the translatability of texts by using MULTILINT is extremely complicated due to the subjectivity of translations evaluated by humans.

The next version of MULTILINT was called CLAT, which stands for Controlled Language Authoring Tool. It relied on the technology developed by the Institute of the Society for the Promotion of Applied Information Sciences at the Saarland University (IAI). Today the tool is known as Congree and is exclusively marketed by a company with the same name, a joint venture between Across Systems GmbH and the above mentioned Institute.

This new rule and style-checking includes a wide variety of checking criteria such as

- Grammar and spelling
- Sentence length
- Use of defined word types and forms
- Conventions for punctuation and syntax
- Word choice
- Sentence structure
- Writing style

Although the look and feel and many of the features of Congree have changed with respect to MULTILNT and CLAT, the linguistic intelligence behind the system has only been slightly modified, and therefore the results obtained from the data checked with MULTILINT and CLAT can be extrapolated to Congree.

With this scenario in mind, I aim at presenting a new approach that can contribute both to the reduction of translation times and costs without neglecting quality, and to objectively evaluate the controlled language behind MULTILINT/CLAT with regards to its effect on translatability with Machine Translation, that is, the translation of a natural language into another natural language by a computer. The approach presented

in this work should help to gain an objective assessment of the quality of the controlled language behind MULTILINT/CLAT with regards to its effects on machine translatability, as well as to prove a methodology for the evaluation of controlled languages in general and the implementation of this kind of technology in authoring processes.

0.2.3 Machine Translation

Machine Translation (MT in subsequent text) can be defined as the transfer from one natural language to another with the help of a computer. Research in the field of MT can be traced back to the 1950s, when it was fostered by the great advances in cryptography and the wish to understand the messages intercepted during the period of the Cold War. The assumed goal was the automatic translation of all kinds of documents at a quality equalling that of a human translation. Soon, it became obvious that this goal was impossible in the foreseeable future. At the same time, nonetheless, it was found that for many purposes, the unedited MT output could be useful to those who wanted to get a general idea of the content of a text in an unknown language as quickly as possible (Hutchins, 2003). I will deal more in depth with Machine Translation and especially on its evaluation in Chapter 4.

Regarding the case of BMW, no track has been found of this technology being consistently implemented in the translation processes. In an internal IAI report, Rita Nübel studied the implementation possibilities of MT for DES-texts (Nübel, 2000). The results of this study, which investigated the MT from German into English, and the translation of machine-translated English into other languages, concluded that MT could not be reliably applied without previous intelligent pre-edition (controlled language) and post-edition processes. Since then, no other experiments have been carried out.

0.3 Hypothesis , Goals and Methodology

The present work, the title of which is “Use and Evaluation of Controlled Languages in Industrial Environments and Feasibility Study for the Implementation of Machine

Translation", presents an investigation into the effectiveness and impact of controlled languages within industrial environments, especially from the point of view of the improvement of machine translatability and the quality of the target texts. With this goal in mind, I will undertake a case study using technical automotive documentation written in German that needs to be translated into English.

This work aims at studying, on the one hand, the effectiveness of implementing controlled languages in the authoring of technical documentation, especially with regards to the improvement of translatability, a concept that will be tackled in this work. On the other hand, the implementation of MT within the translation process in an industrial environment will also be tackled. For this purpose, different aspects will be considered and analysed, such as text typology, cost and time factors, linguistic quality and usability aspects. This investigation aims at proving these two proposals as summarized by the following hypothesis:

- First of all, texts written according to the rules of a controlled language and written with the aid of a tool such as MULTILINT/CLAT for the application of such rules improve their readability, comprehensibility and translatability.

- Secondly, MT represents an objective "evaluator" with regards to the translatability of texts edited in compliance with the rules of a controlled language. In this respect, I expect to discern whether texts edited with a linguistic tool suite are more machine-translatable than others, offering a new approach for the evaluation of this type of authoring tools. Furthermore, this will lead to detect which rules of the linguistic tool are prone to lend more translatability to the text, as well as to identify new rules which could improve both the readability and translatability of the source text.

- Finally, MT represents an alternative technology to tackle the increasing amount of technical documentation and, thus, of translation volume. With well-defined processes, the implementation of MT can save time and costs in the translation processes without compromising on quality.

As a result of this analysis, I expect to offer empirical evidence that controlled languages do indeed bring the claimed advantages (improvement in readability, translatability and comprehensibility of the source text), as well as to establish the elements that might lead to the recommendation of or advising against the implementation of MT. Furthermore, relevant data for an objective evaluation of MULTILINT/CLAT as well as the specification of new identified rules for the improvement of translatability of the source text will be presented.

In order to demonstrate these hypotheses, the first part of this work presents a review of the current literature on controlled languages, the controlled languages industry, technical translation and evaluation methods for natural language processing, specifically controlled languages and machine translation.

The second part of this work presents the methodology carried out and deployed in order to carry out the empirical study, the results of which are presented in part three. This study is divided into three different phases:

- 1st Phase. First of all, a micro-evaluation will be carried out. An analysis will be done of which texts are currently written according to the rules of a controlled language and which texts will be in the near future. Subsequently, with the help of a form especially designed for this purpose, I expect to determine which information type or types are more suitable for machine translation. Finally, a linguistic analysis based on criteria derived from the literature will be carried out and the most suitable information type will be established. Furthermore, different MT-systems will be tested and evaluated to choose the most appropriate one for an industrial context, and for Phase 2.

- 2nd Phase. In the second phase, a corpus of texts checked and written according to the rules of controlled languages and a corpus of texts not following these rules will be compiled in order to carry out a macro-evaluation. Subsequently, after installing and training the MT-system chosen in Phase 1, translations will be carried out and the quality of the translated texts will be evaluated, comparing the

results of both corpora. This will help to draw conclusions about the quality of MULTILINT/CLAT and to take the decision of whether to implement MT or not in the real translation processes.

- 3rd Phase. In a final phase, a feasibility study that analyses the return on investment of implementing MT technology in combination with a controlled language within an industrial environment, as well as the necessary adaptation of workflows and processes, will be presented.

The third part of this work presents the results of this empirical study of machine translatability of technical texts written according to the rules of the controlled language checked with MULTILINT/CLAT. These results include the data of the three phases with an interpretation of their relevance and significance.

0.4 Organisation of the present work

The following work is divided into three parts comprising nine chapters. The first part is devoted to offering an overview of the state of the art in controlled languages and technical documentation, with the aim providing a theoretical framework for the present work. The second part describes the methodology followed during the empirical part of the research and, finally, the third part presents the results of the empirical study, the feasibility study and the conclusions and future prospects. The contents of the nine chapters are defined next.

Chapter 1 tackles the notion of controlled languages (CL) and its conceptual delimitation, putting it up against other concepts such as natural language and sublanguages. It also covers the different definitions that can be gathered from literature. Advantages and disadvantages of using controlled languages are discussed. Furthermore, it offers an overview of the CL typology as well as the areas of control.

Chapter 2 focuses on the application of CLs in industry. It covers a range of examples of the application of controlled languages in industry, especially for the production of

technical documentation. An overview of the different techniques used for controlled language checking is given and the chapter ends with a description of different tools in the market designed to control language automatically in the text production.

Chapter 3 deals with technical documentation. After a short introduction, a historical overview of the development of technical documentation is given. Other aspects, such as the situation of technical writers and their education, are also discussed. A quick overview of the different types of technical documentation and goal groups for which this documentation is written is presented in sections 3.5 and 3.6. Section 3.7 deals with the particularities of technical translation. Finally, a review of the technical documentation at BMW AG is given.

The next chapter gives a general overview of the subject of natural language processing evaluation, in particular of controlled languages and Machine Translation. After a review in language technology evaluation, a number of practical issues are tackled relating to the selection of resources for evaluations where language processing is involved. Subsequently, it deals with evaluating CL rule suites, CL checkers and MT evaluation, concentrating on the notion of quality, previous experiences and different methodological approaches to this issue, with a special focus on human versus automatic evaluation.

Chapter 5 presents the methodology designed for the empirical part of this study. It is a three-base approach methodology with three different goals: the selection of resources, the analysis of a CL rule suite and a workflow and feasibility study. The results of this methodology are presented in chapters 6 and 7, whereas chapter 8 offers the conclusions and some future prospects for further research.

This chapter tackles the concept of Controlled Languages and their application in industrial environments. It starts with overview of a wide range of examples of their application in industry, in particular in the production of technical documentation, both for the English language and other languages.

It is important to distinguish between Controlled language specifications and the software tools used to check these specifications while authors write their texts. Therefore, the second part of the chapter concentrates on controlled language checking, with an overview of the different techniques and different tools available in the market designed to control automatically the text production.

Finally, the chapter ends with a survey of the different CL checkers available, with a special emphasis on the tool MULTILINT/CLAT, which will be subject to analysis in this research work.

Part I
State of the Art in Controlled Languages, Technical Documentation and Evaluation. Theoretical Framework

1 DELIMITING AND DEFINING CONTROLLED LANGUAGES

No one means all he says, and yet very few say all they mean, for words are slippery and thought is viscous.

Henry Brooks Adams, *The Education of Henry Adams*, 1907

1.1 Introduction

Due to the inherent complexity of our current societies, communication needs are very often based on the transfer of expert knowledge. The language used in this type of communication usually differs from the language used in general-purpose communication, where vocabulary is usually unspecific and syntactic rules follow the general rules of language.

Especially important for the communication of expert knowledge is the clear and coherent transmission of information, particularly when this is intended to be localised[8] for different markets that speak different languages. In the past few decades many efforts have been made in order to establish some guidelines[9] for writing expert communication[10] intended for an international audience, since due to its inherent ambiguity, natural language represents very often difficulties for both readers and translators.

Controlled Languages (CLs) address this problem by restricting vocabulary and grammar in a definite domain. They are used to write specialized text, usually technical documentation. It is commonly accepted that texts written according to the rules of a CL

become easier to read and to understand (Nyberg, Mitamura, & Hujisen, 2003). This, in turn, improves the efficiency and accuracy of all tasks related with the processing of technical communication, such as creating it, retrieving information from it or translating it. Furthermore, the "formalisation" of a language helps to smooth the human-machine interaction in applications such as translation memories or Machine Translation (MT). This common belief bases on intuition and on some empirical studies such as those by Adams, Austin, & Taylor (1999); Barthe et al., (1999) and Mitamura & Nyberg (1995) though results cannot be generally applied for all domains and languages. Differences in the structure of different languages and complexity of domains signal that CLs are not always appropriate[11]. This situation is well defined by U. Knops (2000), who points out that

> Generally speaking, there is an urgent need for facts and figures obtained in experimental situations and real-life production environments and relating to the effects of particular CL standards, rules and rule sets on readability and translatability.

This chapter represents an attempt to define CLs, delimiting them from other concepts such as sublanguages or artificial languages. To do this, I start by defining the concepts of natural language and sublanguages (1.2 and 1.3), to focus subsequently on the concept of controlled languages as artificial languages designed to improve the readability and translatability of texts (1.4). Here, different types of CLs as well as the advantages and disadvantages of the application of CLs are discussed. Further, I discuss different CL classification schemas to end up with the different areas involved in the control of a language, such as vocabulary and grammar in 1.4.4). I finish this chapter by adding some concluding remarks on CLs (1.5).

1.2 Natural Languages

The definition of language has been the subject of numerous linguistic and philosophical discussions. However, it is neither my intention nor the goal of this work to deep into this debate.

Generally, I use the term natural language as a communication system made up of conventionalised symbols and rules by which these symbols are governed. This system has been commonly accepted and is used by a community of speakers. Examples of natural languages are at hand. In this world there are plenty of these communities of speakers that are familiar with one or more systems: German, English, Spanish or Japanese are a few examples. The most extensive catalogue of the world's languages, generally taken to be as authoritative as any, is that of the Ethnologue organization, whose detailed classified list currently includes 6,909 distinct natural languages[12].

According to the structuralist linguist Zellig Harris, mathematical linguistics characterizes natural language as a "system of sets of arbitrary objects, the sets being closed with respect to particular operations, with certain mappings of these sets into themselves or into or onto related sets" (Harris, 1968: 1). This definition elucidates that the system of language is formed by arbitrary objects, which turn to be words in the most general sense, and that these objects are organised in sets that can be operated by a limited number of operations, such as coordination, being able to combine among themselves or with other types of related objects. Harris, however, does not indicate which requirements these objects need to fulfil in order to belong to a definite set. Usually, syntactical requirements are applied, so that these objects will be organised in syntactic categories such as noun, adjective, verb etc. However, as we will see later (in 1.3), for sublanguages it will make more sense to organise the sets from a semantic point of view.

For my purposes, I will use the term natural language as opposed to artificial, constructed or planned languages, created to expound a conceptual area (e.g. "formal", "logical", "computer" languages) or to facilitate communication (e.g. Esperanto) (Crystal, 1987: 352). Therefore, in linguistic terms, natural language only applies to a system the components of which have evolved naturally and arbitrarily, while the rules of artificial languages are prescribed prior to its construction and use.

1.3 Sublanguages

If I consider a natural language as a system formed by arbitrary objects such as signs, sounds and rules governing those sings and sounds, I observe that, naturally, that big set tends to be divided in smaller subsets depending on the communicative situation involved. These subsets are characterised for having not only a special vocabulary or lexicon but also particular grammatical and pragmatic and stylistic features. These might be called subject matter varieties, registers, languages of specialization or sublanguages. Indeed, the terminological variety is one of the first problems I am faced with when trying to define the concept of sublanguage[13].

This correlation of what is being written or spoken and the language used for this purpose began to be studied in the 60s[14]. It was the structuralist linguist Zellig Harris[15] in his work on transformations and discourse analysis who developed the idea of sublanguage with the mathematical idea of "subsystem" in mind, the "sub"- prefix indicating not inferiority, but inclusion. Harris defined sublanguages in the following way: "A subset of the sentences of a language constitutes a sublanguage of that language if it is closed under some operations of the language" (Harris, 1988). In this view, he was referring to sublanguages as a subset of the general or natural language. These sublanguages must be closed so that if two members of the subset are operated on, for instance, if they are linked by a conjunction, the resultant belongs also to that subset. However, though in mathematics the definition of a subsystem is relatively easy to define by limiting the elements and the operations among these elements, a sublanguage might allow operations that are not necessarily part of the standard language. Thus generally speaking, sublanguages designate a subset of the natural language that makes itself distinctive for being group or subject oriented. Indeed, "natural sublanguage" would be a more appropriate term and when in the course of this work I tackle the notion of sublanguages I will always refer to the natural variant.

Other definitions of sublanguages can be found in the literature by disciples of Harris. Kittredge & Lehrberger (1982:2) state that "the term sublanguage has come to be used […] for those sets of sentences whose lexical and grammatical restrictions reflect the

restricted sets of objects and relations found in a given domain of discourse". Lehrberger (1982) defines sublanguage as "a language resulting from restriction on and deviation from the standard grammar of a natural language; often a sublanguage grows in a natural way through the use of the standard language, albeit in special circumstance". In the same line, Grishman & Kittredge (1986) define a sublanguage as "the specialized form of a natural language which is used within a particular domain or subject matter". Hirschman & Sager (1982: 27) propose following definition:

> I define sublanguage as the particular language used in a body of texts dealing with a circumscribed subject area (often reports or articles on a technical speciality or science subfield), in which the authors of the documents share a common vocabulary and common habits of word usage. As a result, the documents display recurrent patterns of word co-occurrence that characterize discourse in this area and justify the term sublanguage.

Another definition proposed by Alonso Cortés (1994: 243) is the following:

> Los sublenguajes son especializaciones del lenguaje común para uso de una comunidad humana con fines específicos. Un sublenguaje no se limita al empleo de una terminología específica. También lo caracteriza el uso de ciertas estructuras sintácticas y morfológicas, así como especializaciones semánticas del léxico común.

As we can see in all these definitions, the main particularities of the notion of a sublanguage are lexical and grammar restrictions in contrast to standard language, these restrictions being subject matter or domain-specific. This implies that sublanguages can only develop when there are a number of speakers interested in exchanging specialized information.

As Alonso Cortés (op.cit.) states, the study of sublanguages in Spanish has not been properly investigated, whereas there are numerous works in English, Russian, French or German. In this last language, the term sublanguage correlates the term "Fachsprache" or "Subsprache" in German, for which Lehrndorfer (1996: 36) gives following definition, where the subject-matter specificity is also evident:

> Eine echte Fachsprache/Sublanguage entsteht in einer Kommunikationssituation (überwiegend in schritflicher Form), in der sich der Mitteilungsinhalt (Thema) über ein alltagssprachliches Problem heraushebt und für eine erfolgreiche Kommunikation spezifisches thematisches Wissen und dessen müheloses fachsprachliches Formulieren und Verarbeiten von Sender und Empfänger Voraussetzung ist.

Despite this variety in descriptions, I am mainly interested in the properties of sublanguages so that I can distinguish them of other concepts, particularly from that of CLs. All the definitions above emphasise the following important ideas that Kittredge (1985) summarizes as follows:

Sublanguages are a subset of the sentences from the natural language used in a domain of discourse.

This, in turn, supports the thesis that sublanguages arise when some "community" of expert speakers need to communicate: they use special terminology, common words with specific meaning and certain grammatical expressions in order to express expert knowledge intended to a definite audience, and this in a recurrent situation.

These sentences are formed by a set of objects or classes among which only certain operations and relations are possible, that is, there are lexical and grammatical (syntactic and semantic) restrictions, though they have the "essential" properties of a linguistic system, such as "consistency", "completeness" or "economy of expression".

There might be deviant rules of standard grammar, that is, rules describing sentences which, though quite normal in a given sublanguage, are considered ungrammatical in the standard language, as well as rules describing co-occurrence restrictions within a sublanguage that do not exist in the standard language.

In the mathematical sense, this language subsystem is maximal with respect to the domain, that is, no larger system has the same properties.

However, in order to understand the implications of these properties, I also need to revise Harris' theory regarding sublanguages. Harris (1988) describes the structure of language as a set of sentences (word sequences) that are constructed satisfying three

main restrains: the partial-ordering constraint, the likelihood constraint and the reduction constraint.

The **Partial-order Constraint** determines the sentence structure. It is partial because it is rough, since every word has an argument set based on the probability of certain words occurring next to it. In this respect, there are zero level words, which do not require any argument, first-level operators, which require only zero-level words, and second-level operators, which require first-level operators. This constraint forms word classes and is concerned with syntactic structure.

The **Likelihood Constraint** specifies word meanings and is based on the idea that, depending on their meaning, some words are more probable to appear next to other words than others. The set of words that have a higher-than-average likelihood is called the selection. In general language, the likelihood constrains on operators and their arguments is fuzzy, while in sublanguages the Restrains are generally sharper. In either case, in spite of being a semantic constraint, it affects the syntactic structure of the sentence.

The **Reduction Constraint** consists of the paraphrastic reduction in the phonemic shape of particular word occurrences that have exceptionally high likelihood or a special status in a given position. More concretely, the reductions involve transformation of the sentences from a simple primitive form to a complex form (surface form), which are the actual sentences that appear in documents. This constraint changes the structure of the sentence without affecting the informational content.

This last constraint partitions the set of sentences of a language into two major sets:

- without reduction, they create a base set from which all other sentences are derived;
- with reduction, giving place to reduced sentences.

It is important to highlight that neither the base set nor the other set (the derived or reduced set), is merely a residue of the other. A sublanguage is constituted by a subset of the general language containing both base and reduced sentences. However, a sublanguage usually presents more specialized Restrains than the general language due to limitations of the words and relations of the subject matter. For instance, in the general language, it is permissible to say *John triggers the airbags*, because the syntactic combination of word classes is well-formed. But in the automotive language domain, this sentence is not legitimate because the operator *trigger* permits only certain combinations of the word classes (i.e. a module triggers the airbags, a system triggers the airbags, but a person triggers the airbags is not allowed). In the same way, there will be some structures or operators that only exist in particular sublanguages and not in the grammar of a standard language. The sublanguage operators reflect the salient relations and arguments that are meaningful in the specialized domain.

Another important characteristic of specialized texts is that, when looking at them, sublanguage patters are often interwoven with general language patterns, which makes the process of identifying sublanguage patters difficult and can cause difficulties when processing automatically this type of texts with a sublanguage grammar.

Therefore, if I want to characterize a sublanguage in order to be able to process it automatically or to derive a controlled language from it, it is necessary to study a corpus in order to discover the classes and subclasses and the operators that build up relations among them. As I have said, sublanguages can present non-grammatical sentences from the point of view of the standard language, sentences that are grammatical from the point of view of the general language but not allowed in the sublanguage grammar, as well as omission of information (sentences that miss subjects and verbs) because the information can be recovered from the context.

Based on Zellig Harry's theory, Friedman, Kra, & Rzhetsk (2002) discuss features of languages in specialised domains that have important implications for the development of computerised natural language processing systems. I can group these features, as Nyberg, Mitamura, & Hujisen (2003) do into three main categories: the lexicon —also

including terminology, the syntax, and the text type. These will be discussed in the following sections.

1.3.1 The Lexicon

Here we can find different aspects related to the lexicon used in a sublanguage and the semantic information transmitted by it.

First, in a sublanguage we find a semantic categorisation of words. Relevant words can be categorised into subclasses or types of information where the types form the underlying subject matter of the domain. For example, in the automotive domain, the pieces and functions of cars are divided in engineering groups, such as Lights, Motor, Seats, Gearbox or Communication Systems. These, at the same time, can be divided into subclasses; for instance, Communication Systems can comprise Radio, Navigation or Workshop Telecommunication.

Within these classes and subclasses, I find specialised terminology. This can be constituted either by specialised terms that only occur in a certain domain or everyday vocabulary that is highly characteristic in a sublanguage, being its use specialised. The ambiguity and homonyms of these words (also called semi-technical) is reduced, since the trend is towards univocity and words are used with a preferred sense. For instance, in the automotive domain, *airbag*, *camshaft* and *headlight* are specialized terms, since they only occur in this domain or very close domains. Words, such as *pillar*, *beam* or *inlet* are general words that are used with a special sense. There are even sometimes grammatical hints to indicate that these words are being used with a special meaning. For instance, words that might be grammatically ambiguous because they can belong to different syntactic categories appear predominantly in only one usage: *control* and *signal*, which might be a verb or a noun, are used exclusively as a noun in the automotive domain. In other occasions, these words can sometimes even undergo grammatical changes in order to differentiate them from the "common" meaning, such as the formation of plurals or the gender of terms. For instance, the term *Virus* in

German has a different gender depending on if it refers to a computer virus (*der Virus*) or to a biological virus (*das Virus*).

1.3.2 Syntax

When analysing corpora of specific texts, it can be observed that certain constructions are more prevalent than others. For example, the use of imperative sentences in recipes or instruction manuals is much more common than in weather reports. In the same way, certain constructions are disfavoured in certain sublanguages. For instance, direct questions or tag questions are not common in biomedical reports.

With respect to the general language, there is a reduction in the range of constructions. However, a particular feature of sublanguages is that they permit deviant constructions that under normal circumstances would seem odd. For instance, in job advertisements it might be usual to see sentences that consist of a series of nominal phrases, without a main verb (Buchmann, Warwick, & Shane, 1984)

Friedman, Kra, & Rzhetsk (2002) distinguish certain syntactic features of sublanguages that can be distinguished in a more detailed analysis:

Co-occurrence patterns and Restrains (which can be matched with Harri's likelihood constraint): There are certain classes and subclasses that combine in particular co-occurrence patterns to form the meaningful relations of the domain. In this sense, it is important to indicate that these semantic classes not always match with syntactic categories, but sometimes they can be built from semantic homogeneity. For instance, in the weather sublanguage, <weather condition> might be a class, though it can be formed by a variety of syntactic categories.

Paraphrastic patterns (which can be matched with Harri's partial-order and reduction constraint): A set of patterns represent an equivalence relation where the patters are different grammatically but represent the same underlying operator-argument structure,

that is, there can be in a sublanguage text a combination of reduced structures and non-reduced structures that have the same informational content.

Omission of information: It is characteristic in a sublanguage to omit additional contextual information when the context is known. However, this might cause problems in language processing because a system usually lacks this additional knowledge necessary to recover the implicit information. This is one of the points that controlled languages aim at detecting and correcting.

Intermingling of sublanguage patterns and general language. The sublanguage patters are often interspersed with general language that is not in the sublanguage, and this might be sometimes difficult to detect. Controlled Languages will also try to detect these patterns of general language and adapt them to the characteristics of the sublanguage, especially in very specialized technical text types.

1.3.3 Text-Type

Even within the same sublanguage, I can encounter a great variety of text types. Different text types are determined by the medium (spoken or written, though I will mainly deal with written texts), the author, the content or the function and the goal of the text. For instance, in the automotive domain I can distinguish among marketing brochures, owner's manuals, technical information or repair instructions etc. Each of these types has its own distinctive features, with terminology and syntactical structures particular to it, which will determine a kind of sub-sublanguage (a sublanguage with particular features within the more general automotive sublanguage).

For instance, different vocabulary will be used in a repair manual, which contains mainly technical information addressed to a mechanic, in relation to a marketing brochure intended for final clients that are not always interested in technicalities.

1.3.4 Sublanguages and Machine Translation

The use of sublanguages in combination with MT was pioneered by researchers at New York University in the later 1960s, lead by Naomi Sager, who was indeed a student of linguistics of Harris[16].

Subsequently intensive research was carried out until the 90s, triggered by the successful analysis of different sublanguages in the biomedical domain and the success obtained by the TAUM-METEO system[17]. This system, developed in the late 70s at the Université de Montréal automatically translated weather bulletins from English into French for the Canadian government until 2001. Then it was replaced by a competitor program after an open governmental bid (Canadian International Trade Tribunal, 2002). The system used a controlled sublanguage to improve the MT output quality. Upon the successful completion of the TAUM-METEO MT system, the same group started developing TAUM-AVIATION, an experimental system for English to French translation in the sublanguage of technical maintenance manuals. However, it turned out that the text of these manuals did not constitute a sufficiently limited domain, and AVIATION did not perform as well as the METEO system (Isabelle & Bourbeau, 1985).

Another system that uses a controlled sublanguage for better MT output quality is discussed in Adriaens & Scheurs (1992) and Hutchins & Somers (1992: 322-325). The TITUS system employed a controlled language (Langage Documentaire Canonique) designed at the Institut Textile de France in the 70s for the multilingual treatment of abstracts in an on-line database.

Indeed, there is a clear link between the use of sublanguages and MT. It is well known that MT systems can cope only with difficulties with general purpose texts and that much better results are reached when these texts are somehow restricted. With sublanguages these restrictions occur naturally, and an MT system can adapt to these restrictions or can even be directly developed taking these restrictions into account. For instance, specialized text might only need a shallow parsing if the MT system is

designed on the grounds of a particular sublanguage, requiring thus less resources and computational effort and being therefore faster. Lexical problems such as homograph resolution and polysemy disambiguation can be easily solved since words are usually used with a preferred and unique sense or grammatical category, though a requirement for this is a good terminology management that determines in a precise way the categorization and subcategorization of lexical items. Indeed, usually lexical items are the major bearers of textual meaning: if they are translated properly, there are big chances that the meaning of the text is transferred, even though if syntactic structures are not perfect or do not sound natural in the target language.

However, a sublanguage is not always appropriate for the use of MT. Deviant constructions or omission of information can hinder the performance and output quality of MT and not all sublanguages are necessarily good for MT (Van der Eick, de Koning, & van der Steen, 1996: 66). Kittredge (1985: 159) distinguishes following features that will make sublanguages appropriate or not for MT:

Size of the sublanguage. This will determine the size of the vocabulary and terminology. Indeed, depending on the subject domain, the size of the lexicon can vary hugely. The weather-bulletin sublanguage is reportedly based on a lexicon of less than 1,000 words, excluding place names. A set of aircraft maintenance manuals contained 4052 different entries only for the hydraulics domain (Isabelle & Bourbeau, 1985: 19).

Complexity. A sublanguage can be very big in size but can use predictable sentence structures, making it easy for a MT system to analyze. On the other side, it can have a reduced vocabulary but use complex, unpredictable sentences or structures full of ellipsis that might hinder the performance of MT.

Adherence to systematic usage. The degree to systematicy is given by the systematic usage of the distributional patterns of words that define the sublanguage. The more the sublanguage adheres to systematicity, the more amenable it will be to automatic translation.

1.4 Controlled Languages

1.4.1 Definition of Controlled Languages

Sublanguages arise due to the need to express in a linguistically economical and understandable way certain expert content. In order to maintain this content as informative, precise and unambiguous as possible to maximize its informative purpose, there has been always an interest to control the vocabulary and the grammatical structures used in these sublanguages, especially in written communication. It is not difficult to come across style guides, norms and recommendations for definite expert fields. The idea of Controlled Language arose indeed before the concept of sublanguage, since the first controlled language to be developed was BASIC (British American Scientific International Commercial) English in the 30s. BASIC English based on the idea that 850 words would be sufficient for ordinary communication in idiomatic English.

The term itself, "Controlled Language" (further CL) is used in various contexts with different meanings. Though all of them are based on the conception of a language, be natural or artificial, that undergoes a certain degree of control, there are slight differences depending on the discipline in which the term is used. For instance, in information management and documentation science, the term is often employed as a synonym for controlled vocabularies, a type of documentary language used to index and retrieve information from documents. Likewise, in recent years a new approach to controlled language as a computer processable language has arisen in order to cope with current information processing issues, such as knowledge representation, reasoning or symbolic input to multilingual language generation. This approach uses a controlled language as a basis but goes further in being capable of being completely syntactically and semantically analysed by a natural language processing system (see Chapter 2, 2.2.1, for further information).

However, this is not the focus of this work. I will concentrate here on controlled languages as "languages for practical business" as defined by Sukkarieh, Hartley, &

Scott (2003), that is, a variety of a sublanguage which is restricted with respect to vocabulary and syntax with the aim of minimising the intrinsic ambiguity of natural language and improving the readability and translatability of texts, and in particular of technical documents.

In this regard, it has been in the past 40 years where a special effort has been made to "control" formally specialized sublanguages, resulting in the so-called CLs. One of the main differences between sublanguages and CLs is that restrictions in the former occur naturally, while restrictions in a CL are imposed "artificially" by an author or a group of linguists.

CLs are thus artificially created by defining a set of grammatical, stylistic and lexical restrictions, resulting in advantages compared to the use of a natural language. Authors have to consider these rules when writing, though ideally, automatic-checking tools should support them in this task. Indeed, special tools have been developed to assist the author in the writing process by indicating him which is the right terminology and which are the grammatical structures he should avoid or favour. Examples of these tools and a more detailed description of how they work will be given in Chapter 2 (2.3).

Sometimes, as I will see in AECMA, the term Simplified Language is used instead of CL (Coulombe, Doll, & Drouin, 2005), since the goal of applying it is to obtain a text which is easier to read and to understand. Implementing the rules of a CL, however, does not always mean simplification in the sense of shortening the number of words or sentences, as this example shows when applying the rule 7.3. of AECMA STE: If necessary, add a brief explanation to a warning or a caution to give a clear idea of the possible risk[18]:

Non-SE sentence	THE GRABBER MUST BE ENGAGED BEFORE THE THRUST REVERSER HALVES ARE OPENED.
	BEFORE YOU OPEN THE THRUST REVERSER

SE-compliant sentence	HALVES, MAKE SURE THAT YOU ENGAGE THE GRABBER. IF THE GRABBER IS NOT ENGAGED, DAMAGE TO THE PYLON STRUCTURE CAN OCCUR.

This is the reason why I prefer the term Controlled Language (CL), which I will be using throughout this work.

CLs could be considered as a subset of sublanguages, since they aim at applying restrictions in the language of a definite domain, that is, a sublanguage. Indeed Van der Eick, de Koning, & Van der Steen (1996: 64-65) use the term *controlled sublanguage* and define it as follows: "Controlled sublanguages are derived variants of sublanguages, constructed to impose precise coverage bounds and application-specific additional Restrains such as ambiguity reductions." However, other literature references define CLs as subsets or, better to say, varieties of the standard language (Nyberg, Mitamura, & Hujisen, 2003; Lehrndorfer, 1996). Hujisen (1998: 2) even goes further and defines CLs as part of the natural language with the following definition: "A CL is an explicitly defined restriction of a natural language that specifies Restrains on lexicon, grammar and style". Indeed, all these approaches are somehow right: CLs aim at controlling the vocabulary and syntactic structures of sublanguages, but not only. They also include vocabulary and syntactic structures of the standard language in order to avoid the deviant constructions and the omission of information characteristic to sublanguages that can cause understanding and language processing problems.

Schwitter (1998: 57) presents a graphical representation of controlled languages within the general theory of language, which I interpret here. In this image I see how the standard language is a subsystem of the general natural language used by human beings. Sublanguages are then formed by a subset of the vocabulary and structures of the standard language, but also by a subset of deviant constructions and vocabulary that forms part of the more wide set of natural language (or universal set). CLs are formed by elements of sublanguages, since they restrict them in grammar and vocabulary, but at

the same time they also contain other elements of standard language, trying to avoid the deviant constructions present in the sublanguage that are part from the natural language.

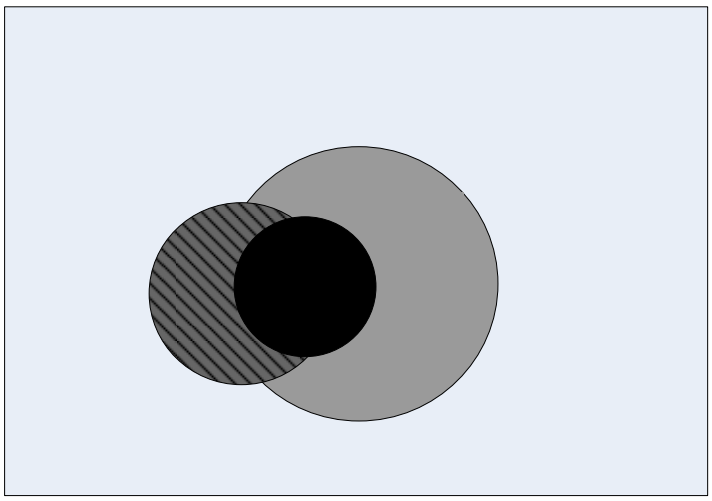

Figure 1: Natural Language, Sublanguages and CLs

I can thus conclude that, generally, I can define CLs as set of elements of language used for specific purposes or domains, aimed at a definite group, with restricted vocabulary, grammar and style. These CLs are primarily used for the authoring of technical documentation. From a more formal point of view and in the same terms as Harris, having in mind the mathematical conception of *sub-* as a part of, I could define a CL as the subsystem containing elements from a sublanguage and from the standard language, being the properties of those elements a restricted grammar and a controlled vocabulary.

1.4.2 Advantages and Disadvantages of CLs

It is generally claimed that the implementation of CLs make the manipulation of texts both for humans and machines much easier. This is achieved by reducing the lexical ambiguity (avoiding homonyms and synonyms as much as possible) and simplifying the syntactic structures in order to improve the comprehensibility and readability of the text. This, in turn, smoothes the processing of the text in any step of the documentation process, from writing to translating, be it by humans or machines.

Especially for big companies dealing with big amounts of documentation, which contain high-precision technical information, it is essential to avoid misunderstandings both for end users and translators. A mistake in the translation or a misunderstanding in the workshop can cause enormous costs due to accidents and other liability risks.

Besides, the uniformity that those texts written in a CL present in word choice, terminology and style also improves the corporate image of the company, which appears coherent and consistent to the customer. The consistency of texts also improves the reusability of the source text. If authors use a standard terminology and sentence structures, the same facts are always expressed in the same way, allowing a text written for a certain purpose to be reused elsewhere if appropriate.

This is also especially important when using translation aids such as translation memory tools. Since these tools detect the repetitions of source text, the more consistent my texts are, the more effective will be the use of these tools and, thus, the more benefits will be achieved. This reduces both the costs of authoring new documentation and the translation costs, because facts are always expressed and translated in the same way.

From the computational point of view, one of the most claimed advantages of CLs is the improvement of machine translatability of texts. It is generally accepted that the restrictions of CLs help to attain a better quality in the translation. However, this is not always the case and it depends on a great deal on the type of texts and the characteristics of the field.

Although the advantages of the use of CLs seem to be obvious, the application of these restrictions has also important drawbacks when creating content. First, the authoring task might become more time-consuming, since the author has to check that what he has written conforms to the rules of the CL. Before language checkers existed, this posed a bigger problem, since this checking had to be realised manually, and authors had to verify their writing every time a doubt might arise. This has been mostly solved by the introduction of automatic language checkers. However, even with the help of language checkers, tasks such as rewriting a sentence that does not conform to the rules of the CL

can be more time-consuming than simply substituting a term that has been considered as deprecated by the terminologist. Nyberg et al. (2003: 249) present the following example from the AECMA Simplified English (SE). The rules of this CL disapprove the use of the phrase *according to* and one is advised to use the verb *refer to* instead. The use of (1a) is this disapproved, and could be rewritten to SE as (1b).

a. Calibrate test set according to manufacturer's instructions.

b. To calibrate the test set, refer to the manufacturer's instructions.

As we can see, the authoring task becomes more complex[19]. Besides, the author might feel limited in his capacity of expression, feeling obliged to express himself according to some rules that are not part of his writing style. Due to these issues, the introduction of CL in an institution or company may encounter some resistance from technical authors and translators. It is therefore recommended that authors and translators are involved in all stages of the creation and deployment process of a CL, so that they have the opportunity to give their input in the language definition process, as well as participating in its introduction and evaluation.

Another aspect which needs to be considered is the integration of a CL into an existing authoring process. Technical issues such as APIs and interaction with other tools might seem trivial, but in some cases can cause real problems. The introduction of a CL checking implies the interpolation of new phases during the process. That usually means an increase in the authoring time, but also a reduction or even the elimination of revision.

The potential initial cost of developing a CL might also be a critical issue. While there are general CLs that can be adapted to the needs of a particular organization, other organizations need to develop a new CL that fits their needs. Depending on the text production volume this might be cost-effective or not. Therefore, before embarking on a

full-scale development, it is useful to complete an initial feasibility study to evaluate the benefits and costs in a particular customer scenario (Mitamura & Nyberg, 2001).

For a company or organization all these advantages and drawbacks have to be clear. Usually, the introduction of a CL involves substantial investment. To summarize, before making this decision, following points have to be considered:

Which is the aim of the CL? The design and the potential benefits of the CL will depend on if it is intended only for humans or also for machine purposes.

If a CL checker is going to be used in order to automate the authoring process, the question of licensing an existing product or the design and development of a new product has to be considered. The second option requires a bigger initial investment, but in this way it is guaranteed that checker strictly complies with the rules of the CL.

CL must be part of an on-going process. It is not enough with the design and deployment of CL and a CL checker. It must be maintained. This implies that new terminology has to be added, new standards taken into account and new structures adapted to the CL. For all these tasks new roles in the authoring processes are needed.

In general, I need to bear in mind that though the initial introduction of a CL can be difficult and expensive, the benefits are mostly only seen on the long-term

1.4.3 CL Classification

1.4.3.1 Human-oriented and Machine-oriented CLs

Generally, the use of CLs has a twofold purpose: on one hand, to improve readability and understandability of technical documentation, particularly for non-native speakers, expecting to reduce translation costs in this way by providing the markets with clear documentation written in the language of the manufacturer. These are known as Human-Oriented Controlled Languages (HOCL) (Hujisen, 1998a). Examples of this

approach are CFE (Caterpillar Fundamental English) (Lockwood, 2000) and AECMA Simplified English, characterized by a higher number of stylistic, text structure and pragmatic rules (O'Brien, 2003a). This would correlate the modular approach presented by Lehrndorfer (1996: 13), focusing on the optimization of the readability, understandability and, eventually, the translatability of technical documentation.

Bernth (1998) makes a further distinction between controlled languages: she distinguishes those that are designed to be more intelligible, also for non-native speakers, but not necessarily to be translated. Indeed, she cites as examples of this type of CL AECMA and GIFAS Rationalized French. However, this seems to contradict what the authors of GIFAS claim when they state that one of the main goals of the development of GIFAS was "to enable authors to write in a French that is easily translatable into SE" (Barthe, 1998).

On the other hand, Machine-Oriented Controlled Languages (MOCLs) have been developed to assist Natural Language Processing (NLP) in applications such as MT or Information Retrieval (IR). Lehrndorfer defines this approach as the machine efficient method, the goal of which is to make technical documents more efficient for the implementation of Machine Translation or any other automatic handling of the text, such as parsing, information retrieval or data mining (Lehrndorfer, 1996:13-16). Bernth (1998) speaks about CLs that are meant for translation, often by a MT system, and includes General Motors CASL (Means & Godden, 1996), ScaniaSwedish (Almqvist & Sångall Hein, 1996), and the KANT system (Mitamura & Nyberg, 1995). Other examples are Controlled English at Alcatel Telecom of Belgium (Goyvaerts, 1996) and EasyEnglish, the guidelines of which having been published by IBM20 for the design and writing of content for the Web that will be enabled for MT by the IBM WebSphere Translation Server for Multiplatforms (WTS). Both approaches have many aspects in common, since many rules and restrictions contribute to increase both human and computer comprehension. Indeed, O'Brien (2005: 6) coins the term "Dual-Oriented Controlled Language" (DOCL) for CLs that should be destined for both human- and machine-processing. However, some important remarks must be done. Not all rules improving "readability", for example, from a human-point of view, help computers and

vice versa. For instance, the AECMA rule "dependent clauses that express a condition on the action in the main clause must precede the main clause" helps humans to understand, but does not make a sentence easier for the computer to process. Conversely, there are writing rules that are of greater benefit to computational processing, such as for example a restriction on the use of pronouns. This can lead sometimes to repetitions, which human translators reject and end up changing or using pronouns instead (Nyberg et al., 2003: 74;100), but can help enormously to reduce ambiguity in certain constructions. Reuther (2003) also tries to distinguish those rules that improve translatability (T-Rules) from those that improve readability (R-Rules), stating that T-Rules are commonly more restrictive. Though when fulfilling R-Rules a better impact is obtained on translatability, it is difficult to state the contrary. In the following table, adapted from Reuther (op. cit., 2003), the goals and applications of HOCL and MOCL are summarized:

Goals of CLs	Readability and Understandability	Translatability
HOCL	Human reader: More clarity Consistency	Human translator: Lack of ambiguity
MOCL	Automated language processing systems (monolingual): Controlled language checking Information retrieval Parsing Data mining	Automated translation systems (multilingual): Translation Memories (CAT Tools) MT systems

Table 1: Differences between HOCLs and MOCLs

From the procedural point of view, I can represent the deployment of HOCL and MOCL in this way (adapted from Mitamura, 2007).

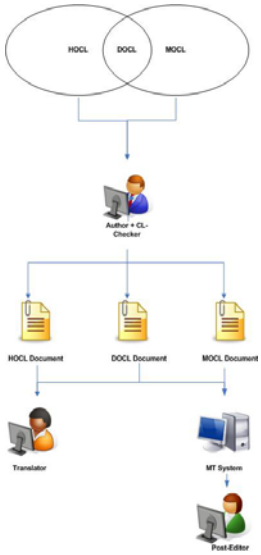

Figure 2: Document creation process with HOCL and MOCL

According to this graphic, HOCL, DOCL and MOCL can be implemented by an author when writing technical documentation, be it with the aid of a CL checker of without, though as I will see in the following chapter the inclusion of this type of tool in the document creation process has become so common and necessary that nowadays is almost unthinkable to implement a CL of any type without the aid of automated checking. The document that is created can be a document intended for human translation (normally created by implementing a HOCL), a document that can be translated either by humans or by a MT system (DOCL), or a document which is adapted to the requirements of automated translation. Though as I have seen the limits between these distinctions are somehow blurry, it is always recommended to bear in mind what are the goals of implementing a CL and what kind of translation process I am going to roll out when designing or introducing a CL in the authoring process, since the rule set will depend on this, being some rules more adequate than others for a certain purpose. Finally, the document can be translated by a human translator or by a MT system. In this case it will probably need light post-editing to attain quality publication.

1.4.3.2 Other Classification Criteria

Though the differentiation between HOCL and MOCL has been generally accepted by the scientific community, CLs can be classified according to other criteria. Hujisen

(1998b: 33-34) classifies CLs regarding their relationship to MT. He distinguishes between loosely controlled languages, such as PACE (see 2.2.1) which aim at improving the quality of the source-language text in order to facilitate subsequent translation by humans or machines, and strictly controlled languages, which are controlled languages with a formally specified syntax, thus constituting an interesting point of departure for automatic translation. An example of strictly controlled languages would be the work at Cap Gemini's Lingware Services (de Koning, 1996; Van der Eick et al., 1996).

Gavieiro-Villatte & Spaggiari (1999) conducted research in order to build an open-ended overview of CLs. To do so, they first divided CLs in two categories: CLs of restricted domain and CLs of grammar, basing their assumptions in the theory postulated by Harris. They define "restricted domain CLs" as "[a sublanguage] composed of sentences which deal with more or less closed subject matter –one of limited vocabulary is used and in which the occurrence of other words is rare" (Harris, 1991). On the other side, "grammar CLs" can be considered as "[a sublanguage] composed of sentences which satisfy certain grammatical conditions that are not satisfied by all other sentences of the language" (Harris, 1991). These definitions, however, are more appropriate for sublanguage, as seen in 1.3. Besides, a sublanguage is always characterized by both a limited domain and a restricted grammar, being these restrictions natural, contrarily to CLs, where restrictions are artificial.

There are notwithstanding CLs which can be classified as general, such as BASIC English, which has served as a base for other domain-specific languages such as CFE (Caterpillar Fundamental English) or MCE (Multinational Customized English) at Rank Xerox.

Further, they also offer a classification of CLs according to their goals: if they are designed to be used in writing guides, if their goal is to improve their performance in computer-oriented applications or if their use should be supported by the implementation of writing-guides. This classification, however, corresponds to the human-oriented and machine oriented CLs classification seen before. Writing guides

would be human-oriented CLs, aimed at producing standardized texts. Computer-oriented CLs would comprehend those that are created with MT in mind, while the implementation of writing-guides refers to CL checking. These two would therefore correspond to the machine-oriented approach.

Further, Allen (1999) distinguishes between two types of CLs:

- Limited vocabulary CLs: here the emphasis is placed on creating a core of lexical items that can be used throughout the document. There are some general writing rules, though its strict enforcement is not usual. The main goals are adherence to vocabulary and overall grammatical correctness.

- Extended vocabulary CL grammar conformance checkers: thanks to a checker, a set of constrained syntactic constructions are controlled. Besides, technical terminology is also automatically checked. The main goal here is the use of a standardized terminology and a controlled syntax and style.

This classification, based on the size of vocabulary, would originally correspond to HOCL, which aim at writing texts with a limited vocabulary in a simplified and correct style, and MOCL, the aim of which is to control the terminology, the grammar and style of texts for better processing (be it human or automatic).

More recently Pool (2006) has established a classification distinguishing between formalistic (a language-like formal notation) or naturalistic CL (a natural language with restrictions) on the one hand, and domain-specific or genre-specific (which overtly or apparently aim for expressivity in a domain or genre) and general[21] (which aim at languages for multiple domains an genres), on the other hand, as Gavieiro-Villatte & Spaggiari (1999) had already done. Tough the author establishes a parallelism between formalistic-MOCL and naturalistic-HOCL, this is not necessarily convenient, since both MOCL and HOCL can be natural languages. In any case, I could consider formalistic

languages as a subset of MOCL. In the following table I can see many of the CLs I will analyse in 2.2.1, classified according to the criteria explained above:

	Restrictive	General
Formalistic	CELT ClearTalk CLIP Common Logic Controlled English MenuChoice PENG	Attempto Controlled English CPL E2V First Order English Formalized English

	Restrictive	General
Naturalistic	Airbus Warning Language ALCOGRAM ASD Simplified Technical English Avaya Controlled English Controlled Automotive Service Language Controlled English (Océ) Ericsson English FAA Air Traffic Control Phraseology Français rationalisé PoliceSpeak ScaniaSwedish SeaSpeak Simplified Technical Spanish Sun Proof TITUS Webtran	Controlled Chinese Controlled Modern Greek DLT Intermediate Language EasyEnglish EasyEnglishAnalyzer interNOSTRUM Controlled Spanish KANT Controlled English MULTILINT Multinational Customized English Perkins Approved Clear English Plain Japanese Siemens-Dokumentationsdeutsch Simplus Universal Translation Language

Table 2: CL Classification

According to this classification, I am mainly interested in naturalistic controlled languages and especially those which are domain or genre-restricted, taking into account, though, that those that are considered as general can also be extended (with specific terminology and writing rules) to work in a specific domain, such as I will see later with MULTILINT.

1.4.3.3 Final remarks on CL Classification

Classifications are important to determine common characteristics of CLs that distinguish them from other types of languages.

As it has been outlined, there are different valid criteria by which CLs can be classified and though the classification of CLs into HOCL and MOCL has been widely accepted, other criteria such as the scope of the language (domain specific vs. general), the language orientation (monolingual-oriented versus multilingual-oriented CLs) , the degree of restriction (loose versus strict CLs) or the size of the vocabulary can be taken into account. All these classifications let space for further subcategorizations, too. For instance, I have observed that some MOCL could be subdivided into formal notations for ontology creation or software specifications and natural languages intended for better automatic processing (MT, information retrieval etc.). Domain specific languages could also be further subdivided into the different domains CLs tend to cover: aviation (e.g. AECMA, BTE), heavy-equipment machinery (Caterpillar), automotive (CASL, ScaniaSwedish), software industry (IBM EasyEnglish Language, Océ Technologies Controlled English) etc.

1.4.4 Areas of control

Different authors divide the areas of control of CLs in different categories[22]. I will use the more general classification by Mitamura (1999) that states that use of controlled language falls into two broad categories: lexical and grammatical control. Lexical or vocabulary control is the most common type of control and a key element in controlling the source language by restricting the authoring of texts so that only pre-defined and validated vocabulary is accepted. In this way, incoherencies and understanding problems are avoided.

Grammatical control is broader and can be subdivided in different categories such as syntax (sentence structure), morphology, orthography and style (pragmatics). The degree to which this features can be controlled varies from language to language.

1.4.4.1 Lexical Control

Lexical Restrains aim at reducing ambiguity of the source text through a restricted lexicon. If homonymy and synonymy are reduced readability, consistency and comprehensibility may improve. Further, if a text is written using standard terminology

and sentence structures, a uniformity of style is achieved and text can be reused more effectively in technologies such as translation memories or machine translation. However, restricting vocabulary and grammar must not hinder the expressiveness of a CL. A limited vocabulary only does not necessarily imply a reduction of input sentences complexity; indeed, it can make authors to write longer, convoluted sentences to express complicated meanings if sufficient terminology is not available. Therefore, the balance between vocabulary size and input complexity is very important for successful CL deployment (Mitamura & Nyberg, 2001).

Lexical control comprises two levels: on one hand, the general or basic vocabulary, that Lehrndorfer calls lexical minima (Lehrndorfer, 1996: 134). On the other hand, the lexicon is made up of the specific terminology of a domain. Therefore, I can distinguish between lexical control and terminology control. I can either store both sets in the same system (a Terminology Management Tool), or establish different control mechanisms for each of them. For instance, specific terminology can be stored in a database while more general vocabulary (usually semi-technical terms) can be supported by specific rules.

Lehrndorfer considers the lexicon as a weak point within the specification of a CL due to the cognitive difficulties implied by its learnability and implementability by the authors. However, though these considerations are justified from a theoretical point of view, the development of automatic checkers leaves little doubt that lexical control is a necessary and applicable process within language control.

The first step when building up a controlled vocabulary and a controlled terminology is to analyse a big corpus of pre-existing documents in order to define a starting status of vocabulary. There are different methods for analysing a corpus and extracting terminology. Usually, however, the steps comprise tokenization, stemming and creation of a word and phrase list. This initial vocabulary will be then further refined with the help of a human terminologist in charge of organising these words in concepts, coding the different entries and defining closed classes (word types –multiword, abbreviation, acronym etc.) to which the concepts belong. Once the domain vocabulary is fixed and,

if necessary, divided between specific terminology and semi-technical vocabulary, an ongoing process is necessary in order to capture the new terms arising in new texts created by authors. These terms will be then organised in the existing concepts or new concepts will be created in glossaries where the preferred terms and the preferred vocabulary are recorded together with deprecated terms and variants, all cross-referenced, so that an author writing a text can be easily referenced to the right term when he writes a wrong one.

Obviously, the design and degree of lexical control will depend on the aim and type of the controlled language: terms will be encoded and stored differently depending on their final purpose (controlled authoring, monolingual language processing, human translation, MT etc.). For instance, for MOCLs, the lexicon may include with each lexical entry many other pieces of information that are needed for the computational processing of the text or for administrative purposes such as detailed information on the syntactic properties and semantic categories of the word at issue, and the date of creation or latest modification of the relevant lexical entry (Hujisen,1998b: 24).

In any case, when defining a controlled vocabulary, there is a guiding principle that rules over all the other principles and guidelines: the univocity principle, or "one word per meaning". It must be pursued that, for every concept or meaning, there is only one denomination. From the terminology theory I know that the combination of a concept and a denomination forms a term. Therefore, the main goal is to obtain univocal terms in my lexicon. This is valid both for the basic vocabulary and the specific technical terms and it is the best way to avoid ambiguity.

However, this is not always as straightforward as it might sound. Though the univocity principle might help to avoid ambiguity, it also implies an increase in the lexicon size, since I will need a different denomination for every concept. Furthermore, sometimes synonyms or variants are unavoidable, or a word that is coded as a deprecated term in a concept, might be coded as a preferred term in another concept. For instance, the term *Abdeckkappe* (*tapa* in Spanish) is correctly used as a preferred term when talking about body equipment, seats or wheel and tires. However if we are talking about the front

axle, the right term will be *Verschlussklappe* (*caperuza de cierre* in Spanish), and *Abdeckklappe* (*tapa*) will be a deprecated term. Another case might be following: if I have a defined CL for a domain but for different types of texts, I might need different denominations for the same concepts depending on the type of text. For instance, a repair manual in the automotive industry might use the term *Rückspiegel außen* for literature intended for the workshop, while literature intended for the end clients will use the term *Außenrückspiegel*. These terms, in turn, might have a different or the same translation in the target languages, such as for instance Spanish, where both terms would be translated as *retrovisor exterior*. As we can see, problems can be numerous and the more complex the sublanguage I am dealing with, the more complex it will be to control its terminology.

To overcome these hurdles, different possibilities exist. For instance, in order to avoid ambiguity, the first step might consist of limiting the part of speech of every denomination that can only be noun, or verb, or adjective. This might not be relevant in all languages: for instance, the number of words in English that share the same form but can belong to different grammatical categories is much larger than in Spanish. Another method consists of restricting the valency of verbs to the subcategorizations that are sensible in the domain.

A way of delimiting words semantically is to assign them a semantic field. For instance, the term *Stuhl* in German usually means a seat with a back on which a person sits, usually having four legs and often having arms. Therefore, I will usually translate *Stuhl* as *chair* in English and *silla* in Spanish. However, if I am talking about bowel cancer and I encounter the word *Stuhl*, it might probably be translated as *stool* in English and *heces* in Spanish, and not as *chair* or *silla*. Therefore, assigning semantic information to the terms might be crucial, especially when dealing with MT. This semantic information can be added to the dictionary but can also be introduced in form of mark-up language interactively while the author writes, as it was done in the KANT Project (Mitamura & Nyberg, 1995).

It is also important to notice that in order to process in an appropriate way all the words that appear in a text, it will not be sufficient by encoding full terms. In technical texts I will encounter other types of terms such as:

Multiwords: These are terms that are formed by more than one word. The meaning and syntactical behaviour of these terms cannot be usually derived from the meaning of its single components, hence it is advisable to encode them as a single unit. Such multiwords include noun phrases such as *distributor-type ignition system* or *tread wear pattern,* where the word *pattern* has no separate domain meaning. Phrasal verbs or verb-particle constructions are also easier to analyze if taken as a unit.

Technical Words: Technical terminology is not only made of full single terms. As I have seen, there can be multiwords that need to be treated separately. Besides, I can find other types of terms such as acronyms and abbreviations. Depending on the technical domain I am dealing with, new types of lexical items might be needed, such as wire colours or controller identification codes. It is important to encode all these terms in the right way so that a CL checker can analyse them correctly.

Technical Symbols: Technical texts are characterized by the special use of numbers, numerals, units of measure, letters of the alphabet, etc. All these uses must be encoded so that a correct analysis is possible.

Reuther (2003) also points out certain aspects that are important when designing the vocabulary and terminology of a CL, especially if it is going to be processed automatically or if it is going to be the base of MT. Though these aspects might vary from language to language, they represent a general approach that is worth considering.

Spelling variants: while these might not affect human processing, they can have dramatic consequences for a MT system. For instance, the use of hyphen versions such as *low beam headlight* and *low-beam headlight.* If only one of the versions is stored in

the dictionary of my system, the other one will not be recognised and, possibly, not translated or wrongly translated.

Morphological variants: as well as spelling variants, this might not hinder a human reader from understanding the text, but a translator might doubt if it is the same concept or not. For instance, *reinforcing plate* versus *reinforcement plate*.

Synonym variants such as *interior mirror* and *inside mirror* can also cause understanding problems for humans and for machines, since it is not possible to discern if they refer to the same concept or if they designate two different pieces.

There are many other aspects that might vary depending on the natural language to be controlled. For instance, in English it might be reasonable to standardize the meaning of modal verbs or the use of participial forms. However, I consider that the above mentioned guidelines cover the main necessary aspects to create a consistent and robust terminology.

1.4.4.2 Grammar Control

Though lexical control contributes greatly to gain terminological consistency in texts, it is also necessary to control the constructions used when authoring technical documentation.

Indeed, these are already somehow restricted since they belong to a certain sublanguage. However, sometimes these restrictions are not enough, and sometimes these restrictions are not the most appropriate to improve the readability and translatability of the texts. In these cases, it will be necessary to apply certain writing rules to improve and, especially to standardize the language used in the texts.

The KANT project defines two general types of grammar restrictions: those on the phrase-level, made to avoid the formation of complex phrases and those on sentence-level, made to avoid ambiguous sentence structures (Mitamura, Baker, Nyberg, &

Svoboda, 2003;(Mitamura & Nyberg, 1995) Mitamura, 1999). These are divided as follows:

Phrase-level Restrains

- Verb + particles. Particles can be prepositions, adverbs of other parts of speech. It is desirable, when possible, to use single-word verbs instead.
- Coordination of verb phrases should be avoided to prevent the ambiguity of arguments and modifiers.
- Conjoined prepositional phrases should be made explicit by repeating the preposition.
- Determiners should be used in noun phrases.
- Nominal compounding should be avoided to reduce ambiguity.
- Quantifiers and partitives should be made explicit and it must be clear which nominal head they are modifying.

Sentence-level Restrains

- Coordinated sentences should be of the same type.
- Subordinate sentences must contain a subject and a verb, so that they are able to stand on its own if the conjunction is removed.
- In general, ellipsis should be avoided. If there are any necessary elliptical constructions, these should be defined in the controlled language as a “closed class”.
- Relative clauses should always be introduced by the relative pronouns that or which.
- The use of WH-questions should be avoided.
- Rules for consistent and unambiguous use of punctuation should be specified in the controlled language.

Apart from these Restrains that are applied as rules, Mitamura & Nyberg (1995) also discuss the disambiguation using SGML tags in order to indicate the desired choice among ambiguous structures. In general, grammar restrains can be summarized as follows:

- Avoid every construction that might result ambiguous by avoiding prepositions that can modify the meaning of the verb.
- Avoid the coordination of verb, prepositional phrases or phrases with any other complements that might result ambiguous (for instance, partitives). If sentences are conjoined, these should be of the same type.
- Limiting the number of conjunctions since they might increase the potential ambiguity of syntactic analysis.
- Express meaning more precisely by using determining adjectives to make the referential nature of the noun they modify more precise.
- Avoid anaphora in the form of any kind of pronouns, quantifiers, reflexives and partitives.
- Avoid elliptical constructions so that neither humans nor machines have to reconstruct the missing elements.
- Avoid complexity in sentences by keeping the sentences as simple as possible. This, apart from improving readability, will grow the number of Translation Memory leverage.
- Maintaining the standard order of elements within a phrase or sentence.
- Following general stylistic recommendations with regards, for instance, to the use of passive, the use of future tense, the use of negation, the use of personal references in texts etc.

The more concrete definition of rules derived from these general Restrains will depend on every language. For instance, in German it will be necessary to define rules in order to control the case governed by a certain preposition. Indeed, rules are not as easy to transfer from one language to the other, as Kathleen Barthe (1998) describes when

defining controlled French for the aerospace industry. Another example for German is presented by Hernandez & Rascu (2004: 75), that explain how the tool MULTILINT works. Here the grammar checking component detects misplaced commas, capitalisation mistakes, misspelled compounds, agreement mistakes, misplaced relative clauses, word repetition, etc.

Besides, general grammar rules might be necessary too in order to detect ungrammatical structures. Even though one should not expect ungrammatical structures in texts written by native speakers, corpus studies show that these are far more than common. Indeed, as Bernth (2006) states: "Controlled Languages have been invented to solve the problems associated with readability and translatability, with slight regard to ensuring grammaticality".

1.4.4.3 Style

Style includes all aspects that cannot be defined either by means of lexical Restrains or grammatical rules, identifying phrase structures that are ambiguous or difficult to understand. Examples of style can be punctuation and layout rules, parentheses, slashes etc. Rascu (2006) also points out that company specific stylistic requirements can also be included under the style checking component. These can be newly defined or found in company style guides or writing rules for technical documentation. These usually include both requirements characteristic of technical writing in general and company specific regulations. Some examples of stylistic requirements can be: the avoidance of jargon, the avoidance of complex, conjoined sentences, the use of positive constructions or the use of parallel structures. For instance, the style module of MULTILINT/CLAT addresses following issues: layout, lexical problems, ambiguity, ellipsis, complexity, order of sentence constituents as well as other stylistic problems.

Style would correspond to textual rules (text structure and pragmatics) in the classification by O'Brien (2003) which comprises rules that control aspects such as sentence length, punctuation and verb form usage. Bernth (2006) points out that CLs usually neglect rules that affect paragraph length and structure, being the sentence most

of the time the maximum scope of analysis. As requirement for style she includes "good rhythm", which is treated as a requirement for proper sentence variation[23]. However, this is not always possible in certain text types, which are usually required to be written in purely narrative format (no questions, no fragments etc.).

1.5 Summary and final remarks

In this chapter I have attempted to delimit the concept of Controlled Languages in contrast with other concepts such as natural languages and sublanguages. After defining the different concepts of natural languages and sublanguages with respect to controlled languages, I have dealt with different aspects of CLs. Starting with a definition, the advantages and disadvantages of using such constructs have been exposed. Subsequently, different ways of classifying CLs have been presented. Finally, the different areas of control that CLs usually cover have been presented.

2 CONTROLLED LANGUAGES IN INDUSTRIAL ENVIRONMENTS

Language is by its very nature a communal thing; that is, it expresses never the exact thing but a compromise - that which is common to you, me, and everybody.

Thomas Earnest Hulme, *Speculations*, 1923

2.1 Introduction

As we have seen in the previous chapter, CLs experienced a great development thanks to their adoption in industrial contexts. First, CLs were mainly developed for human purposes, that is, to improve the readability and understandability of texts. However, as natural language processing applications gained more popularity (data mining, information retrieval, MT), rules oriented to automatic processing started to being developed. Further on, the development of computational linguistics gave place to the development of automatic language checkers that could indicate the authors if their texts complied with the CL specification.

At this point it is necessary to distinguish between CL checkers and general language checkers, which many text processing systems are equipped with. The latter generally offer spell and grammar checking, such as MS Word orthography facility. Though it is possible to create domain specific dictionaries, these tools only check if the word is correctly written with respect to the dictionary, and not if terminology is consistently used or if there are terms that should not be used and which would be the right one.

CL checker software is designed to assist the writer in the authoring phase (also defined, especially in the MT context, as pre-editing phase) to write the text according to certain predefined syntactic and style rules as well as with a validated and, as much as possible, unambiguous vocabulary and terminology of a certain domain. Further, some of them

even offer rewriting features that provide the author with suggestions of how the text should be modified to comply with the rules of the CL. This pre-editing effort enables texts to be processed by other natural language processing systems with more accuracy. Usually, their assistance consists of a series of indications with respect to the language used in the texts: for instance, if the right terminology has been used or if the sentence structure conforms to the rules. If not, a recommendation will be given so that the author can correct the sentence so that it complies with the CL specification.

In this chapter, I will discuss some of the issues related to CL in industrial contexts. First an overview of some of the most remarkable examples of CLs in research and industry is given. Subsequently, I will deal with the issue of CL checking: technical aspects that are challenging when implementing such tools will be discussed, as well as which approaches can be considered to solve these hurdles (2.3). After that, a short overview of the different techniques for giving feedback to the author will be reviewed (2.3.3). Section 2.4 will cover some aspects of the use of CL checking in the authoring process, focusing especially on issues such as the deployment of a CL checker among authors, the maintenance of the checker and the interaction of a CL checker with other tools, especially with MT systems, within the multilingual documentation creation process. After presenting a survey of CL checkers that have been developed either as a result of a research project or as a commercial product (2.5), a more detailed insight into MULTILINT/CLAT will be given (2.6), to end up with some final conclusions about this issue.

2.2 Controlled Language Examples: Initiatives in Research and Industry

Most research and industry initiatives in the field of CLs have been carried out for the English language, though other languages where efforts in this respect have been made include Swedish, French, Greek, Spanish, Japanese, Chinese and German. In this section I will concentrate on English, as the auspices of the research initiatives, and German for the interest of this work. Nevertheless, examples of controlled languages in other languages will also tackled.

There have been, as far as my knowledge goes, a few attempts of gathering and classifying the CLs that have been developed since Odgen created BASIC English in 1930s. One was carried out by Scheurs & Adriaens (1992), who examined the roots of CLs in their article. Gavieiro-Villatte & Spaggiari (1999) undertook the realisation of a database of CLs as a pre-work of a PhD research, though the complete database is nowhere available. Another contribution has been made by Pool (2006) that reviews different available CLs to evaluate them for knowledge representation purposes. Based on these two collections, I go on to expose the most meaningful CLs that have been developed throughout the last decades. An overview of the existing CLs to date can be seen in 0[24]. This overview resembles the cards designed by Gavieiro-Villatte & Spaggiari (1999) intended to be a work aid for further research.

2.2.1 Controlled Languages for English

The first attempt[25] in this direction for the English language was called BASIC, which stands for "Basic American Scientific International Commercial" and was developed in the 30s by Charles K. Odgen, an English man who hit upon the idea of an 850-word Simplified English vocabulary. The goals of this "experiment" were to simplify the English vocabulary both to facilitate the communication for scientific and commercial purposes, and to make it easier to learn to give everyone a second, international language[26]. However, BASIC English has never been widely used for the purposes it was developed for.

Some years later, industry also noticed the need to control the language in the creation of technical texts due to the always increasing amount of documentation and its internationalization. Besides, the use of a CL was seen as a competitive edge, since products that are easier to operate and service are usually more prone to be successful. Starting in the late 60s and until nowadays, with a boom during the 90s, more and more companies have turned to CLs to make their documents easier to understand and to optimize their document production and translation processes.

In 1970, Caterpillar Inc., in Peoria, Illinois, created an initially 800-word vocabulary based on BASIC English and limited to a specific discourse domain. The result was CFE (Caterpillar Fundamental English), which included, apart from the BASIC vocabulary, technical terminology, in which each term had only a univocal defined meaning. This CL was intended as a form of English as a Second Language for non-English speakers, who would be able to read the service manuals written in CFE after some basic training and would therefore eliminate the need to translate. It was thus conceived as a HOCL. This language was used for slightly over ten years (approximately from 1971 to1982). Nowadays Caterpillar does not use CFE anymore, but CTE (Caterpillar Technical English), developed during five years (1991-1997) and successfully used in a combined authoring and translation system for the creation of technical documentation and as source language for MT (Hayes, Maxwel, & Schmandt, 1996: 88 and ff.; Kamprath, Adolphson, Mitamura, & Nyberg, 1998). Though CFE was abandoned in 1982, CFE continued to be used outside Caterpillar. Based on CFE, other CLs were developed, among them: Smart's Plain English Program (PEP), White's International Language of Service and Maintenance (ILSAM) and Case's Clean and Simple English (CASE). Based on ISLAM arose Perkins Approved Clear English (PACE), which consisted primarily of a single wordlist (2500 words in 1990, 10% of them being verbs) plus ten very general writing rules (Newton, 1992: 46-47). PEP gave place in turn to other CL versions for companies such as Clark, Rockwell International and Hyster (Hyster's Easy Language Program or HELP). The further development of ILSAM resulted in the creation MCE (Multinational Customized English), a system developed within the Xerox Corporation involving a controlled vocabulary and a set of writing standards, Ericsson Telecommunications, BSO/DLT (as a formalistic language) and IBM EasyEnglish (Bernth, 1998). Alcatel-Bell also developed COGRAM, based on the specifications of IBM EasyEnglish, AECMA and Ericsson (Scheurs & Adriaens, 1992).

All these CLs were originally developed as HOCLs, since at that time the automatic processing of texts was not as regular as nowadays. Applications such as data mining, MT or terminology extraction were developed subsequently and some of these CLs had to be adapted to the new requirements of MOLCs.

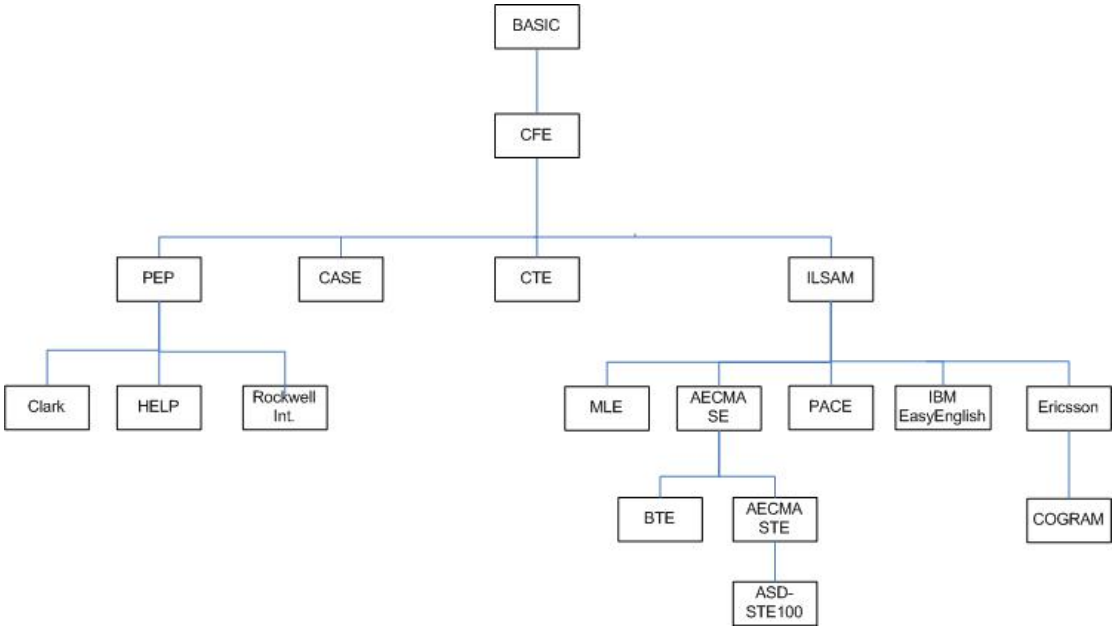

Figure 3: The evolution of industrial CLs

Special mention deserves ASD-STE100 Simplified Technical English due to the fact that usually CLs are developed following the guidelines and idiosyncrasies of a particular company. However, AECMA has been widely accepted and is currently used by all the aerospace industry. It was in 1979 when the Association of European Airlines (AEA) asked the European Association of Aerospace Manufacturers (AECMA) to investigate the readability of maintenance documentation in the civilian aircraft industry. Simplified English (SE) was originally developed at Fokker, primarily by John Kirkman. AECMA asked then the Aerospace Industries Association (AIA) of America to assist in this project. In the subsequent years, Simplified English was modified and developed and finally officially adopted by AECMA for application to technical documentation in the aerospace industry (Hoard, Wojcik, & Holzhauser, 1992). Procedural texts and maintenance manuals were analysed and, at the end, the AECMA Simplified English Guide was released. AECMA Simplified English was primarily a HOCL, since it was developed to help the users of English-language documentation (both native and non-native speakers) in the aerospace sector understand what they read, particularly in multinational programs. In 2004 AECMA merged with the European

Defence Industries Group (EDIG) and the Association of the European Space Industry (EUROSPACE) to form the Aerospace and Defence Industries Association of Europe (ASD). As a result, AECMA Simplified English was renamed to ASD Simplified Technical English or ASD STE[27].

ASD STE is described in a document known as the "Specification ASD-STE100", the Issue 5 having been released in January 2010[28]. ASD-STE100 is a required part of the S1000D documentation specification (for military projects) and is primary used for maintenance manuals, data modules and service bulletins for the commercial and military aircraft. Besides, it has received European Community Trademark, which means the recognition of this standardization work that has been going on now for nearly 30 years. It is currently being used by almost all companies that produce aircraft maintenance procedures, including Aerospatiale Industry, The Boeing Company, British Aerospace, Deutsche Aerospace, Fokker, General Electric, Lockheed, McDonnell Douglas, and Pratt & Whitney, Airbus SAS, BAE SYSTEMS, Cobham Plc, Dassault Aviation, Diehl Avionik Systeme, EADS, EADS CASA, Finmeccanica S.p.A., Rheinmetall AG, Rolls-Royce plc, Saab AB, SAFRAN and Thales and Thales Alenia Space.

ASD STE was conceived as a pure HOCL. The SE standard consists of a core vocabulary and a set of writing rules that govern grammar and style. There are also guidelines for company-defined technical vocabulary. The 1500-word core vocabulary consists of verbs, prepositions, conjunctions, adjectives, adverbs, and nouns. Words approved for the core vocabulary were chosen for their simplicity and commonality with other European languages (Hoard et al., 1992). In most cases, a given word is restricted to one meaning (to reduce lexical ambiguity), and a given meaning is represented in the vocabulary by only one word (to reduce synonymy). For example, "follow" can be used only in the meaning "to come after" and not in the meaning "obey"; and "start" is a legal SE word, but "begin" and "initiate" are not allowed (Polvsen, Underwood, Music, & Neville, 1998).

Though as I will see in 2.5 a number of STE automatic checkers have been developed, there are rules that are difficult if not impossible to check automatically. For instance, the SE guide rule 1.13 in the section Words, or Rules 6.2. and 6.8 in Section Descriptive Writing are nearly impossible to check automatically:

RULE 1.13 Make your instructions as specific as possible

RULE 6.2. Try to vary sentence lengths and constructions to keep the text interesting

RULE 6.8. Present new and complex information slowly

Further, AECMA grammar reveals a remarkable degree of lexical flexibility: "Besides the words in the dictionary, the writer can also use those words which he decides belong to one of two categories: either Technical Names or Manufacturing Processes" (AECMA, 2004: 1-1-1).

An Extension of AECMA SE is represented by Boeing Technical English (BTE) which, contrarily to AECMA SE that is confined to the aerospace industry, aimed at being a general-purpose CL writing standard for technical documents. In their paper, Wojcik, Holmback, & Hoard (1998) explain that BTE consists of a set of rules that meets the demands for clear descriptions of systems and processes. It is remarkable that they already include procedural aspects of CL maintenance: BTE is planned to include not only a base lexicon, but also a process for adding and validating additions to technical vocabulary. Further, it was also conceived to provide users with a thesaurus of alternatives for non-approved vocabulary.

Apart from these two examples, there are other companies and institutions that have developed their own CLs. Specialists in applied linguistics at Wolfson College in Cambridge and in Plymoth as well as specialists in maritime and air traffic communication developed two restricted languages for cross-boarder communication: Airspeak for Air Traffic Control (Civil Aviation Authority, 2006)[29] and Seaspeak, created in 1982-83. PoliceSpeak was developed subsequently during the 90s within the PoliceSpeak project, funded by the British Telecom, the Home Office (Police

Requirements Support Unit, PRSU) and the Kent Conty Council to enable fast and accurate communication with the French counterparts when the Channel Tunnel opened in 1993. Six months after the delivery of the PoliceSpeak results, a successor project with the wider brief of addressing the communications challenges of inter-agency communications was launched. PoliceSpeak is integrated within the LinguaNet system and allows multilingual cross-frontier communication. The INTACOM project worked on much more varied languages, work practices, conventions and plans of the entire range of British and French emergency services working at the Tunnel (fire, ambulance and medical services). However, this project was deemed unfeasible since the operational procedures and language differed too much to establish a standard restricted language, and ended up in a report with recommendations (Johnson, 2000).

Companies from the telecommunication, automotive and heavy machinery areas have developed proprietary CLs, among them Diebold Controlled English (Moore, 2000), Kodak English (Muldoon, 1999), NSE Nortel Standard English, General Motor's Controlled Automotive Service Language (CASL) (Godden, 1998; Means & Godden, 1996), Controlled Language at Alcatel Telecom or COGRAM (Adriaens & Scheurs, 1992; Goyvaerts, 1996; Scheurs & Adriaens, 1992), Avaya Controlled English, Sun Controlled English (O'Brien, 2006), Controlled English at Océ (Cremers, 2003; Cucchiarini, 2002) or the Standard Language at Ford Motor Company (Rychtyckyj, 2002, 2006a).

Another natural general controlled language is the CLOUT™ rule set developed by Uwe Muegge[30] specifically for the purpose of helping authors write source text for subsequent MT. CLOUT stands for Controlled Language Optimized for Uniform Translation. However, no evidence that this language is being implemented in any authoring process is available.

As it was mentioned in 1.4., there are also CLs that have been created not to write technical documentation for human readers, but to cope with current information processing issues. These languages are not used for writing technical documentation

intended for reading or translation, but for knowledge representation (e.g. in form of ontologies), reasoning or symbolic input to multilingual language generation[31].

Despite the great variety of CLs for English that could be devised, one would expect that their content is more or less common, since they aim at reaching the same goals. This is indeed the view exposed by (Tablan, Polajnar, Cunningham, & Bontcheva, 2006) who compared the rule sets of PACE (Pym, 1990) and Bull Controlled English (Lee, 1993) and discovered that the rules of these sets can be encapsulated in more or less ten higher level rules which appear virtually identical. They also considered these rules to be consistent with AECMA SE rules.

A different approach is presented by (O'Brien, 2003) who, after analysing eight different CL rule sets (AECMA SE, Attempto Controlled English, Alcatel's COGRAM, IBM's Easy English, GM's CASL, Océ's Controlled English, Sun Microsystem's Controlled English and Avaya's Controlled English), unveiled that although there is some similarity of rules across same rule sets, there is only one common rule, which is the one limiting the sentence length.

It is likely that both views are somehow true and that reality lies somewhere in the middle. Though all these sets do not probably contain exactly the same rules, at an abstract level I can say that all rules aim at attaining the same goals (e.g. elimination of ambiguity by banning the use of clusters of more than three nouns, use of coherent and univocal terminology, avoidance of certain stylistic constructions etc.). Differences might be due to the different speciality domains and discrepant corporate styles.

2.2.2 Controlled Languages for other languages

With regards to other languages, different controlled natural languages have been developed in the industry or for the industry with the help of academic institutions. These languages are specially tailored to the needs of the company, with a specific terminology and restraining grammar and style rules.

In Sweden, Scania developed, together with the Institute for Linguistics at the University of Uppsala, the controlled language ScaniaSwedish. This language is used for the production of owner's handbooks and as source language for the machine generation of multilingual documents (Almqvist & Hein, 1996: 159 and ff.)

In France, the French aerospace industries associations (GIFAS), which had participated in the development of ACEMA SE since the beginning of that project, decided in 1985 to set up a working group whose prime task was to develop a form of controlled French called *français rationalisé* (Rationalised French) based on AECMA SE. Rationalised French has been under development for approximately 16 years and is currently used by some French manufacturers such as Dassault Aerospace (Barthe, 1996, 1998; Barthe, Bès, Escande, Pinna, & Rodier, 1998; Barthe et al., 1999). Another project, LARA, was triggered at the beginning of 2000 by the French public administration together with the Centre de Linguistique Apliquée (CLA), the Université de Franche-Comté at Besançon, the Dicctionnaires Le Robert and Vivendi Education. The goal of the project was to improve the communication with the citizens (Coulombe et al., 2005: 24) and its results and applications can be downloaded from the site of the Ministère de la fonction publique et de la réforme de l'Etat[32].

Remedios Ruiz, in supervision by Richard F. E. Sutcliffe, at the University of Limerick, studied in her doctoral dissertation the implementation of a Simplified Technical Spanish (STS), developing a set of rules based on AECMA SE and on a corpus of maintenance documents from Construcciones Aeronaúticas Sociedad Anónima (CASA). However, no evidence has been found that this variant of controlled language is being currently used in the industry (Ruiz Cascales, 2002; Ruiz Cascales & Sutcliffe, 2003).

Other languages that have also made efforts in creating controlled languages are Greek (Markantonatou, Vangelis, & Maistros, 2002; Vassiliou, Markantonatou, Maistros, & Karkaletsis, 2003), Italian (Fellet, 2011) or Chinese (Zhang, Zhou, & Yu, 1998).

In Germany, apart from some studies on controlled German as a scientific and pedagogical concept for technical documentation (Lehrndorfer, 1996), no concrete implementation existed until the development of Controlled Siemens Documentary German (CSDG) or Siemens-Dokumentationsdeutsch (SDD), a Machine Oriented Controlled Language (MOCL) developed by the company Siemens AG. In this case, the efforts to write in a CL were mainly aimed at the implementation of MT with the system TopTrans, also developed by Siemens: "Therefore, the focus of CSDG doesn't lie on the generation of texts that are simple (e.g. Caterpillar Fundamental English) and intelligible for the reader in the first place. The most important aim of CSDG is the increase in effectiveness of machine translation components." (Schachtl, 1996). However, SDD exists only as a research prototype and has not been further developed.

In 1995, the Institute of Applied Information (IAI) in Saarbrücken started developing, within the framework of a project supported by the BMWi (Bundesministerium für Wirtschaft und Technologie; Federal Ministry of Economic Affairs) and with the help of BMW a tool to support authors writing technical documentation called MULTILINT that was later developed as CLAT and is nowadays marketed in the form of a tool called Congree. Though no formal definition of a controlled German underlies, the tool controls grammar, style and terminology of the text on the base of general writing rules and controlled terminology, adapting them every company to their needs. A more detailed description of this tool will be given in Chapter 2 (2.6).

2.3 Controlled Language Checking

The definition of a CL can be the first step for the creation of consistent, readable and translatable documentation. However, depending on the extent of the CL, its application can be difficult from the cognitive point of view, since most authors are not able to retain the proper use of thousands of words and the application of a number of rules. Indeed, writing texts in a CL can represent a big burden for authors, since if restrictions and writing rules are going to be consciously considered, this might prevent them from concentrating in their thinking. Besides, sometimes it is difficult to judge if a text conforms to the CL and if it does not, it can be hard to find an alternative expression.

Therefore, an automatic way of controlling the language is needed in order to assure the application of the CL. Thanks to the development of the field of Computational Linguistics and especially the advances in language parsing technology in the past years, it has been possible to design applications that control the proper use of a CL. These can be defined as a "specialized piece of software which aids an author in determining whether a text conforms to a particular CL" (Nyberg, Mitamura, & Hujisen, 2003: 251). As I already mentioned in the introduction, there is a difference between CL checkers and general language checkers. While the former seek the conformance to the CL definition, the emphasis of a general grammar checker is to ensure that the text is not ungrammatical. Authors such as Hernandez & Rascu (2004) and Bernth (1997: 160; 1998: 31) account for this difference: "The main object of a checker for this type of controlled language is to ensure that the text stays within the language defined by the grammar rules and vocabulary Restrains. This is in contrast to a grammar checker, whose main object is to ensure general grammatical correctness for the full natural language." Bernth, however, points out that this might be a problem since this lack of attention to the issue of grammaticality makes the author the only responsible of the grammatical correctness of the text.

The aid offered to the author to determine the CL compliance can vary depending on the degree of accurateness of the CL checker and the degree of control aimed at the authoring process. It can go from a soft warning message to a detailed diagnostic message with an alternative paraphrasing of the structure that conforms to the CL. Most applications offer thus a checker, but others aim at offering as well a corrector which either automatically corrects the text or suggests rewrites to the user.

2.3.1 Design Issues in CL checkers

There are a number of things that need to be taken into account when designing and implementing a CL checker in the authoring process: Is the checker going to be based on an existing pre-defined CL, or are the rules going to be derived from a corpus? How are the rules going to be implemented algorithmically? What kind of parsing is necessary, swallow or deep? How is terminology going to be managed?

Surely, the definition and settings of the checker will depend on the purpose of the CL: it is not the same if texts are intended to be more comprehensible and readable, than if those texts will be subsequently translated, and especially if it will be done with a MT system.

Three main levels at which CLs are checked can be distinguished: Lexicon, Grammar and other rules (these can be either pure grammatical rules or style rules). Generally, a CL is composed by a lexicon containing the terms that are accepted in the CL together with those deprecated terms that are not allowed. This lexicon can contain either basic vocabulary and specific terminology, or only one of them. Besides, a parsing with either a proscriptive or a prescriptive approach is carried out in order to detect common grammatical mistakes (non-concordance of genre and number, case mistakes etc.). Finally we find some rules aimed at detecting style problems or some lexical preferences that cannot be captured by the lexicon. Checking levels are an important aspect of CL Checking: what should be checked first and why? Should different control areas be revised separately, or simultaneously? For instance, solving a terminology problem might directly solve a grammar problem or, contrarily, create one. Besides, we should be asking ourselves if interactive disambiguation should be supported to solve certain problems (Mitamura & Nyberg, 2001).

The automatic support of a CL checker seems to bring a series of advantages, some of them (1-4) already mentioned by Reuther (2007: 21):

- Validation criteria remain the same and are objective and psychologically neutral, without the subjectivity that might imply human revision.
- More consistency is achieved since the CL checker always checks texts with the same criteria.
- If appropriate, different criteria can be applied in different scenarios.
- It is possible to integrate checking as a compulsory process before texts are released by applying a meta attribute “checked” or “non-checked” to documents.

- Authors do not need to memorize CL rules nor approved or deprecated terms to start writing.

- After a learning curve by the authors, the time used for writing and reviewing the text with the checker is usually less than the time needed for writing and human revision. Thus, the human revision step can be eliminated.

I analyse now the three levels for CL checking.

2.3.1.1 Terminology Management and Lexical Control

One of the main components of a CL checker is terminology. It is necessary to design an application that offers the possibility of storing and managing terms, with interfaces to the different systems that might need access to them. This application might part of the CL checker or an independent program. In any case, communication with other systems such as the Translation Memory System (TMS) or a Machine Translation System (MTS) will most probably be necessary too.

As we have seen, CL definitions usually do not include specific terminology in the specifications, but a list of general semi-technical terms with preferred usages. However, one of the most interesting aspects of controlling automatically documentation is to ensure consistency in the use of technical terms. Therefore, to obtain a terminology database to work with, the first step will consist on retrieving all the specialized terms that we want to control through automatic checking. These should include preferred terms as well as deprecated terms or variants that should refer to those preferred terms. In this way we guarantee not only terminological consistency, but also detecting other possible real new term candidates that are not yet in my system. A way of making a first retrieval of the terms commonly used in the technical documentation is to construct a corpus and use a tool such as a terminology extracting tool to obtain a list to start working with. The next step will consist of a human validation of the list[33] to determine which terms are really specialized non-ambiguous terms of the domain, as well as to gather information about them (grammatical information, contextual information, usage information etc.), establish the proper relationship among them

(preferred terms and their deprecated terms, or terms that are allowed in a specific domain but not in the rest) and introduce them in a terminology management system. This information is essential for automatic language checking. There can be many different types of technical terms: single-word terms, acronyms, abbreviations, proper names (brands), measures, multi-word terms or even captions etc. All these types must be properly codified so that the CL checker can parse them adequately. For instance, multi-word terms or captions must be parsed as a single, atomic unit of meaning rather than trying to analyse them as a compositionally-derived structure (Nyberg & Mitamura, 1996).

As to the control of more general vocabulary or semi-technical terms, this can be achieved in different ways: with help of the terminology management system; by means of grammatical rules that for instance limit the allowable parts of speech of a term or restrict the valency of verbs to the subcategorizations that are sensible in the domains; or through the use of some words in detriment of the others.

2.3.1.2 Grammar and Style Management

Regarding grammar control, there are different possibilities: if the CL checker is going to be based on an existing CL definition, the algorithmic representation of these rules will be necessary for automatic checking. However, when adapting a CL to an algorithmic formalism, not all the rules of CL specifications are appropriate for their automatic processing. For instance, a rule such as “Do not write sentences with more than 20 words” is easy to check, while others, more vague rules such as “If possible, use an article before a noun phrase” are harder to check. Besides, there are some rules that are impossible to check automatically, such as “Make your instructions as specific as possible”.

If there is no previous definition, there are two possibilities: either acquiring a commercial CL checker which usually includes certain standard CL rules and adapt them to my needs, or to create the rules from scratch and design a CL checker ourselves. The latter is obviously the most challenging, but also the method that might obtain the

best benefits. For this it will also be necessary to compile a corpus and analyse the different syntactical structures present in the documentation, choose which are preferred and which are not, and decide an approach (see 2.3.2 for the different approaches to grammar checking) to implement the rules algorithmically. The chosen rules should achieve reduced parsing complexity and increased translation accuracy by reducing ambiguity as much as possible.

To manage the rules, the CL checker should have a special module where rules can be customized and parameterized: for instance, the rule about sentence length could have different length parameters (15 words, 20 words, 25 words etc.), depending on the text type. Further, this module is necessary to add, activate, deactivate or simply delete existing rules.

These were examples of terminological, grammar and style design issues. Their inclusion in a CL checker will depend on the language to be controlled and when these techniques are useful. All these components will need to communicate appropriately, and the CL checker will most of the time be integrated within the authoring tools. As I will see, it is necessary to bear in mind that a CL checker rarely works alone, but in a complex authoring process where a myriad of systems coexist and communicate.

2.3.2 Approaches to Grammar Checking

Depending on the depth of the grammatical analysis, the degree to which writing rules can be checked automatically can vary greatly. In general, grammar rules in CL checking can either be proscriptive or prescriptive.

2.3.2.1 Prescriptive approach

The controlled language is implemented by a grammar which describes all allowable sentences. Any sentence which cannot be parsed by the grammar is considered outside the CL, and must be rewritten. This approach is more labour-intensive, since the developers must work very carefully to define all of the allowable sentence structures in the domain. This approach is taken by systems like Caterpillar Technical English CTE

(Kamprath et al., 1998) and the Controlled Automotive Service Language CASL (Means & Godden, 1996). Thanks to the exhaustiveness of the analysis, this approach is less prone to give inappropriate feedback, though, especially at the beginning, initial tuning and extension of the rules is necessary, since there might be structures that have been overlooked during the design phase that need an appropriate rule.

2.3.2.2 Proscriptive approach

The CL is implemented by a set of patterns which will match any sentence that should be rewritten. Only sentences which match one of the patterns must be rewritten. This approach typically requires less work, since the developers may limit their attention to only those sentence patterns which are considered unacceptable. I can consider this approach to be "partial checking", since there may be other problems with a sentence which are not detected by the existing set of patterns. This approach is taken by systems like Diebold's controlled language checker or EasyEnglish at IBM, where CL checking is restricted to the detection of structural ambiguity, complexity and violations of vocabulary restrains. However, the proscriptive approach can overlook certain problems, and is more likely to give inappropriate feedback (for example, when a pattern is matched by an exceptional sentence which is perfectly acceptable).

2.3.3 CL Feedback: Correction and Rewriting

When an author is writing in a CL environment with a CL checker, he needs some feedback from the latter in order to take a decision. The type of feedback will depend on the depth of the morphological, syntactical and semantic analysis. The feedback can be limited to simple reminders of a rule when one of the negative patters of the grammar is matched in the text (proscriptive approach) or when one of the sentences cannot be parsed (prescriptive approach), or can reach the level or rewriting, that is, the system will automatically rewrite the wrong sentences or terms into correct constructions.

With regards to the binomial checking and correcting (Fouvry & Balkan, 1996) made a classification of controlled language checkers into:

- Checkers which flag mistakes but do not make any suggestions for error correction.
- Checkers which flag mistakes and make suggestions for error correction.
- Checkers which flag mistakes, make suggestions for error correction and actually perform some amount of automatic correction in case error detection is fairly straightforward.

According to these authors, automatic lexical correction is potentially more feasible to carry out than syntactic or stylistic correction, being in these cases more appropriate to suggest a correction and let the author introduce the changes.

With regards to the checking process, Allen (1999) distinguishes two opposite ways of facing CL authoring: on one hand there are stop-and-go authoring systems, where the author first writes the entire text and then submits it to the conformance checker, which will analyse sentence by sentence and will notify the author of potential CL violations. On the other hand, there are interactive authoring systems that assist the authors while writing, with optional correction suggestions.

With regards to compliance with the lexicon, depending on the language and on the type of feedback the CL checker is intended to give, it can consist of a simple pattern-matching or it will need determining the syntactic category of words in their context and a morphological analysis. If a rewriting of a wrong pattern is intended, it will be necessary to determine the meaning of a word, since many words in the CL lexicon can be either approved or unapproved depending on their meaning (especially in languages such as English or German). If the CL checker is just intended to alert the author about the different meanings of the word in different contexts, such a deep semantic analysis is not necessary.

Rewriting might improve user acceptance since using a CL checker would not be so time-consuming. However, usually only a direct indication to the author that the sentence should be rewritten is given and, indeed, automatic rewriting is a complex

issue that still needs further research. Rascu (2006) deals with the issue of rewriting trying to extend the style module of the controlled authoring tool CLAT (see 2.5), so that it not only prompts inappropriate structures but also provides a concrete proposal of reformulation. She argues that, when CL checkers are paired with a corrector, it is necessary a hybrid architecture that employs both prescriptive and proscriptive methods. The proscriptive rules indicate which structures are not compliant, while prescriptive rules help to indicate how to reformulate the identified items. However, automatic rewriting is scarcely applied in CL checkers. Mitamura & Nyberg (2001) discuss CL rewriting issues in their KANT System. Such architecture can be found at the last version of KANT CE Checker that includes correction.

Figure 4: The KANT interactive correction module

As far as I know, no other CL checker includes a correction module to date and, therefore, the KANT CE Checker is unique in its architecture.

2.4 CL Checking in the Authoring Process

When a CL is introduced in an authoring process, this causes a series of modifications not only in the workflows, but also with regards to the resources (economic, human) needed to set up and maintain it. New tasks appear and new roles have to be defined. Besides, new communication interfaces among tools must be specified. Among the new tasks that will arise I can mention terminology maintenance, user administration, bug fixing, rule parameterisation etc. It must be decided which roles are going to take part in the process and who is going to be in charge of new tasks, if it should be done internally or rather outsourced. Very often companies which have another central business than documentation or translation tend to externalize this kind of tasks. This might be good since they save internal resources for core business tasks. However, in order to obtain

the maximum benefit of the processes and to keep the control over them, a good quality control mechanism and communication protocols must be established so that both parts are satisfied with the work done.

2.4.1 CL Maintenance

When a CL is being applied in an industrial context within an authoring process with the help of a CL checker, in most cases this CL has to be maintained. That is, terminology needs to be updated –because of new terms that need to be added but also because of terms that turn obsolete or illegal due to patent issues. Besides, rules have to be constantly reviewed to detect any precision or recall problems and try to fix them up.

I could not find many references in the bibliography that deal with this aspect of CL. However, in an authoring process where a more or less numerous group of authors work simultaneously creating documentation in a CL, rather than a single person, it is essential that the maintenance process is a well-defined one and is well implemented. For terminology processes, this has to include automatic recognition of new terminology and a process for the manual proposal of new terminology, as well as change requests, terminology processing and automatic updates in the LC Checker and any other related systems (for instance a MT system or a Translation Memory). For language quality processes it is necessary to make use of a problem reporting database, change requests, process monitoring and quality control through periodic reviews.

2.4.2 CL Training

The introduction of a CL and the use of a CL checker within the authoring processes of a company might not always be accepted by the authors that have to use it in their everyday work. They might feel it as a burden to their freedom of expression. Here, two notions are important: on one side, change management, and on the other side, training.

Change management is about managing changes within a company. When authors are used to writing texts in their own style, it might be difficult for them to change their

writing habits. Here it is important to show them the advantages they will achieve accepting the change. Besides, a comprehensive training is necessary so that they feel comfortable using the new system, especially because they will need more time than before to create their texts.

Learning how to use properly a CL checker should not just be learned by doing. Though many authors might be able to command the basic functionalities of the tool, they will not profit from all its potential and they might incur in inadequate use if they are not properly trained. Therefore, as much as maintenance, a training plan should be carefully planned. As Mitamura & Nyberg (2001) put it, "Since author usability and productivity are essential for success, providing comprehensive training with a supportive CL checker is crucial".

Kamprath et al. (1998) report how they organized periodic seminars one year in advance to gather requests by the authors and, subsequently, they trained the authors and offered them updates to training materials. Further, they also introduced an internal publication, CTE Author, which documented updates to CTE and gave CTE writing tips.

2.4.3 Controlled Automated Translation

I will deal in depth with the concept of translatability in 4.4.1. At this point I will simply point at some special aspects that have to do with MT and the relationship of this technology with a CL checker.

It is well known that MT output is rarely perfect in its raw state and especially when it is intended for dissemination it usually needs to be post-edited. However, different measures can be taken in advance in order to facilitate MT work: if the text is properly pre-edited many parsing and translation problems can be avoided. Different approaches exist in order to improve MT output quality: Interactive MT (IMT) in order to control interactively the analysis of the input; Machine-Aided Human Translation (MAHT), which aims at correcting interactively the target text; annotations in form of mark-up languages such as SGML or special standards in order to resolve ambiguities; and

Controlled Languages, a set of pre-defined rules intended to improve the translatability of texts.

The relationship between CL and MT has always been a direct one, especially when thinking of MOCL. With the term controlled automated translation I denote a scenario where CL and MT are tuned or have been simultaneously developed to obtain the best possible results[34]. It can refer both to the control of the source language and to the control of source language and translation output, though usually the former is meant.

In industry, many CLs have been designed with the purpose of making their texts more readable and understandable, and eventually, more translatable. However, the features that make a text more translatable are not always as straightforward as it might seem at the beginning, and understandability and readability do not always have to go hand in hand with translatability. As I will see in 4.4.1, both scopes do not always match. Besides, it is necessary to distinguish between what is translatable for a human translator, and what is translatable for a computer. Simple orthographic rules that have little or no impact for a human translator can produce a totally wrong parsing of a sentence and, thus, a mistranslation.

Probably one of the most known scenarios where CL has been successfully deployed in conjunction with a tailored MT system is CMU's Kant Knowledge-based MT system (Mitamura & Nyberg, 1995; Nyberg & Mitamura, 1996), which uses Caterpillar Technical English (CTE) (Nyberg, Kamprath, & Mitamura, 1998). Besides, the METAL MT system (Slocum & Bennet, 1985) was used in conjunction with General Motor's Controlled Automotive Service Language (CASL) (Means & Godden, 1996). Recent research[35] has also addressed the issue of controlled translation in conjunction with MT systems. Following this research line, Way & Gough (2005) explore different approaches of generating controlled translation environments and conclude that EBMT (Example based Machine Translation) fares better on the controlled translation task than RBMT (Rule Based Machine Translation).

However, this is not always necessarily the case, since requirements for a controlled automate translation setting will depend on the type of MT system that is being developed (RBMT -transfer and interlingua-, SMT, EBMT). For instance, for transfer-based systems, all three stages of processing are required for controlled translation: the source language, the transfer routines as well as the generation component. If any one of these stages remains uncontrolled, then it is not guaranteed that a high-quality controlled translation is produced. Other systems such as SMT (Statistical Machine Translation[36]) or EBMT will need a controlled bitext to generate analogies and a controlled language model. Since these bitext corpora are not wide available, some related research explores the multilingual generation of controlled language texts (Hartley, Scott, Bateman, & Dochev, 2001; Power, Hartley, & Scott, Donia, 2003).

2.5 Survey of CL Checkers

Coulombe et al. (2005) divide CL checkers into three generations; they distinguish a first generation where CL checkers are limited to lexical control, using rather statistical methods to find non-compliant structures, recognising patterns and identifying the words by looking up character strings. The main application of these tools is the lexical verification but, due to its poor or inexistent lexical analysis, it is not possible to offer a sophisticated correction system. The second generation tools already introduce linguistic knowledge. Very often, the syntactic analysis modules of these tools are based on an MT system. An example of this generation is SECC (Simplified English Checker Corrector) (Adriaens, 1996), based on the technology of the METAL system. Here I could also include nearly all CL checkers that are currently applied and available in the market. Finally, the authors envisage a third generation which includes semantics for the analysis. These tools will not only limit the analysis to syntactic structures, but will be able to offer feedback with regards of the content written by the author.

Most CL checkers have been developed internally by companies in order to check also a CL that is only applied within the company. Sometimes these checkers have been developed using in-house staff and sometimes in collaboration with other companies devoted to language processing issues in the computer services industry. In certain cases

there are also commercial CL checkers that offer a general infrastructure that can be adapted to the rules and terminology determined by a certain CL.

Generally, these tools aim at meeting all or a subset of following requirements:

- Linguistic analysis of CL compliant text
- Generation of useful critiques to authors
- General morphosyntactic and spelling correction[37]
- Support for interactive transformation of general sublanguage expressions into the CL.
- Integration in standard DTP environments.

I will now have a look at some checkers that have been developed in-house. These can be divided in more or less clear industrial areas.

In the aerospace industry there is a variety of checkers that have developed or deployed to check AECMA SE (in all its variants). Boeing developed the Boeing Simplified English Checker (BSEC) in 1989 and 1990 which has been used in Boeing since then. Based on the Generalized Phrase Structure Grammar (GPSG), the Boeing checker is built around a syntactic analyzer containing a tokenizer, a lexicon, a parser, and grammar containing more than 350 syntactic rules of English. The checker can distinguish between procedural text, descriptive text, and notes. The checker counts with a basic general vocabulary of about a thousand selected words. Besides, Boeing has defined a company-specific technical vocabulary of about 2700 words (Language Industry Monitor, 1993). During its runtime, BSEC was modified to meet the requirements of Boeing Technical English and to add semantic and pragmatic language checking capabilities, resulting in the EGSC checker (Enhanced Grammar, Style, and Content Checker) (Wojcik & Holmback, 1996: 27). Though at the beginning the checker was only used internally and not offered to third parties since it was regarded as a strategic asset, now it can be purchased[38].

Other examples of checkers based on the AECMA SE are EUROCASTLE (Barthe, 1996), the GSI-Erli's AECMA Checker[39], the short-lived Oracle's CoAuthor AECMA Checker[40], and Carnegie's ClearCheck (Andersen, 1994; Language Industry Monitor, 1995; Nyberg, Mitamura & Carbonell, 1997: 3). All these checkers are only used in-house or not on the market any more. Some commercial tools for SE that can be currently acquired commercially are HyperSTE[41], MaxIt Controlled English Checker[42], Language Manager LM[43], and Simplus[44].

In the automotive domain, I can mention the case of CASL (General Motors Controlled Automotive Service Language), where the need for a special checker for this CL can be stated in Godden (1998) and Means & Godden (1996).

In the telecommunication industry I find the example of ALCOGRAM (Adriaens & Scheurs, 1992; Scheurs & Adriaens, 1992) for the Alcatel-Bell company, based on the COGRAM paper grammar. ALCOGRAM (algorithmic controlled grammar) is a further development and the strict algorithmic representation of the COGRAM paper grammar. ALCOGRAM has a different organization of the rules and consists of four modules ranging from "conciseness" over "extra-textuality" to "layout and punctuation". The next step was the development of a computer-program to guide the authors through the algorithm, which lead to the development of SECC (a Simplified English Grammar and Style Checker/Corrector) within the context of an LRE-2[45] project that ran from November 1993 to May 1996 (Adriaens, 1994; Adriaens & Macken, 1995). EasyEnglish Analyzer, developed for IBM (Bernth, 1997, 1998, 1999a, 2006) is another example of a CL checker developed within the telecommunications industry. One of the particularities of this checker is that they are making efforts to analyse texts not only at sentence level, but at text and discourse level, something that most checkers do not [46].

CL checkers have also been used in other industrial domains, such as heavy machinery. One example would be the Checker for PACE (Perkings Engines Ltd) (Douglas & Hurst, 1996). Another example is the Multinational Customer and Service Education (MC&SE) organization within Xerox Corporation that used a system of writing, called Multinational Customized English (MCE), that involves a controlled dictionary and a

set of writing standards, to facilitate both machine and manual translations (Adams et al., 1999). Other examples are Diebold's controlled language checker (Moore, 2000) and AutoPat, an authoring system for patent claims (Falkedal, 1994; Sheremetyeva, 2007). With AutoPat, authors do not write the text directly, but they fill in predicate and argument structure templates which AutPat uses to build a deep content representation and then transform it into a final surface claim text.

Finally, there are some other checkers that have been developed either for general purposes or for other languages. One example is LANTmaster CL checker, developed by the company Xplanation[47], together with Pulsar Consulting, that was assigned the project to develop an application, resulting in the development of this tool. It is based on the METAL MT engine and nowadays is commercialized by Pulsar Consulting and can be customized to any specialized domain[48]. More information on the development of LANT can be found at (Caeyers, 1997a, 1997b) and (Knops, 1999). Another tool is Acrolinx IQ[49], developed by the German company Acrolinx, based in Berlin. As they define it, Acrolinx IQ is "enterprise client-server solution that promotes quality and efficiency during content development". The product contains the Acrolinx IQ Lingware and Terminology which can be fine-tuned to promote controlled authoring practices.

Finally, I will mention MULTILINT/CLAT[50], a tool which was originally developed by IAI to help authors to create consistent and high quality documents. The tool not only checks from a proscriptive approach the texts with respect to correctness (orthography and grammatical correctness), but also with regards to company specific rules (controlled language rules) and terminology. This tool has been applied in a number of different domains such as automotive (BMW), printing machines (Heilderlberger Druckmaschinen) or high technology (Siemens). Since this system will be object of my study in this dissertation, a more detailed overview of MULTILINT/CLAT and Congree, its development and structure, will be given in the next section.

2.6 MULTILINT, CLAT and Congree

In 1995, the German Federal Ministry of Economy fostered the project MULTILINT. BMW AG and the Institut der Gesellschaft zur Förderung der Angewandten Informationsforschung e.V. an der Universität des Saarlandes (IAI) were, among others, the main partners in this project. Its goal was to develop an intelligent linguistic system for the production and administration of multilingual technical documentation (Schütz, 1996). The subsequent project, TETRIS (starting in 1999 and lasting until 2002), resulted in the development of the tool MULTILINT, a sophisticated language checker (Haller, 2001; Reuther & Schmidt-Wigger, 2000).

The approach of MULTILINT deviates slightly from the traditional approach and definition of a controlled language, since there is no previously defined controlled language. Rather, MULTILINT aims at "controlling" the language by helping authors to write technical documentation according to a definite set of style, spelling and grammar rules (general language correctness). These rules belong to the core of the system. The style rules represent an exception. These are given by the system, but the author or linguistic resources manager can add new rules or adapt them to the style of the company where the checker is being deployed. Besides, authors are required to use a controlled vocabulary and a controlled terminology (corporate language correctness). The latter is defined by the user (Reuther, 1998).

Figure 5: MULTILINT Front-end

In 2002, MULTILINT was upgraded by CLAT (Controlled Language Authoring Tool). Though the linguistic intelligence behind MULTILINT and CLAT is the same, both systems present some differences. These include, among others, the front end, which is implemented in Java in CLAT, in contrast to the tcl tk implementation of MULTILINT. Besides, the interaction of the different modules is different: CLAT presents an editor where the author can undertake the corrections based on suggestions of the system, whereas MULTILINT the author needs to switch between the application and its document, where he introduces the necessary changes. MULTILINT and CLAT are in use by important industrial companies in Germany, such as Heidelberger Druckmaschinen, Sun Mycrosystems Inc. (for English) and BMW AG.

Further information on the MULTILINT/CLAT architecture can be found in Carl, Haller, Horschmann, Maas, & Schütz (2002) and Carl, Hernandez, Preuß, & Enguehard (2004). These authors describe how terminology is managed and controlled. Hernandez & Rascu, 2004 and Rascu (2006) deepen in the issues of style control and rewriting through paraphrasing.

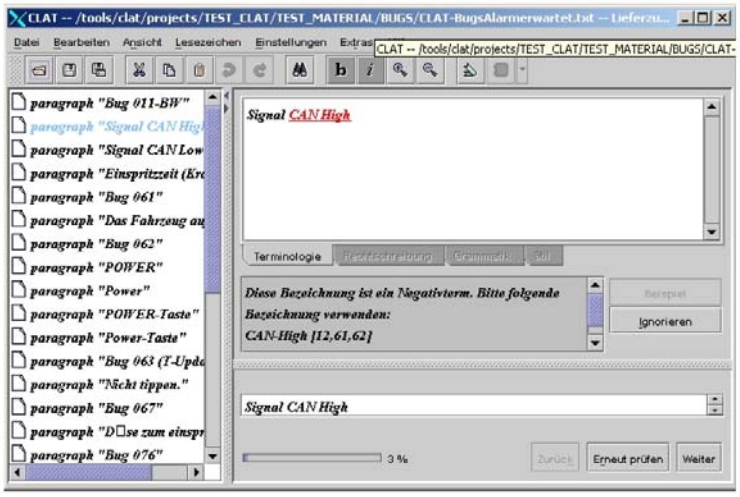

Figure 6: CLAT Front-end

CLAT stands for Controlled Language Authoring Tool and it relies on the technology developed by the IAI

Since September 2010, CLAT is exclusively sold by Congree Language Technologies GmbH. This company also markets a tool called Congree Personal Edition which is integrated within MS Word and can conduct authoring quality assurance of texts written in English.

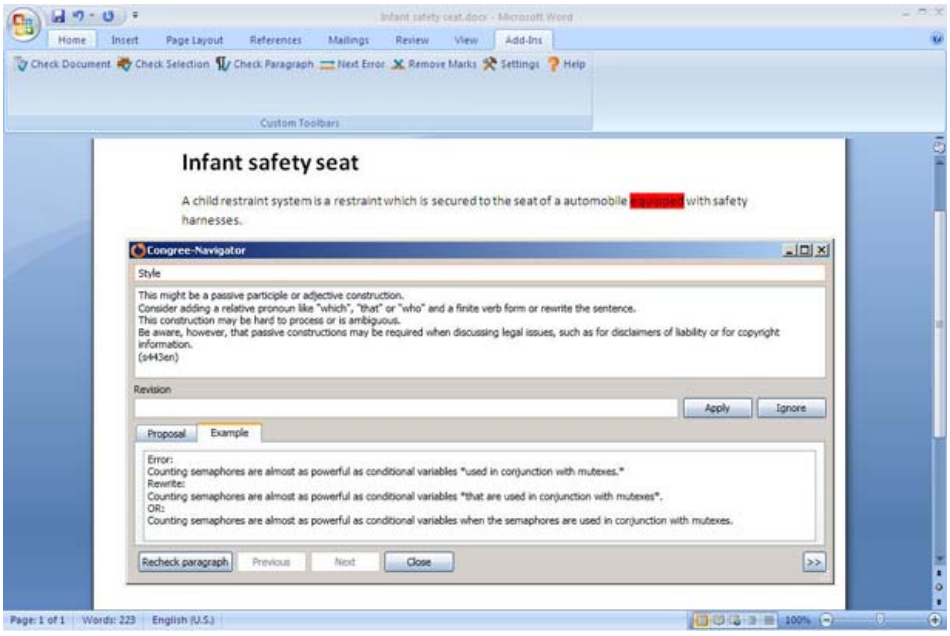

Figure 7: Congree Front-end

2.7 Summary and final remarks

In this chapter I have tackled the concept of Controlled Languages and their application in industrial environments. First of all I have made an overview of a wide range of examples of their application in industry, in particular for the production of technical documentation, both for the English language and other languages.

It is important to distinguish between Controlled language specifications and the software tools used to check these specifications while authors write their texts. Therefore, the second part of the chapter concentrates on controlled language checking, with an overview of the different techniques and different tools available in the market designed to control automatically the text production.

Finally, the chapter ends with a survey of the different CL checkers available, with a special emphasis on the tool MULTILINT/CLAT/Congree, which will be subject to analysis in this research work.

3 TECHNICAL DOCUMENTATION AND TRANSLATION

Wir leben technisch, der Mensch als Beherrscher der Natur, der Mensch als Ingenieur, und wer dagegen redet, der soll auch keine Brücke benutzen, die nicht die Natur gebaut hat.

Max Frisch, *Homo Faber*, 1957

3.1 Introduction

Since the 1950s the development of science and technology has experienced a dramatic revolution. Nowadays the frequency with which new products and new releases of older models are launched daily to the market is extremely high. Furthermore, there is a tendency towards the complexity of these products to increase exponentially (Westendrop, 2003). Consequently, the need and importance to communicate the expertise, operation, functionality etc. of these products to all kind of different groups, from maintenance staff to end users, is of vital importance not only for the products to be successful in the market, but also to avoid damages and high costs caused by an inadequate documentation.

Besides this need, technical documentation is also mandatory in most countries by the law, and Europe and USA are no exceptions. Everyone manufacturing a product or releasing one that needs explanation with regards to its functionalities is legally obliged to include the appropriate documentation[51]. The protection of the user derived by the legal normative is also a clear advantage for the manufacturer: accidents are avoided and the elimination of potential malfunctions reduces the risk of damage compensation. Another important factor is the supporting effect in the pre-sales and after-sales areas:

catalogues and lists of pieces contribute to improve the image of the company and act as commercial advertisers.

Generally, a product is considered as everything that is produced or manufactured, from a new vaccine against a lethal disease to the newest computer model. This definition of product comprises fields so differentiated as biology, technology, medicine or social sciences. In all these fields documentation is needed in order to pass on the expert knowledge and enable further development. Indeed, my modern societies base their creative capacity above all on expert knowledge (Friske, 1996).

This chapter will deal with some aspects of technical communication and technical documentation relevant to the use of CL and MT. After explaining some terminological discrepancies among the terms technical communication, technical documentation and technical writer in 3.2, I give a short overview on the history and current situation of technical communication as well as technical writers (3.3 and 3.4). Subsequently, the features of different types of technical documentation are reviewed in 3.5, dedicating special attention to the documentation in the automotive industry as central part of this work (see 3.6). The last part of the chapter is dedicated to examining the particularities of translating this type of documents and their relation to the implementation of CLs and MT (see 3.7).

3.2 Technical communication and technical documentation

Terminologically there are some discrepancies, especially in the German literature, about the denominations “technical communication”, “technical documentation” and “technical writer”. Some authors, like Friske (1996) discuss that the direct translation of the terms technical documentation and technical writer in English as *Technische Dokumentation* and *Technischer Redakteur* in German, as coined by the members of *Tekom* (see 3.4.2) are terminologically not correct, since a technical writer must not be necessarily technical. Gabrielle Bock (1993) goes further and suggests the terms *Technikredakteur* and *Technikdokumentation* for German, rejecting the attributive use of the adjective “technisch”, since neither the documentation nor the author writing it

are technical. Rather *Technik* is the main subject of the documentation, or the author writes about *Technik*.

In English there is a clear preference for the term "technical communication" or "technical writing" to the detriment of "technical documentation", as the title of numerous manuals for technical writing evidence[52]. Technical Communication is a much broader term including not only the writing of manuals in form of documents, but any form of communicating and exchanging information on technology. Indeed, the German Society that gathers technical writers is called *tekom* or *Gesellschaft für Technische Kommunikation*.

Technical communication can generally be described as the process of conveying usable information about product or a process within a specific technical domain through any communication channel (writing, speech, audiovisual etc.) to an intended audience. With this general definition in mind, authors of technical communication might include not only technical writers or authors, but also illustrators, translators, graphic designers, scientists, professors, trainers, engineers etc.

Therefore, when I use the term "technical communication", I will refer to any type of information transfer where technology is involved, whereas the term "technical documentation" will especially refer to written documents[53], both in paper and in electronic form, and is mainly authored by technical writers. The *tekom* propagates following more detailed definition:

> Der Begriff technische Dokumentation umfasst verschiedene Dokumente mit produktbezogenen Daten und Informationen, die für verschiedene Zwecke verwendet und gespeichert werden. Unter verschiedenen Zwecken ist zu verstehen: Produktdefinition und Produktspezifikation, Konstruktion, Herstellung, Qualitätssicherung, Produkthaftung, Produktdarstellung, Beschreibung von Funktionen und Schnittstellen, bestimmungsgemäße, sichere und korrekte Anwendung, Instandhaltung und Reparatur eines technischen Produkts sowie gefahrlose Entsorgung.

According to this definition, I understand technical documentation as every document produced during the life span of the product, from its conception to the further production, maintenance, service, final disposal and, eventually, recycling.

Friske (1996) gives also his own definition of technical documentation:

> Eine Technische Dokumentation ist die strukturierte Sammlung aller notwendigen und zweckdienlichen Informationen über ein auf technischem Wege hergestelltes Produkt und über seine Verwendung.

In this definition Friske stands out the fact that the information presented must be structured in some way. As we will see in more detail in 3.5, Gamero Pérez (2001) and Göpferich (1998: 80 and ff.) also refer to the importance of the macrostructure of certain text types within technical documentation.

3.3 Historical Background and Current Situation

Though the origins of technical documentation can be traced back to the ancient Greeks and Egyptians, the Renaissance[54], and the mid-19th century, it was in the last half of the nineteenth century that technology in science and industry began to grow continuously to end up boosting with World War II. The research and development on fields such as the military, medicine, engineering, telecommunications and science gave place to many of the technological advances from which we, nowadays, still profit.

"Sad to say, but many of the benefits of science that we enjoy today (e.g., air travel, antibiotics, high-performance materials, computers, and telecommunications) would be in a primitive state of development, if extant at all, if it were not for the exigencies of war". (O'Hara, 2001: 2).

As a consequence, technical documentation was a beneficiary of belligerence, since there was a pressing need for clear, concise and understandable proposals, reports, manuals, and instructions for military, industrial and civilian personnel.

In spite of this, the rapid growth on the technology and scientific fields made that the industry efforts concentrated over all on the product. When documentation was created it was generally made without taking into account the processes and the quality behind it. It has been in the past thirty years where quality of the technical documentation has

received the necessary attention and has been recognized as an important factor in the sales and after sales areas[55].

Similar to the industrial revolution experimented fifty years ago, in the last twenty years, a technological and electronic revolution has taken place with the development of computers and the booming of telecommunication technology, dramatically influencing the work practices of technical communicators and writers. Nowadays the day-to-day work of technical writers includes work with the computer, templates based on schema languages, knowledge- and terminology databases and sophisticate editors that very often include quality checking modules such as CL checkers.

3.4 Technical Writers

Technical writers are professionals that act as communicators between the developers of products and the people who use them. Their education includes both knowledge of technical matters as well as rhetorical and publishing skills. But this has not always been the case.

For a long time, the profession of technical writer was not officially recognized. Technical documentation such as handbooks or product descriptions were written either by the engineers who had developed them, or by other members of the development team. Companies assumed that these were the people who best knew the product and, therefore, could best describe its functionalities, neglecting the fact that, usually, people receiving a technical education not always master their communicative skills. Technical writers developed their skills gradually by "training-on-the-job", that is, by learning the necessary skills in their everyday jobs and visiting some specialized courses offered by the company itself.

The profession of technical writer is relatively recent. In Germany[56], for instance, the profession was not officially recognized until April 1989 by the Tekom and the German Federal Labor Office (Bundesanstalt für Arbeit) (Bock, 1993). On the contrary, the profession in the USA has gained recognition for a longer time, fostered by the

foundation of the Society of Technical Communication (STC) in the 50s. USA also counts with a large number of institutions and Universities offering education and studies in this specialty, whereas the first education initiatives in Germany began in 1987 with the conception of a fulltime training program (Fritz, 2003). The reasons for this advance in the situation of technical writers in Germany are, as mentioned before, the always increasing global competition and the higher technical complexity, as well as more exigent clients and the inclusion of owner's manuals as part of the marketing strategy of the companies.

In their daily work, technical writers gather information from different sources (libraries, product descriptions by engineers etc) and elaborate all types of technical documents, such as manuals or instructions, both for lay users and experts, always taking into account the legal requirements imposed by laws and quality norms. They are responsible for choosing the most appropriate communication channel in order to achieve an optimal interaction between machine or product, on the one side, and client, on the other. For this task they use a series of applications and aids such as DTP-Programs, authoring systems, multi-media applications etc. Further, there are also technologies that help to structure documents as well as linguistic technologies that assist them in creating terminologically coherent and syntactically correct documents, such as the deployment of CL checkers (Lehrndorfer, 1996: 96-101).

3.4.1 STC: Society for Technical Communication

The German Tekom counts with two counterparts in the USA: on the one hand, the Society for Technical Communication (STC), on the other hand, the IEEE Professional Communication Society, the 26th Organization of the Institute of Electrical and Electronic Engineers.

In 1957, the Society of Technical Writers and the Association of Technical Writers and Editors merged to found the Society of Technical Writers and Editors. In 1971 the organization's name was changed to the current denomination: Society of Technical

Communication. Nowadays, STC is over 50 years old, with 150 chapters (regional associations) and 25,000 members worldwide.

In their web page (<http://www.stc.org>) STC defines itself as "an individual membership organization dedicated to advancing the arts and sciences of technical communication". Among their members one can find technical writers and editors, content developers, documentation specialists, technical illustrators, instructional designers, academics, information architects etc.

The strategic goals of the society can be summarized as follows:

- Define the profession of technical communication
- Communicate the value of technical communication
- Establish and expand strategic partnerships
- Globally improve the practice of technical communication
- Ensure the long-term viability of the organization

3.4.2 TeKom: the German association of specialists on technical communication and information development

During the last thirty years, the Tekom has played an essential role by promoting the task of technical writers as a profession in Germany and is a symbol of the recognition of the technical writer profession in this country. This institution was founded on the initiative of Brigitte Beuttenmüller when she attended the first conference of the INTERCOM (International Council for Technical Communication). Inspired by the positive and creative atmosphere of this conference, Beutenmüller organized a meeting in 1978 with a group of colleagues in Stuttgart to convince them of the advantages of founding such an institution. Seven members attended this first meeting. Since then, the number of members has multiplied up to over 4800 and today the Tekom, including the annual conferences as well as its monthly published magazine with the last news on

technical documentation, constitute an indispensable resource of information and networking for anyone working on this field.

These two examples speak for the growing importance of the profession and the role of the technical writers, who are, all in all, responsible for creating the types of texts that I am dealing with and also in charge of applying CLs to them. Therefore they must be specially taken into account when designing authoring and translation processes where controlled language is going to play an important role.

3.5 Types of technical documentation

As mentioned before, technical documentation can be defined as every document produced during the life span of the product, from its conception to the further production, maintenance, service, final disposal and, eventually, recycling. This definition includes a wide range of document types, from handbooks to memoranda and marketing reports. Further, depending on the language they are written in, documents types can have very different characteristics, since there is not always a univocal equivalence of text typology among different languages. Therefore, a universal classification of text types is extremely complex and, indeed, questionable.

My goal with this review of possible classifications is not to establish or propose a model, but to characterize the types of text I will be working with in the second part of this work: repair instructions and service information.

Reiss (1983: 95) developed, based on the basic functions defined by Bühler (*Darstellung* or representation of states, *Ausdruck* or expression of the sender's feelings and *Appell* or appeal to the receiver), a universal text typology, where three different types of text can be distinguished:

- The informative text. The goal of this kind of text is to present information in an objective way. Its main function is the representation and some examples of this type are essays, comments, certificates, instructions and operating manuals.

- The expressive text is focused on the sender of the message and has an expressive function. Theatre plays, biographies or novels could be examples of this type of text.
- The operative text is behaviour oriented and the recipient stays in the centre of the communication act. The text has an appeal function, such as in comments, pamphlets, propaganda etc.

According to this classification, which is based both on the content of the documents and their communicative function with respect to a potential user, documents such as training documentation, repair instructions, tightening torques, service instructions, maintenance instructions, installation instructions etc., would be informative texts, since they are presenting information to the recipient. Usually, no expressive texts can be found in the technical documentation, since the sender and his personal way of expressing him or herself must be as objective and neutral as possible. Examples of operative texts could be catalogues or brochures from the marketing department. In this case, the recipient is on the focus of the message. Nevertheless, I find this classification too rigid since it does not contemplate the fact that, mostly, texts are multifunctional as Gamero Pérez (2001) states. A training document may very well be an informative text, but it might also contain expressive elements that the sender uses in order to motivate the readers so that they take as much cognitive profit as possible. Similarly, instructional texts are not only informative (description of the product), but they also present operative features (instructions), since they aim at influencing the recipient behaviour to move him to do something.

This aspect is well treated by Gamero Pérez (op.cit.) who attempts a classification of technical texts with translation in mind. She starts by defining what is a technical text (ibid: 38), taking into account the user roles implied in the process, the communicative situation, the textual function, the communication channel, the specialized field and its intertextual features. She combines all these factors to define the technical text as a concrete communication act where the senders are engineers, technicians or professionals, whereas the recipients can be represented either by other engineers, technicians and training professionals, or by general users; the communicative situation

is related to industry, farms, the manufacturing of products or the offering of services; the main textual functions are exposition (informative) and exhortation (operative or conative). The channel is usually written, the specialization field is usually technical and it scarcely displays any variation as to temporal, geographical or social dialects.

This characterization is of great importance for my purposes since, as I will see in 3.7., some of these features will make (some) technical texts more appropriate for automation in translation than other text types.

Lehrndorfer (1996: 82-83) discusses the ambiguity of the term *Technische Dokumentation* as regards to content. In this respect, she proposes a systematization that tries to express the nuances in meaning of the different denominations and suggest a functional classification that comprehends both product descriptions (e.g. technical data, lists of pieces, function descriptions) and process descriptions. She depicts a two dimensional classification, where the two main axes are content (comprising product and process descriptions) and a series of functional factors comprising different types of devices, goal groups and text characteristics such as extent, structure and function. According to this classification, generally instructions are texts that describe processes (how to do something), both of consumer durables and assets invested. The goal groups of this type of text can be either experts or lays, depending on the type of instruction (for instance, assembly instructions of a car engine are addressed to expert users, while the operating manual of a TV is addressed to a standard user). Finally, there can also be different types of instructions as to their extent (short instructions) or function (learn instructions).

It is obvious that, depending on the type of product we are considering and on the interests and education of the consumers of the documentation, the goal groups of the documentation can vary considerably. But we can distinguish generally among different expert groups who are the recipient to different types of technical communication. Depending on whether the documentation is for internal or external use, we can draw the following distinction:

- Internal (producer): service, design, marketing, sales, production, technical inspection, authorities, administrative homologation etc.
- External (user): user, buying, service and maintenance staff, recyclers, planners, fire department and emergency service, authorities, employers' liability insurance associations, magazine editorials, consumer organizations etc.

Depending on the user and the type of text, textual structure and content (degree of detail and terminology) will vary. Therefore, before starting a technical documentation project it is important to carry out a goal group analysis, so that the information regarding the interaction with the product is presented in an understandable way for the end user. Aspects like the homogeneity and level of education of the group, their experience with technical devices, their learning capacity or their language level must be taken into account.

Göpferich (1998: 89-136) dedicates one chapter in her book to elucidate the problem of technical text typology. She argues that technical writers usually work on instructive text types (instructions and tutorials), as well as descriptive text types (technical descriptions). However, she includes many other types in her systematization, which is based on a hierarchical structure where each level devises a criterion to differentiate the texts (the table can be found in page 90 of her book). The first level deals with the general specific text types and uses the communicative function criterion to classify them. She distinguishes among legal-normative texts, texts including advancements or novelties, didactic-instructive texts and knowledge gathering texts. The second level classifies texts as to the type of content that is transferred and the type of relation that is established with the reader: unidirectional, with theoretical contents, or bidirectional, with interactive, practical contents. The third level considers how information is presented (oriented to facts, intended for advertising, organized mnemonically, aimed at awakening interest, encyclopaedically or in sentence fragments). The fourth and fifth levels already present real examples of text types divided in primary texts (fourth level) and secondary texts (fifth level), derived from primary texts. Instructions are characterised as didactic-instructive texts with practical information that establishes a bidirectional relationship. There is no characterization as to the way they are presented.

Finally, Gamero Pérez, (2001: 69) suggests a multifunctional classification for the different genres within technical texts, considering following parameters: genre, main contextual focus or function, secondary contextual focus or function and recipient. She argues that this classification is very flexible because it is open and thus allows the inclusion of new genres at a later stage. But undoubtedly, the most novel feature is that it considers the multifunctionality of texts, making it possible to account for the mixed character of some genres. For instance, a Technical Description is a text addressed to a specialized reader and with a unique expositive or informative focus, whereas instructions can be either addressed to a general reader or to a specialized reader, but always with an exhortative or conative focus. This contrasts to what Reiss (1983) exposed with regards to instructional texts[57]. In her classification she does not include content, communication channel (she only considers written texts) or text structure as decisive criteria, such as other authors do (for instance Lehrndorfer).

Seewald-Heeg (1998) describes technical texts as an heterogeneous class of texts that comprise, among others, instruction manuals for technical instruments, assembly and installation manuals, repair instructions as well as recipes, game rules or enclosure notes of medical products. Though all these text types have distinct features they all share a set of common features:

- They provide action knowledge: They are not autonomous texts, since they are always bound to the existence of the product they are delivered with.
- They provide information to the reader who simultaneously has the role of the user. The information is given in the form of the instructions which allow him to handle the product i.e. to use it the way intended by the producer.
- They are utility texts where contents are usually presented in chronological sequence.
- They include directive text parts which leave the reader no room to move.

As we have seen, authors use a series of different criteria in order to characterize and classify technical documents. These criteria range from the most obvious features, such

as content or end users, to other variable elements such as text length, text structure, communicative function, presentation channel etc.

3.6 Technical documentation in the automotive industry

In this section I aim at giving a general overview of the types of documents I can find in the automotive industry, especially focusing on Service Literature, that is, documents produced in the departments devoted to the after-sales areas. This can include documentation intended for the subsidiaries, the authorized dealers, the workshops and the final customers.

Subsidiaries and workshops need information to assist their clients: this might comprehend installation instructions to install new or additional components, information on operating fluids, techniques, special tools and appliances, tightening torques, diagnosis encoding etc.

Within workshops, car mechanics need technical data to be able to carry out the reparations in the right way. Repair literature describes repair processes for the mechanics. This information is provided in form of instructions that represent, with the help of graphics, inspection sheets and technical data, the necessary work steps.

Another type of service literature is represented by diagnosis documents which contain information that helps the mechanic to find defects quickly and surely. Diagnosis documents can have different information chunks: function descriptions, instructions for function evaluation, positions of pin-outs and terminal sides, target values etc.

Further, information to code, individualize, program, install or uninstall control devices in the automobile is needed by mechanics and electricians in form of a database which includes specific data of the automobile, process definitions, texts as well as menu and navigation structures (as HTML examples and PHP scripts), which can be modified by the user.

Training documents are another common type of literature within the after-sales area. These are provided to the trainer and the participants during after sales training.

It is also frequent to count with an interactive information medium for significantly faster fault recognition, recording and reporting in the entire after-sales area. These applications enable direct communication from the dealer organization to the specialists in the central service department and from the headquarters of the company across the wholesale level (sales subsidiaries, regional offices) to the dealer organization. In addition relevant cases and reports are evaluated with trend analyses. The results are integrated in the fault eliminating process in order to increase product quality and customer satisfaction. The information gained in this way can be taken into account for production starts and face-lifts. The central task of these applications is to provide the dealer organization proactively and quickly with all the information available regarding problems in the service area in a clearly laid out form. In addition to pure data collection, these systems may also offer functions such as the attachment of multimedia files, an optimized search function, interface extensions, as well as —resulting from the direct connection to the dealers and elimination of manual data interchanges from the wholesale level to the headquarters and vice versa— a drastic shortening of both retrieval times in the reporting phase and provision times for solutions and measures for targeted improvements going out to the dealer organization. However, it must be taken into account that the type of information generated in this type of systems is generated spontaneously by the mechanics and the people in charge within service as the faults arise. It is therefore characterized by a higher degree of informality compared to other text types which are created by technical writers.

Finally, the headquarters might be interested in informing the subsidiaries of news and technical issues. This might include a letter for the selling point and service information with the content of the technical campaign itself.

As we can observe, technical information within the automotive industry and, more concretely, within the sales and after-sales area can be very heterogeneous with regards to content, support (text documents, videos, pictures, websites, etc.) and structure (lists,

running text etc.). This will require a wide range of different strategies when facing translation processes, as I will see in the next section.

3.7 Translation of technical documentation

3.7.1 Particularities of technical translation

Technical translation, similarly to any specialized sort of translation, renders an important number of differences with respect to literary or general translation. Therefore, some aspects have to be taken into account.

First of all, the translation of technical texts requires expert knowledge. When translating complex technical issues the translator must be able to understand its content, that is, the functionalities and devices described in the document. Schmitt (1994) argues: "Zusätzlich zu ausgangs- und zielsprachiger Sprachkenntnis ist bei Fachtexten auch Sachkenntnis erforderlich; - nicht nur, um Übersetzungsprobleme zu meistern, sondern zunächst, um potentielle Probleme überhaupt zu erkennen."

This expert knowledge is linked to a specific terminology. It is thus not only necessary to understand the text, but to use the appropriate terminology. By employing the right terminology, in the best cases coined and maintained by the client himself, the translator assures the satisfaction of the client and contributes to the acceptability of the text by the end user (Horn-Heft, 1999: 106).

Contrary to literary texts, marked by the personal expression of the author and usually with a high linguistic quality, technical texts must adjust to certain formal conventions and must aspire toward a clear, consistent and univocal expression of ideas and facts. Unfortunately, this is not always the case, and the translator of technical texts must always be aware that the quality of the source text is not always the most desirable for translation purposes.

Each genre of technical texts has a uniform macro or superstructure that usually remains the same in all exemplars and facilitates the user to recognize a document as pertaining to a certain genre. For instance, all instructions have a title page and a front matter, a table of contents, a text part and a subject index if the instructions are long. They might also contain other variable elements such as a glossary, a list of abbreviations and a table with remedies for potential failures (Göpferich, 1998: 100 and ff.). All these elements have to be considered when implementing translation processes and, especially, when considering translation technologies. Certain parts of the text, such as indexes, glossaries or tables which are organized alphabetically, might pose a problem for full automation, since these must be rearranged afterwards. Sometimes, this can be done automatically, but sometimes the intervention of humans will be necessary. Other elements, such as screenshots or crossed references must also be appropriately handled.

From the linguistic point of view, Lehrndorfer (1996: 89 and ff.) argues that quantitative stylistics still assigns technical documentation a series of stylistic means that distinguishes it from standard language:

- Long, convoluted sentences.
- Phrases instead of subordinate clauses.
- Constructions with semantically weak verbs such as *in Gefahr bringen* or *in Betracht ziehen*.
- Passive and impersonal constructions.
- Ellipses.
- Multiwords with a fixed order and compounds.
- Nominalization, technical terms.
- Neologisms.

All these constructions are rather functionally oriented and aim at achieving conciseness (for instance through ellipses) and impersonality (through the use of passive constructions). To have a better understanding of the importance of linguistic elements

and text functions with regards to translation, it is necessary to give a short overview on the theory of speech acts.

3.7.1.1 The theory of speech acts

The theory of speech acts can be useful to explain some of the difficulties I can encounter when translating technical documents from one language into another. This theory captures the idea that when I say something I do it with the intention of provoking actions. The intention with which I say something is called illocutionary force or simply illocution. These speech acts, also called illocutionary acts, have an influence on translation, since not all intentions are expressed in the same way in all languages.

For instance, the use of the English modal verb "should" in these two examples depict two different illocutionary acts:

[1] Your driver maintenance chart, shown here, lists important items which you should check regularly.

[2]These instructions should not be faxed or reproduced on a digital copier.

The first one is a recommendation, whereas the second one is rather an obligation to do something. Therefore, the same verb must be translated in different ways to express the different intentions. Usually, illocutionary acts have no formal equivalent in another language and idiomatic ways of expressing the same illocution must be found. Otherwise, the resulting text might result unidiomatic, incomprehensible or even misunderstanding.

In technical documentation it is thus recommended to use as less indirect speech acts[58] as possible, since these might cause problems to translation, especially when trying to automate the process. Intentions should be expressed directly, with performative verbs or constructions.

Searle (1970) set up the following classification of illocutionary speech acts:

- **assertives** = speech acts that commit a speaker to the truth of the expressed proposition
- **directives** = speech acts that are to cause the hearer to take a particular action, e.g. requests, commands and advice
- **commissives** = speech acts that commit a speaker to some future action, e.g. promises and oaths
- **expressives** = speech acts that expresses on the speaker's attitudes and emotions towards the proposition, e.g. congratulations, excuses and thanks
- **declaratives** = speech acts that change the reality in accord with the proposition of the declaration, e.g. baptisms, pronouncing someone guilty or pronouncing someone husband and wife

Although there can be a great variety of illocutionary acts in texts, some sorts stand out with respect to others in specialized texts. For instance, on human-machine interactive texts the most common are assertive and directive illocutionary acts.

3.7.1.2 Problematic constructions for the language pair German-English

Though there might be general problematic constructions for translation, I concentrate on the language pair German-English, since this is the language pair I will analyze in the experiment presented in the second part of this work.

Seewald-Heeg (1998) analyzes the morphosintactic features of instructional texts, concentrating on verbal requests from English to German. This illustrates how the commissive speech acts can be expressed in different languages in very different ways.

For verbal requests in German she distinguishes following forms:

- Imperative plural and "distance form of request" for the second person (singular and plural): stellt, stelle. However these forms are intended to control directly one or several present addressees and are therefore not used in instructional texts which are characterized by the physical distance of the communication partners.
- Distance form of request: imperative combined with a deictic personal pronoun: Stellen Sie, Entfernen Sie. It can include the marker of politeness "bitte", which is placed before the request sentence or after the syntactic group of verb form and deictic pronoun.
- Requests by the infinitive: Kühlsystem ausfüllen, Dichtung entfernen.
- Requests expressed by a modal auxiliary verb combined with an infinitive: Sie dürfen die Dichtung nicht entfernen. Sollen would also be included in this group, though the function of advising is prevalent to the function of requesting.
- Requests expressed impersonally by the indefinite pronoun "man": Man muss die Dichtung entfernen.
- Passive constructions: *Die alte Dichtung muss komplett entfernt werden.*
- Use of lassen, both as distance form of request as well as by the infinitive: Das Lösungsmittel ablaufen lassen. Lassen Sie Ihr Mobilteil anschliessend vom Service überprüfen.
- Declarative sentences in the active or passive voice are also used to express instructions: Mit einem Strahl reinigen Sie die Zähne….
- As for English, she distinguishes following forms:
- Imperative mode. There is no distinction between singular and plural nor by person. Sometimes they are followed by an exclamation mark, but it is not common in instructional texts. The use of the 2nd person pronouns does neither occur in written distance texts Imperative sentences can be introduced by the politeness formula "please" in initial or final position to tune down their effect. The use of do before the imperative verb also aims at a less abrupt and more persuasive effect.
- Declarative sentences with the pronoun in subject position.

- Requests in interrogative form extended by will you or won't you, through they are not normally found in instructional texts.
- Use of do before the imperative verb.
- Use of should or ought.
- Verbs such as recommend.
- Modal verbs (must).

As we can see, sometimes there are direct equivalences from German into English. However, it is not only necessary to take into account the lexical and the morphosyntactic level, but also the context in which the corresponding forms occur and their function. For instance, instructions for operational texts arranged in lists are expressed in English with the imperative, while in German the infinitive form is normally used.

Finally, Seewald-Heeg (1998) proposes a series of formal correspondences of English and German request forms which occur in written instructional texts:

English	**German**
Imperative	Distance form of request/infinitive
Negated imperative	Negated distance form of request/negated infinitive
Please, request sentence	*Bitte*, request sentence
Should / Ought to	Sollte(n)
Modal verb construction, e.g. *must*	Modal verb construction, e.g. *müssen*
Declarative sentence	Declarative sentence

Let	Lassen Sie .../ … lassen

Table 3: English and German request forms in instructional texts

Göpferich (1998: 153-156) also deals with potential translating difficulties for the language pair German-English, distinguishing the following:

- German expressions that do not have any formal equivalent in English:
  - Infinitives in an imperative function (*Getriebeölwanne aus- und einbauen/abdichten oder austauschen*...)
  - *man* + Konjunktiv I (*Man beachte*...)*
  - The third person plural of present indicative (*Sie schalten nun das Gerät ein, und warten, bis die rote Diode leuchtet*)*
- German expressions, the formal equivalent of which in English cannot generally be used as directives:
  - *man* + present indicative (*Man gibt die Wäsche in die Maschine und*...)*
  - *bleiben* (*Hydraulikleitungen bleiben angeschlossen*.)
- German expressions which have a formal equivalent of which in English as directives, although they are not conventionally used for expressing instructions or interdictions in human-machine communication texts:
  - ist/sind zu + Infinitif Aktiv (*Dabei ist zu beachten, dass die Gummilager nur einmal gewechselt werden dürfen*!)*
  - Indikativ Präsens Aktiv + *dabei* (*Der Pfeil zeigt dabei nacht recths = muß nach rechts zeigen*)*?
  - *Sollen* (*Diese Punkte sollen nicht berührt werden*)*
- German expressions that have a formal and functional equivalent in English and that are also used in human-machine interactive texts, though they can have a different

meaning in the English language system, or another formulation is used for which there is no usual German equivalent, such as for example:

  - Indikativ Präsens Pasiv (*Antrieb der Bremsenrolle wird beidseitig abgeschaltet*). This is used in German for instructions, but the direct translation into English only denotes an assertion.

- German prepositional phrases (*durch Einlegen einer Originalscheibe...*).

The best way to transfer these expressions into English will be through the use of imperative. The Indikativ Präsens Aktiv with directive function can be expressed in English with *make sure* or *ensure that*, whereas *ist/sind zu* + Infinitiv Akiv can express either obligation or possibility. In each case, it must be translated differently, either by an imperative or by a formulation such as *can*, *may*, *able* etc. Indeed, since these expressions might result tricky when translating, many of them are covered by the grammar and style rules of MULTILINT in order to avoid potential ambiguities and mistranslations (sentences marked with *).

3.7.2 Technical documentation and MT

The general assumption that technical texts, and in general all kind of specialized text with a technical terminology, are good candidates to consider the application of MT, is widely widespread. These texts are claimed to be potentially non-cultural, and therefore, universal, and to contain a univocal terminology and a clear and non-ambiguous syntax, making possible to obtain good output results of MT, given, of course, that the MT has been appropriately trained. As Schmitt (1994: 252) remarks:

> Die überwältigende Mehrheit der Sprachwissenschaftler, literarischen Übersetzer und nicht-technischen Fachübersetzer scheint sich in dessen einig darin zu sein, daß man, wenn überhaupt irgendwo, dann in der Technik von Äquivalent im Sinne einer 1:1 Ensprechung zwischen den Begriffen verschiedener Sprachen ausgehen könne – und daß dort mithin auch die Zukunft der maschinellen Übersetzung liege.

Besides, as Seewald-Heeg (1998) points out, there are also commercial reasons to use MT for this type of text. Technical documents usually accompany products that are marketed around the world, what makes translation a necessity. There are, however,

numerous products with an enormously short version cycle that cannot wait for human translation. Thus, the need to accelerate the translation process while maintaining an acceptable quality[59] makes MT technology one of the few possible solutions.

However, I must keep in mind that MT does not really translate[60] texts from one language into another, but makes a linguistic transfer from one linguistic system to another. With this assertion I want to point out at the fact that when the final goal is translating one text in one language into another language, all the issues that arise from not simply transferring words in one linguistic system to another linguistic system, will become problematic for MT.

Indeed, there are many difficulties when translating technical documentation since it is not really true that all expressions have univocal equivalents in other languages. As I have seen in the previous point, it is necessary to take into account contextual factors to translate properly request forms from German into English. However, usually MT systems do not consider these factors and only recognize and translate structures on a formal basis. Besides, sometimes concepts and terms are "culturally" marked and their meaning in one language does not match 1:1 the meaning of their, theoretically, equivalent, in another language, as Schmitt (1994) illustrates in his very informative article about the alleged unambiguousness of technical texts.

As Maillot points out in his book *La traduction scientifique et technique* (1968, cited in Gamero 2001: 30), linguistically, technical texts can present many different translation problems, such as equivalence of terms and concepts, synonyms, false friends, terminological gaps, syntax, word building, multiword terms, style, cultural references, proper names, nomenclature, transcription, transliteration, measure units, symbols, abbreviations, acronyms, punctuation and typography. Bedard (1986, cited in Gamero 2001: 32) also deals with this issue and argues that univocity, accuracy and uniformity are myths of technical translation, since technical vocabulary is as imperfect as general vocabulary. It is therefore obvious that the mere look-up (be it automatic or manual) in dictionaries is not enough to obtain a good quality translation. Indeed, terminology is

not the only important factor when translating technical texts: terms might be right, but many other factors can influence on a poor or nonsensical translation.

Despite all these remarks, technical texts are still the only text type that is successfully deployed in automated translation processes. This is due to the fact that these processes include all kind of steps to control the input that is sent to the MT system. Multilingual terminological control aims at avoiding the above mentioned problems of ambiguity, synonymy, false friends etc. Univocal terminology is thus implemented through controlled language checkers that also try to avoid difficult syntactic structures, such as ambiguous or indirect illocutionary expressions (modal verbs, particles etc.) which might have no direct formal equivalent in the target language. Indeed, many of the mistakes that can found in technical translations, be it human or automatic, are due to a poor source text. Therefore, the act of controlling the language to improve the source text will necessarily improve translation.

Further, MT is applied in texts which keep the same macrostructure or a very similar one in origin and target languages, to avoid excessive formal adaptation after linguistic transfer from one language to another.

3.8 Summary and final remarks

Technical documentation has been present in society since ancient times. However, it was during the war periods in the 20th century that technology suffered a rapid development and, with it, the creation of technical texts grew. In this chapter I have discussed the concepts of technical communication and technical documentation as well as their historical origins.

Further, I have deepened into the role of the technical writer to gain a better insight of the creators of these documents. An overview of the different document types that can be produced for technique is also analysed, with special attention to the documents that can be produced within the automotive industry. This sets the ground to explain the difficulties associated with translating technical documentation and the particularities of

this type of transfer, focusing in the language pair German-English. Finally, some notes on the relationship between technical documentation and MT are given.

4 EVALUATING CONTROLLED LANGUAGES AND MACHINE TRANSLATION

I think we have to understand that there are millions of evaluations you can do, all kinds of things you can measure, and what evaluation you put together depends on what you want to get out of the evaluation.

Eduard Hovy, in Vasconcellos, 1992

4.1 Introduction

As it has been explained in previous chapters, CLs are claimed to bring advantages in the authoring and translation processes. However, clear evidence that this is true, especially in industrial contexts where costs savings are critical, is needed in order to keep on deploying this language technology.

In this chapter I fathom out evaluation issues concerning CL and MT. Rather than evaluating CL checkers, stress is laid on evaluating CL rule suites. In the first case, the evaluation of checkers usually aims at evaluating the system as a software application with regards to precision (proportion of the number of correctly flagged errors to the total number of errors flagged), recall (proportion of the number of correctly flagged errors to the total number of errors actually occurring) and convergence (proportion of the number of automatically corrected sentences that are accepted when resubmitted to the total number of automatically corrected sentences) (Nyberg et al., 2003). Though this is not the main line of my work, I will review some of the efforts made on this matter.

In the second case, the goal of the evaluation is to establish if CL rules are effective in improving mainly the understandability and readability of texts. Further, since one of the claimed advantages of CL (especially Machine Oriented Controlled Languages) is that they improve translatability, and especially machine translatability, another possibility of evaluating the effectiveness of a CL is to perform an evaluation of the MT output. Indeed, if it is true that implementing a CL improves the quality of the source text, then the quality of texts that are machine translated should be better and there should be a correlation between the quality of the source text and the quality of the translation. This effect can be measured as a function of post-editing cost or effort, though it is also possible to measure if comprehensibility and readability have improved both in the source and the target texts. Some studies that have favoured this hypothesis are those by Aikawa, Schwartz, King, Corston-Oliver, & Lozano (2007), Reuther (2003; 2007), Roturier (2004) and Vassiliou et al. (2003). I will indeed follow this line of research since I am mainly interested in the use of CL for later implementation of MT.

This chapter introduces the topic of language technology evaluation for language processing systems (see 4.2) with an overview of the evaluation types and a historical sketch, as well as the difficulties that arise when setting up an evaluation plan regarding the selection of resources (test materials and test subjects or evaluators) (see 4.3). Further, different methods for evaluating CL rule suites and CL checkers are reviewed (see 4.4 and 4.5), to go on with an overview of different approaches to evaluating MT output as a way of validating the effects of CL implementation (4.6). Here I introduce the FEMTI Framework which I will use as a methodological starting point for my evaluation effort. I go on discussing the notion of translation quality and its implications to define a standard evaluation methodology (4.7). Finally human and automatic measures are outlined (4.8).

4.2 Evaluation of Language Technology

Evaluation has always been a subject of interest within the language technology community. This interest has been fostered due to the need to determine the

improvements made in the development of this technology. Evaluation of software consists mainly of three steps: measurement, rating and assessment (The EAGLES MT Evaluation Working Group, 1996). The first two steps are intuitively straightforward: in measurement, the selected metrics are obtained and, subsequently, for each measured value, the rating level is determined. Assessment is the final step of the software evaluation process, and the result is a summary of the quality of the software product.

However, evaluating certain language technologies such as CL and MT has some added difficulties: on the one hand, evaluation aims at measuring some attribute of something against a standard for that attribute, as White (2000) points out. However, language is not an exact science and there is not always a univocal expected correct or best result or a golden standard or behaviour against which to compare results in order to obtain an objective assessment. On the other hand, the literature on evaluation, particularly on MT evaluation, is so extensive that it is hardly impossible to give a comprehensive overview. Indeed, it has been remarked that more has been written about MT evaluation than about MT itself[61]. Besides, not all references are readily available since many of the evaluation efforts that have been carried out have been published within private institutions and corporations. This fact makes them difficult to obtain, as King, Popescu-Belis, & Hovy (2003) remark regarding the report written by Van Slype (1979), which was made publicly available shortly before the publication of their article.

All these factors have contributed to a lack of a standard methodology in NLP evaluation, in particular with translation tools, in spite of some intents to counteract this. Therefore, there have been various and multiple approaches depending on the users, context or even planned budget for the evaluation. Initiatives such as EAGLES or FEMTI, which will be further discussed in this chapter (section 4.6.3), try to tackle these shortcomings.

Before I give an overview of the historical milestones of language technology evaluation (4.2.3) and before I deepen on CL and MT evaluation in 4.4, 4.5 and 4.6, it is necessary to offer an sketch of the different approaches to evaluation of language technology as well as of the various stakeholders that can take part in an evaluation.

Though this can be applied to all kind of language related software products, I concentrate on MT and CL evaluation.

4.2.1 Evaluation Types

There are mainly two dimensions of evaluation: depending on the focus of the evaluation and the depending on its purpose and the stakeholders (context-based evaluation). In the first case, the relationship between the input and the output is alluded to as the difference between black-box and glass-box evaluation. According to King, Hovy, White, T'sou, & Zaharin (1999), "for the former, the system –however it may work internally, and whatever its output quality– is evaluated in its capacity to assist users with real tasks. For the latter, some or all of the system's internal modules and processing are evaluated, piece by piece, using appropriate measures". According to White (2003: 215), the main advantage of the black-box approach is its portability (the methods and measures are independent of the design of the system). This makes this method more amenable for the comparison of systems and to determine the current language coverage of a particular system. Contrarily, the glass-box view is more focused on determining the extensibility of the system. There are some types of evaluation according to the purpose that are more appropriate for the black-box method, whereas some of them are more appropriate for the glass-box method.

With regards to the evaluation purpose and the stakeholders behind the evaluation effort, I now review some of the main evaluation types based on White (2003) and FEMTI (King et al., 2003). White (op.cit) based his work on Arnold, Sadler, & Humphreys (1993); Church & Hovy (1993); King, Wilks, Allen, Heid, & Albisser (1993), who assumed that "as there can be no single general purpose machine translation system, so there can be no single purpose evaluation methodology", and augmented it by the models of Van Slype (1979) as well as his previous work (White, 1994, 1992, 2000; White, O'Conell, & O'Mara, 1994; White, O'Conell, & Carlson, 1993):

- **Feasibility evaluation** studies the possibility that a particular approach has any potential for success after further research and implementation. Especially indicated for researchers and the sponsors of research. According to White (2003), this type of evaluation is highly automatisable.

- **Internal evaluation** tests whether the components of an experimental prototype, or pre-release system work as they are intended. The main goal of this evaluation is to show that the system is actually improving as a result of development. This type of evaluation can be carried out as black box and glass box. This type of evaluation is also known as **progress evaluation**.

- **Declarative evaluation's** purpose is to measure the ability of an NLP system to handle text representative of actual end-use. Other names to designate this evaluation type are **adequacy evaluation** or qualitative evaluation. It purports to measure the actual performance of a system external to the particulars of the feasibility of the approach or of the development process. This evaluation type generally tests for the attributes of intelligibility and fidelity, which generate results with a high degree of subjectivity.

- **Operational evaluation** generally addresses the question of whether an NLP system will actually serve its purpose in the context of its operational use, being the cost-benefit factor the main one. According White (op.cit.), the more fundamental question to ask for operational use is whether the NLP system enhances the effectiveness of the "downstream" task, or whether the end-to-end process is better off without it.

- **Usability evaluation's** purpose is to measure the ability of a system to be useful to people who are actually going to operate it. Usability for an language technology application will measure such things as the time to complete a task, the number of steps required naturalness of navigation, how easy it is to learn etc.

FEMTI also adds **requirements elicitation**, which is "often an iterative process in which developers create prototypes in order to elicit reactions from potential stakeholders", and **diagnostic evaluation**, the purpose of which was purpose was "to discover why a system did not give the results it was expected to give".

As it can be observed, a classification of different types of evaluation will depend on the criterion used in order to establish the differences among the different types of approaching the evaluating task. Regarding the specific case of MT, evaluation methodologies are usually classified from the point of view of the context. It is obvious that not all users need to know the same things about an MT system or approach. A researcher needs to know if the system he is developing is improving, while a venture capitalist who wants to get involved financially might be more interested to know how profitable the system is in terms of market sales. In any case, although these evaluation types were born in the context of MT evaluation, they can also be applied to the evaluation of other language technologies.

With respect to black-box and glass-box evaluation methods, the former scenario is usually associated with the system user, while the system developer is obviously associated with the latter. Since I face evaluation from a user perspective, I will mainly use black-box evaluation methods.

4.2.2 Evaluation Stakeholders

Once the purpose and thus the type of evaluation is defined, other factor that can be controlled is the people that are going to carry out the evaluation and the people interested in the results of such an evaluation. White (2003: 209) reviews the different stakeholders in the MT evaluation process and divides them into end-users, managers, developers, vendors and investors.

End-users include translators, editors, monolingual information consumers and office automation users. **Managers** comprehend operational and procurement managers.

Researchers and *productizers* make up the group of **Developers**, whereas **Vendors and Investors** can be either research organizations or venture capitalists.

All of them will have different interests in the evaluation of a certain tool and their needs and requirements must be taken into account to carry out a well-designed evaluation plan.

4.2.3 Historical Sketch

The book written by Spärck Jones & Galliers (1995) book offers an extensive review on NLP evaluation, including both a thorough analysis of what it involves and a comprehensive review of what has so far been done. One of the first attempts in the field of language technology evaluation was carried out by the Automatic Language Processing Advisory Committee (ALPAC) in order to evaluate the state of the art on MT research (ALPAC, 1966). Focus was put on speed, cost and quality of MT compared to human translation[62]. The most used method was rating scales aimed at measuring readability, fidelity and comprehensibility. Emphasis was laid on demonstrations, with little attention to developing a comprehensive methodology for different scenarios. The result of this evaluation was extremely negative as to what could be hoped from MT systems in the short or medium term and although the results and the methodology of this report can be discussed, the ALPAC evaluation effort must be considered a pioneering effort if only because it emphasised the importance of good evaluation methodologies.

Another remarkable language technology evaluation effort was the evaluation of TAUM-AVIATION, a machine translation pilot system based at the University of Montreal. The system was evaluated in 1980 by the Canadian Secretary of State Department and its Translation Bureau and as a result the project was discontinued. SYSTRAN also undertook an evaluation carried out for the US Air Force in 1979-80, which was basically a diagnostic evaluation to measure the impact of improvements made in the MT engine (Isabelle & Bourbeau, 1985).

Finally, it is worth mentioning the JEIDA Report and the JTEC Panel Report, which reflect the efforts being made for Japanese and include some general comments on evaluation. The JEIDA report recognises that for operational evaluation the environment factors (document type, intended output use etc.) and the encompassing setup as opposed to the system alone are extremely important, as well as the economic perspective. This is thus one of the first initiatives considering contextual factors for setting up an evaluation plan and constitutes a precursor of initiatives such as FEMTI that will be analysed in section 4.6.3. of the present chapter.

A detailed account of all these evaluations can be found in Falkedal (1991) and Galliers & Jones (1993: 78-81). This perspective already sets the ground of context-based evaluation, which I will deal with in 4.6.3.

Since then, and especially in the last 20 years, evaluation has experienced a renaissance, as the US DARPA (now ARPA) initiatives and the projects conducted within the LRE[63] show. The conferences sponsored by DARPA (the US Defence Advanced Research Projects Agency) during the 90s consisted normally of a specific evaluation exercise and reporting meeting and covered topics such as Message Understanding (MUC Conferences), Spoken Language Systems (SLS), Dialogue Systems and Speech Recognition, Text Retrieval and Machine Translation. Also in the 90s, the European Commission co-funded a series of projects under the LRE Framework for projects on Language Technology Evaluation. One of the most important and largest projects related to evaluation was EAGLES[64] (Expert Advisory Group on Language Engineering Standards). The EAGLES-I[65] project, based on the ISO/IEC 9126 standard for the evaluation of software, was initiated by the European Commission within the DG XIII Linguistic research and Engineering programme with the aim of providing means and recommendations for de facto standards for:

- Creating and manipulating very large-scale language resources (such as text and speech corpora or computational lexicons);
- Manipulating knowledge;

- Assessing and evaluating resources, tools and products.

The project, which ran from February 1993 to May 1996, was formed by five groups: Text Corpora, Computational Lexicons, Grammar Formalisms, Evaluation and Spoken Language. Due to the nature of this work, I am mainly interested in the results of the Evaluation Working Group.

A further phase of this project, EAGLES-II, was built on the experience and methodology of EAGLES-I, and concentrated in disseminating results and finding real-life applications and evaluation needs. It spanned over 1997-1998 and ended in spring 1999.

The two final reports of both projects (The EAGLES MT Evaluation Working Group, 1996, 1999) constitute a valuable and very complete source of information on general principles about software evaluation in general and, in particular, with language technology evaluation. These reports intended to establish a framework for evaluating NLP systems, in terms of a hierarchically structured classification of features and attributes, where the leaves of the hierarchy were measurable attributes, to which specific metrics were associated (Hovy, King, & Popescu-Belis, 2002: 47). Taking as a base the standard ISO/IEC 9126 published in 1991, an international standard for the evaluation of software quality, the EAGLES-I group worked mainly with three types of system: writer's aids, translator's aids and knowledge management systems. Further, the work distinguished between progress (or internal evaluation), adequacy and diagnostic evaluation, putting the focus on adequacy evaluation, understood as "the activity of assessing the adequacy of a system with respect to some intended use of that system" (The EAGLES MT Evaluation Working Group, 1996: 7). This is made from a "consumer report (CR) paradigm" perspective, with a costumer wondering which of a group of market products are good buys for what he or she wants, comparing a translation system with, for instance, a washing machine.

The final report of EAGLES-II establishes the theoretical and methodological grounds for natural language processing systems evaluation, aiming at giving concrete

guidelines for designing test-beds for the different systems that have been mentioned before. The report deepens in the issues of software quality evaluation, establishing a function of three components as the basis for a comprehensive evaluation: products (as the objects of evaluation); descriptions of classes of users (as the customers of evaluation); and descriptions of attributes of systems potentially of interest to classes of users coupled with metrics which are measured with a value. An attribute-value pair represents a feature. When applied to a product, each feature provides a value for that product. Comparing the features list of a product, we can see if it fits the needs of the user or not.

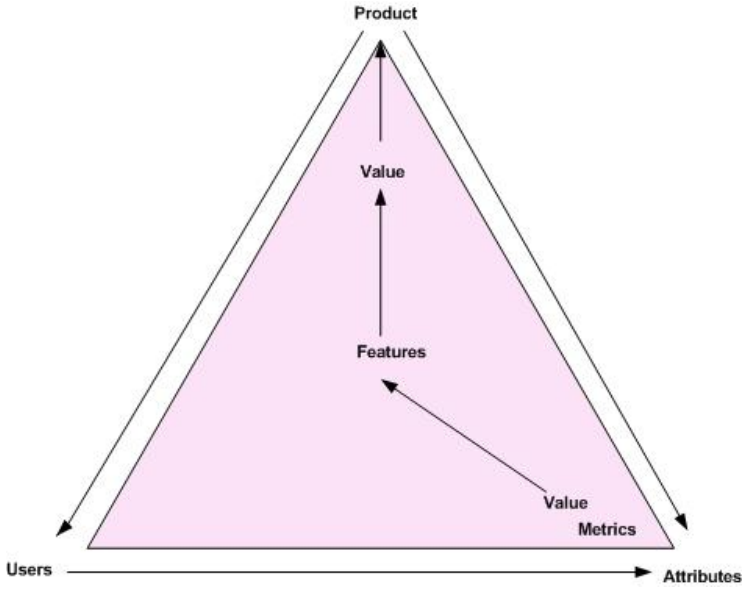

Figure 8: Evaluation parameters

Indeed, as a general framework, the report considers the design for an evaluation as involving four steps:

- Defining the relevant (product) quality characteristics;
- Defining the attributes pertinent to each characteristic;
- Defining the measures to provide values for each attribute; and
- Defining the methods for applying the measures to determine actual values.

For instance, a quality characteristic of a writer's aid system could be functionality interpreted as not flagging errors when no such error is present, or detecting misspelled words which do not correspond to a legitimate form of the language. An attribute of this characteristic could be "no false errors are flagged", or "all misspelled words are detected", which can be measured by means of precision and recall techniques. Obviously, the definition of attributes at an appropriate level of granularity can be very complex, since they may be of different types (facts, features, tests, judgements) and have values of different types such as quantitative, (absolute or relative), qualitative, Boolean etc.

A series of appendixes complete the report. Appendix D is dedicated to Evaluation of Writer's Aids and Appendix E deals with the evaluation of translation aids. In these appendixes, the methodology for evaluation is explained, detailing the description of the user models, text types, languages used etc. Annex F deals with different user profiles that can be interested in language translation technology evaluation. Different aspects are tackled, such as languages involved, text types, nature or amount of the activity where translation is involved.

The TSNLP (Test Suites for Natural Language Processing) project is also an LRE project. It started in December 1993 and ended in October 1995. The partners of the project were researchers from the University of Essex, DFKI GmbH (Saarbrücken) ISSCO (Geneva) and Aérospatiale (Paris) and it was concerned with the design and use of test suites for NLP processing. More information on the project can be found in (Balkan (1994); Balkan, Arnold, & Meijer (1994); Balkan & Netter (1994); Balkan, Netter, Arnold, & Meijer (1994); Fouvry, Balkan, & Arnold (1995); Lehmann et al. (1996).

One of the inheritors of the TSNLP project was the DiET project[66], the aim of which was to develop "a comprehensive software package for the construction, annotation, customisation, and maintenance of structured reference data for the evaluation of NLP applications" (Netter et al. 1998), since one of the main difficulties of evaluations is the lack of structured and classified test and reference data. Klein, Lehmann, Netter, &

Tillmann (1998) offer an extensive review of how this method could be applied to MT evaluation. Some other projects contemporary or predecessors of EAGLES, DiET and TSNLP were TEMAA[67] or COBALT[68], which also dealt with issues concerned with language technology evaluation.

4.3 Selection of resources

An important step when designing an evaluation plan is the selection of resources: metrics and tools, materials and evaluators have to be carefully planned and selected so that results are as objective and representative as possible. In this section, I examine some of the factors that need to be considered when selecting the needed resources that establish a good evaluation plan.

4.3.1 Evaluation Tools

There are a number of tools to carry out evaluations in the NLP domain. King & Falkedal (1990: 212) distinguish three kinds of approaches. First, one of the most extended are quizzes and scales to obtain ratings on aspects such as intelligibility, fidelity or clarity. A second approach is to count the number of errors by counting the number of corrections made by a post-editor. Thirdly, to weight these errors according to pre-established classification schemes. According to these authors, all these approaches suffer from two major drawbacks: on one hand, the results do not really give relevant information for an assessment of the actual acceptability of the translation quality to the final users and, on the other hand, the metrics do not provide the necessary data on how to improve or modify the system. Moreover, one of the most claimed deficiencies of these human evaluation methods is their subjectivity, which can deduct credibility to the results.

However, and despite all these disadvantages, I think that the results that can be obtained from these types of evaluation are far more illuminating than those obtained by automatic evaluation measures, though the latter also involve certain advantages. The

differences between human and automatic evaluation methodologies are discussed in 4.8.

Another type of evaluation procedure are performances or tasks, where the user is asked to perform a relevant task either with the software to be evaluated, or with the output from it. It is more difficult and probably more costly to set up such an experiment, but it is a more realistic experiment, since procedures are written to be performed, not to be quizzed. However, this type of tests require a more complex setup, since the evaluator has to replicate the situation in which the tasks need to be performed, and this is not always feasible.

4.3.2 Test Materials

Another of the main issues when setting up an evaluation plan is to decide what kind of material is going to be used as the basis of the evaluation. There are two traditional approaches for language technology: test suites and text corpora. On the one hand, test suites consist of a more or less systematic collection of specially constructed linguistic expressions (with optional associated annotations or descriptions) with the aim of tackling specific phenomena. For instance, the developers of a MT system might have a test suite containing problematic sentences that they test in every release to see if the system has improved or not. Balkan (1994) gives a detailed overview on test suites and test suite construction within the TLSNP project. Some examples of real use of test suites can be found in Fouvry & Balkan (1996), who concentrate on test suites for controlled language checkers. The paper describes the particularities of testing language checkers, focusing on syntactic phenomena but also adding test items for semantics, lexicon and punctuation. The goal of this experiment was to carry out a diagnostic evaluation of the test checker with regards to its functionality (progress evaluation). King & Falkedal (1990) also talk about using test suites as well as their advantages and disadvantages.

Text corpora, on the other hand, consist of naturally produced texts, where linguistic phenomena are not as controlled as in test suites, but give a more realistic view of the quality of the system.

Test suites are more appropriate for testing specific phenomena, such as anaphora resolution in MT, while test corpora are ideal to test how the system performs with real life texts. Therefore, test suites are particularly well-suited for diagnostic evaluation, while test corpora are necessary for adequacy evaluation, where the overall performance of the system is tested.

The EAGLES MT Evaluation Working Group (1996: 36-38) distinguishes not only between test suites and text corpora (also called test sets), but includes test collections as test materials. A test collection is defined as "a set of inputs associated with a corresponding set of expected outputs". Typically, the elements of a test collection are divided into training sets and test sets. Despite the great effort needed to construct such test collections, this type of test material is very common and has been used in the evaluation of parsers. This reference-based evaluation approach is used with automatic metrics to estimate the quality of MT output (the input) comparing it with human references (expected output).

4.3.3 Recruiting Subjects and Raters

An important issue when designing an evaluation plan which will entail human evaluation is the recruiting of subjects and raters.

On the one hand, human evaluation is difficult and time consuming, and results are not always reproducible due to human subjectivity. The problem of subjectivity can be diminished by offering training sessions or giving specific directions as how to evaluate or rate. On the other hand, it is usually necessary to pay these subjects and raters, something that can result problematic if I have limited funds for the project. An alternative solution can be to engage students, though these are not always the most appropriate users, since they lack the experience and know-how of expert subjects. The

best option is usually to involve real users of the end product, such as professional technical translators and technicians (Houlihan, 2009: 14). Apart from the user background, some issues related with human evaluation are the number of evaluators needed to obtain statistically significant results, their experience in previous evaluation efforts, the time left for the evaluation, their bias towards a certain technology etc.

4.4 Evaluating CL Rule Suites

CLs are widely used within the aircraft industry and increasingly in other areas. This fact seems to underline their practical relevance. However, it is not easy to determine the effects of CL rules, especially if they are HOCL, and though the number of empirical studies on this subject is growing, there does not seem to be a standard methodology to assess their validity. I will thus begin by discussing the difficulties of evaluating CL rules and then I will sum up some of the available studies.

Nyberg, Mitamura, & Hujisen (2003) point out a series of variables that must be taken into account when carrying out an integral evaluation of CL rule suites. I classify them according to the resource type they are:

	Tools	Materials	Evaluators
The number of texts and test persons used in the evaluation		X	X
The amount of time available to the test persons to execute the test	X		X
The complexity of the texts and their subject matter		X	
The degree to which the test persons are familiar with the subject matter and the CL			X
Whether they prefer the CL texts to the uncontrolled ones			X
In how far they are more inclined to use the texts			X
And whether they are native speakers or not			X

Table 4: Variables for CL rule suite evaluation according to Nyberg, Mitamura & Huijsen (2003)

As it can be observed, most factors have to do with the characteristics of the evaluators, whereas the materials and the tools are considered to be not as problematic. Besides, Holmback, Shubert & Spyridakis (1996) mention following elements and again, evaluators and their features are the main issue of concern:

	Tools	**Materials**	**Evaluators**
The language ability of the testers both for native and non-native			X
The place of residence of testers (especially non-native speakers living in a different country)			X
Time used for understanding and translating	X		X

Table 5: Variables for CL rule suite evaluation according to Holmback, Shubert & Spyridakis (1996)

However, it is not enough to identify all these different elements. One of the main problems is how to quantify these variables, since either they are very difficult to measure (e.g. "the degree to which the test persons are familiar with the subject matter and the CL") or the results can be very subjective (e.g. "whether they prefer the CL texts to the uncontrolled ones"). Another critical factor is that many CL rules are loosely defined in a very informal way and it is often unclear which part of the definition of the CL should be applied to determine conformance. Owing to this, it is very often unclear what the contribution of each individual writing rule is to the overall effect of the CL, unless they are formalized and applied computationally with the help of a CL checker.

The first known evaluation efforts of CL rule suites were carried out at The Boeing Company (Shubert, Holmback, Spyridakis & Coney, 1995; Spyridakis, Holmback & Schubert, 1997). The first experiment consisted of comprehensibility[69] tests with questions about the content of four documents (two SE compliant and two non-SE compliant), suggesting that "… using SE significantly improves the comprehensibility of more complex documents. Further, readers of more complex SE documents can more easily locate and identify information within the document." The second experiment consisted of testing the effects of the CL by translating two documents in two different versions (SE/non-SE) by native speakers. The documents had the following

characteristics: no less than 450 words, no more than 1000 words, and no more than 15% passive voice. The users were university students native in one of the experiment languages. Furthermore, a baseline translation made by Boeing employees was also provided for evaluation. Translations were then assessed by native speakers on several parameters: accuracy of the translation, style match with the original document, ease of comprehension, number of major and minor mistranslations, and number of major and minor omissions. The article does not specify if the students or the evaluators of the translations had language or technical-related studies or degrees, which leaves a relative uncertainty on the quality of the translations and their assessments. These experiments are also described by Holmback et al. (1996) who concludes that, through the translatability results were less clear-cut than the comprehension results, a significant improvement for languages similar to English, such as Spanish (in contrast with Chinese) could be stated.

As it can be observed, the first CL evaluation attempts already use translation as an evaluation methodology, including reference translations, which will subsequently give place to the idea of automatic evaluation of MT, as we will see in 4.8.

Furthermore, there is a number of studies concerning the evaluation of AECMA SE. A study carried out by Chervak, Drury, & Ouellette (1996) compared comprehensibility of SE and non-SE versions of work cards by 175 aircraft-maintenance technicians and stated that complex documents (according to general readability measures and type of task) written with SE obtained clearly superior accuracy than easier documents.

All these studies concluded that the use of a CL can significantly improve comprehensibility, especially among non-native speakers, and improvements in the ease and quality of human translation can be observed too. However, none of these studies provided information as to which specific rules made texts more comprehensible or translatable than others. Furthermore, the studies concentrate on human translatability and no reference to machine translatability is yet made.

Though there might be more research regarding CL rule suites evaluation, due to the rather industrial application and, thus, the private character thereof, they have not all been made publicly available. I have tried to offer an extensive review of all the publications available, though I cannot guarantee that there are not other case studies relevant to this work.

4.4.1 Metrics: Readability, Understandability and Translatability

As we have seen in the previous sections, the main goals for deploying a CL in the creation of texts are, on one hand, to improve readability and understandability [70] and, on the other hand, to improve translatability. However, the definition and limitation of these concepts might result somehow ambiguous and, therefore, it is important to give some guidelines as to what these concepts imply.

4.4.1.1 Readability and Understandability[71]

The concepts of understandability and readability are closely related and they belong to the cognitive sciences. DuBay (2004) gives a comprehensive overview[72] of the principles of readability and defines it as "what makes some texts easier to understand than others". Dale and Chall (1949, in DuBay 2004: 3) provided a definition that details what readability is for them:

> The sum total (including all the interactions) of all those elements within a given piece of printed material that affect the success a group of readers have with it. The success is the extent to which they understand it, read it at an optimal speed, and find it interesting.

There are indeed many different definitions of this concept and most of them include also the fact that if a text is readable, readers will succeed in understanding it. Therefore, readability seems to be a prerequisite for understandability, though this is not always applicable: a text might contain punctuation and grammatical errors, making the text difficult to read and be, however, understandable because the terms with semantic load have been correctly translated.

There are a number of formulas that have been developed to measure readability and though each of them has its particularities and many of them were designed to measure literacy levels rather than to test the readability of technical texts[73], I can state that, generally, the variables that can be measured to calculate the readability of a text are:

- Density of one-syllable words, sentences, pronouns, prepositions and modifiers density in a text (usually per 100 words)
- Word and sentence length
- Vocabulary complexity

Indeed, as we will see in the next section, CLs usually contain a number of rules intended to control these variables, such as "Do not use sentences longer than 20 words", which controls sentence length, or "Do not make sequences of more than four nouns", which affects modifiers density.

However, most of these formulas are based on the characteristics of the English language, with more one-syllable words than other languages such as German or Spanish. Therefore, they are not directly applicable to measure the readability of texts written in other languages than English. Other problems that these formulas present is that they do not account for the structure of the elements in a text, being the result the same for a well-constructed text than for a text where all sentences have been switched. Further, the value that the formula gives does not indicate where the problems are, though this is a common problem when evaluating language.

Further, there might be also other factors related to the reader, such as his reading level, his literacy level or subject matter knowledge. However, I consider that these factors are rather subjective and belong to the cognitive capacity of understanding the text. As I have said before, readability is closely related to understandability, but they are not synonym concepts. While readability is the property of a text, understandability has more to do with the cognitive process of comprehending the information contained in the text. Making an analogy with the information theory and the concepts of data and information, I could say that readability has to do with how the data are presented, while

understandability is related to the fact of how these data are interpreted and converted into information.

Of special interest for this work is the contribution made by Göpferich (1998: 198-251) regarding understandability. She dedicates chapter 8 of her book to this issue, especially with regards to technical documents. She deals with the new theories of cognitive constructivism, where the reader assimilates information not only through bottom-up processes (from text into the cognitive structures of the reader), but also through top-down processes, that is, relating the new information gained by bottom-up processes with previous information stored in his long-term memory. The different approaches presented are propositional models, network models, semantic macrostructures, schema-theoretical approaches and mental models. Besides, she introduces the instructional psychology approaches, including the Hamburger Modell[74] and the comprehensibility concept by Groeben. The Hamburger Modell presents four dimensions of text understandability:

1. Simplicity
2. Arrangement-Structure
3. Brevity-Conciseness
4. Stimulating elements

These four dimensions are measured with the help of attribute pairs (e.g. simplicity versus complexity) with a scale that goes from +2 (indicating a positive aspect of the attribute) to –2 (indicating a negative aspect). It is important to notice that, in the third dimension, a punctuation of +2 is as negative as –2, since an extremely concise and brief text does not guarantee a better understandability. With stimulating elements, the authors of the Hamburger Modell refer to measures taken by the author to call the reader's attention, such as exclamations, rhetorical questions, colloquial language etc. Though all these measures might seem logical when applying understandability to texts intended for pupils, they are clearly out of scope in technical documentation. Therefore, these dimensions, though generally applicable for all kind of texts, must be adapted to the particularities of each text type. The authors propose the following model to

measure the understandability of a text. The following example would show an ideally understandable text:

SIMPLICITY ++	ARRANGEMENT STRUCTURE ++
BREVITY CONCISENESS 0 or +	STIMULATING ELEMENTS 0 or +(+)

Table 6: Hamburger Model for Understandability

However, this model, as I have mentioned, is too general and subjective. It needs to be adapted for technical documentation, since it does neither take into account different text functions (and thus different text types) nor the previous knowledge of the readers. The former aspect, the function of the text, might have a dramatic influence on the comprehensibility of a text: an informative text needs different resources than a text with a conative function. With regards to the previous knowledge of the reader it is essential to know what are the interests and background of the reader. As a consequence, this model is mainly based on the text and does not consider the connection with the its receiver.

Adapting the Hamburger Model to technical texts, Göpferich (1998: 238-247) presents some guidelines for all of the dimensions with a special focus on instructional texts. For the simplicity dimension there are guidelines regarding the typographical layout, the nominal style, which should be avoided, and the position of attributes and modifiers, which in German usually happen to be before the noun. Other aspects such as anaphora, the use of slang, univocal terminology, compound terms or ambiguous structures such as conditional sentences without conjunction and accumulation of genitives, prepositions or conjunctions are covered under this dimension. Under the second dimension, arrangement and structure, guidelines with regards to the order of the elements are given. The dimension conciseness and brevity comprises guidelines for the avoidance of redundant structures and semantic weak verbs. Finally, Göpferich gives

general recommendations to increase the motivation of the reader for the dimension dealing with stimulating elements. It is interesting to observe that many of the recommendations and guidelines given to improve comprehensibility are part of the specifications of many CLs[75]. This in a certain way accounts for the fact that controlled languages aim at improving understandability.

Another study that tackles readability in CLs is that by Cadwell (2008), who suggests a conceptual terminological framework to establish some rigour in the different metrics, since these have been used randomly in the literature. Cadwell uses as a distinguishing criterion the focus of the different metrics: the text itself, the reader and the results or consequences of the text. He deals with the relationship between CLs and readability, proposing a set of rules to improve it in English. These rules are divided in four categories: textual/pragmatic, syntactic, grammatical, and lexical. But he also adds extra-linguistic variables: motivation, reading ability, interest in the topic, relevance of the topic, familiarity, prior knowledge, and testing conditions. Indeed, DuBay (2004: 39) points out that many experiments in the field of CL do not achieve the expected results because they fail to control such variables.

4.4.1.2 Translatability

It is generally claimed that one of the benefits of CLs is the improvement of translatability. From a general point of view, the issue of translatability has been used in translation theory since the nineteenth century together with the birth of language as a science. There have been three main approaches: the universalist one, claiming that the existence of linguistic universals ensure translatability; the monadist one, that states that each linguistic community interprets reality in its own way and therefore pure translatability is not possible; and the deconstructionalist one, that questions the notion of translation as transfer of meaning (de Pedro, 1999)[76].

Many of the theoretical works dealing with the notion of translatability deal with literary texts and do so in a rather philosophical way. At present, there is a tendency to presuppose that most texts are translatable.

The Encyclopaedia of Translation studies (M. Baker, 1998: 273) defines translatability as "the capacity of some kind of meaning to be transferred from one language to another without undergoing radical change". But the concept of what is translatable and which criteria constitute exactly translatability highly depends on the context. It seems undeniable that some texts are more easily translatable than others. In general, texts with an aesthetic function are more difficult to translate than texts that are purely informative or conative, which are the ones that I will be dealing with in the course of this work. Technical texts, in general, do not contain many cultural elements and are, therefore, easier internationalisable and therefore, easier translatable. It also seems obvious that translation between close languages might be easier than between languages that have completely different linguistic roots. Translatability will thus also depend on the languages involved, though it is desirable, when defining this concept, to be as language independent as possible.

As a consequence it seems reasonable that different requirements are needed for the translatability of different text types and eventually different languages. Furthermore, as it happens with the pairing readability-understandability, translatability is not only a quality of the text, but it also depends on the cognitive abilities of the person dealing with it, that is, the translator. Indeed, when defining translatability, it must be clear who or what is going to transfer it from one language into another: a human being or a machine translation system. In either case, requirements might be different, some complementary but some certainly divergent. In this work I am especially interested in machine translatability, though requirements for human translatability will also be considered.

4.4.1.3 Index-based Approaches to Translatability

Some authors have dealt with the concept of translatability regarding controlled languages and technical texts, as well as machine translatability[77].

Gdaniec (1994) introduces the concept of a "translatability index" developed at Logos Corporation for the Logos translation system. This index aims at automatically

assessing the suitability of a given original text for the MT engine. This index is based on statistical properties of the text, such as sentence length or degree of syntactic complexity, rather than on syntactical parsing of the sentences. The index is therefore based on parameters that are measured in readability indexes too and suffers from the same weaknesses: the index is a numerical value that does not give us information about the real issues that might cause problems in translation or the output quality which is to be expected. Besides, the experiment also included a Quality Index, manually obtained through the evaluation by human translators that measured the quality of the raw MT output. Both indexes were compared to see if there was any correlation, taking into account aspects that might affect the index, such as differences in the source languages or dictionary up-to-dateness. Bernth (1999b) and Bernth & McCord (2000) contributed to translatability by developing the "Translation Confidence Index", which is described as "a function that assigns to each source language segment a number that estimates the confidence that the MT system can translate that segment well". For both indexes the authors warn that they measure the translatability by a particular MT system, and do not represent a general translatability measure for any source text and any MT system. The difference between Bernth & McCord's index and that of Gdaniec is that the former takes into account linguistic phenomena (rather than only statistical counts as Gdaniec's index does), including also a number of factors such as language pair and language distance that might influence the quality of the translation.

Underwood & Jongejan (2001) also developed a tool to assess the machine translatability of English source texts. This approach distinguishes itself from the former two in that it assesses the translatability of whole texts, but also of single sentences within the text. Further, the tool designed by these authors assesses translatability by a shallow and rapid analysis, leading to a trade-off between robustness and speed on the one side, and accuracy on the other. The tool does not only outputs a Translatability Index (TI), but it also does the analysis of each sentence so that the user can interpret the score.

In all of these three approaches, the translatability is calculated on the grounds of so-called translatability indicators which are considered to have a negative effect on the

quality of MT. A numerical scale is taken and points are subtracted if translatability indicators appear in the text, which might have different weightings depending on the relative effect of the indicator on the translation process. In any case, it is clear that, the fewer translatability indicators, the better is the text suited for MT. The TI and QI by Gdaniec are based on a scale of 1-7, where the numbers indicate different actions to be taken (Gdaniec, 1994: 105). In the Translation Confidence Index by Bernth & McCord (op. cit) the index reports a value between 10 and 0. Underwood and Jongejan (op. cit) apply following formula, where m_{ik} are the occurrences of an indicator i in sentence k. A value between 0 and 1 is obtained depending on the number of translatability indicators found in one sentence:

$$I^{ik} = \frac{m_{ik}}{1 + m_{ik}}$$

In general these approaches are justified by the fact that MT output quality can be very poor, so sometimes it is faster to apply a more traditional method of human translation. Such indexes aim at indentifying in advance which of the two options might be more adequate.

4.4.1.4 Other Approaches to Translatability

This group of works dedicated to translatability aim at giving some recommendations or guidelines to write for translation, rather than trying to measure if a text is translatable or not.

Jan H. Spyridakis et al. (1997) report the results of an experiment carried out to check the human translatability of Simplified English (SE) compliant texts versus non-compliant texts. Translatability is defined as "quality and ease of translation", a rather neutral and general definition that does not contribute too much to this concept. The type of texts used for the experiment was maintenance procedures in the airline industry. For this purpose they translated SE-texts and non-SE texts into Spanish, Chinese and Japanese and they rated the translations with regards to accuracy, style and comprehension. They also counted major and minor errors, as well as major and minor omissions. The results were that SE-compliant texts achieved indeed better translation quality results than non-SE texts, at least for Spanish and Japanese (for Chinese no major difference could be found).

Kohl (1999) deals with the concept of syntactic cues to improve readability and translatability. These are defined as "elements or aspects of language that help readers correctly analyze sentence structure and/or to identify parts of speech" (ibid: 149). These syntactic cues include suffixes, articles, prepositions, auxiliary verbs and word order. It must be taken into account that the syntactic cues that Kohl presents are based on the English language and should be adapted for other languages with different syntactic cues. Very often technical writing and controlled languages are associated with conciseness and brevity. However, writing clearly and for translatability is often related with adding more words to eliminate ambiguity, enabling thus readers, translators and MT systems to analyze sentence structure more quickly and accurately. The difference between the approach presented by Kohl and CL is that the former does not impose restrictions on vocabulary nor on the range of grammatical constructions that are permitted. I could thus classify them as stylistic recommendations for the sublanguage used in technical writing. However, the author also states that syntactic cues are not always the best solution, but rather restructuring the sentence completely is needed. Therefore, his proposal is also a way of trying to control the language. Indeed, when talking about the syntactic cues procedure he author reckons that "I wouldn't be surprised if controlled language tools such as the Carnegie Group's ClearCheck or Cap Gemini's CLarity search for some of the same things that the syntactic cues procedure draws attention to".

Bernth & Gdaniec (2001) explore different ways of improving translatability related to MT, what they coin as MTranslatability. Among the methods exposed in their paper, they deal with basic awareness of how to write for MT, user guidance during translation (or Interactive MT) and CLs. Their aim is not to define rules for a CL, which are, according to them, usually tightly tied in with a specific MT system, but to give general recommendations. There are 26 rules that are divided in five groups:

- Grammatical structures: This is due to the fact that most MT systems are based on syntactic analysis. Therefore, ensuring that texts are grammatical, it is more likely to obtain a better output.
- Ambiguous structures: There are a number of ambiguous constructions that can be avoided. Usually, these ambiguities result from using a telegraphic style and can be removed by deploying what Kohl (op. cit., 1999) calls "syntactic cues". Some of the structures that can cause ambiguity and can be avoided are coordination, postnominal modifiers, pronouns and, especially for English, *ing*-words.
- Stylistic issues such as sentence length, metaphors and other idiomatic structures (idioms, slang, dialects), ellipsis, the use of passive voice and segment independency.
- Orthographic issues include both the use of punctuation, and other issues such as the use of brackets to indicate plural or gender variations, the use of symbols such as / and, though they include it in a different group, spelling issues.
- File Characteristics: This is the last group and includes rules intended to revise the format and characteristics of documents (e.g. content in graphics, use of mark-up).

Following the ideas of these two authors, Reuther (2003; 2007) explores the relationship between readability and translatability. The idea that these concepts are related might seem logical: if a text is easier to read and thus, to understand, it will be easier to translate. However, Reuther states that, though there are rules that might contribute to improve both readability and translatability, other rules only help improving translatability, especially MTranslatability, even worsening the readability of texts.

There were 7 rules that were considered a must for translatability, but irrelevant for readability:

- Avoid complete sentences in brackets
- Avoid unambiguous genitive constructions
- Avoid parenthesis starting with d.h. (corresponding to i.e.)
- Avoid additional plural forms in brackets
- In a condition/action sentence the condition part should precede the action part
- Avoid passive constructions (without by-agent)
- Avoid double negation

This incompatibility confirmed the assumption that rules that apply to readability also apply to translatability, but not vice-versa. Further, she also concluded that T-rules were more restrictive than R-rules.

These findings correlate to some extent with the conclusions of Bernth and Gdaniec (op. cit., 2001), who also argue that human readability does not always match MTranslatability scores. Shorter words and segments might be easier to read, but they are also more ambiguous and difficult to translate with an MT system. Nevertheless, the results of both studies can only be compared on a more abstract level, since some rules describe language specific phenomena.

Wells-Akis & Sisson (2002) present a case study at Microsystems Inc, where authors used the tool SunTM Proof, created by the Institute of Applied Information Sciences (IAT) and based on the tool CLAT (Controlled Language Authoring Tool), to improve the translatability of their written material. The authors use the terms translatability-checking application and controlled language system as synonyms, what advocates for the idea that the use of a controlled language improves translatability. However, this article does not offer any major contribution to the concept of translatability apart from

the statement that the rule that limits the sentence length to maximum 25 words is the most effective.

Finally, it is worth mentioning O'Brien (2003b) who makes a review on contributions to translatability, comparing in detail the linguistic features used in the calculation of Translatability Indexes proposed by Bernth & McCord, Gdaniec and Underwood & Jongejan. She also analyses the contributions by Bernth & Gdaniec, Kohl, Spyridakis et al. and Wells-Akis & Sisson, concluding that the generic approach by Bernth & Gdaniec is the most suitable for her study. The goal of this work, however, is to evaluate a given controlled language with existing rules. Therefore, here I aim at giving a general overview on the issue of translatabiltity, and not to analyse the different approaches to determine which suits best to my needs.

Later on, O'Brien (2005) deepens in this issue and proposes a methodology to measure the translatability of documents by calculating the post-editing effort needed for MT output. She picks up the idea of translatability indicators proposed by Underwood and Jongejan renaming them as "negative translatability indicators" or NTIs, which are defined as "a linguistic feature, either stylistic or grammatical, that is known to be problematic for MT". After discussing the rather inadequate appropriateness of the TAP methodology (Think Aloud Protocols) to study the cognitive processes when post-editing, the author identifies two alternative methodologies: the keyboard-monitoring software Translog and CNA (Choice Network Analysis). With this method, the cognitive effort needed to post-edit MT output is measured, relating it to the translatability of a text: the more cognitive effort, the more negative translatability indicators and thus, the less translatable is the text. However, this method does not prove to be completely effective, since there are also source-text elements that would not normally belong to the NTIs but caused increased processing. Besides, other elements that would usually be identified as NTIs did not put demand on cognitive effort. Therefore, the correlation between both aspects (presence of NTIs and translatability degree) cannot be confirmed.

4.4.1.5 Conclusions on Translatability

I have argued that, depending on the language, text type and translation agent (human or machine), there can be different translatability problems. The question is if general translatability issues can be transversally identified. There are indeed certain common translatability problems across most languages that can be reflected in the following four dimensions. These must be consequently implemented in CL rules if translatability is to be claimed as one of the advantages:

- Lexical ambiguity, which includes the phenomena polysemy and homonymy, synonymy and compounds (orthographical variants). Tuggy (2006: 167) distinguishes between ambiguity and vagueness. As ambiguity he refers to homonymy, where two completely different concepts coincide in the form, for instance *bank*. Vagueness is polysemy in which a word can have different interpretations within the same concept, so that both meanings are related. For instance, in English the term *aunt* can mean "mother's sister" or "father's sister", being these meanings intuitively united into one, "parent's sister".
- Syntactic ambiguity, which includes phenomena such as categorical ambiguity, prepositional phrase attachment, modifiers scope, word order, anaphora, ellipsis etc.
- Contextual ambiguity, which includes the phenomena of connotation. This is indeed one of the most difficult ambiguities to resolve, since it depends on the relationship among the different constituents of a sentence
- Formal pitfalls: punctuation, orthography, structuring (for instance marking a title to differentiate it from running text) etc. This kind of problems might not be very relevant for a human translator, who can infer which is the right construction. However, an MT system can produce completely non-sense if only a full stop is wrongly placed.

4.5 Evaluating CL checkers

The aim of CL checkers is to check the correctness of texts, being able to handle correct input, but also detecting the errors and correcting them or at least making suggestions. They are complex tools containing many different modules, such as parsers, grammars, sets of rules and terminological databases, which must interact with each other to produce the desired results. If I want to test a CL checker regarding its intrinsic features, I will have to take all these factors into account. Usually, formal evaluation studies on the properties of CL checkers include precision, recall, and convergence. Nyberg, Mitamura, & Hujisen (2003: 258) describe these concepts as follows:

> **Precision** is the proportion of the number of correctly flagged errors to the total number of errors flagged; **recall** is the proportion of the number of correctly flagged errors to the total number of errors actually occurring; and (for automatic correction) **convergence** is the proportion of the number of automatically corrected sentences that are accepted when resubmitted to the total number of automatically corrected sentences.

A good working CL checker will thus not report inexistent errors (100% precision), will flag correctly all real errors (100% recall) and will suggest corrections that eliminate all errors and do not introduce new ones (100% convergence).

Besides, Fouvry & Balkan (1996) add that it should be checked if any critiques are incorrectly reported (this would correspond to noise in Information Retrieval terms) and if the system failed to identify errors according to the CL definition (which would correspond to silence). Other aspects that can be evaluated are the quality of the system prompts – if they are vague critiques or indicate clearly what the problem is or if the system can provide auto-correction or correction examples.

It is, however, very complex to obtain reliable results with this kind of testing, since data can be biased by subjective factors and full coverage of precision and recall results is not always possible. Besides, the success of this type of testing does not necessarily indicate that the application of a controlled language indeed results in any of the effects pointed out before (better understandability and readability as well as translatability), but that computationally all modules interact correctly. Indeed, Wojcik & Holmback

(1996) defend the point made by Adriaens & Macken (1995) and Wojcik, Harrison & Bremer (1993), who state that "one ought not only evaluate Controlled Language checkers on the basis of precision and recall, but also on the basis of how well writers can use them to 'converge' on a compliant revision".

Since I concentrate on the evaluation of a CL rule suite rather than on the software itself, I refer the reader to the EAGLES reports, especially the part where the evaluation of writing aids is described The EAGLES MT Evaluation Working Group (1996: 25) and The EAGLES MT EValuation Working Group (1999: 116). Other studies that deal with CL checkers evaluation are those by Adriaens & Macken (1995), Barthe et al. (1998), Fouvry & Balkan (1996) and Wojcik, Hoard, & Holzhauser (1990).

4.5.1 Evaluation of MULTILINT

As already detailed in Chapter 2 (2.6.), MULTILINT was first developed in the frame of the project MULTILINT, which was developed from 1995 to 1998[78]. It was followed by the project TETRIS (1999 to 2002), the goal of which was to further develop a prototype system to support technical writers when creating their documents. Nowadays, it is a commercial product with the name of Congree which is being constantly developed and improved.

With regards to evaluation of MULTILINT/CLAT, I find a first reference in Haller (1996) who rather than describing an evaluation methodology, he brings up a request of requirements, pointing out that an exhaustive evaluation will take place in 1997. Therefore, in this first phase of the project, I witness the development of the tool taking into account the preferences of the users, with some technical testing in order to keep the development ongoing.

In the project final report, Reuther, Schmidt-Wigger, & Fottner-Top (1998) dedicate a paragraph to evaluation. First, it is stated that during development all control functions were tested regarding the correctness of their application for different document types. However, no specification is given with regards to what kind of tests were undertaken.

Further on, a cyclical and interactive test period was established, where the application was validated by pilot users regarding different aspects, such as functional, linguistic or usability factors. In a third evaluation phase, the final users tested the real possibilities of implementation by comparing the output of the system with the proposals of a human corrector. The conclusions of this last phase were that, though much had been achieved, functionality and integration had to be improved and style checking had to be added as one of the control functions.

Once the style module was incorporated, Schmidt-Wigger (1998) mentions the tests that were carried out to validate the style rule set. They used the comparison of system output and human corrections. Further, a convergence test of the system was also undertaken. This test consisted of submitting to the system a manually corrected example to be corrected again. This was done not only to test the tool, but also to check the consistency of the CL definition. Shmidt-Wigger also speaks about performance testing, in which the style and grammar checkers were judged on the basis of recall and precision, with about 92% precision for the style checker on a corpus of about 750 complex sentences and 65% recall. For the grammar checker, precision on the test corpus was about 81% and recall about 57%.

The subsequent project, TETRIS, also made important efforts on evaluation. Indeed, one whole chapter in the TETRIS[79] project documentation dealt with this issue and it was divided into two parts: “Proof-Reading” and “Hit Rate in Translation Memory Systems”. The goal of the first evaluation scenario was to determine the average cost saving potential gained by using MULTILINT in contrast to human proofreading. The tests included a statistical macro evaluation, where factors such as different scenarios for creation of content, usability of the system and general program behaviour were tested. A dynamic microevaluation was also carried out, focusing on texts verified with MULTILINT. In this case, the results had to be evaluated regarding the information retrieval measures precision and recall, that is, how many mistakes were recalled and, from them, how many of them were indeed correctly recalled (precision). The conclusion of this first evaluation scenario was that MULTILINT, although it assists the

technical writer to an important degree, could not completely substitute an experienced and specialized human proof-reading.

The second evaluation scenario, "Hit Rate in Translation Memory Systems", intended to prove that the use of MULTILINT could increase the hit rate in translation memory systems by assuring more consistency in the source texts. Though this scenario was repeated twice, the results were not meaningful enough due to subjective factors such as the learn effect on MULTILINT and the differences on the writing skills of the different authors.

All in all, it was not possible to assess and prove the quality of MULTILINT in a meaningful way.

There is no record of further evaluation efforts for CLAT or Congree, though they might be internal and therefore not available for general research. In any case, I propose a new evaluation approach to test the extrinsic features of a CL checker, that is, whether the application of CL rules by means of a checker carries the advantages pointed out in the first section, and, if so, under which conditions. The methodology of this new evaluation approach will be tackled in Chapter 5.

4.6 Evaluating MT

In general, evaluating translation quality is a complex task since there is not an absolute way of stating when a translation is good or correct. Due to the non-existence of this golden standard, many correct answers are possible and, as Vashee (2009) states "there is no entirely objective way to measure the quality/accuracy of automated translation software, or of any translation for that matter".

This lack of standard measures to evaluate translation quality happens both with human translation and MT. Indeed, the evaluation of MT has been an issue of interest almost since the origins of this technology. Experiments such as the one carried out by Pfaffin (1965) or the ALPAC report itself, which constituted the first big evaluation effort to

sound out the state of the art in the development of MT (Hutchins, 1986: chapter 8), support this statement. There is indeed nearly in every book on MT a chapter dedicated to evaluation and the number of articles on the issue does not stop growing. Indeed, (Hovy, King, & Popescu-Belis (2002b) state that "it is impossible to write a comprehensive overview of the MT evaluation literature".

Since then, the wide interest for automatic evaluation has lead to an outbreak of publications, confirmed by initiatives like the 2009 issue of the journal *Linguistica Antverpiensia on Evaluation of Translation Technology* (Daelemans & Hoste, 2009) or the last special issue of the prestigious journal *Machine Translation* (Way, 2010) on Automated Metrics for MT Evaluation. Further, most workshops and conferences on MT include a few contributions on this issue, such as the CWMT2009 Machine Translation Evaluation Workshop that was held on October 2009 in Nanjing, China[80]. Currently, a number of evaluation campaigns are also being carried out to assess the quality of different MT systems, especially those which are statistically based, such as the NIST Open Machine Translation (OpenMT) Evaluation Plans, which were carried from 2001 to 2009[81].

There is also a great deal of overviews on the issue of machine translation evaluation which have been carried out in doctoral dissertations (Giménez Linares, 2008; Schäfer, 2002: 173), or book parts or chapters (Arnold, Balkan, Meijer, Humphreys, & Sadler, 1994; Lehrberger & Bourbeau, 1988), though there the number of monographs on evaluation of MT is rather scarce, as Schäfer (2002: 173) states.

Due to the big amount of literature dedicated to MT, it is not my intention to be exhaustive here, but to emphasize the most important aspects on MT evaluation that set the grounds of the empirical part of this work. Since I already dealt with the first initiatives of MT evaluation in section 4.1 and 4.2, I proceed to highlight the work by Church & Hovy (1993), Lehrberger & Bourbeau (1988), Van Slype (1979) and White (2003) as well as the MT Evaluation project that was developed by the FEMTI Framework, to continue with the notion of translation quality and a discussion about the metrics.

4.6.1 Evaluation of MT according to Van Slype

One of the first methodological approaches to a classification of evaluation types was proposed by Van Slype (1979) who established a three-dimensional evaluation framework. In his work he distinguishes two main conceptual levels: the first conceptual level discerns between evaluation itself, market research and system development. Evaluation itself comprehends macro and microevaluation. Macroevaluation implies the complete evaluation of the system, whereas microevaluation is a detailed evaluation aimed at assessing the improvability of the system or establishing an improvement strategy. Finally, regarding the evaluation methodology, Van Slype distinguishes three types of evaluation:

- Superficial evaluation: when a new version of an MT system has to be approved on delivery.
- In-depth evaluation: for "turning points" in the development of an MT system, distinguishing between evaluation of acceptability and market research and evaluation of improvability and development of the system.
- Pinpoint evaluation: to see the impact of specific changes to the system.

Macroevaluation is carried out at four levels: cognitive, economic, linguistic and operational. Each level comprehends a series of evaluation parameters, with different methods to measure them:

- Cognitive level: Intelligibility, fidelity, coherence, usability, acceptability.
- Economic level: Reading time, correction time, translation time.
- Linguistic level: Reconstruction of semantic relationships, syntactic and semantic coherence, "absolute" quality, lexical evaluation, syntactic evaluation, analysis of errors.
- Operational level: Automatic language identification, verification of the claims of the manufacturer.

With regards to Microevaluation, the method that Van Slype proposes can also be classified in different levels as follows:

- Grammatical symptomatic level: Analysis of the grammatical errors detected in the translated texts.
- Formal symptomatic level: Tally of the deletions, additions, modifications, shifts and replacements of words by the revisers and post-editors (i.e. revision and post-edition rates).
- Diagnostic level: Analysis of the causes of errors input, analysis of the source language, dictionary, etc.
- Forecast level: Analysis of the improvability of the system.
- Therapeutic level: Detection of the improvements to the system following an upgrading operation.

Van Slype also introduces the concepts of the importance of text typology, sampling of evaluators and translation quality when running an evaluation. He then reviews the different evaluation typologies and metrics developed until then for measuring the different aspects of each of the levels discussed above: fidelity, intelligibility, acceptability, correction and reading time etc.

Therefore, he sets the grounds for further MT evaluation research works aiming at establishing a classification or methodology, such as that by Rinsche (1993) or, as I will see in 4.6.3, the FEMTI Framework.

4.6.2 Evaluation of MT according to Lehrberger and Bourbeau

Lehrberger & Bourbeau (1988) focus on the linguistic evaluation by the user. Their contribution also takes into account context details, which must be specified in detail in order to determine what to measure. These details include elements such as the type of texts to be translated, the linguistic processing model, the planned level of automation, Restrains on the quality of translation of the raw output, Restrains on the quality of

translation of the final version, mechanisms for dealing with errors and the word processing system used. Indeed, according to these authors, evaluation only makes sense if carried out within a given context (op.cit.: 192):

> The question now is not whether MT (or AI, for that matter) is feasible, but in what domains it is most likely to be effective. The object of an evaluation is, of course, to determine whether a system permits an adequate response to given needs and Restrains.

They distinguish three approaches to a detailed evaluation:

- Evaluation by the system designer.
- Cost/Benefit evaluation.
- Linguistic evaluation by the user.

It is also remarkable that, when identifying user's needs, the authors highlight three factors: the characteristics of the texts, the desired level of automation in the translation process and the quality of translation acceptable to the user. Indeed, the type of text and the degree of quality expected or acceptable are factors that appear in a recurrent way in MT evaluation literature.

Lehrberger and Bourbeau (op cit., 1988: 186) suggest the following process for evaluating a system:

1. Identification of the user's needs.
2. Choice of texts.
3. Identification of type of use.
4. Performance requirements.
5. Cost and benefit study.
6. Linguistic evaluation.
7. Linguistic performance.
8. Linguistic capability.
9. Preliminary use.

10. Final judgement.

As it can observed, certain patterns such as the selection of texts, the type of use (dissemination, information gisting[82] etc.) and the cost and benefit study are repeated in different evaluation methodology proposals, which confirm them as cornerstones of MT evaluation.

4.6.3 The ISLE Project and Context-based Evaluation: the FEMTI Framework

Church & Hovy (1993) analyzed what requirements a good niche application for MT should meet. They suggested six desiderata: (i) it should set reasonable expectations, (ii) it should make sense economically, (iii) it should be attractive to the intended users, (iv) it should exploit the strengths of the machine and not compete with the strengths of the human, (v) it should be clear to the users what the system can and cannot do, and (vi) it should encourage the field to move forward toward a sensible long-term goal. These principles were further discussed and extended by the Evaluation Working Group of the ISLE Project (1999-2002).

The ISLE Project (International Standards for Language Engineering), which was carried out from 1999 to 2002 was the successor of the EAGLES. It was funded by the US Government National Science Foundation and the Swiss and Danish Governments. Three working groups participated in the project, one of which was devoted to Evaluation (EWG). This group[83] focused on Machine Translation Evaluation and the main result of it was FEMTI[84]. The project, and the resulting framework, worked on the idea of context-based evaluation, which, as I have seen in my review, was based on previous evaluation efforts.

FEMTI is described in detail by Estrella, Popescu-Belis & King (2009) and Hovy, King & Popescu-Belis (2002a). FEMTI aims at gathering all previous MT evaluation efforts and establishes a methodology to evaluate MT systems taking into account the intended context of use of a system when designing its evaluation. Based on the ISO/ISEC 9126 and ISO/IEC 14598 standards, which are domain independent guidelines for the

evaluation of software products, FEMTI comprehends six top-level quality characteristics proposed by the standard: functionality, reliability, usability, efficiency, maintainability and portability, which conform the overall quality of a product and can be decomposed in further features. Further on, the FEMTI model was extended with an additional top-level characteristic, namely Cost.

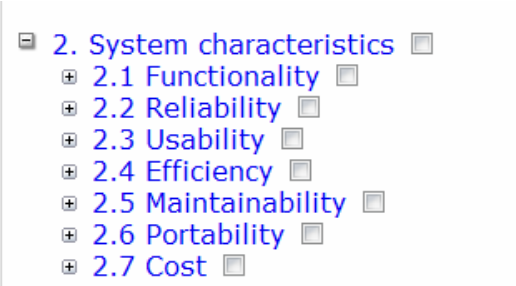

Figure 9: FEMTI external top-level quality characteristics

FEMTI is devised as a function of two interrelated taxonomies. The first taxonomy can be used to define a potential context of use for the MT system to evaluate. The second taxonomy presents the relevant quality characteristics as well as a set of metrics linked to the different situations which can be described using the first taxonomy. Therefore, once the evaluation context is depicted, FEMTI suggests the relevant quality characteristics and appropriate metrics to be used by the evaluator.

For the measurement of these quality characteristics, however, FEMTI only offers a listing of different metrics, without assessing any standard. The user must decide, according to the context defined, which metrics from the literature are most appropriate to measure the features chosen or he must develop new metrics according to his needs. Besides, there might be quality characteristics that the framework does not include, such as post-editability effort or return on investment.

I will offer an extensive review of the FEMTI methodology in Chapter 5, where I will explain the methodological approach of my work.

4.7 The notion of translation quality

One of the main problematic issues of translation evaluation and, hence, of MT evaluation, is the notion of translation quality. Indeed, the notion of quality varies depending on the given needs or context and, for instance, academia and industry have different conceptions of what quality can be. Since I am working in an industrial context, I will concentrate on the latter. As Schnitzlein (2003: 4), who offers an extensive review of industrial norms on quality, points out, I must distinguish between process and product quality. Process quality refers to the quality of the processes from the assignment of a translation project to the delivery of the desired product, while product quality designs the quality of the translation itself. This can be divided between formal product quality and linguistic product quality. Within formal product quality there are aspects such as layout, formatting, desktop publishing etc.

When dealing with software, different dimensions of quality can be considered when evaluating a system. This is called a quality model and consists of external quality requirements (user needs that become a set of specifications) and internal quality requirements, which refer to the characteristics of the system itself (E. Hovy, King, & Popescu-Belis, 2002a). The relation between external and internal qualities is not always straightforward, especially in MT, where external quality requirements do not always predict all the results of using the software before it is completely operational. Indeed, E. Hovy, Margaret King, & Popescu-Belis (op.cit.) follow the norms ISO/IEC-14598 (Information technology – Software product evaluation) and ISO/IEC-9126-1 (Software engineering – Product quality) to establish a quality model within the FEMI Framework.

With regards to linguistic quality, Van Slype, 1979 (31-37) dedicates one whole section to the notion of translation quality, where he includes contributions by different authors on the concept of translation, the quality of translation and the link between translation quality and evaluation criteria. In accordance to the ideas postulated by Lehrberger and Bourbeau (1988), one of the most appropriate definitions of quality is the one postulated by A.J. Petit, who states that a product is acceptable only if it meets the requirements of

its users. Therefore, it will be most important to be very specific in defining user's needs in order to establish, further on, if these are fulfilled or not. For utility technical texts (such as maintenance or user manuals), which are the ones I will be working with, he distinguishes following principal requirements: no errors; homogeneity; clarity, without ambiguity or gibberish which might obscure the sense of the message; simple correct style, without extravagances or recherché elements; use of the terms recognized in the relevant sector.

He also states that quality requirements can vary depending on whether the text is going to be revised (in the case of MT post-edited) or if it is going to be supplied directly to the final user.

Van Slype (op cit.) summarizes the different author's contributions on translation quality stating that quality has to be assessed according to the aims of the final user of the translation and that the expected quality cannot be the same for human translation and for MT. Therefore, evaluation criteria will have to be chosen according to the specific aims defined in a pre-evaluation phase, and these will need to be varied in order to reflect the multidimensionality of the translation task, which cannot be measured in absolute terms.

As I can see, the definition of the linguistic quality is a rather delicate matter. This is due to the fact that translation is an open natural language processing task, that is, given a certain input, there are different possible solutions, and the set of potential solutions is not closed. Further, the notion of quality might be different in different target languages and different text types, what makes very difficult to standardize evaluation methods. As Lehrberger and Bourbeau (1988: 186) conclude, the acceptability of a given translation will depend on the particular needs and Restrains of the user: a methodology can be general, but the results apply to a specific situation and context. This lack of standards has lead to a chaotic situation, as the organizers of a 1999 conference on translation quality in Leipzig noted:

> There are no generally accepted objective criteria for evaluating the quality both of translations and of interpreting performance. Even the latest national and international standards in this area—DIN 2345 and the ISO-9000 series—do not regulate the evaluation of translation quality in a particular context. […] The result is assessment chaos. (Institut für Angewandte Linguistik und Translatologie: 1999, in Williams, 2001)

4.8 Human versus Automatic Evaluation

Lavie (2010b: 11) distinguishes four different dimensions of MT evaluation:

- Human evaluation vs. automated metrics.
- Quality assessment at sentence (segment) level vs. system level vs. task-based evaluation.
- "Black-box" vs. "Glass-box" evaluation.
- Evaluation for external validation vs. target function for automatic system tuning vs. ongoing quality assessment of MT output.

I have already dealt with the black-box vs. glass-box dimension in 4.2.1. The second and the fourth dimension can be assigned to the different context-based evaluation types that were also tackled in that section. Now I will consider the human evaluation vs. automatic metrics dimension.

Since evaluation has been an issue in MT research and development, human evaluation has been the classical method to assess the quality of a system. This is usually done by means of scales, where the evaluator grades a translation from best to worst, or with questionnaires about the text to check if he understood it. However, this type of evaluation has three main pitfalls:

- First, it is costly and time consuming, since usually external evaluators have to be hired to do the job, and it takes a while and many evaluators to obtain statistical significant results.
- Second, the results of such an evaluation are hardly reusable, since every time an evaluation takes place, the whole procedure has to be repeated.

- Finally, the results of a human evaluation are subjective, since two evaluators can assess a sentence in a different way depending on many factors such as their education, experience, background information etc.

Therefore, in the past years new *n*gram-based intrinsic metrics have been developed to automatically score system-outputs against human-produced reference documents. These are the so called reference-based MT Evaluation methods and are mostly used to compare performance of two or more different MT engines/technology for the same language pair (Lavie, 2010). Though the interest on these methods arose already in the 90s (Shiwen, 1993; Thompson, 1991, 1992) it was not until the beginning of the 2000 that they began to be broadly used. One of these methods is BLEU, a corpus-based metric based on the assumption that "the closer a machine translation is to a professional human translation, the better it is" (Papineni, Roukos, Ward, Zhu, & Heights, 2001). Thus, to assess the quality of a machine translation, the numeric closeness between two translations (a candidate machine translation and one or more reference translations) is calculated, though overgeneration of correct word forms is penalised in order to avoid erroneous results. A brevity penalty that penalises test sentences found to be much shorter than the reference sentences is also included. NIST was the next following important measure to appear (Doddington, 2002), also using *n*gram co-occurrence statistics.

Automatic evaluation represents a cost-effective method to carry out quick and frequent evaluations. These methods are also useful for contrasting the relative frequency of different MT outputs. However, the results are not always reliable and it is difficult to make any statements about the real quality of the system. What does, for instance, a BLEU score of 0.326 mean? As Koehn & Monz (2006) state:

> While automatic measures are an invaluable tool for day-to-day development of machine translation systems, they are only an imperfect substitute for human assessment of translation quality.

Hence, it is always recommendable to cross-check the results with human evaluation data. The following table summarizes the advantages and disadvantages of human and automatic evaluation:

	Human Evaluation	Automatic Evaluation
Advantages	Easy to interpret More informative	cost-effective objective
Disadvantages	Subjective Costly and time-consuming Non-reusable results	Difficult to interpret Not always reliable Need of meta-evaluation

Table 7: Advantages and Disadvantages of Human and Automatic Evaluation

4.8.1 Human Judgment

One of the first experiments using human evaluation methods was carried during the elaboration of the ALPAC report. It was conducted by John B. Carrol (ALPAC, 1966: 67-75) and measured fidelity and intelligibility using nine-point scales that had been established based on the method of equal-appearing intervals.

Indeed, human judgement usually takes place in form of rating scales for quality aspects such as intelligibility, fidelity, comprehensibility or readability; another way of evaluating has been to count the number of errors, either by analysing the output or by counting the number of corrections made by post-editors. Sometimes, these errors have been classified according to their importance or severity, leading to an scheme of translation errors. Finally, another way of carrying out human evaluation is the so called procedural evaluation, where the evaluator has to undertake the task the translated text is describing, to see if comprehension and fidelity are kept. Lavie (2010b: 15) summarizes the main types of human evaluation as follows:

- Adequacy and fluency scores.
- Human ranking of translations at the sentence-level.
- Post-editing measures: Post-editor editing time/effort measures HTER: Human Translation Edit Rate.

- Human editability measures: can humans edit the MT output into a correct translation?
- Task-based evaluations: was the performance of the MT system sufficient to perform a particular task?

As we have seen before, human evaluation entails a series of disadvantages: subjective results, time and cost and not reusable. In order to diminish these disadvantages, it is necessary to design the evaluation task in a way that it promotes high agreement. This can be achieved by defining strict scales and giving clear instructions to the evaluators as well as designing easy to use applications or interfaces to carry out the evaluation. Besides, as Lavie (2010b) points out, it is important to pay special attention to the qualifications of the human raters. In this way I will reduce the subjectivity of tests and diminish the cost and the time, since results will be more meaningful and thus, less evaluators will be needed.

There is a vast amount of evaluation studies based on human evaluation metrics. As I have mentioned before, one of the first studies was the one carried by John B. Carrol within the ALPAC evaluation campaign. Later on, Van Slype (1979) compiled most of the evaluation methods carried until then and established an evaluation framework. This report outlines many different kinds of scales used to measure features such as fidelity or intelligibility, gathering scales that ranged from 2 to 3 points up to 25. Other subsequent reviews on MT evaluation methods is those by (Falkedal, 1991; Giménez Linares, 2008: 25-27). Besides, FEMTI includes in one of the web pages of the project an extensive bibliography on MT evaluation[85].

Another effort which is worth mentioning is the SAE J2450 Translation Quality Norm (SAE, 2001), which became SAE Recommended Practice in October 2001, was specially designed to measure the quality of automotive service information and comprehends seven error categories metrics for language translation, as I can see in this table extracted from (Secară, 2005):

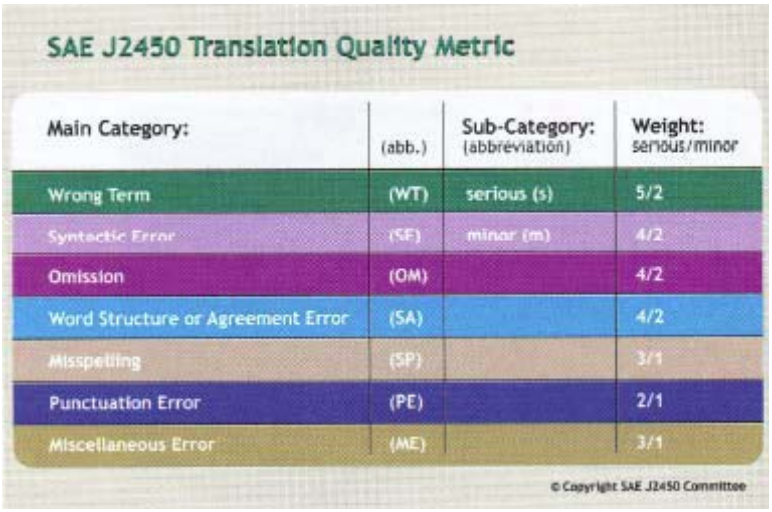

Figure 10: SAE J2450 error categories

Schütz (1999) reshapes this quality metric to adapt it to the evaluation of machine translations as well as to embed the whole evaluation process into an object-oriented quality model approach to account for the established business processes in the acquisition, production, translation and dissemination of automotive service information in SGML/XML environments. This new form of the metric has 8 classes:

1. Wrong or unapproved term, abbreviation and acronym. In contrast to the J2450, I restrict this class entirely to the terminological level in its genuine sense, i.e. I do not include function words (WT).
2. Omission of text and of graphics with text elements remains as defined in the J2450 class (OM).
3. Superfluous text remains as defined in the J2450 (SF).
4. Morphological error regarding word structure, orthography, etc. (MO).
5. Grammatical error regarding word order, agreement, punctuation (GE).
6. Style violation of a specific set of writing rules including controlled language use, honorifics and localization issues (SV).
7. SGML structure error (SS).
8. Miscellaneous error (ME).

Secară (2005) discusses further recent developments of human based metrics, such as the BlackJack, developed by the British translation agency ITR; TransCheck or the MeLLANGE error annotation matrix.

4.8.2 Automatic Metrics and Measures

In recent years, a number of automatic metrics for the evaluation of MT have been developed due to the drawbacks involved with human metrics as well as the move towards data-driven MT systems, especially statistical MT. These metrics have been designed exclusively for MT development, that is, MT developers need them in order to check the improvements and drawbacks of their systems and to get an orientation of what should be the next steps to follow.

Before Statistical Machine Translation experimented its revival, the first automatic metrics used for the evaluation of MT during 1990s were metrics from the speech community, such as WER (Word Error Rate) or PI-WER[86]. In 2002, BLEU, developed by (Papineni et al., 2001), appeared and, since then, a number of metrics have emerged as a result of the new statistical developments in MT and the consequent growing necessity for evaluation.

The past decade has given birth to a number of initiatives, which currently add up to around 30 metrics, some of which are variants of the original. Giménez Linares (2008: 28-49) offers a good overview on automatic evaluation metrics. The dominant approach to automatic MT evaluation is based on lexical similarities. These metrics are therefore also called *n*gram based metrics and they can be classified according to the type of computed measure:

4.8.2.1 Edit Distance Measures

- WER. Word Error Rate, described by Nießen, Och, Leusch, & Ney (2000). It is defined as the "edit distance *d(t,r)* (number of insertions, deletions and substitutions) between the produced translation *t* and one predefined reference translation *r*".

- PER. Position-independent Word Error Rate, described in Tillmann, Vogel, Ney, Zubiaga & Sawaf (1997) is a similar measure with WER, but the positions of the words in the sentence are ignored.
- TER. Translation Edit Rate which, according to "measures the amount of editing that a human would have to perform to change a system output so it exactly matches a reference translation".

4.8.2.2 Precision-Oriented Metrics

- BLEU. Bilingual Evaluation Understudy, defined by Papineni et al. (2001), calculates precision by comparing "n-grams of the candidate machine translation with the n-grams of the reference translation" and this is done in a position-independent way.
- NIST, a metric developed by the National Institute of Standards and Technology, is an improved version of BLEU (Doddington, 2002). NIST uses an arithmetic mean of co-occurrences over N (whereas BLEU uses a geometric mean) and it weights more heavily those N-grams that are more informative, i.e. those that are less frequent.
- WNM or weighted N-gram model, combines BLUE with weights for the statistical significance of lexical items (Babych & Hartley, 2004).

4.8.2.3 Recall-Oriented Metrics

- ROUGE. Recall-Oriented Understudy for Gisting Evaluation (Lin & Och, 2004) is divided between ROUGE-L and ROUGE-S. ROUGE-L "measures sentence-to-sentence similarity based on the longest common subsequence statistics between a candidate translation and a set of reference translations", whereas ROUGE-S "computes skip bigram co-occurrence statistics".
- CDER, Cover/Disjoint Error Rate (Leusch, Ueffing, & Ney, 2006). A recall-oriented measure that models movement of word blocks as an edit operation.

4.8.2.4 Measures Balancing Precision and Recall

- GTM. General Text Matcher. GTM generalizes precision, recall, and F-measure to measure overlap between strings, rather than overlap between bags of items (Melamed, Green, & Turian, 2003; Turian, Shen, & Melamed, 2003).
- METEOR. It is currently called METEOR-NEXT and it scores machine translation hypotheses by aligning them to one or more reference translations (Banerjee & Lavie, 2005).
- BLANC is a family of dynamic, trainable evaluation metrics for machine translation (Lita, Rogati, & Lavie, 2005).
- SIA or Stochastic Iterative Alignment, is a metric based on loose sequence alignment but enhanced with alignment scores, stochastic word matching and an iterative alignment scheme (Liu & Gildea, 2006).

Apart from this approach based on the lexical similarity, there are also some initiatives that have given place to metrics based on syntactic similarity, metrics that use shallow semantic analysis and combinations of metrics at different levels to obtain a comprehensive global measure of the quality of the system (Giménez Linares, 2008: 30).

4.8.2.5 BLEU and NIST

As we have seen before, some of the claimed advantages of these metrics are that they are cost-effective, since you do not need human resources to undertake them, and they are objective, since they seem to treat all systems alike[87]. However, there are certain disadvantages too. First of all, the results of these measures are very often difficult to interpret, since they only give us numerical values that do not state anything about the real quality of the system. Second, they are not always reliable, since the more reference human translations, the better and more accurate scores. Therefore, depending on the evaluation set and the resources available, results can vary considerably. Finally, in order to assess the validity of automatic metrics, there is usually a need for to compare results with human evaluation results.

Two of the most popular *n*gram metrics up to date are BLUE (Papineni et al., 2001) and NIST (Doddington, 2002), which are based on the idea that the highest the similarity of a translation *n*grams' distribution with regards to a good translation (understood as a human translation), the better will be that translation. More precisely, this means that for each segment of the text which must be evaluated, the corresponding aligned segments from the reference translations are analysed, *n*gram counts are extracted and compared among them. Therefore, these metrics deal with similarity rather than with quality, based on the assumption that a good translation will be similar to other good translations of the same texts. However, this is not always the case, and these metrics try to compensate these exceptions by using an adequate number of reference texts in order to cover all different translation variants.

As I have pointed out, BLEU is still one of the most established metrics in the field of automatic MT evaluation. BLEU relies on the idea that, "the closer a machine translation is to a professional human translation, the better it is" (Papineni et al., 2001). For this purposes, the primary task that this metric accomplishes is to compare *n*grams of the candidate translation (the MT translation) with the *n*grams of reference translations (human translations) and count the matches. The more the matches, the better the candidate translation is. The metric is then adjusted with the following precisions:

- **Modified unigram and n-gram precision**: a reference *n*gram should be considered exhausted after a matching candidate word is identified in order to avoid inflated precision scores. According to the authors, this modification accounts for two aspects of translation quality: adequacy and fluency. Besides, it is not only applied to unigrams and *n*grams on a sentence level, but for the entire test corpus. To illustrate the value of this modification, Papineni et al. (op.cit.) give the following example:

Candidate: the the the the the the the.
Reference 1: The cat is on the mat.
Reference 2: There is a cat on the mat.

Without the modification, I would have a precision of 7/7 (there are seven words in the candidate translation and all of them match with words in the reference translations). However, with the modification, the real score is 2/7, since the word *the* happens only twice in the reference translations.

- In order to take into account the exponential decay of the modified n-gram precision, the **geometric mean** of the modified *n*gram precisions is introduced.

- In order to avoid that too short sentences recall more words from the reference translations than a longer sentence, a sentence **brevity penalty** is applied. In this way it is avoided that, phrases such as *of the* obtain high scores because the *n*gram appears in the reference corpus, though they are inadequate translations. Thus, a high-scoring candidate translation must also match any of the reference translations in length, in word choice and word order.

The NIST score, a metric developed by the National Institute for Standards and Technology is a variation of the BLEU score. It was first presented by Doddington (2002) and it has been used in all the evaluation campaigns carried out by this institute since then until 2009. NIST uses the same algorithm as BLEU but using the arithmetic mean (whereas BLEU relies on a geometric average)[88] and weighting more heavily those *n*grams that are more informative, that is, those *n*grams that appear less frequently. As (Coughlin, 2003) points out "this difference is significant when dealing with very low-quality translations". Further, the two algorithms also calculate their respective brevity penalties in different ways.

As I have mentioned before, these metrics were designed with developer scenarios in mind. They can be very useful for comparing alternative systems on the same benchmark data-set, or for contrasting two versions of the same system (Lavie, 2010b). However, their use in translation consumer contexts is not very widespread. Developers can work with the same resources (same benchmark data) over and over again to see the improvements of their system (internal evaluation, diagnostic evaluation, feasibility evaluation and requirements elicitation), whereas final users (the real consumers of the

translations) usually need to carry declarative, operational or usability evaluations, where different resources (different corpora, number of users etc.) are needed every time. Therefore, when using BLUE or other automatic metrics for evaluation, it is necessary to design a very strict evaluation set (same language pairs, same data) in order to avoid false interpretations of the results. To my knowledge, there are only a few studies that implement these metrics in real contexts of use, such as the research by Aranberri Monasterio (2009), who studies how to improve the MT output for -ing words in RBMT in the localisation industry.

Furthermore, BLUE is difficult to interpret. The results range from 0 to 1. The closer to 1, the more overlap with human references. Lavie (2010b) suggests following interpretation scale for BLEU scores:

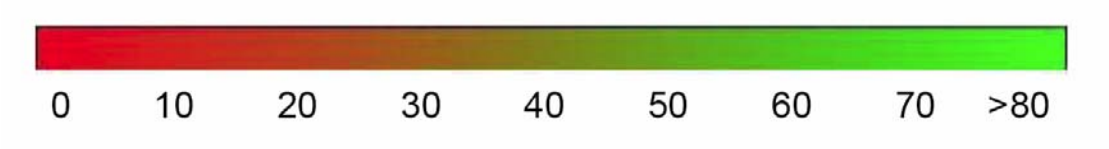

Figure 11: BLEU Interpretation according to Lavie (2010b)

This kind of interpretation might be helpful to compare systems in a declarative evaluation, in order to assess which of the systems perform best with the corpus used during the experiment. However, from the user point of view, it is still difficult to surmise what a score of 0.66 means in terms of, for instance, post-edition effort or language correctness.

With regards to NIST, Zhang et al. (2004)

Apart from interpretation problems, some of the other critiques that have been raised against BLEU is that it renders better results in a document level rather than in a sentence or phrase level (Blatz et al., 2004; Kulesza & Shieber, 2004). Owczarzak, van Genabith, & Way (2007a; 2007b) criticize that BLEU and NIST only compare strings at a morphological level, penalizing any divergence from them. This means that "candidate translation expressing the source meaning accurately and fluently will be

given a low score if the lexical and syntactic choices it contains, even though perfectly legitimate, are not present in at least one of the references." According to Callison-Burch, Osborne, & Koehn (2006) the reasons shy BLEU not always correlates with human judgements are that BLEU allows for a great n-gram variation for hypotheses with identical scores. This results in identical scores for sentences that have the same n-grams regardless of their position in the sentence, as long as they appear in the reference translations. This makes BLUE and NIST useful measures for comparison purposes.

Other measures that have gained more space in automatic evaluation during the last years are TER (Snover, Dorr, Schwartz, Micciulla, & Makhoul, 2006), METEOR (Agarwal & Lavie, 2008; Banerjee & Lavie, 2005) and the F-Measure (Turian et al., 2003) , which balances precision and recall y calculating the weighted harmonic mean of the two.

4.8.3 Interpretation of results

There is a fundamental problem with the interpretation of translation evaluation results, both in human and in automatic evaluation set-ups.

It is necessary to clear up how is the relationship between the number and gravity of errors and the real quality and acceptance of the translation. I can say, e.g., that a translation contains has an error rate of 0.209, according to the example presented in SAE J2450 Translation Quality Norm (SAE, 2001). But what does this really mean? Is the text acceptable for publication, are the errors only minor errors, which do not necessarily need post-editing? What if the error rate would be 0.1, but there were two serious errors? How does all this relate with the length of the text?

This norm also claims to be applicable regardless of the language pair and of how the translation is performed, if by a human or by a machine. However, as (Schütz, 1999) points out, we believe that some error categories must be weighted differently depending on their author. Further, different weightings should be given depending on the translation type: human or automatic. One of the big advantages of MT, e.g., is the

application of correct and consistent terminology (provided there is a complete glossary) over big amounts of texts. The machine has no preferences, no stylistic predilections, and will translate therefore the same term always in the same way. If, for example, a "wrong term" error occurs in a Machine Translation despite of being correctly coded in the glossary, this should be scored more severely than if it were produced by a human. On the other side, consider morphology errors: a human translator, if translating in his mother tongue, is rather unlikely to produce morphologic or agreement errors, only because of his language instinct. However, machine translations lack this instinct and usually produce many more errors of this type. Therefore, these should be scored with a lesser scale for machine produced output.

Another consideration for the scoring question could be this: in translation courses the translations made by students are corrected according to an error classification scheme. This scheme depends on the professor or teacher and from faculty to faculty, but usually includes lexical, grammatical and style errors. It is true too, however, that good translations and smart ideas, are usually awarded with extra points. Why should I do not do the same for MT, especially when I know that a perfect translation occurs so rarely?

Some authors claim that "the connection between translation phenomena and the attributes of MT (e.g. fidelity, intelligibility, etc.) is not straightforward" (S. Corston-Oliver, Gamon, & Brockett, 2001; J. S. White, 2000, 2001). In particular, it is presumed, but not demonstrated, that the apparent fluency of an MT output (measured, perhaps, by counting structural errors) will allow us to predict its usefulness in information-intensive tasks such as information extraction (J. S. White et al., 1994).

4.8.4 Metaevaluation and correlation

When human evaluations there is always a shadow of a doubt about the reliability of judges and their fluctuation (Hamon, 2010; Hamon, Mostefa, & Arranz, 2008).

In order to counteract this, one of the claimed advantages of automatic metrics is that they represent a fast and cost-effective method for evaluating translations, in contrast to human evaluation methods. Therefore, these automatic metrics need to be as precise as possible so that they render reliable results and are able to predict human judgements.

However, in order to test their reliability, results have to be tested as well against human evaluation tests in what is called a meta-evaluation, establishing correlations between automated and human assessments. This will help the evaluator to decide if the metric that has been used is representative or not. According to Amigó, Giménez Linares, Gonzalo, & Márquez (2006) and Giménez Linares (2008: 19-20) there are two main meta-evaluation criteria: human acceptability and human likeliness. In the first case, the quality of automatic metrics is measured according to their ability to capture how acceptable are automatic translations to humans. In order to establish this acceptability, correlation between automatic metric scores and human assessments is calculated by means of coefficients such as Pearson, Spearmann or Kendall. According to Giménez Linares (op.cit.), meta-evaluation based on human acceptability "presents the major draw relying on human evaluations, which are expensive, not reusable, subjective, and possibly partial".

In the second case, human likeliness bases on the assumption that good automatic translations should resemble human translations. Usually, human likeliness is measured in terms of discriminative power, that is, "the metric ability to capture the features which distinguish human from automatic translations" Giménez Linares (op.cit.)[89]. The main advantage of this technique is that there is no need of human assessments and the subjective factor is thus eliminated. However, their reliability depends strongly on the heterogeneity/representativeness of the test beds employed.

When BLEU and NIST appeared, it was claimed that they correlated well with human judgements. Papineni et al. (2001) calculated the linear regression of the human monolingual evaluation group and compared it with the BLEU score for five systems using two reference translations. A correlation coefficient of 0.99% indicated that BLEU tracked human evaluation well. For the bilingual group, the correlation

coefficient was 0.96%. Coughlin (2003) states that these metrics can be highly reliable even when only one reference translation is available. She uses the Pearson product-moment correlation (PMCC) 90 coefficient. The result for BLEU was 0.811, whereas the NIST correlation coefficient was 0.796. As to the language pairs, the highest correlation coefficient was achieved for English-German for a small number of evaluations (14). It is also suggested that BLEU correlates better with human assessments when data sets are larger than 500 sentences.

However, some studies have shown that this is not always the case. Callison-Burch, Osborne, & Koehn (2006) claimed that "the translation community is overly reliant on BLEU" and already warned about the unreliability of BLEU, especially when comparing systems with a different approach (rule-based and statistic). Their conclusions were based on the NIST Machine Translation Evaluation exercise that took place in 2005 as part of DARPA's TIDES program. Within the framework of the shared tasks of the ACL Workshops on Statistical Machine Translation, Callison-Burch, Fordyce, Koehn, Monz, & Schroeder (2007) carry out an evaluation of statistical MT systems within the Euromatrix project[91], which fosters research in statistical and hybrid machine translation between all European languages. They apply eleven different automatic evaluation metrics, and conduct three different types of manual evaluation. With this broad evaluation methodology, they aim at discovering the consistency of human evaluation (among evaluators and of each individual evaluator), how it can be improved and to which automatic evaluation metrics correlate most strongly. Regarding the human evaluation, they simplified it to adequacy and fluency 5-point scales, which were developed by the Linguistic Data Consortium (LDC, 2002):

FLUENCY How fluent is the translation?	ADEQUACY How much of the meaning is expressed?
5 = Flawless English	5 = All
4 = Good English	4 = Most
3 = Non-native English	3 = Much
2 = Disfluent English	2 = Little
1 = Incomprehensible	1 = None

Table 8: Scales for adequacy and fluency developed by LDC (2002)

In order to state this correlation, they used the Kappa coefficient to measure agreement among evaluators as well as correlation with different automatic metrics. The Kappa[92] coefficient allows to measure the agreement between n judges with k criteria of judgment and is defined as

$$K = \frac{P(A) - P(E)}{1 - P(E)}$$

where P(A) is the proportion of times that the annotators agree, and P(E) is the proportion of time that they would agree by chance.

Regarding the correlation with human judgments, Callison-Burch, Fordyce, Koehn, Monz, & Schroeder (2007) opt for the Spearman correlation coefficient ρ, since it makes less assumptions about the data than Pearson's. The highest correlation was for the metric semantic role overlap (Giménez & Márquez, 2007), followed by ParaEval measuring recall (Ye, Zhou, & Chin-Yew, 2007) and METEOR (Banerjee & Lavie, 2005). The correlations of BLEU were higher when translating into other languages than into English:

	ADEQUACY	FLUENCY	RANK	CONSTITUENT	METEOR	BLEU	1-TER	GTM	PARAEVAL-REC	PARAEVAL-PREC	DEPENDENCY	SEMANTIC-ROLE
German-English News Corpus												
adequacy	1	0.900	0.900	0.900	0.600	0.300	-0.025	0.300	**0.700**	0.300	**0.700**	**0.700**
fluency	—	1	1.000	1.000	0.700	0.400	-0.025	0.400	**0.900**	0.400	**0.900**	**0.900**
rank	—	—	1	1.000	0.700	0.400	-0.025	0.400	**0.900**	0.400	**0.900**	**0.900**
constituent	—	—	—	1	0.700	0.400	-0.025	0.400	**0.900**	0.400	**0.900**	**0.900**
German-English Europarl												
adequacy	1	0.893	0.821	0.750	0.599	0.643	**0.787**	0.68	0.750	0.643	0.464	0.750
fluency	—	1	0.964	0.537	0.778	0.858	0.500	0.821	0.821	0.787	0.571	**0.93**
rank	—	—	1	0.500	**0.902**	0.821	0.393	0.714	0.858	0.643	0.464	0.858
constituent	—	—	—	1	0.456	0.464	0.714	0.18	**0.750**	0.250	0.214	0.43

Table 9: Correlations between human evaluation and automatic metrics into English

	ADEQUACY	FLUENCY	RANK	CONSTITUENT	METEOR	BLEU	1-TER	1-WER-OF-VS	MAX-CORR-FLU	MAX-CORR-ADEQ
English-German News Corpus										
adequacy	1	0.943	0.83	0.943	0.187	0.43	**0.814**	0.243	0.33	0.187
fluency	—	1	0.714	0.83	0.100	0.371	**0.758**	0.100	0.243	0.100
rank	—	—	1	0.771	0.414	0.258	**0.671**	0.414	0.414	0.414
constituent	—	—	—	1	0.13	0.371	**0.671**	0.243	0.243	0.13
English-German Europarl										
adequacy	1	0.714	0.487	0.714	0.487	**0.600**	0.314	0.371	0.487	0.487
fluency	—	1	0.543	0.43	**0.258**	0.200	-0.085	0.03	**0.258**	**0.258**
rank	—	—	1	0.03	-0.37	-0.256	-0.543	-0.485	-0.37	-0.37
constituent	—	—	—	1	0.887	**0.943**	0.658	0.83	0.887	0.887

Table 10: Correlations between human evaluation and automatic metrics into German

Due to the low Kappa correlation scores, one year later, Callison-Burch, Fordyce, Koehn, Monz, & Schroeder (2008) reconsidered the design of human evaluations. More than 100 people participated in the evaluation, with a collective of 266 hours invested. Translations were evaluated in three different ways:

- Ranking translated sentences relative to each other
- Ranking the translations of syntactic constituents drawn from the source sentence
- Assigning absolute yes or no judgments to the translations of the syntactic constituents.

Besides, judges had to evaluate syntactic constituents by deciding if they were acceptable or not clicking *Yes*, *Not*, or *Not sure*, instead of using adequacy and fluency scales like in the previous years.

4.9 Summary and final remarks

In this chapter I have offered a general overview on the subject of natural language processing evaluation, with particular attention to CLs and MT. I first approach some general concepts related to evaluation types and stakeholders and I make a historical sketch of the evaluation of language processing tools that dates back to the 60s. Then I

concentrate on important aspects to be considered when designing an evaluation plan: tools, materials and evaluators. These elements can heavily influence the results of the evaluation and must be carefully chosen in order to guarantee the validity of our study.

The second part of this chapter concentrates on evaluating CL rule suites, CL checkers and MT. I discuss the metrics and some examples of CL evaluation. The concepts of readability, understandability and translatability and their relation to CLs are tackled. Then I concentrate on MT evaluation, an issue that has generated an intense debate and research due to its complexity. I review some of the research approaches intended to standardize MT evaluation and then I concentrate on the different metrics and measures, both human and automatic, developed with the aim of optimizing MT evaluation.

Part II: Methodology

5 METHODOLOGY FOR EVALUATING A CONTROLLED LANGUAGE. A THREE-PHASE APPROACH

Show not what has been done, but what can be. How beautiful the world would be if there were a procedure for moving through labyrinths.
Umberto Eco, *The name of the Rose*, 1980

5.1 Introduction

In the theoretical part of this work I have analysed the four milestones that set the grounds of this research work: the definition of controlled languages, their use in industrial environments, the relationship between controlled languages, technical documentation and translation and, finally, how to evaluate language processing applications, specifically controlled languages and machine translation.

My aim now is to set out the methodology of the empirical part of this work. As I have mentioned in the previous chapters, my goal is to analyse the effectiveness of implementing controlled languages in the authoring of technical documentation, especially with regards to the improvement of translatability and, more concretely, machine translatability for the eventual deployment of this technology within the translation process. Nevertheless, other aspects such as understandability and readability will also be considered. Further, I am also interested in studying the deployment of Machine Translation (MT) technology within the translation process in an industrial environment.

The two hypotheses that my study plans to test are, on the one side, if MT can be a neutral evaluator for the assessment of controlled language proofed texts, especially with regards to translatability. On the other side, I hypothesize that MT represents a reliable technology to confront the increasing amount of technical documentation and, thus, of translation volume. This is the reason why MT has been used to undertake the experiments.

The demonstration of these two hypothesis is designed to offer empirical evidence that controlled languages bring the claimed advantages that have been discussed in chapter 1 (1.4.2), as well as to establish the elements that might lead to the recommendation or dissuading from the implementation of MT. Furthermore, this will lead to detect which rules of the linguistic tool are prone to render more translatability to the text as well as to suggest new rules which could improve both readability and translatability of the source text.

In order to carry out the empirical part of this project, first I needed to select resources that allowed me to carry out the evaluation, as it was described in Chapter 4 (4.3). This implied choosing an MT system and building a corpus of texts to retrieve relevant data and information about the best text type for my purposes. Besides, I needed another corpus for the evaluation of CL rules in order to assess the appropriateness of implementing them together with MT technology for certain types of automotive literature. Finally, I also wanted to analyse if such an implementation was cost-effective, for which types of texts it was most suitable and under which conditions it should be implemented. For this purpose, I collected economic data that allowed me to carry out an ROI (Return on Investment) analysis. My empirical approach is thus divided into three different phases:

1. Phase 1: Selection of resources. First of all, I conducted a microevaluation to detect which types of texts were most appropriate for my study on the grounds of three main factors: the implementation of a controlled language in their creation; their suitability for MT and their linguistic characteristics. Further, I also evaluated different MT systems to choose the one that best met my needs.

2. Phase 2. Evaluation. Secondly, I compiled a real corpus of texts of the best-suited type as it was stated in Phase 1. It was a comparable monolingual corpus containing, on the one hand, texts checked written following the rules of the controlled languages and, on the other hand, texts not following them. Subsequently, the MT system chosen during Phase 1 was installed and trained and texts were translated with the MT system chosen in Phase 1, thus building two parallel comparable bilingual corpora[93]. The quality of the translations was cross-checked and the data of both corpora were compared.
3. Phase 3: Workflow and ROI. In a final phase, I undertook a feasibility study that analysed the return on investment of implementing MT technology in combination with a controlled language within an industrial environment as well as the necessary adaptation of workflows and processes.

In this chapter I present the methodology of the first two phases, whereas the methodology of Phase 3 will be presented, together with the results, in Chapter 7. Phase 1 is designed according to FEMTI (see 4.6.3), which offers a framework for evaluating MT systems[94], whereas Phase 2 has its own distinctive features.

5.2 Phase 1. Framework

As I have just mentioned, FEMTI is divided into two sections: the first section contributes to the definition and description of a context in which the evaluation is going to take place. Features such as the purpose of the evaluation, the input characteristics or the role of the MT system within a translation workflow are taken into account. The second section concentrates on the MT internal and external characteristics, meaning the software architecture and the quality of the output.

FEMTI bases on the principles of context-based evaluation (Balkan, Netter, et al., 1994; Hovy, King, & Popescu-Belis, 2002a; Klein et al., 1998). This methodology postulates that, before the evaluation starts, it is important to define the context in which it is going to take place. This description contributes to choose subsequently the appropriate features to be evaluated and the appropriate metrics to evaluate these features. It has

been widely discussed that only context-based evaluations in a well-defined domain offer relevant data that fulfil the needs of the evaluator or end-user (King & Falkedal, 1990; Popescu-Belis, Manzi & King, 2001). An evaluation must be designed based on all the factors that might contribute to define the context: language pairs, goals of the translation task, characteristics of the MT system etc. Only after analysing all these factors it is possible to interpret the results of the evaluation in the appropriate way.

Once the context is analysed, it is necessary to choose the features most appropriated to be evaluated, adapting and expanding the framework to the needs of the context, in my case an industrial environment. For the measurement of these features, however, FEMTI only offers, if any, a listing of different metrics from the literature, without assessing any standard. The user must decide, according to the context defined, which metrics (either human or automatic) are most appropriate to measure the features chosen or he must develop new metrics according to his needs.

This research work seeks to establish a standardized methodology for similar industrial contexts where MT comes as a technology into question. White & Taylor (1998) state that an ideal MT evaluation method "should be readily reusable, with a minimum of preparation and participation of raters or subjects". Goals of this work are, thus, to employ standard and objective metrics and to make the evaluation design re-usable for future potential evaluations within similar contexts.

5.3 Phase 1. Evaluation requirements

First I will subsume the specification of user needs along with other aspects preliminary to evaluation, that is, I will describe the context in which the evaluation will take place.

5.3.1 Purpose of the evaluation

FEMTI distinguishes seven types of evaluation, each one of them corresponding to one purpose: feasibility, requirements elicitation, internal evaluation, diagnostic evaluation, declarative evaluation, operational evaluation and usability evaluation. These were

already explained in Chapter 4 (4.2.1). I am mainly interested in declarative evaluation for both Phase 1 and 2. According to White (2000: 104), "the purpose of declarative evaluation is to measure the ability of an MT system to handle text representative of an actual end-user".

In Phase 1, I pursued two goals: First, to determine which information type within the automotive literature was most appropriate and representative for the actual end-user and, thus, for Phase 2; second, to determine which MT system was most suitable to handle this type of texts and therefore would be most adequate for Phase 2.

5.3.2 Object of the evaluation: the MT system

In a previous non-dynamic version[95], FEMTI suggested the evaluation of MT as three different objects: the test of a component of a MT system, MT system considered as a whole, and MT considered as a component of a larger system. I focused my analysis on Machine Translation as a whole. Indeed, in Phase 1 I was only interested in how well the system performed as a stand-alone software tool and not how it could be integrated in a workflow. Despite of this, essential aspects such as existing interfaces and export/import facilities were addressed, since the results of the selection and the evaluation would deliver a recommendation of the implementation of MT for a future translation workflow in an industrial environment.

My purpose was thus to determine which system could perform best for a certain type of text and would be therefore most appropriate for the evaluation afterwards. In order to assess the best MT system for my purposes, it was necessary to evaluate the output quality of machine translations and to choose one system for further tests, carrying out an horizontal evaluation in terms of Rinsche (1993a: 267-268)[96].

First of all, I conducted an Internet and literature inquiry and considered criteria to pre-select three commercial systems:

- Language pairs. As Bennett & Gerber (2003) point out, one of the three key factors of commercial use of MT is the language direction. This is indeed a feature that FEMTI puts in the second part of the current framework (2.1.2.4.1. Linguistic resources and utilities, Languages). Since the source language of my documents is German, the language pairs selection metric was based on the greatest number of language-pairs from German and into German. However, since the tests were going to be carried out with German texts that should be translated into English, I prioritized those systems that had the language pair German ↔ English.

- Terminology. Bennett & Gerber (2003) indicate as a second key factor the dictionary coverage. This feature is presented in the current dynamic FEMTI in the second part, 2.1.2.4.2 Dictionaries. Two aspects were considered for this characteristic: specialized dictionaries and the possibility to create user dictionaries for corporate terminology. Further, it was considered if the systems offered the possibility of integrating user dictionaries for specific domains. This is of utmost importance for my project since usually automotive companies manage their own terminology and strive for a corporate univocal terminology in all languages. Terminology is one of the crucial points where quality of MT can extremely vary (El Haidi et al., 2004; El Haidi, Timimi, & Dabbadie, 2001). Petit (1977) already noticed this fact when he concluded that "correct translation of 'grammatical' words, frequent non-technical words and technical words and expressions is INDISPENSABLE". Therefore, it is very important to maintain a controlled terminology so that the MT process runs smoothly and texts are produced consistently. In order to import this terminology in the MT systems, these must offer an interface that accepts different import formats, such as Excel, TXT, TBX, CSV, Trados Multiterm XML or Martif, among others. Another important aspect is if these dictionaries are easy to maintain and if there are special tools to do so.

- Status of Vendor: As mentioned before, I conducted a literature and Internet research to check if the systems had successfully carried out projects with relevant clients, especially if they were also industrial and even automotive company customers. This aspect was inspired by Arnold et al. (1994: 158) who stated that

"Buying an MT system is a considerable investment, and the stability and future solvency of the vendor is an important consideration". Despite the fact that the goal of the selection of one system is not to integrate MT in the current translation processes, it makes sense to consider this aspect since, if MT is ever going to be embedded in translation processes, a network solution with server/client architecture is simply indispensable.

- Evaluation studies: I also checked if the systems have been evaluated in other studies as well as the results obtained compared with other systems.

5.3.3 Characteristics of the Translation Task

This point refers to the information flow intended for the output, from the point of view of the agent (human or otherwise) who receives the translation. FEMTI quotes the work by E. H. Hovy (1999) who suggested dividing the purposes of a translation tasks into three main groups "to make the taxonomization of features to people who do not already know much about MT and do not wish to become experts in evaluation". These three groups are:

- Assimilation, the aim of which is to use translated texts produced by people outside the organization to sort, extract, summarize or search for relevant information.
- Dissemination, aiming at delivering to others (internal or external users) a translation of documents produced inside the organization.
- Communication to support multi-turn dialogues between people who speak different languages.

Applied to my context, the main purpose is to disseminate documents produced inside the organization among internal users –in this case the evaluators and ourselves, who share aspects of the culture, terminology, and domain knowledge to some extent, though they speak in different languages.

5.3.4 Input Characteristics: Selection of a text type

Input characteristic embrace two main aspects: the properties of the source document, ranging from its form, format, topic domain to its linguistic characteristics; and the author factors, such as proficiency in language and domain and use of authoring tools. FEMTI also distinguishes characteristics related to sources of errors and defines them as "the errors that are likely to be in the unchecked text. Errors are defined as the difference between the unchecked text and the subsequent proofed text."

5.3.4.1 Document type

Bennett & Gerber (2003) presented three essential factors for the commercial use of an MT system. The first two, language direction and dictionary coverage, have been mentioned in 5.3.2, and will be tackled in depth in Chapter 6. The last one, suitability of the text, will be addressed here. It is generally accepted that certain types of text are more appropriate for MT than others, such as technical documentation. Numerous references underscore this view (Bernth & Gdaniec, 2001: 175; Church & Hovy, 1993; Lehrberger & Bourbeau, 1988: 192).

My evaluation focuses on technical documentation. I concentrate on service texts, that is, texts that are produced within the Service and After-Sales processes of an automotive company. The selection of one of these information types will provide a basis for Phase 2.

Since my evaluation is mainly a declarative evaluation where the performance of an MT system is tested in a given context, I will use text corpora for my experiment. In this respect, I follow the recommendations by Holmback, Hubert, & Spyridakis (1996), who state that constructed documents would bias the conception of ideal CL texts, whereas the use of documents occurred naturally ensures the relevance of the study with regards to the application of CL in industry.

To choose the most appropriate type of text, I checked following requirements:

- Integration within an authoring system. First of all it was checked if the information type was likely to be included in an authoring system. This means that it is produced regularly and the contents need to be managed. The fact of being included in such an environment means that within that environment, different technologies could be integrated, such as CL or Machine Translation.
- Controlled Language Application. A second important criterion is the existence of a language quality process with the application of a controlled language. Since one of the goals of this work is to empirically determine if language quality checking with MUTLILINT/CLAT brings any advantages and to which extent, it makes sense to choose an information type which is being already checked by MULITLINT/CLAT. In this way, I will be able to determine which MT-system best interoperates with texts produced with MULTILINT/CLAT. A detailed analysis of the CL compliance is presented in 1. This analysis is also linked to the linguistic characteristics of the documents, since it includes data about the grammaticality, the use of terminology and style issues.
- External characteristics. These criteria consider process and context related characteristics that could make a text type appropriate or not for the technology MT. Among them I could distinguish security aspects (if the texts are security relevant it is not advisable to translate them with MT since the degree of accuracy is not very high and could lead to accidents or cause damages); the degree of experience of the authors with a CL application; goal languages (this aspect will also influence the decision for a MT system depending on the language pairs it offers) and publication volume (the bigger the amount, the more is it worth deploying MT).
- Linguistic characteristics. Important aspects are the terminology, the structure of the sentences, the length of texts (neither too long nor too short), as well as the translatability indicators studied in Chapter 4 (4.4.1.2). After a detailed study, I grouped these criteria into 4 main groups:
    - Formal Rules. This group includes criteria regarding punctuation, formatting, layout and orthography. This is an essential category for Machine Translation since output quality can enormously suffer if

segmentation is not carried out properly. Aspects in this category include the use of punctuation marks, parentheses, lists, spacing and the spelling.

- Grammar. This group includes syntactic indicators such as ambiguous or too complex structures, subordinate and coordinate clauses, order of elements, use of pronouns, prepositions and articles and sentence length. Other aspects refer to the use of certain verbal forms and tenses, the structure of noun phrases and the presence of ungrammatical constructions.
- Terminology. The restricted use of variants (spelling variants, compound variants, synonyms), abbreviations and acronyms, as well as the usage of a consistent and standardised terminology constitute the main focus of this group.
- Style. This group concentrates on elliptical and passive constructions, the use of metaphors, slang or dialect variants and application of negation.

According to this classification and the data analysed in 0, the most recurrent rules are those related with avoiding long sentences and elliptical constructions. Other language-specific recommendations are related to the use of pronouns, the imperative form and the use of the passive voice. Further, there are also some general rules regarding formatting.

Chapter 6 will depict the results of the analysis of these recommendations applied to the type of text selected.

5.3.4.2 Author Characteristics

Author characteristics are defined by FEMTI as a set of characteristics that cover writer attributes that are relevant to the writing task, which influence the text that is produced.

Authors of automotive texts within an automotive company usually are both internal and external (through agencies or consultancies). They are mostly educated native speaker individuals with a background either in technical writing or in engineering. Very often they have received further training for the special task of writing technical

documents for the automotive company. Some of the information types that can be created within the automotive industry are Service Information, Repairing Instructions, Tightening Torques, Inspection Sheets, Technical Data and Training Documents, Diagnosis, Technical Campaigns, Flat Rates or Programming data.

In the writing process, some authors are supported by an authoring tool when checking and proofreading texts in the case of some of the texts that were analysed for this study. The tool that was used when the analysed texts were written is MULTILINT/CLAT. This tool contributes to check the terms used in the text, to the creation of short and intelligible sentences, and to apply abbreviations correctly. However, it must be taken into account that, depending on the experience of the author, the degree of application of the controlled language as well as the time used to write with this kind of authoring support can vary considerably. Further, different authors have different styles. A CL authoring tool aims at diminishing these variances as much as possible. Furthermore, I decided to include in this study only experienced authors with a definite and coherent writing style.

5.3.4.3 Characteristics related to sources of error

According to FEMTI, these errors fall into three categories:

- Intentional errors. Errors in this category include dialect differences between the writer's language and some standard language, second language errors such as wrong prepositions in prepositional phrases and genuine misconceptions.
- Medium-related source errors. Considering sources of writer errors during the writing process, this characteristic includes: errors from speech recognition, from OCR, cut/ copy and paste slips, etc.
- Performance-related error sources. Depending on the writer model, this type of errors includes concentration lapses resulting in "derailed" sentences (for example slips through tiredness), planning fault errors (for example failed agreement between noun phrase determiner and header) and other performance errors.

Since the texts I will use in Phase 1 have been proofed with regards to the rules of a controlled language, the number of errors should be minimized as much as possible, especially intentional and performance-related errors. Medium-related errors, however, are less prone to appear since speech recognition or OCR are not methods used for creating the texts of my study.

5.3.5 User Characteristics

In this section, user needs are specified. According the definitions in the FEMTI Framework, the end-user must not be necessarily understood as the final recipient of the translation. A user is always a human and can be either the person who interacts with the machine translation system (either for evaluation or for tuning up the system), or the end user of the final product, or the organisation deploying the machine translation system.

The different users can have different motivations for using this type of technology. For instance, in my case the end users of the translations are, on the one side, the evaluators who are going to assess their quality according to different aspects, and, on the other side, I as final assessors, who are going to evaluate the results of the pre-selection to draw conclusions.

FEMTI proposes three types of end-users, which are defined and can be adapted to my situation as follows:

5.3.5.1 Machine Translation user

Definition. This refers to the translation producer who interacts directly with the machine translation system or with the raw output produced by the machine translation system. This user may be, on the one hand, an administrator or translation project manager in charge of carrying out the translations; on the other side, a translator or a post editor could play this role as recipient of the MT raw output.

Adaptation. I as the precursor of this study will interact directly with the machine translation system in order to make the appropriate configurations and to carry out the translations.

5.3.5.2 Translation consumer

Definition. This refers to the person or organisation to whom the translation product is delivered. Subsequent use of the translation is intimately related to characteristics of the translation task.

Adaptation. The translation product is delivered to the evaluators for assessment. A group of translators will be in charge of assessing the quality of the translations. The evaluators are native speaker translators with a long experience in translating automotive texts and a high computer literacy; hence their assessments are of great value.

5.3.5.3 Organisational user

Definition. an organisational user of MT may be a corporate user, a translation service, a translation agency or other provider of translation.

Adaptation. I act as corporate users and am therefore end-user of this category.

To sum up, my users for Phase 1 are divided in two groups:

- Translator/post-editor evaluators. These will evaluate the translation quality regarding the features comprehensibility, readability and fidelity. Besides, they should assess the translations as to "post-editability". This should deliver the degree to which the quality of the translations is acceptable for a future translation process with post-editing. Through the degree of post-editability and some directed questions, I can infer the degree of acceptability of this group with respect to this task.

- Organisational user as surveyor. This role analysis the translation output with regards to terminology and wellformedness, stating if there is any relationship between formal errors and comprehensibility and readability difficulties, and fidelity problems. To end with, this end-user will have to analyse the evaluation results and draw conclusions from the data. This is my role.

FEMTI proposes to take into account some contextual or environmental factors such as the proficiency in the target and source languages and the computer literacy for the evaluation. This will be done in part with a questionnaire in order to catch up factors which could influence or deviate the statistical results of the survey.

As with the number of individuals for each group, it is important to maintain the groups as homogenous as possible and as J. S. White (2000: 104) suggests, to gather a fairly large sample of evaluators for each group in order to counteract the extremely subjective nature of attributes such as intelligibility and fidelity. Since the evaluation group was not so big as desirable, an additional measure to guarantee the reliability of results was introduced: the inter-rater agreement which is calculated through the Kappa statistic (Callison-Burch et al., 2007; Carletta, 1993). The Kappa coefficient measures "the difference between how much agreement is actually present ('observed' agreement) compared to how much agreement would be expected by chance alone ('expected' agreement)" (Viera & Garrett, 2005):

$$K = \frac{P_{obs} - P_{exp}}{1 - P_{exp}}$$

where P_{obs} is the observed agreement and P_{exp} is the expected agreement. P_{obs} is calculated as

$$P_{obs} = \frac{\text{number of agreements}}{\text{number of agreements} + \text{number of disagreements}}$$

and P_{exp} is calculated as

$$P_{exp} = \sum_{i=1}^{n} (p_{i1} + p_{i2})$$

where:

n = number of categories

i = category number

p_{i1} = proportion of occurrences of category i for evaluator 1

p_{i2} = proportion of occurrences of category i for evaluator 2.

Kappa scores can vary from -1 to 1, where 1 is perfect agreement, 0 is agreement due to chance, and -1 is perfect disagreement. Here is one possible interpretation of Kappa.

- Poor agreement = Less than 0.20
- Fair agreement = 0.20 to 0.40
- Moderate agreement = 0.40 to 0.60
- Good agreement = 0.60 to 0.80
- Very good agreement = 0.80 to 1.00

5.4 Phase 1. Customized Quality Model

The new dynamic version of FEMTI offers evaluators the possibility to generate a customized quality model with the relevant quality characteristics according to the specified context of use.

In order to do so, once the context of use is defined, evaluators have to select the quality characteristics and metrics of interest. I first selected the different aspects of the context that has been previously described:

First of all, it is a declarative evaluation with internal dissemination as the main translation task. The input is a concrete type of text, with a definite genre and domain

(technical documentation of the automotive domain), thus the MT system can be customized with the right terminology and certain grammar rules.

With regards to the authors, they are all proficient in the source language with a superior-level of performance. They have all received professional training and have experience producing the type of texts object of this analysis. Further, I do not consider sources of errors since texts are proofed and written by professional technical writers.

With respect to the user characteristics, FEMTI distinguishes three types of users: the end user who will interact with the machine translation system; the end user of the final product of the translation process which may include for example, post-editing; the organisation deploying the machine translation system. In this study, the end user who interacts with the MT system is me, since I installed, customized and carried out the translations. I hold a formal linguistic education, with a distinguished level in the source language (English) and a superior knowledge of the target language (German). Further, I also have a high level of computer literacy. The translator consumers are the evaluators who will assess the output of the MT system. They have both a distinguished level in the source and target languages. Finally, the organisational user is simulated by me. Data regarding the volume of translation, the number of personnel and the time allowed for translation is simulated according to information from the literature and personal interviews.

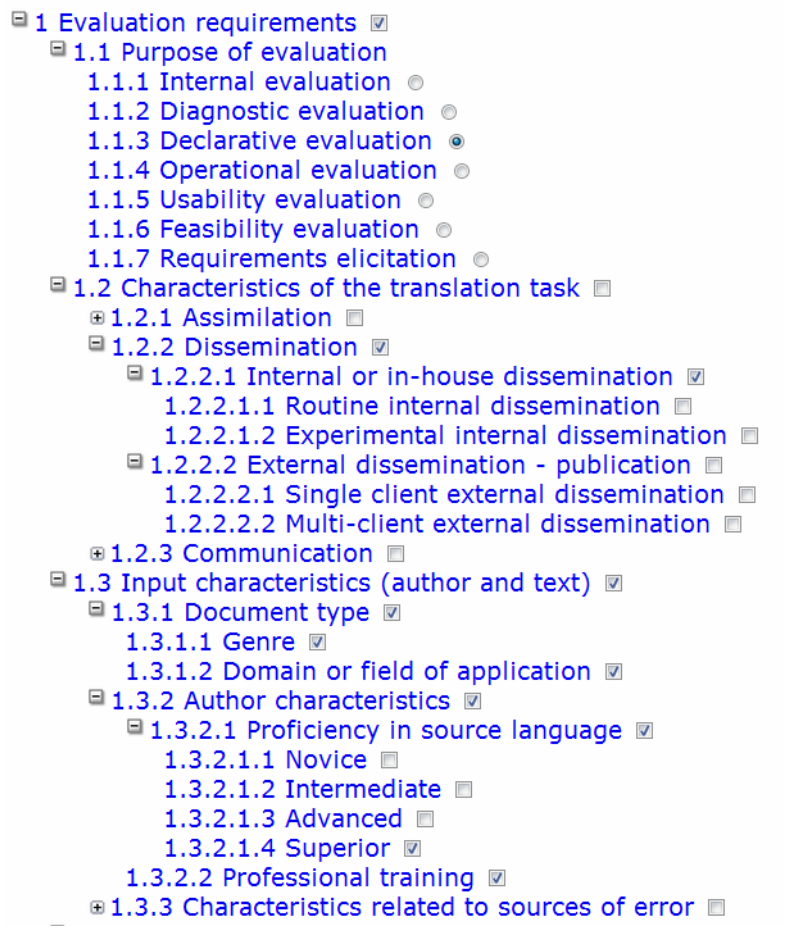

Figure 12: FEMTI evaluation requirements

After describing the context and following the instructions in order to generate my customized quality model, I had to decide which system characteristics were to be evaluated and which metrics were going to be used. FEMTI first proposed a series of system features, as can be seen in Figure 13.

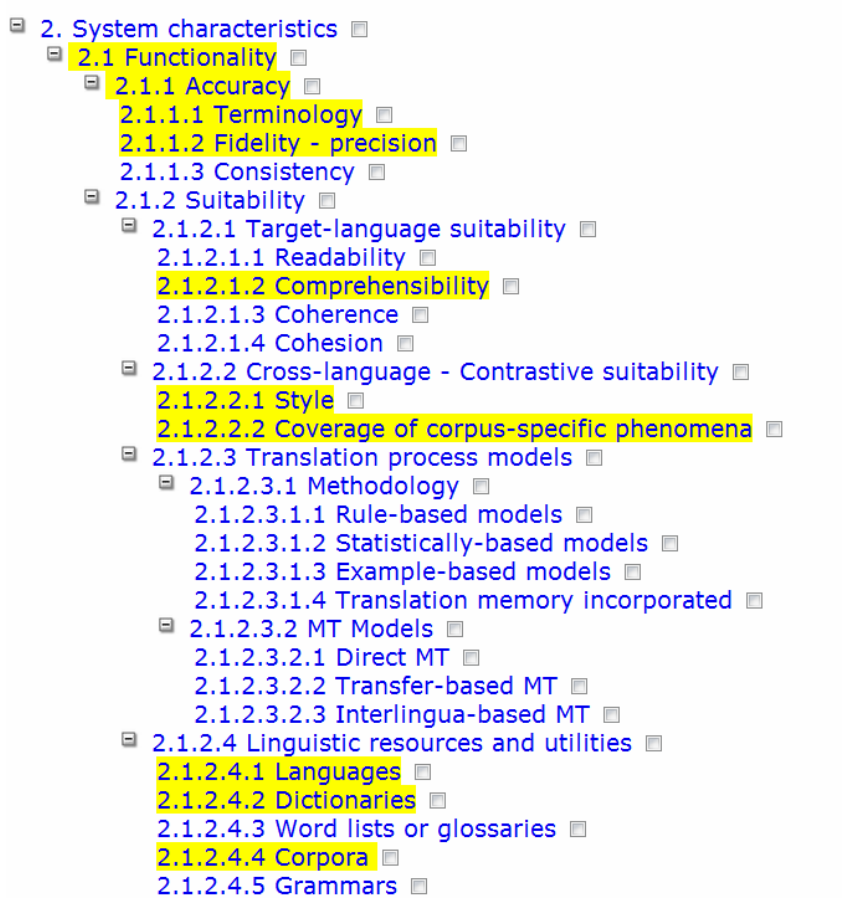

Figure 13: FEMTI proposed system characteristics

Figure 14: FEMTI selected characteristics and metrics

As we can see in Figure 13, our customized quality model suggested the following features to be evaluated:

- Functionality. Terminology and fidelity-precision under accuracy; comprehensibility, style, coverage of corpus-specific phenomena, languages, dictionaries, corpora and dictionary updating under suitability and, finally, well-formedness.
- Reliability.
- Usability.
- Efficiency. Overall production time and input to output translation speed under time behaviour and memory usage under resource utilisation.
- Maintainability. Ease of dictionary update under changeability, also included, and stability.
- Portability. Adaptability and installability
- Cost: Other costs.

All these features will be detailed in 5.4. Once I chose the relevant quality characteristics and metrics for my evaluation plan, I could save it as a PDF document, which summarizes my evaluation plan and is available in Annex IV.

In the next section I explain the particularities of this plan, as well as the deviations that were necessary in order to make the model feasible for us.

5.5 Phase 1. System characteristics

The previous FEMTI distinguished between MT-system-specific characteristics and system external characteristics, based on the distinction made by ISO 9126. The former pertained to the internal static properties of the software and the latter are the characteristics that can be observed when the system is in operation. There is some connection here with the notions of glass box and black box evaluation. These internal

characteristics are included in the current dynamic version of FEMTI as part of the Functionality top-level characteristic, as it can be seen in Table 11.

Old FEMTI Classification	New dynamic FEMTI Classification
2. System characteristics to be evaluated	2. System characteristics
2.1 System internal characteristics	2.1 Functionality
2.1.1 MT system-specific characteristics	2.1.2 Suitability
2.1.2 Translation process models	2.1.2.3. Translation process models
2.1.3 Linguistic resources and utilities	2.1.2.4. Linguistic resources and utilities
2.1.4 Characteristics of process flow	2.1.2.5. Characteristics of the process flow

Table 11: Comparison of system characteristics

As we learned in Chapter 4 (section 4.6.3), FEMTI took as a starting point the ISO/IEC 9126 and ISO/IEC 14598 standards, which are domain independent guidelines for the evaluation of software products and are, therefore, intended to be applicable to all kinds of software.

ISO/IEC 9126 defines quality as "the totality of characteristics of an entity that bear on its ability to satisfy stated and implied needs". The goal of the ISO/IEC quality model is to represent the overall quality of a product as a result from six-top level characteristics: functionality, reliability, usability, efficiency, maintainability, portability. Each characteristic is further decomposed and certain attributes are assigned, which are the terminal nodes of the hierarchy and represent the measurable features of the software product. In order to measure these attributes or features metrics need to be associated.

FEMTI represents an adaptation of this model to a particular domain, defining new attributes and metrics appropriate for that particular domain. According to Estrella et al. (2009) "In FEMTI the ISO/IEC generic quality model was tailored to the MT domain, maintaining its top-level structure and extending it with an additional top-level quality characteristic, namely Cost, and with sub-characteristics specific to MT systems".

5.5.1 Functionality

Functionality is defined as "the capability of the software to provide functions which meet stated and implied needs when the software is used under specified conditions".

With regards to MT translation, functionality embraces two groups of features: some of them are related to the general functioning of the software, whereas some other refer to the quality of the translation output. In the first group interoperability, functionality compliance and security can be included. In the second group, the following characteristics are to be found: accuracy, suitability and well-formedness.

According to our quality model, all features to be evaluated within the Functionality quality characteristic relate to quality[97].

FEMTI distinguishes two modes in which quality of a translation can be evaluated: without and with adjustment. In the first case, the system is evaluated before the dictionary and/or grammar is adjusted. In the second case, dictionary and/or grammar are adjusted, in order to obtain the best possible results. Of course, the more adjustments are realised, the more severely the evaluation has to be made.

Since I am interested in achieving the best possible translation quality in order to choose the most appropriate system for an industrial environment, I opt for the second option. Indeed, the correct translation of domain-specific terms is, for the texts intended for this Phase 1, of utmost importance. It must be then checked if these terms have been correctly translated and their percentage (Voss & Van Ess-Dykema, 2000). If the term has not been correctly translated, this can have two grounds: the term has not be translated or badly translated because there is no entry in the dictionary for it. In this case, the dictionary has to be maintained and actualised. The second ground can be that the term has been badly translated due to a failed analysis. In this last case, it must be checked if, through modifications in the dictionary and/or configuration, a correct translation can be achieved.

Thus the systems were filled with automotive specific terminology for the language pair German-English and settings were reviewed according to the style of the documents. For instance, one common rule when translating instructions from German into English is that the verbal forms ended in *–en* should be interpreted as imperatives and not as infinitives. However, this depends on the context, since the same sentence used as the title of a paragraph could be translated as an infinitive. Thus, the sentence *Signal prüfen* could be translated as *Check signal* or *Signal check*.

A human and an automatic evaluation were carried out in order to cross-check results from both tests and determine if automatic evaluation was rendering reliable results. This had a twofold purpose: on the one hand, to cross-check automatic evaluation results with human assessment; on the other hand, to prove if only automatic evaluation methods could be meaningful enough to carry out an evaluation and to make decisions on their basis. This would without doubt constitute an approach towards the ideal MT evaluation method suggested by J. S. White & K. B. Taylor (1998): "readily reusable, with a minimum of preparation and participation of raters or subjects". However, it's necessary to bear in mind that this type of evaluation only renders data regarding how good is a system compared to others or if the system has improved during a development process. Information on the types of mistakes, the need for post-editing or the linguistic quality of the text for dissemination is rarely available when carrying out automatic evaluation methods[98]. Therefore, depending on the purpose of the evaluation, automatic methods can be useful or not.

Within functionality, relevant qualities for declarative evaluations are translation process models, linguistic resources and utilities, suitability, accuracy and well-formedness. Our customized quality model included terminology and fidelity-precision under accuracy; comprehensibility, style, coverage of corpus-specific phenomena, languages, dictionaries, corpora and dictionary updating under suitability and, finally, well-formedness. I decided to exclude style as part of the evaluation because I considered this feature to be too subjective and not so relevant for our scenario. Further, under suitability I included the translation process models, giving a description of rule-

based direct MT models, since all the systems I included in the evaluation were rule-based.

I also included a new feature which is not directly represented in FEMTI, post-editability, which I adapt from Roturier (2006).

I will now detail all these features and will explain the methods and metrics used to evaluate them.

5.5.1.1 Accuracy

Accuracy is defined in FEMTI as "the capability of the software product to provide the right or agreed results or effects with the needed degree of precision." According to Margaret King (2005), who offers an interesting discussion around the dichotomy of the concepts accuracy and suitability, "this leads to an interpretation of accuracy as something very close to conformity to specifications: a software is accurate if it produces the results or effects that its specifications say it will."

Under accuracy we find terminology and fidelity-precision. I did not consider it was necessary to evaluate terminology directly, since, as I have mentioned before, I imported the relevant automotive terminology into the systems. Therefore, I assume that most terms are correctly translated.

Contrarily, I did chose to evaluate fidelity-precision, which can be defined by Van Slype (1979: 72) as the "subjective evaluation of the degree to which the information contained in the original text has been reproduced without distortion in the translation". In automotive literature fidelity is an essential factor, since information must be transferred accurately and free of content mistakes. This is especially important when a misinterpretation or false information can cause damages or negligence in security aspects.

To measure fidelity I developed a simple 3-point scale that aims at measuring to which point the information is transferred from one language to the other:

1	Totally faithful	All relevant information is correctly transferred
2	Fairly faithful	Despite of minor sense nuances, the main information is transferred. There are silences (word not translated) or noises (word added by the system)
3	Totally or almost completely unfaithful	The information has not been accurately transferred. There are important sense errors such as: – C.S.: contrary sense: the translation says exactly the contrary of the source text – F.S.: false sense: the translation transmits a different sense as the one intended in the source text – N.S.: non-sense: the translated sentence is nonsensical and is therefore, not accurate.

Table 12: Fidelity scale

I plead for short scales in order to avoid a too much the excessive granularity of the evaluation, which it makes it difficult to draw clear conclusions about the results of the evaluation. Indeed, with scales that rate more than 5 points results are usually perverted since most systems usually score the most points in the middle areas, making a decision for a system very difficult, or nearly impossible.

I also applied the BLUE automatic score to see if results correlated with fidelity and could be thus used for further automatic evaluations.

5.5.1.2 Suitability

Under Suitability FEMTI suggests the following relevant characteristics to be evaluated: comprehensibility, style, coverage of corpus-specific phenomena, languages, dictionaries, corpora and dictionary updating. Suitability is more linked to user specific needs.

Comprehensibility

Comprehensibility is defined as the extent to which the text as a whole is easy to understand. The tests to measure this feature normally consist of multiple-choice questionnaires of content-related questions. However, this metric is usually applied at a text level.

As mentioned in Chapter 4 (4.4.1), readability and comprehensibility are closely related. Usually, a text that is intelligible can be well understood. On the contrary, if a text must be read repeatedly to make sense of it, comprehensibility will consequently suffer.

In order to evaluate this characteristic, I first considered a questionnaire with content-related questions. However, I ended up discarding it due to the following reasons: it is not the most economical (both with regards to time and money) method to check comprehensibility: too much time is needed to prepare the questionnaires, which are always corpus-specific. Besides, it is also necessary to spend too much time to carry out the tests, which can be tiresome for the evaluator. Furthermore, there are other factors that might influence the results, such as the previous knowledge of the evaluator, which can interfere when asking certain questions: even if text has not been understood because of a bad translation, if the evaluator knows the answer, he will tip the correct answer based on his previous knowledge rather than act only on the basis of the knowledge acquired by the translation.

Therefore, I opted for a 4-point intelligibility[99] and instructed the evaluators to be as objective as possible. In this way I expected to obtain neutral results that reflect the real quality degree of the translation.

1	Totally intelligible	The meaning of the sentence is perfectly clear. It is grammatical and reads like ordinary text.
2	Very intelligible	The sentence has minor mistakes, but is generally clear and intelligible. It is possible to understand (almost) immediately what it means
3	Intelligible	Sense can only be understood after repeated reading.
4	Non-intelligible	Sentence is unintelligible.

Table 13: Intelligibility scale

Style

Though style was suggested as one of the features to be evaluated, I decided to ignore it due to its extreme subjectivity and because I did not consider it to be among the most relevant factors of the user needs.

Coverage of corpus specific phenomena

As I mentioned in 5.3.4.1, I compiled a representative corpus with the linguistic phenomena that normally occurs in the chosen text type.

Translation Process Models

Though this feature was not included within the FEMTI suggested characteristics, I opted for including a description of the process models of the systems used for the evaluation. All systems selected for my pre-selection are rule-based systems and present a transfer MT approach. These are the most common methodologies and methods in commercial MT systems. Other approaches, such as knowledge-based and statistical-based models can achieve better results in domain-specific fields. The disadvantage, however, is that great amounts of parallel data are needed at the beginning to train the system.

Recently, the emergence of new hybrid systems combining both rule-based methodologies with statistical algorithms has burst into the scene of MT. I will discuss these advances in the final chapter, when I outline the future prospects of my research.

Languages

I considered the range of languages that the MT systems claimed to support. I was especially interested in language pairs from and into German, being this the preferred source language for automotive literature in Germany. However, I also considered language pairs from and into English as a pivot language, especially for Asiatic languages, since language pairs combining German and Asiatic languages are not that common.

Dictionaries

Another factor I analysed was the availability of general and specific dictionaries and, more specifically, if the systems included specialized automotive dictionaries. I also examined the format of the dictionaries to ascertain the possibility of importing external glossaries and terminology to the MT system.

Corpora

Corpora are one of the characteristics that were suggested by FEMTI to be evaluated. However, since at the time of the evaluation the systems I tested were all rule-based, no corpora were included. This would be a feature, nevertheless, that would be included in case statistical-based systems, the so-called SBMT, formed part of the experiment.

Characteristics of process flow

Though this aspect was not included by FEMTI, I decided to add a description of the customisation facilities offered by the systems. These include:

- Translation preparation activities**.** Especially important for Phase 1 are the translation preparation activities, which include text format aspects, performance aspects (how long can texts, sentences etc. be), lexical aspects (can terms be marked as not-to-translate, e.g. proper nouns) and configuration aspects (e.g. How to translate the German imperative in English).
- Interactive translation activities. First it must be cleared if systems offer this feature and, if so, if it is desirable to apply it. Indeed, this can either speed up or slow down the translation process, depending on where it takes place and who operates the system. In Phase 1, translation will take place directly, without any interaction.
- Post-translation activities. In this case, only the functionalities offered by the system are important for us. Questions such as "which post-editing functions are offered by the system? (E.g. can ambiguous words be disambiguated by mouse click?)" will have to be answered.

Dictionary updating

This feature includes the facilities to assist users in researching and entering terminology which the machine does not recognize into the system's dictionary. I also considered the ability of the system to include specialized or customized glossaries.

5.5.1.3 Well-formedness

This characteristic refers to the degree to which the output respects the reference rules of the target language at the specified linguistic level.

After the other characteristics have been evaluated, an analysis of the not correctly translated sentences will be made to list the errors produced by the MT system. A typology of errors will be used to classify these. It is important to check, afterwards, if these errors can be solved by applying CL rules, or if they are system dependent. This can also render where the most mistakes take place (vocabulary, grammar…) and how they relate with the readability, comprehensibility and fidelity features.

FEMTI includes four error categories, which are the most frequent: punctuation, lexis or lexical choice, grammar/syntax and morphology. Further, SAE developed a translation quality metric for service information, where the target customer of the translation is the service technician. This metric is described in the norm J2450 (SAE, 2001; Schütz, 1999), issued in December 2001. This norm distinguishes 7 error categories: wrong term, syntactic error, omission, word structure/agreement, misspelling, punctuation and miscellaneous. All these categories are scored with different weights, among them and depending on the gravity of the error (serious or minor). The translation quality is then calculated by adding all these scores and dividing them among the number of words evaluated. Other machine translation error classifications are presented by Asensio (1999) and Flanagan (1994).

In this study I work only with one target language, English. However, for multilingual studies it must be considered that "Although some error categories may apply to many languages, a unique category set should be developed for each language pair to reflect the error types that actually occur" (Flanagan, 1994).

Metric: list of errors by categories.

5.5.1.4 Post-editability

Apart from the metrics suggested by FEMTI, I also decided to add a new metric which would be especially important for the evaluation and later processes with MT. Based on the work by Roturier (2004) who designed a single metric which focus on the usability (in the sense of task-performance) of the MT output for the post-editor. He defines a scale where he conjoins readability, comprehensibility and fidelity features focusing on the subsequent post-editing process. For me it is still important to separate these features, since I want to find out what each end-user group thinks of MT output depending on the task assigned to them. It is generally assumed that there is a direct relationship between all the features: normally, a text which is highly readable, understandable and accurate will not need much post-editing.

I adapted the scale of Roturier (op. cit) for post-evaluators in the following way:

1	No post-edition needed	Read the MT output first. The text must not be modified for publication. Then read the source text (ST). The text must still not be modified for publication. Requirements for publication are: grammatically correct and proper terminology. It can be stylistically poor, but it fulfils the main objective, i.e., transferring all information accurately.
2	Minimal post-edition needed (a)	Read the MT output first. The text must be modified for publication. Only "superficial" modifications such as morphological dependencies, punctuation, accents or articles must be modified. Then read the source text (ST). No further modifications are needed.
3	Minimal post-edition needed (b)	Read the MT output first. The text must or must not be modified for publication. Then read the source text (ST). The text must be slightly modified for publication due to ellipsis, over generation or a false sense.
4	Total post-edition needed	Read the MT output first. The text must be modified for publication, but you need the source text to make sense of it. Then read the source text (ST). The text must be partly or totally modified (retranslating from scratch) for publication due to significant errors in the MT output (textual and syntactic coherence, textual pragmatics, word formation etc.).

Table 14: Post-editability scale

Apart from evaluating the sentences with regards to their post-editability, evaluators were also asked to correct them so that they were readily publishable. We base this methodology in White & Taylor (1998), who conducted an experiment for the publication task, in which evaluators had to judge texts depending on if they were publication-ready or if they had to be corrected. They recorded the number of texts "given up on" as well as the number and type of corrections made by the evaluators to texts. Following rules were given to the post-editors:

- Goal of the post-edition is to transfer all information accurately. For this purpose following options are possible:
- Rectify what is grammatically (morphological or syntactical errors) deviant from an output of commercial quality.
- Rectify what is lexically essential for the understanding of the target text (wrong or unintelligible words or phrases).
- Correct terminology only if this is wrong. Do not correct terminology in order to avoid redundancies or to improve the style.
- Try to use the words used by the system and do not use synonyms of these words to improve the style.
- The stylistic quality of the document is not as important as its accuracy and intelligibility.

Table 15: Post-editability rules

5.5.2 Reliability and Usability

FEMTI suggests reliability and usability as features to be taken into account. These features, however, refer to the quality of the software as a product, and in Phase 1 I am more interested in the output of the software as in the software itself.

However, in further phases where MT is going to be integrated within a translation workflow, these features should be considered evaluating aspects such as setting the level of access, setting up directories and file preparation and obtaining customized printouts.

5.5.3 Efficiency

Under *efficiency* FEMTI suggests the following aspects to be assessed: overall production time, input and output translation speed and memory usage. Since the corpus to be translated was rather reduced and therefore no significant time was needed in order to translate, the time difference between the human and the automatic versions was dramatic. Therefore I also considered the time needed to carry out the post-editing, in order to state the real difference between one process and the other.

5.5.4 Maintainability

Under *maintainability*, FEMTI suggests the following aspects to be assessed: Changeability, Ease of Dictionary Update and Stability. I did not include any of these features in my evaluation.

5.5.5 Portability

Under *portability*, FEMTI suggests the following aspects to be assessed: Adaptability and Installability. I did not include any of these features in my evaluation.

5.5.6 Cost

As additional information, the prices for the test-versions were considered in this phase. The cost factor is also important, since all these investment factors have to be taken into account when analysing the return on investment of implementing MT in the translation processes of a company.

5.6 Phase 2: A parallel evaluation

5.6.1 Introduction

Once Phase 1 was accomplished and I had chosen an information type and a MT system, I conducted Phase 2. I defined the goal of this phase as follows: to analyse the effectiveness of implementing controlled languages in the authoring of technical documentation, especially with regards to the improvement of translatability and, more concretely, machine translatability for the eventual deployment of this technology within the translation process. Further, I was also interested in studying the deployment of MT technology within the translation process in an industrial environment.

The two hypotheses that my study planned to test were, on the one side, if MT can be a neutral evaluator for the assessment of controlled language proofed texts, especially with regards to translatability. On the other side, I hypothesized that MT represents a reliable technology to confront the increasing amount of technical documentation and, thus, of translation volume. This is the reason why MT has been used to undertake the experiments.

5.6.2 Corpus characteristics

The first step consisted in compiling a real corpus of texts of the best-suited type as it was stated in Phase 1, that is, service texts from the automotive area.

Roturier (2004) proposes a method for creating a corpus with natural language examples and CL examples, which is adapted from the procedure described by King & Falkedal (1990) and consists of the following steps:

- Find an example from the corpus that does not conform to the rule.
- Edit this example to make sure that it conforms to all the other rules under study (this example will be referred to as example A).
- Reduce even further the linguistic complexity of the example to a minimum to make sure that no extra problems are introduced.
- Apply the CL rule under study to turn the example under study to turn example A into what will be referred to as example B.
- Repeat this procedure twice so as to obtain 3 test examples A and 3 test examples B per rule.

The method I implemented consists of using a natural occurring corpus and letting authors rewrite this corpus following the rules of the CL. Then, the cases in the corpus which have been edited following the directions of MULTILINT/CLAT are extracted and each sentence is stored with the following information:

- Example A (not checked)
- Phenomenon and rule applied (rule code)
- Example B (checked)

The result was a comparable monolingual corpus containing, on the one hand, texts checked written originally written without taking into account the rules of the controlled language and, on the other hand, texts proofed with the CL checker MULTILINT/CLAT. Subsequently, the MT-system chosen during Phase 1 was trained and all texts were translated with the MT system chosen in Phase 1.

Then I extracted the sentences that were affected by the controlled language rules and built a test suite containing the following data:

- A set of 149 sentences in German. There were two versions for each sentence: the first version as the author wrote it originally and the second version as it was corrected by the author following the indications of MULTILINT. In some cases the previous and next sentences were attached to help the evaluator with the context.
- A set of 149 sentences in German machine translated into English. There were two versions for each sentence: the first translation is from a German text as the author wrote it originally; the second translation is from a German text that has been corrected by the author following the indications of MULTILINT.

The result was two parallel bilingual corpora composed by two monolingual comparable corpora, as well as two parallel bilingual subcorpora composed by two monolingual comparable subcorpora, as it can be seen in Figure 15:

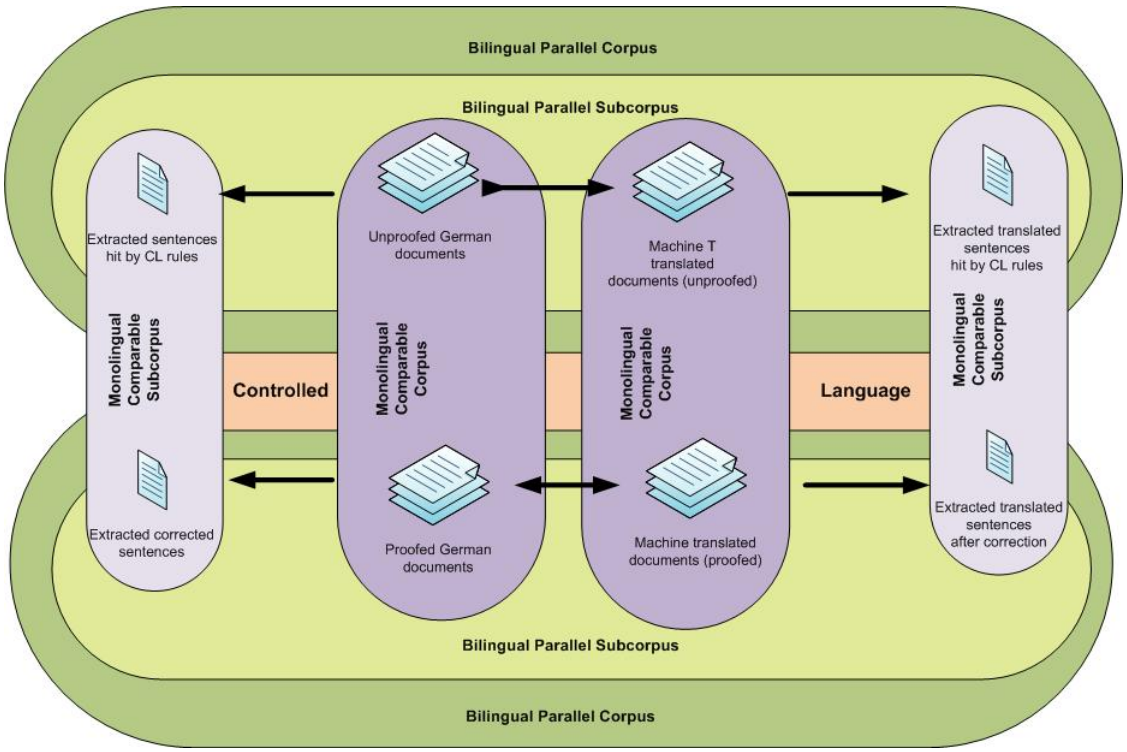

Figure 15: Design of the corpus for the parallel evaluation

The characteristics of the corpus with regards to the number of tokens and types were the following:

- Whole German corpus before MULTINT: 16410 tokens, 2786 types. After applying the stoplist: 10449 tokens and 2423 types. Most frequent word: Night-Vision.
- Whole German corpus: after MULTILINT: 16434 tokens and 2585 types. After applying the stoplist: 10468 tokens and 2384 types. Most frequent word: Night-Vision.

This corpus was then translated into English with the MT system chosen in Phase 1 (Personal Translator). From them, only 149 sentences were affected by MUTLILINT rules, containing:

- German corpus:
  - Before MULTILINT: 1940 tokens and 800 types. After stoplist: 1180 tokens and 693 types. Most frequent word: FLA (Fernlichtassistent).
  - After MULTILINT: 1949 tokens and 786 types. After stoplist: 1043 tokens and 630 types. Most frequent word: FLA (Fernlichtassistent).
- English corpus (machine translated):
  - Before MULTILINT: 2503 tokens and 725 types. After stoplist: 1348 tokens and 586 types. Most frequent word: (high-beam) headlight
  - After MULTILINT. 2544 tokens and 687 types. After stoplist: 1366 tokens and 557 types. Most frequent word: (high-beam) headlight

As we can see in the following table, before applying the stopword list[100], only 11.82% and 11.86% of the words were affected by CL rules. However, after applying the stoplist, this quantity amounted up to 28.60% and 26.43%. This might be due to the fact that the filtered stoplist includes a great amount of specific terminology which is more likely to be controlled by the CL checker.

	GERMAN					
	Before Multilint			After Multilint		
	Whole Corpus	Reduced Corpus	Percentage of words affected by CL	Whole Corpus	Reduced Corpus	Percentage of words affected by CL

Tokens	16,410	1,940	11.82%	16,434	1,949	11.86%
Types	2,786	800	28.72%	2,585	786	30.41%
Tokens (stoplist)	10,449	1,180	11.29%	10,468	1,043	9.96%
Types (stoplist)	2,423	693	28.60%	2,384	630	26.43%

Table 16: Types and tokens of the corpus for Phase 2

In both tests, I marked the position where the controlled language rule had signalled a problem, as well the position in the version after MULTILINT that indicated how the rule was applied, in order to direct the evaluator's attention to these fragments.

5.6.3 Evaluators

In order to evaluate the quality of both corpora, we conducted a parallel evaluation among two groups: a group of German native speakers, with automotive background knowledge, and a group of English native speakers, with automotive background knowledge.

There were a total of six evaluators for German and 3 evaluators for English. We are aware that this numbers are not enough to achieve statistical significance. However, the characteristics of the evaluators counteract this fact and make the results highly valuable, since they were all native speakers who worked within the automotive industry and therefore mastered the terminology and the background knowledge necessary to understand the texts. Furthermore, one of the goals of this research was to establish an ecological and reusable evaluation methodology and not to focus only on the results themselves.

This methodology contrasts with many of the studies carried out to evaluate CL and MT technologies, where students usually carry out the assessments (Babych, Hartley, & Sharoff, 2009; Spyridakis et al., 1997). In these cases it is easier to get more evaluators, but their detachment with a real context of use make them less representative than what would be desirable.

5.6.4 Metrics

In order to be able to state if there had been any improvement when applying a CL in the source language as well as in the translations of these texts, I developed a simple scale to test the CL effectiveness that evaluators had to apply in both monolingual comparable subcorpora.

This scale intends to bring together comprehensibility and terminological aspects, being the former one of the benefits CL are claimed to provide. The scale was applied both for English and German. Further, a comment field was added so that evaluators could add any relevant information

Improvement ++	The sentence is more comprehensible or the terminology is more appropriate after MULTILINT.
Worsening --	The sentence is less comprehensible or the terminology is less appropriate after MULTILINT.
No influence +	The sentence is as comprehensible and correct as before.
No influence -	The sentence is as incomprehensible and wrong as before.

Table 17: Evaluation of CL effectiveness

Once the tests were carried out, it was necessary to apply a correlation coefficient to see the relationship between both monolingual comparable corpora and to state if there was a cause-effect relationship within the parallel comparable subcorpora.

Apart from the numerical coefficient, this would result in a table randomizing all the possible assessments of the corpus and drawing the following conclusions applied to the CL rules that were signalled when proofing the texts as well as in the resulting translations:

Source	Target	Explanation	Result
++	++	There is an improvement both in the source and in the target text	Source text: positive rule Target text: positive rule

Source	Target	Explanation	Result
		⟶ MULTILINT has a positive impact.	
--	--	There is a deterioration in the source text and also a deterioration in the target text ⟶ MULTILINT has a negative impact	Source text: negative rule Source text: negative rule
+	+	The source text is as good as the target text ⟶MULTILINT shows no effect, though the quality remains good	Source text: neutral rule (it does not have any effect) Target text: neutral rule (it does not have any effect)
++	--	There is improvement in the source text, but a deterioration of the target text. → MULTILINT is effective in the source text, but not in the target text	Source text: positive rule Target text: negative rule
--	++	There is deterioration in the source text. Contrarily, MULTILINT causes a positive impact in the target text. ⟶ MULTILINT causes a negative effect in the source text, but it is effective in the target text	Source text: negative rule Target text: positive rule
++	+	There is improvement in the source text. The target text does not present any changes, but the quality is still good. ⟶ MULTILINT is effective in the source text, but has no effect in the target text	Source text: positive rule Target text: neutral rule (it does not have any effect)
++	-	There is improvement in the source text. The target text does not present any changes, and the quality remains bad.	Source text: positive rule Target text: neutral rule (it does not have any effect)

Source	Target	Explanation	Result
		→MULTILINT is effective in the source text but has no effect in the target text	
--	+	There is deterioration in the source text. The target text does not present any changes, and the quality remains good. →MULTILINT causes a negative effect in the source text, but has no effect in the target text	Source text: negative rule Target text: neutral rule (it does not have any effect)
--	-	There is deterioration in the source text. The target text does not present any changes, and the quality remains bad. →MULTILINT causes a negative effect in the source text, but has no effect in the target text	Source text: negative rule Target text: neutral rule (it does not have any effect)
+	++	The source text does not present any changes, but the quality remains good. There is an improvement in the target text. →MULTILINT shows no effect in the source text and has a positive effect in the target text	Source text: neutral rule (it does not have any effect) Target text: positive rule
+	--	The source text does not present any changes, but the quality remains good. There is deterioration in the target text. →MULTILINT shows no effect in the source text and has a negative effect in the target text	Source text: neutral rule (it does not have any effect) Target text: negative rule
+	-	The source text does not present any changes, but the quality remains good. The source text does not present any changes, and the quality remains bad.	Source text: neutral rule (it does not have any effect) Target text: neutral rule (it does not have any effect)

Source	Target	Explanation	Result
		→MULTILINT shows no effect neither in the source text nor in the target text	
-	++	The source text does not present any changes, but the quality remains bad. There is an improvement in the target text. →MULTILINT shows no effect in the source text and has a positive effect in the target text	Source text: neutral rule (it does not have any effect) Target text: positive rule
-	--	The source text does not present any changes, but the quality remains bad. There is a deterioration in the target text →MULTILINT shows no effect in the source text and has a negative effect in the target text	Source text: neutral rule (it does not have any effect) Target text: negative rule
-	+	The source text does not present any changes, but the quality remains bad. The source text does not present any changes, but the quality remains good →MULTILINT shows no effect either in the source text or in the target text.	Source text: neutral rule (it does not have any effect) Target text: neutral rule (it does not have any effect)
-	-	The source text does not present any changes, but the quality remains bad. The source text does not present any changes, but the quality remains bad →MULTILINT shows no effect either in the source text or in the target text.	Source text: neutral rule (it does not have any effect) Target text: neutral rule (it does not have any effect)

Table 18: Parallel Evaluation Scale

5.7 Summary and final remarks

In this chapter I have settled down the methodology that was applied to carry out the empirical part of this work. This methodology consists of a three-phase approach that covers different aspects: first of all, the selection of results; second, the evaluation of the CL rule suite and, finally, the study of the integration of this technology and MT into a workflow, with the economic implications that this might bring.

The first phase is based on FEMTI, a theoretical framework that sets the grounds of MT evaluation. First of all, I described the context of use where the evaluation was going to take place. This step is essential in order to choose the most appropriate characteristics to be evaluated. Further, I discuss the different characteristics and metrics proposed by FEMTI and I adapt them to my own study. These results are discussed in next chapter. In the second phase I present a methodology that aims at comparing the different results of CL-proofed texts and CL-non-proofed texts, with the goal of stating if the application of a CL brings any real advantages.

Although I present a three-phase approach, the methodology and results of the last phase, the workflow and ROI analysis, are entirely presented in Chapter 7.

Part III: Results, Conclusions and Future Prospects

6 ANALYSIS OF RESULTS

However beautiful the strategy, you should occasionally look at the results.
Sir Winston Churchill. British politician (1874 - 1965)

6.1 Introduction

In this chapter I present the results of the analysis and evaluation of Phase 1 and 2, as well as the conclusions of both evaluation phases. The methodology and results of Phase 3 will be presented in Chapter 7. Phase 1 constitutes a declarative evaluation the aim of which was to determine which MT system best handled representative texts. The goal was thus twofold: to choose a text type which was representative for the end-user and to choose the best MT system. The goal of Phase 2 was to determine if there was any improvement between texts written with and without the aid of a CL, especially with regards to their translatability.

6.2 Phase 1: selecting resources

6.2.1 MT system

In order to choose a MT system, I undertook an Internet research and considered the following features to pre-select three systems: language pairs, the ability to manage terminology, the status of vendor and previous evaluation studies.

With regards to the language pairs, the following results were obtained:

- Personal Translator[101] offered 4 language directions at the time of this research: English ↔ German and French ↔ German. Currently they have added five

more language: Spanish, Italian, Brazilian Portuguese and Chinese. However, the number of language pairs with German as a source/target language has not varied.

- Comprendium[102] offered 23 language directions:

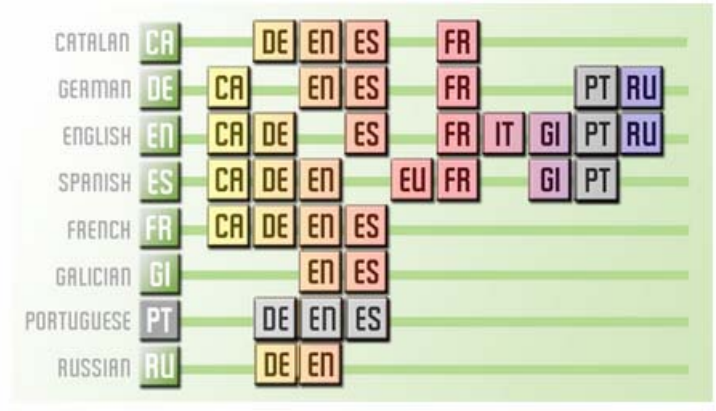

Figure 16: Comprendium language pairs

At the time of the empirical study, the language pairs German ↔ Catalan, English ↔ Galician and Spanish ↔ Galician were not available. Currently they are preparing the language pairs German ↔ Portuguese, English ↔ Portuguese and Spanish ↔ Portuguese.

- Systran[103] offers the greatest number of language pairs with up to 52 combinations, though the languages available depend on the version of the software:

52 ***available language pairs***	English ⇔ Arabic English ⇔ Chinese English ⇔ Dutch English ⇔ French English ⇔ German English ⇔ Greek English ⇔ Italian English ⇔ Japanese English ⇔ Korean	English ⇔ Polish English ⇔ Portuguese English ⇔ Russian English ⇔ Spanish English ⇔ Swedish French ⇔ Dutch French ⇔ German French ⇔ Greek French ⇔ Italian	French ⇔ Portuguese French ⇔ Spanish German ⇔ Italian German ⇔ Portuguese German ⇔ Spanish Italian ⇔ Portuguese Spanish ⇔ Italian Spanish ⇔ Portuguese

Figure 17: Systran language pairs

Since the source language of the documents of this research work is German, the language pairs selection metric was based on the greatest number of language-pairs from German and into German. In this respect, Comprendium was the system offering most language pairs from and into German (English, French, Spanish and Russian)[104]. Systran and Power Translator both offer two language directions from and into German, with English and French as source or target languages respectively[105]. Besides, one of the advantages of the Systran system is the high number of language pairs with English as a source language. Since there is no system that translates from German into all target languages that a big automotive company such as BMW or Mercedes Benz would translate into, this aspect is interesting when considering a translation workflow where English, and not German, could be the source language of the documents.

Terminology. All three systems provide technical or automotive dictionaries. Comprendium offers a Common Technical Vocabulary with 13,700 entries and different specialized dictionary modules such as Electrical engineering with 15,200 entries or Mechanics with 3,736 entries. Systran offers 20 specialised dictionaries, with a dictionary for the automotive domain apart from Electronics and Mechanical Engineering glossaries. Personal Translator also offers an automotive dictionary. Further, it was considered if these systems offered the possibility of integrating user dictionaries for specific domains. This is of utmost importance for this research since companies using a CL strive for a corporate univocal terminology in all languages. Terminology is one of the crucial points where quality of MT can extremely vary (Dabbadie, El Hadi & Timimi, 2004: 19; El Haidi et al., 2001). Therefore, it is very important to maintain a controlled terminology so that the MT process runs smoothly and texts are produced consistently. In order to import this terminology into the MT systems, these must offer an interface that accepts different import formats:

- Systran offers Excel, TXT, Trados Multiterm XML; Martif[106] can be implemented on request.

- Comprendium offers LIF, TXT and CSV; conversion from XML and other on demand.
- Personal Translator offers XML and TXT.

Since the only common standard is txt, an export from the current terminology in plain text format will be made.

Another important aspect is if these dictionaries are easy to maintain and if there are special tools to do so. Comprendium with its Dictionary Manager and LexShop, as well as Systran (Dictionary Manager) offer this feature. Personal Translator does not offer any special tool for dictionary management; rather dictionaries are administrated in the MT program itself.

Status of Vendor: as mentioned before, literature and Internet research (Flanagan, 2002; Maier, Clarke & Stadler, 1999; Morland, 2002; Nübel, 2000; Rychtyckyj, 2002, 2006b) have confirmed that all of these three systems, each to a different degree, have successfully carried out projects with important clients. As Arnold et al. (1994: 158) state "Buying an MT system is a considerable investment, and the stability and future solvency of the vendor is an important consideration". Representative examples are Comprendium providing the German companies SAP and Volkswagen with MT in their translation workflow and Intranet[107]; Systran, which delivers translation services to the European Commission and DaimlerChrysler; and Personal Translation, which has been chosen by Siemens for an Intranet application. All of these examples and references show the network feasibility of these systems. In spite of the fact that the goal of Phase 1 is not to integrate MT in the current translation processes, it makes sense to consider this aspect since a network solution with server/client architecture is indispensable if MT is ever going to be embedded in the translation processes.

Evaluation studies: all of these systems have been evaluated in other studies (Bohan, Breidt & Volk, 2000; Nübel, 1998; Seewald-Heeg, 1998; Seewald-Heeg, 1998) and have obtained the best general results compared with other systems or were pre-selected for the evaluation on the basis of favourable characteristics. This is an important

argument to assess that the quality of these systems is above the average. The hypothesis of this work is that a higher linguistic quality should result in a better task performance.

The selection of one (or maximum two of these systems) does not mean in any way that this system is the best to be implemented within the translation processes of an automotive company. As it has been mentioned before, the main goal of this phase is to assess if, in general, the translation output quality of MT is satisfactory enough so that the technology can be considered for the future. It could also be thinkable, with regards to the language pair issue, to conceive a process where English is used as a pivot language if it is demonstrated that the quality dramatically deviates from one system to the other.

Once the systems were selected, they were tuned to achieve the best possible results. Indeed, FEMTI distinguishes two modes in which quality of a translation can be evaluated: without and with adjustment. In the first case, the system is evaluated before the dictionary and/or grammar is adjusted. In the second case, dictionary and/or grammar are adjusted, in order to obtain the best possible results. Obviously, the more adjustments are realised, the more severely the evaluation has to be carried out. Since I was interested in achieving the best possible translation quality and my scenario was thought to use MT with adjustment with a standardised terminology, I opted for the second option. The terminology import process was monitored in order to avoid "lexical noise", as described by King & Falkedal (1990) and Roturier (2004).

The systems were filled with corporate specific terminology for the language pair German-English and configuration possibilities were revised according to the style of the documents. For instance, one common rule when translating instructions from German into English is that the verbal forms ended in *–en* should be interpreted as imperatives and not as infinitives. However, this feature also depends on the context, since the same sentence used as the title of a paragraph could be translated as an infinitive. Thus, the sentence *Signal prüfen* could be translated as *Check signal* or

Signal check. Therefore, it was necessary to check manually the correct implementation of the rule.

6.2.2 Text type

For my research I had access to different text types from the automotive company BMW AG, which I considered for the evaluation. These were texts from AWKat (flat rates catalogue), RA (repair instructions), SBT (Service Bulletin Technique), SI (Service Information), Technical Campaigns (OSCAR), PuMA and Schulungsunterlagen (training documentation).

In order to find out which was the best suited text type for the evaluation, I carried out a detailed analysis of all these types of texts following the methodology explained in Chapter 5 (5.3.4.1). The criteria used to evaluate the texts were the use of a CL application, external characteristics such as the translation volume, and linguistic characteristics which were further developed in the following aspects: Integration within an authoring system, CL-Compliance (Translatability), Translation Languages and Volume, and Text length. Annex II shows the methodology and results of a detailed analysis of the compliance of the analysed text types with the CL specification that lies behind MULTILINT/CLAT.

These criteria were weighted, being the use of a CL application the most relevant one with 3 points, followed by external and linguistic characteristics with 2 points and finally the integration within an authoring system with 1 point. Then each text got a punctuation ranging from 3 to 0, being 3 100% compliance, 2 50% compliance and 1 25% compliance. The results can be observed in the following chart:

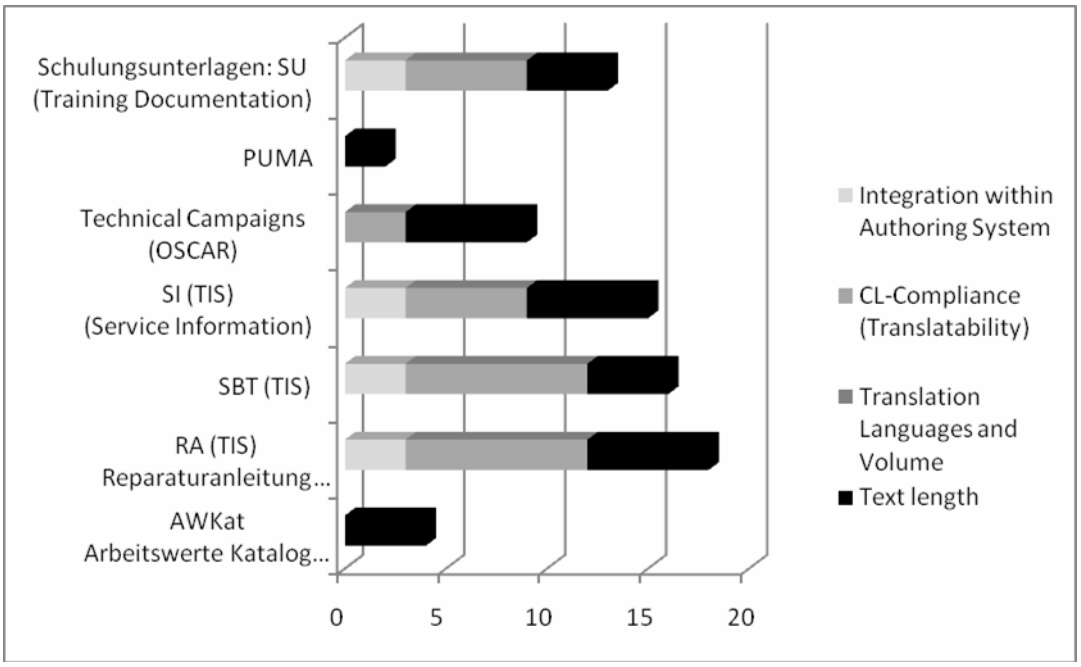

Figure 18: Text types and their suitability for MT

The diagram shows that the info type RA, together with SI, are, for my purposes and according to the requirements exposed above, the most appropriate information types for the pre-selection. Other TIS Documents as well as the STZ-RS Documents (Training documentation) could be also good candidates for the implementation of MT. Other information types such as AWKat, OSCAR or PUMA are less appropriate for the integration of MT technology due to various factors such as the small translation volume, the non-compliance of CL rules or the linguistic inadequacy (e.g. PUMA contains familiar expressions).

As we saw in Chapter 3 (3.5), authors use a series of different criteria in order to characterize and classify technical documents. These criteria include the content itself and the end users, as well as variable elements such as text length, text structure, communicative function, presentation channel etc. In this respect, Repair Instructions and Service Information can be characterized with the following features:

Classification criterion	**Repair Instructions**	**SBT (TIS)**
Text function	Exhortative (operative or conative), main focus Expositive (or informative), secondary focus	Expositive (or informative), main focus

Sender	Expert	Expert
Recipient	Expert	Expert
Text extent	short (average)	short (650 words per document)
Product or process oriented	Process oriented	Product oriented
Type of product	Automotive pieces and elements	Automotive pieces and elements
Structure	Fixed structure	Fixed structure

Table 19: Characteristics of Repair Instructions and SI

For my pre-selection, I will work with a small corpus of RA and SBT texts, the output quality evaluation of which should indicate which system is most appropriate for the evaluation.

6.2.3 The test corpus

As Elliott, Hartley & Atwell (2003) point out, there are two ways of assessing the quality of a MT system: a test suite and a text corpus. A test suite is usually artificially created and is designed to test specific linguistic phenomena. This kind of resource is especially used by MT developers to check where the system fails and where can it be improved (a glass-box evaluation approach). On the other hand, a text corpus is composed by real texts and is therefore more useful for a potential end-user of MT, such as the language department in a company. The corpus typically comprises an original version of the source text, different MT translations (especially if the goal of the evaluation is to compare different MT systems for acquisition) and, possibly, a human reference translation. This depends on the features evaluated and the metrics applied.

Once I selected a text type, I needed to build a text corpus composed by real documents in order to carry out the evaluation. For this purpose, I could access different document types from the automotive company BMW, which included texts from the flat rages catalogue, repair instructions, technical information, service information, technical campaigns, user support messages and training documents.

Finally, repair instructions and SBT (Technical Information) proved to be the best suited document types for my evaluation, since they were going to be integrated in the authoring system, MULTILINT/CLAT had been applied for at least 3 years and it was planned to continue with the checking, the translation volume into English was enough

to consider the implementation of MT and they were compound documents written by professional technical writers, which was one of the translatability rules seen in Chapter 4.

With these documents I built a corpus which comprised an original version of the source text which was translated by the three different MT systems that were chosen. For the automatic evaluation, a human reference translation and post-edited reference translations were added.

The whole text corpus contained over 3,000 different segments of real texts verified with the CL checker. For human evaluation I used a reduced version that included 250 segments divided into two halves that had to be evaluated with respect to comprehensibility, fidelity and post-editability. In order to make this reduced corpus as representative as possible, I analysed the whole corpus to find common grammatical patterns, such as infinitive constructions, imperatives or pre-modifying participial attributes. This reduced corpus aimed at reflecting the content of the bigger corpus, that is, the 250 segments chosen represented the segments in the larger corpus.

The segments were extracted from RAs (Reparaturanleitungen) and SBTs (Service Bulletins Technik) from 2004 and 2005. Each translator proofed 750 segments, resulting from the three different translations of 250 German originals. Segments belonging to the same original document were identified thanks to a sentence ID.

The test for human evaluation was built containing following parts:

- Detailed instructions of how to undertake every part of the evaluation.
- A questionnaire before the evaluation that should provide general information about the situation of the evaluator, his experience with the types of text evaluated etc. This information should explain occasional statistical deviations in the results.

- Test 1 that contained 125 segments that should be evaluated according to a scale of comprehensibility.
- Test 2 that contained another 125 segments that should be first evaluated according to their fidelity and post-editability
- A final questionnaire to check the impressions of the evaluator and his disposition to do post-editing work instead of pure translation.
- Test 1 and Test 2 segments were tested alternatively for comprehensibility, on the one hand, and fidelity and post-editability on the other hand by two different translators, so that they would be double-checked, assuring in this way more objectivity: four translators received the instruction of validating Test 1 for comprehensibility and Test 2 for fidelity and post-editability, while the other four evaluated Test 1 for fidelity and post-editability and Test 2 for comprehensibility.

For the automatic evaluation, I used both the big corpus, with 3,262 segments as well as the reduced corpus, with 228 segments. I also conducted different analysis with subsets of the corpus: on the one hand, I tested only the RA segments (529) from the big corpus and, on the other, the SBT segments (2,733). Then I also conducted a subset analysis of the reduced corpus: 121 RA segments and 107 SBT segments. This reduced corpus was analysed once with a single human reference as well as with four references extracted from the post-edited versions obtained during the human evaluation.

	Whole Corpus			Reduced Corpus		
	Whole Corpus	**SBT**	**RA**	**Reduced Corpus**	**SBT**	**RA**
Tokens	19,659	16,493	3,166	2,330	1,441	889
Types	3,342	2,770	796	915	560	423
Tokens (stoplist)	11,932	9,845	2,087	1,313	768	545
Types (stoplist)	3,121	2,565	700	801	470	348

Table 20: Tokens and types information in the corpus for automatic evaluation

	Reduced Corpus		
	Reduced Corpus	**SBT**	**RA**
Tokens	2,300	1,441	889
Types	915	560	423
Tokens (stoplist)	1,313	768	545
Types (stoplist)	801	470	348

Table 21: Tokens and types information in the corpus for human evaluation

6.2.4 Evaluation setup

The human and an automatic evaluation were carried out in order to cross-check results from both tests and to determine if automatic evaluation was rendering reliable results. This had a twofold purpose: on the one hand, to cross-check automatic evaluation results with human assessment; on the other hand, to prove if only automatic evaluation methods could be meaningful enough to carry out an evaluation and to make decisions on the basis of the results. This would without doubt constitute an approach towards the ideal MT evaluation method suggested by White & Taylor (1998: 22): "readily reusable, with a minimum of preparation and participation of raters or subjects". However, it is necessary to bear in mind that this type of evaluation only renders data regarding how good a system is compared to others or if the system has improved during a development process. Information about the types of mistakes, the need for post-editing or the linguistic quality of the text for dissemination is rarely available

when carrying out automatic evaluation methods[108]. Therefore, depending on the purpose of the evaluation, automatic methods can be useful or not.

6.2.5 Human Evaluation

For Phase 1, the evaluation team was composed of 8 professional translators with English as a mother tongue who had at least 3 years experience translating complex technical texts. In this way, we wanted our results to be as homogeneous as possible. The amount of time available for the experiment was one week. Due to performance questions, we considered that no more than 4 hours a day should be dedicated to evaluate the segments. All in all, translators needed an average of 13.9 hours to evaluate the whole reduced corpus. These data were collected in a questionnaire that was made available to all raters, the answers of which can be seen in Annex VI.

Following criteria were evaluated, according to the methodology outlined in Chapter 5: comprehensibility, fidelity and post-editability. Further results were cross-checked applying the Kappa coefficient in order to check their consistence. The three systems that were evaluated were:

- Comprendim 2.0 (System A)
- Personal Translator PT 2004 Office Plus (System B)
- Systran Enterprise 5.0 (System C)

6.2.5.1 Comprehensibility

As it can be seen in Figure 19, system B leads in the categories “Totally and very intelligible” and occupies a middle range in the “non-intelligible” category for RA texts. System A has a middle score in “Totally intelligible”, but the highest score in “non-intelligible”, while system C has the lowest score in “totally intelligible”, a middle score in “very intelligible” and the highest score in “intelligible”. Besides, system C occupies the lowest score in “non-intelligible”.

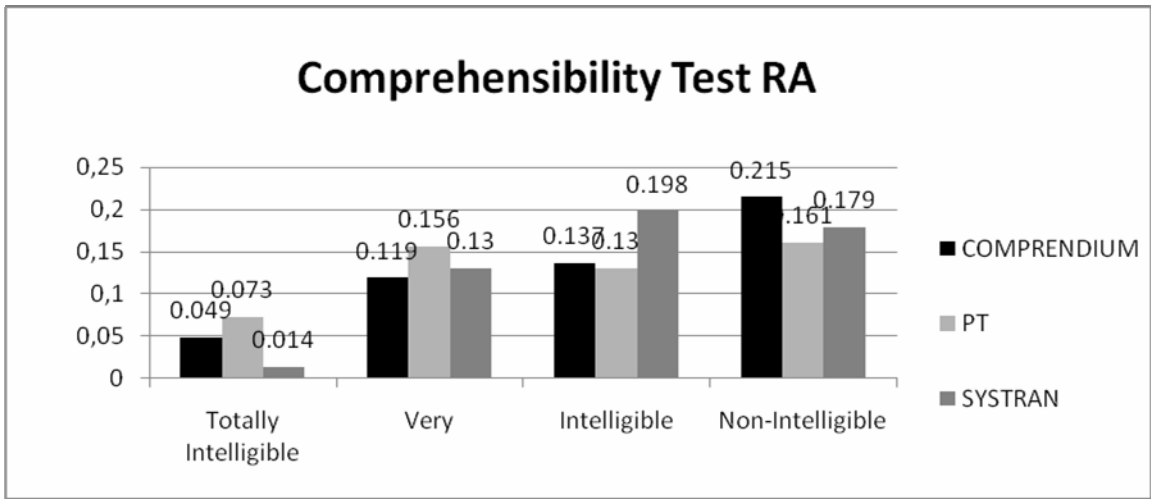

Figure 19: Comprehensibility test for RA

In order to make the classification clearer, I grouped the scale as showed in Figure 20. This grouping makes clear that system B leads the classification, while systems A and C fall behind. They have, on the one hand, a lower number of "totally to very intelligible" sentences, as well as a high number (C slightly more than A) of "intelligible to non-intelligible" sentences.

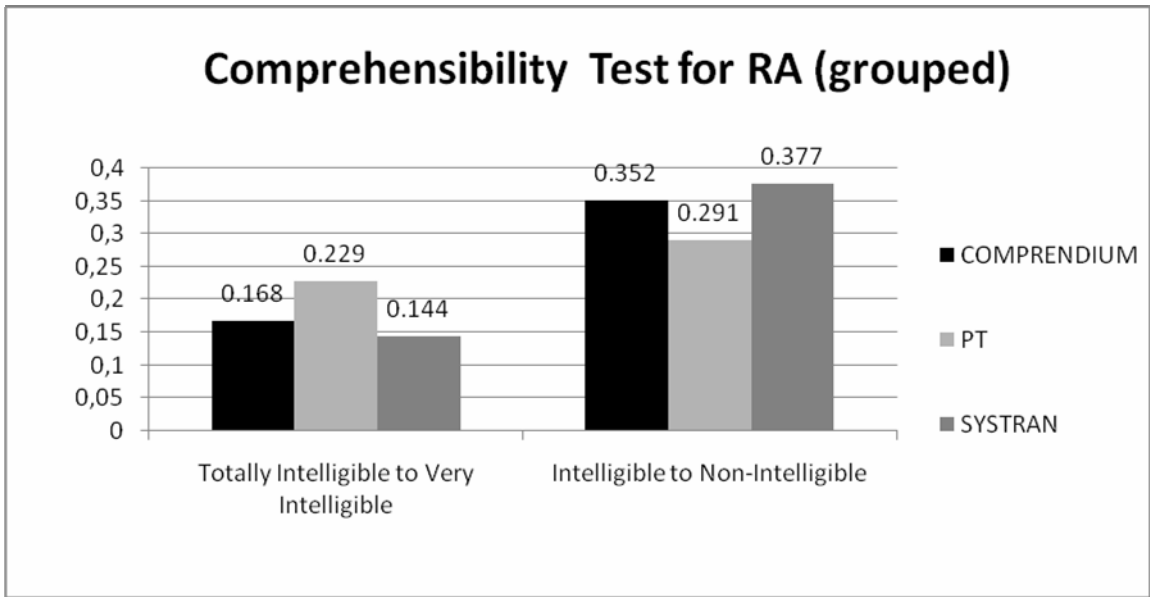

Figure 20: Comprehensibility test (grouped) for RA

With regards to SBT texts, Figure 21 and Figure 22 show that Figure 19 system C leads in the categories "Totally to intelligible", but has the lowest score in the "non-

intelligible" scale. In this test, system B falls behind with the worst results in the "totally intelligible, "very intelligible" and "intelligible" categories and the highest result in the "non-intelligible" category.

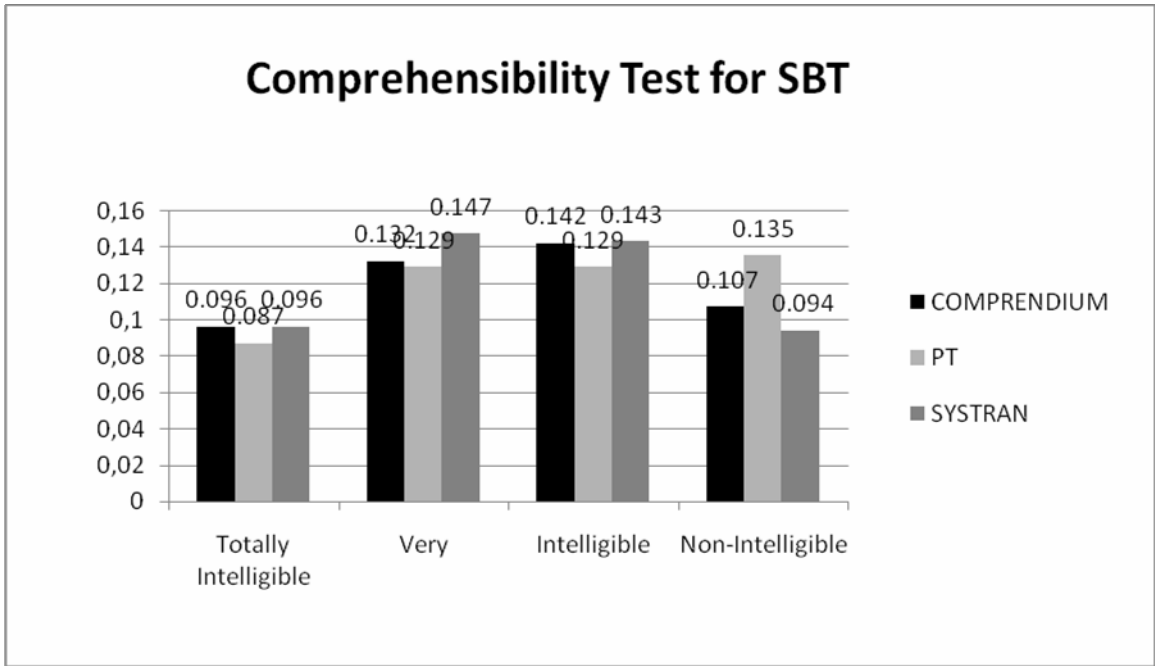

Figure 21: Comprehensibility test for SBT

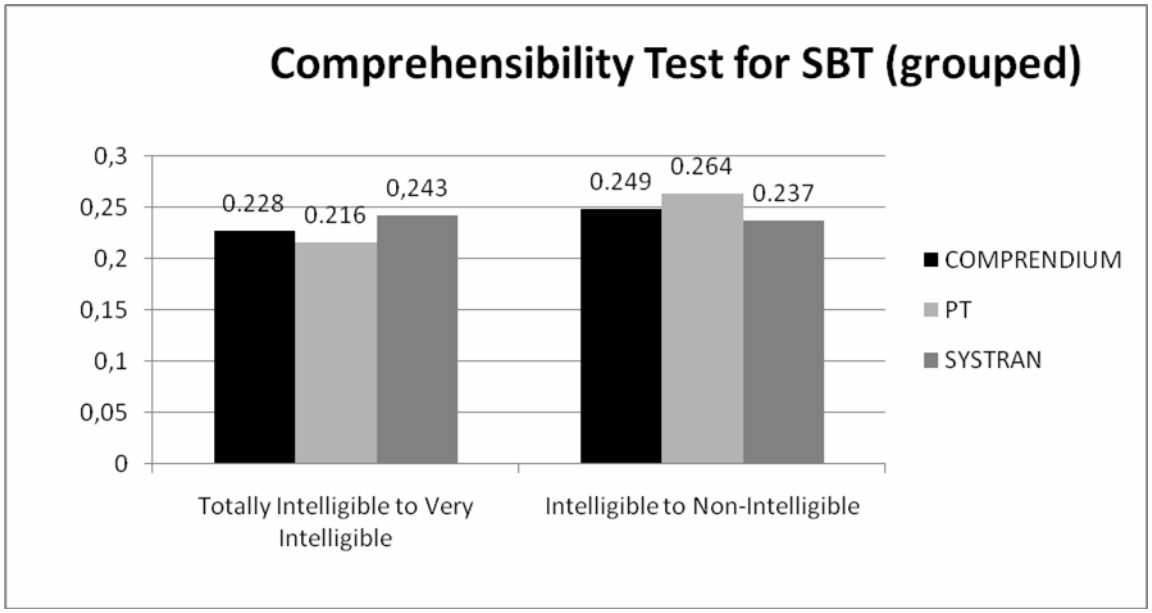

Figure 22: Comprehensibility Test for SBT (grouped)

As Figure 23 shows, the average punctuation of all the assessments of the eight evaluators was the following: System A got 2.22; System B got 2.33 and System C, 2.25.The higher the value, the better the result. Therefore, as we have discussed above, system B received the best score, closely followed by system C and then by system A.

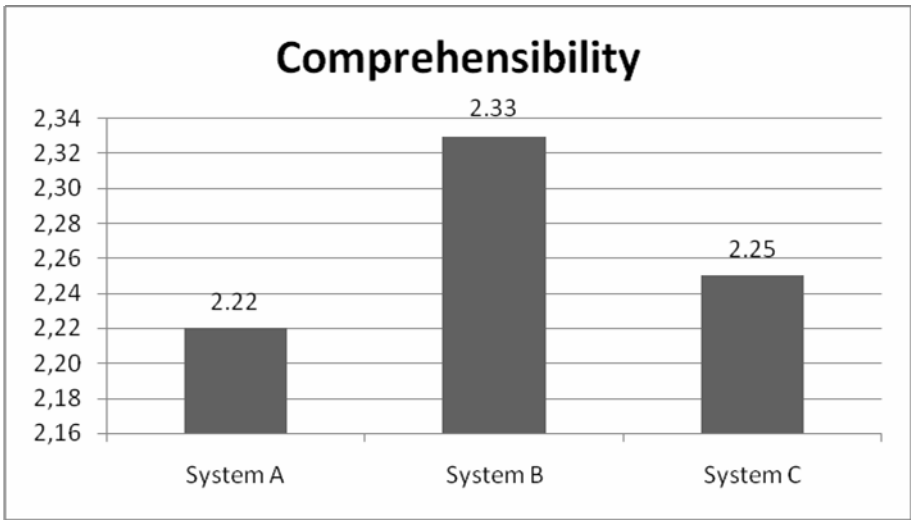

Figure 23: Comprehensibility average scores

6.2.5.2 Fidelity

The fidelity (Figure 24 and 25) evaluation shows that, despite all systems have a similar number of "totally faithful" sentences, systems C and B are stronger in the middle range and, thus, have less unfaithful sentences. In this case, system C would lead the classification with the highest number of "totally to fairly faithful" sentences and the lowest number of "unfaithful" sentences. Systems B and A would follow.

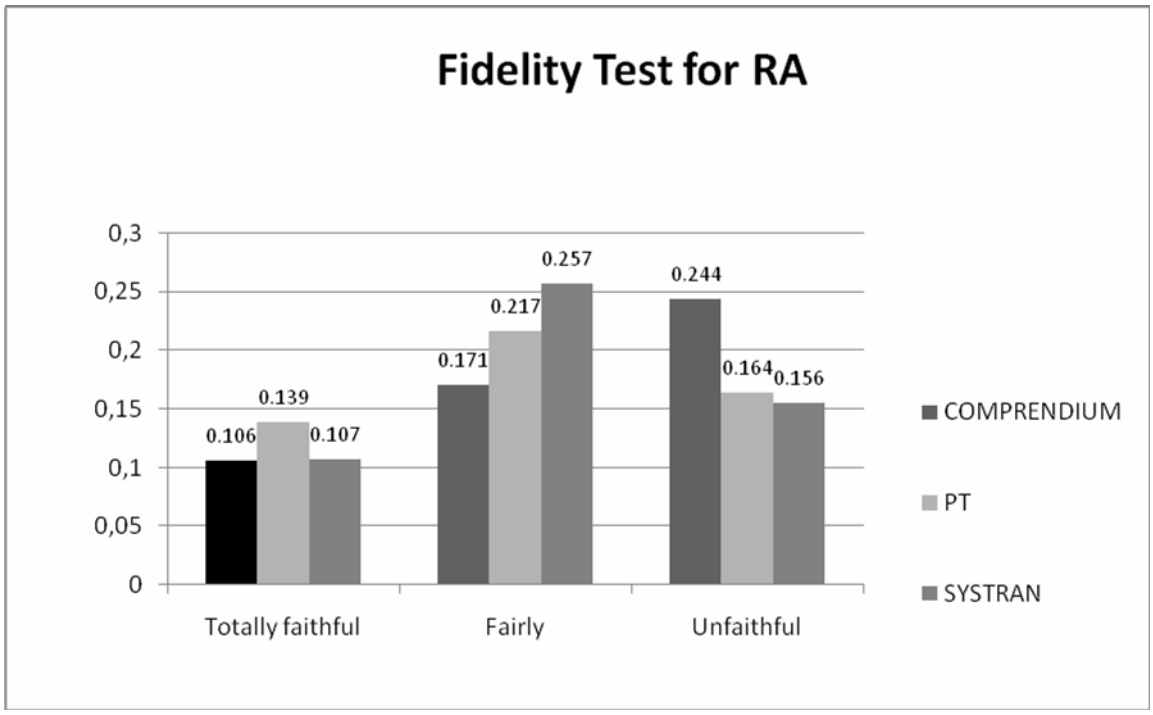

Figure 24: Fidelity test for RA

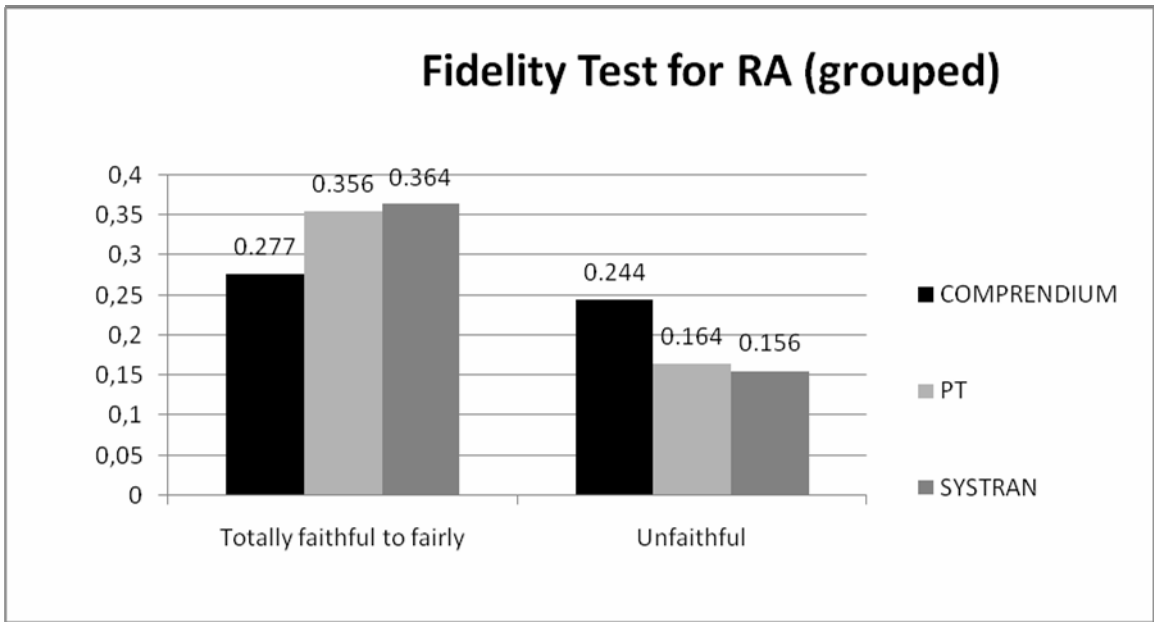

Figure 25: Fidelity test for RA (grouped)

Figure 26 and 27 show the results for SBT texts. Despite all systems have a similar number of "totally faithful" sentences, systems C and A are stronger in the "totally faithful" category and system C would also lead the classification with the highest number of "fairly faithful" sentences and the lowest number of "unfaithful" sentences. Systems B and A would follow.

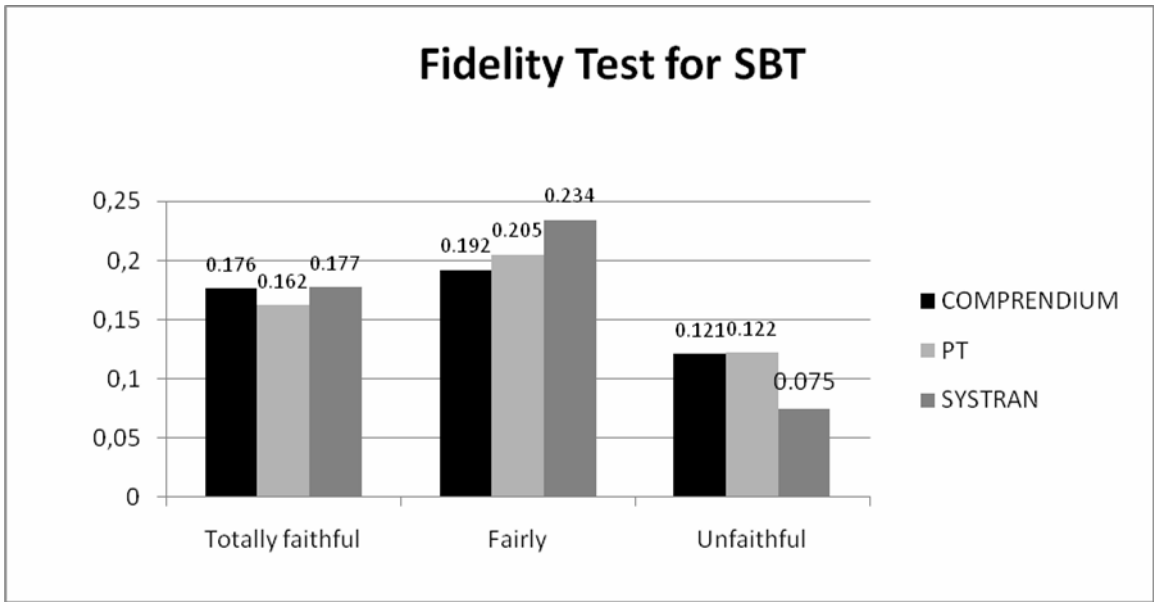

Figure 26: Fidelity Test for SBT

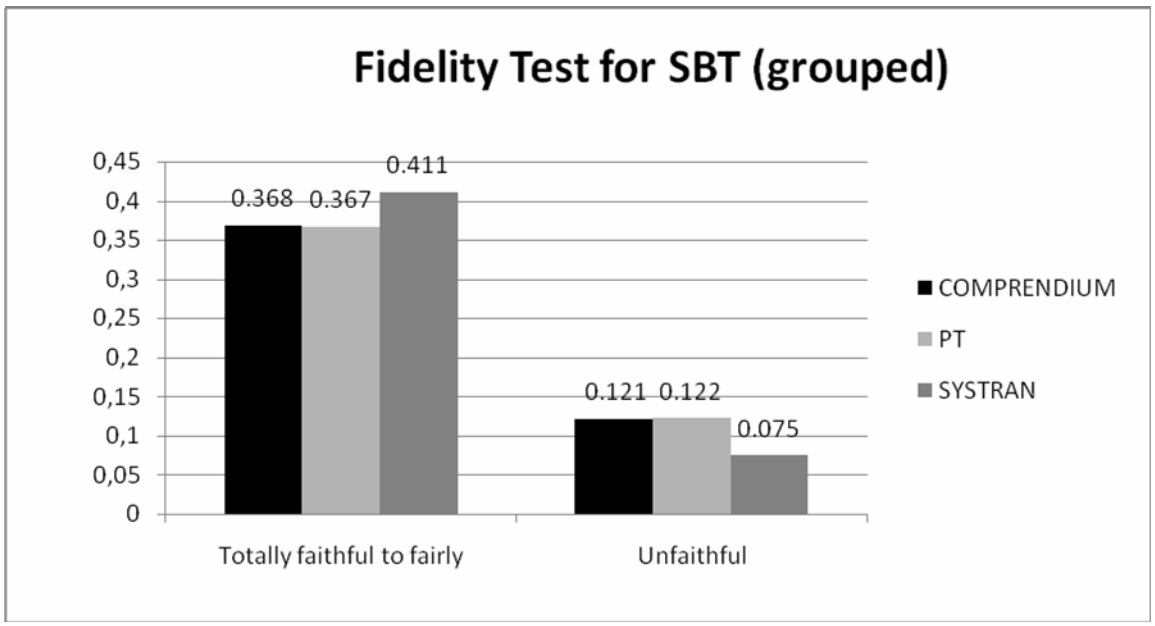

Figure 27: Fidelity Test for SBT (grouped)

The average punctuation of all the assessments of the eight evaluators was the following: System A got 1.95, System B 2.02 and System C 2.05. The higher the value, the better the result. The results therefore confirm that system C receives the best fidelity scores, while system B and system A closely follow behind:

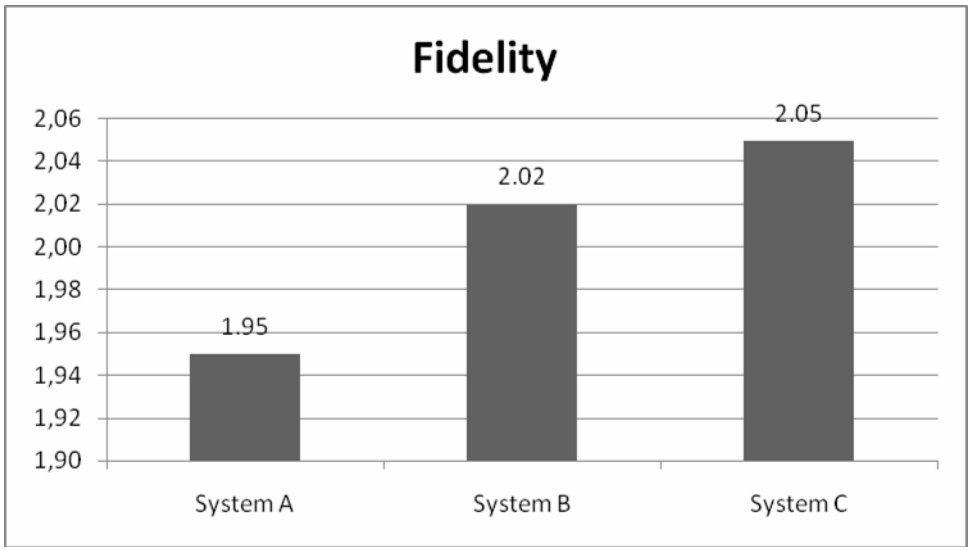

Figure 28: Fidelity average scores

6.2.5.3 Post-Editability

Post-Editability (Figure 29 and Figure 30) evaluates how "useful" (usability aspect) the translations produced by the MT systems were in the case that these had to be improved later for publication. This index was intended to indicate the real effort that would be needed to transform machine translated segments into publishable ones.

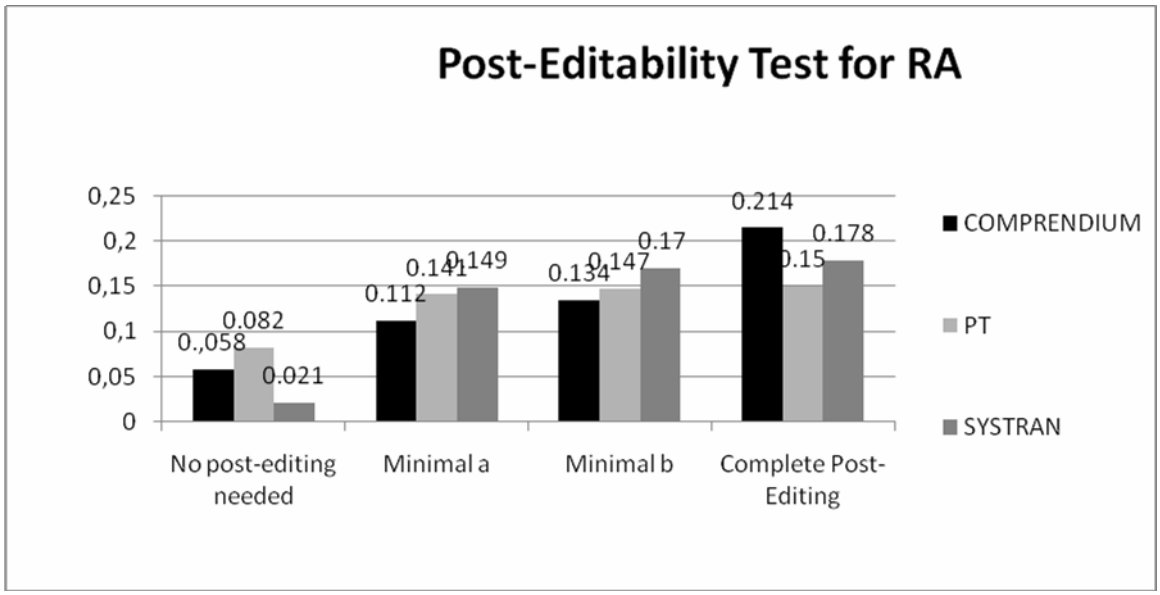

Figure 29: Post-editability test for RA

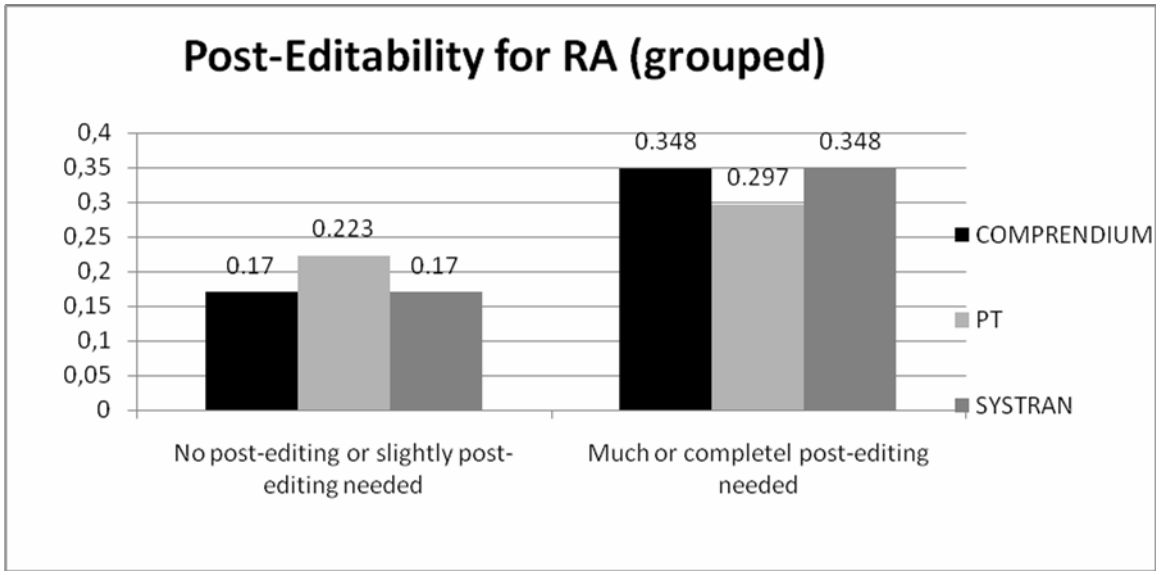

Figure 30: Post-editability test for RA (grouped)

System B offers again the highest result in "non post-edition needed", and middle results in the resting categories (pretty low in "total post-edition needed"). System A offers the highest number of total-postedition and, despite the middle range in "no post-edition", the low ranges in minimal post-edition make it fall behind. System C offers the lowest "no post-edition" needed result and also the lowest "total post-edition" (though very close to system B). System C also scores best in minimal post-edition. The grouped chart shows that system A falls behind systems B and C, that are very close together.

Post-Editability evaluation (Figure 31 and Figure 32) in SBT assesses System B offers again the highest result in "non post-edition needed", and middle results in the resting categories. System A offers the highest number of total-postedition and, despite the middle range in "no post-edition", the low ranges in minimal post-edition make it fall behind. System C offers the lowest "no post-edition" needed result and also the lowest "total post-edition" (though very close to system B). System C also scores best in minimal post-edition (a).

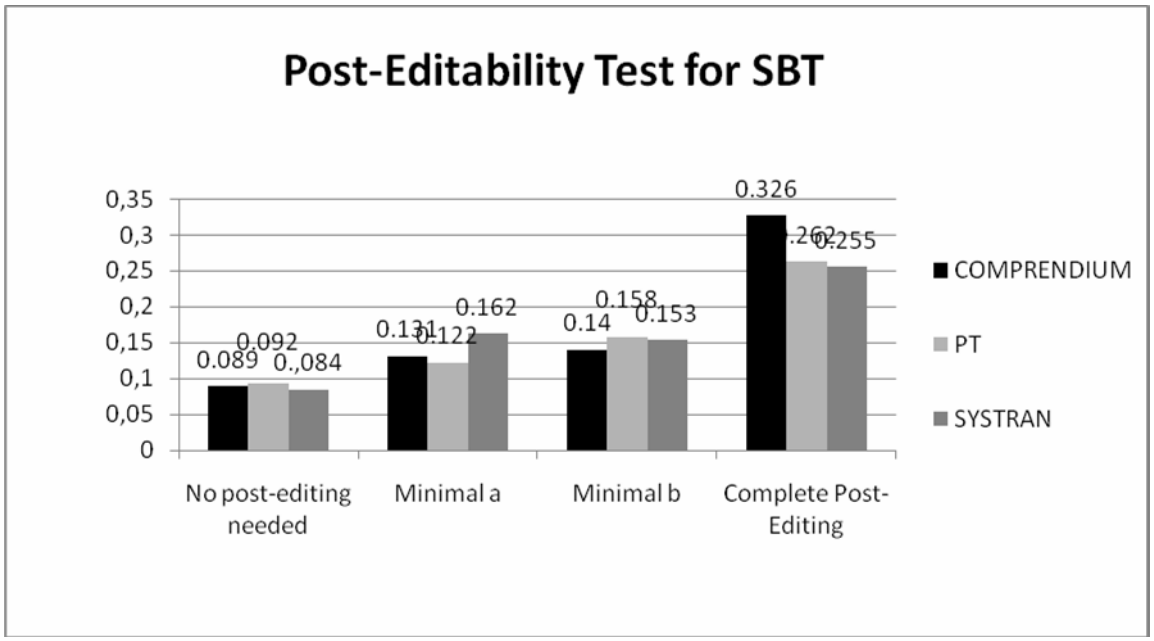

Figure 31: Post-editability test for SBT

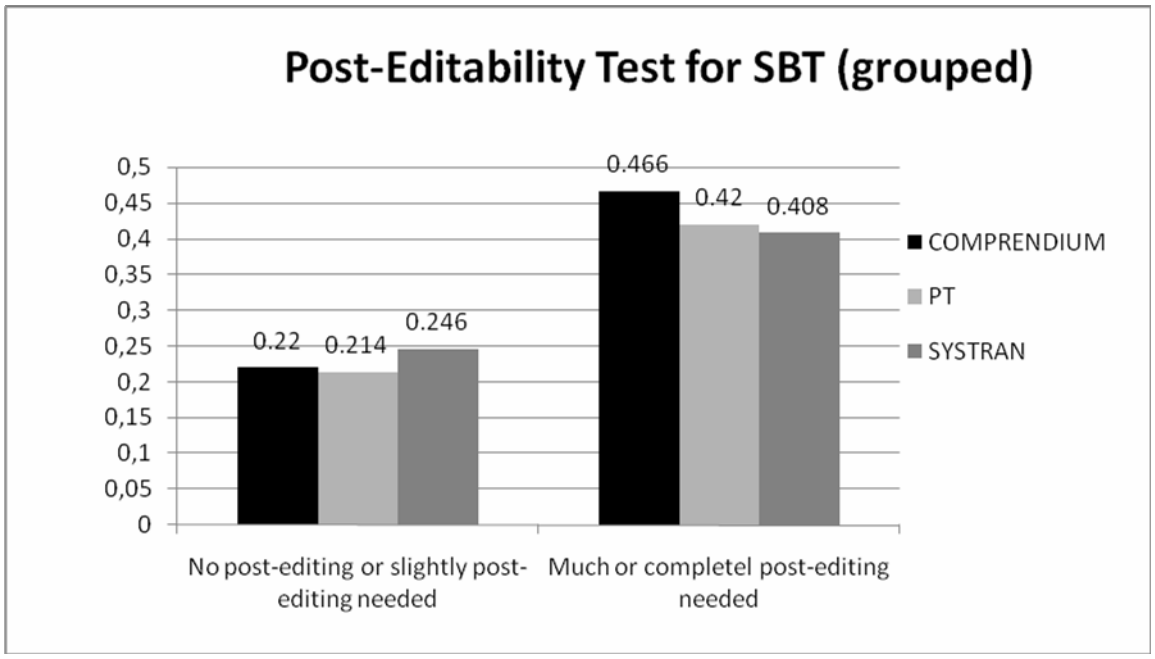

Figure 32: Post-editability test for SBT (grouped)

The average punctuation of all the assessments of the eight evaluators was the following: System A got 2.24, System B got 2.30 and System C got 2.20. These results confirm that system B leads the classification followed by system A and system C:

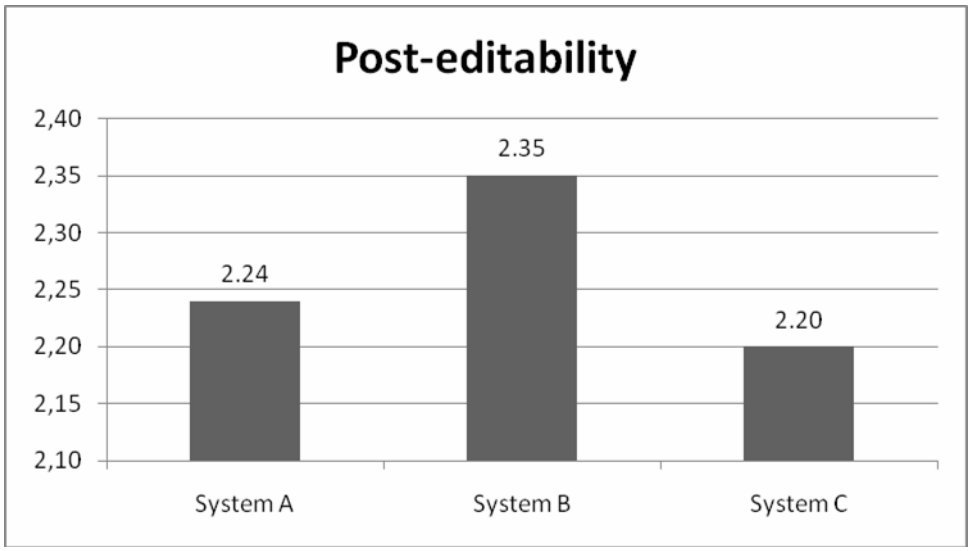

Figure 33: Post-editability average scores

In Annex VI we can find the resting average results grouped by test (comprehensibility, fidelity and post-editability for each of the systems).

As we saw in 4.8.4, the Kappa coefficient is used to measure agreement among evaluators. The Kappa coefficient allows measuring the agreement between *n* judges with *k* criteria of judgment. In order to calculate this coefficient, we use the online Kappa calculator developed by researcher Justus Randolph[109]. This calculator provides two variations of kappa: Siegel and Castellan's (1988) fixed-marginal multirater kappa and Randolph's free-marginal multirater kappa (see Randolph, 2005 and Warrens, 2010). Brennan and Prediger (1981)[110] suggest using free-marginal kappa when raters are not forced to assign a certain number of cases to each category and using fixed-marginal kappa when they are.

As we can see in Annex VII, in all cases Kappa values are between 0 and 1[111], indicating a level of agreement among raters better than chance. The highest agreement is found in fidelity in test 1 and 2 in systems A (average 0.40) and B (average 0.36), followed by post-editability (average 0.29 and 0.28). Intelligibility gets lower scores (0.26 and 0.23 respectively). In System C, post-editability gets the highest score (average 0.27), followed by fidelity (average 0.22) and intelligibility (0.20). The

average agreement for System A is 0.32, 0.29 for System B and 0.23 for System C, indicating fair agreement in the three cases and with the highest agreement for System A.

These data confirm that human evaluation can be subjective and, thus, it is difficult to reach almost perfect agreement. A higher number of raters might solve this problem, but this would involve higher costs and a difficulty to make the evaluation strategy sustainable.

6.2.6 Automatic Evaluation

In order to complete the human evaluation and to check if automatic metrics can be a reliable way to conduct evaluations, I carried out an automatic evaluation. The metrics chosen to conduct this evaluation were BLUE and NIST, which have been already introduced in 4.8.2.5.

6.2.6.1 Complete Corpus: 3,262 segments (mono-reference)

For the first evaluation, I used 3,262 segments extracted from real texts. For this evaluation, a unique human reference was available, which was the official translation that had been published by BMW.

6.2.6.1.1 BLEU

We can see the BLEU results in

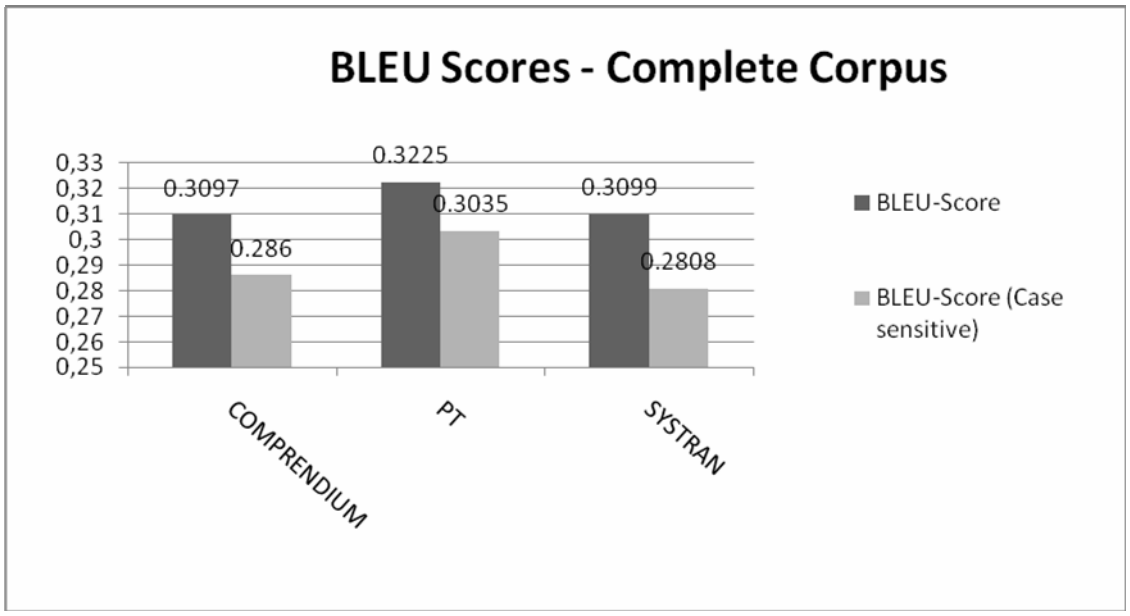

Figure 66, Figure 67, Figure 68 and Figure 69 in Annex VII.

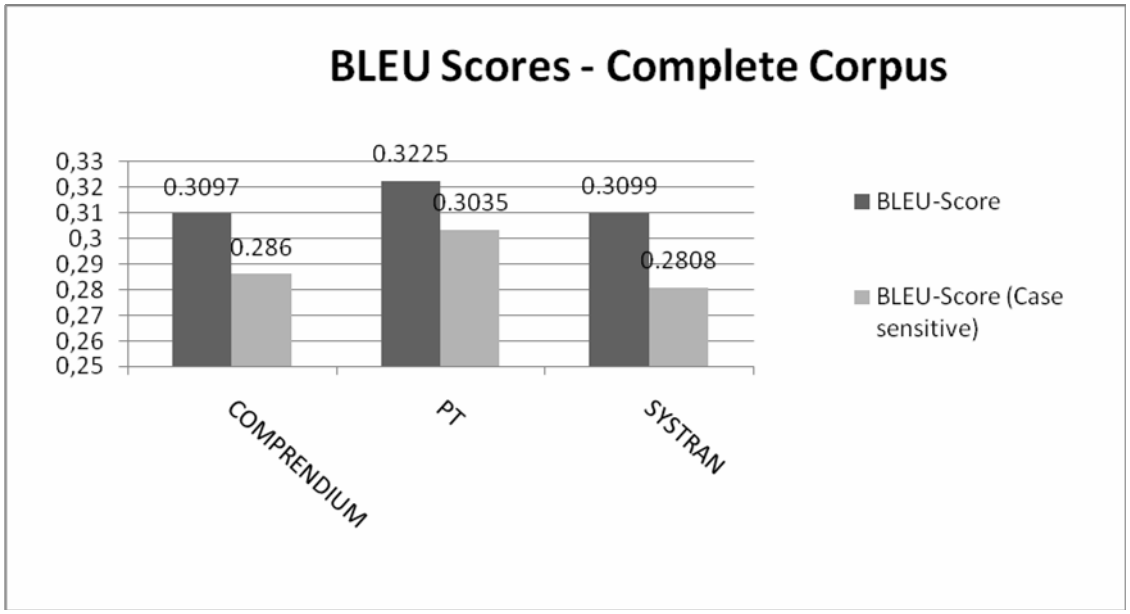

Figure 66 presents the overall results, while Figure 67 and Figure 68 present the results for RA and SBT respectively. System B leads the classification with both settings (with and without case-sensitive configuration[112]) and in the three cases (complete corpus, RA and SBT).

Results of systems A and C are very close together, with a slight advantage for C in the overall corpus and the RA corpus and a slight advantage of System A in the SBT Corpus.

6.2.6.1.2 NIST

NIST scores show similar results, as we can observe in Figure 74, Figure 75, Figure 76 and Figure 77 in Annex VIII. Systems B and C are close together, though B leads the overall classification. Indeed, the case-sensitive analysis stresses the differences between all systems: B leads, whereas C and A follow. In the text type analysis, System B clearly leads the classification for RA, while System C leads it for SBT, closely followed by Systems B and A.

6.2.6.2 Reduced Corpus: 228 segments (mono-reference)

Apart from the complete corpus, I also conducted an automatic evaluation with the 228 selected segments that had been used in the human evaluation. The slight difference in number (250 for the human evaluation) was due to the fact that some segments were not translated due to technical problems.

In this second analysis, I used the different references that had been created by the evaluators when post-editing the texts.

6.2.6.2.1 BLEU

Figure 69 to Figure 73 show the results. Again it is System B that leads the overall classification. However, there are differences in the text type analysis. System B also leads the RA BLEU test, followed by Systems C and A, whereas System C clearly leads the SBT test and System B falls behind.

6.2.6.2.2 NIST

Figure 77 to Figure 81 show the NIST results for the reduced corpus with a single reference. The overall classification shows an outstanding position of System B, which is also reflected in the RA test. However, the SBT test shows better results for System A, followed by System B and C.

6.2.6.3 Reduced Corpus: 228 segments (multi-reference)

In these tests, the 228 selected segments were evaluated with the metrics NIST and BLEU using 5 references: the official translation plus 4 post-edited versions that had been created during the human evaluation.

6.2.6.3.1 BLEU

In the multireference analysis that is depicted in Figure 72 and Figure 73 in 0, the test for RA shows better results for System B, whereas System C leads the classification for SBT texts.

6.2.6.3.2 NIST

Figure 80 and Figure 81 show the results of the multireference analysis. Here, the same pattern is repeated. System B leads the classification for RA texts, while System C clearly leads the classification for SBT Texts.

6.2.6.4 Correlation of human and automatic metrics

To establish a correlation between the human and the automatic evaluation in order to state if there is any relationship between them, I collected the data of both tests. For the human evaluation, I calculated the averages for the comprehensibility, the fidelity and the post-editability test.

	HUMAN EVALUATION			AUTOMATIC EVALUATION							
				WHOLE CORPUS (3262 segments)				REDUCED CORPUS (228 segments)			
	Comprehensibility	Fidelity	Post-editability	BLEU-Score	BLEU-Score (Case sensitive)	NIST-Score	NIST-Score (Case sensitive)	BLEU-Score	BLEU-Score (Case sensitive)	NIST-Score	NIST-Score (Case sensitive)
System A	2.22	1.95	2.24	0.3097	0.2860	6.9614	6.6395	0.3035	0.2944	5.8757	5.7231
System B	2.33	2.02	2.35	0.3225	0.3035	7.1161	6.8276	0.3311	0.3236	6.1528	6.0403
System C	2.25	2.05	2.20	0.3099	0.2808	7.1137	6.6650	0.3083	0.2891	5.9743	5.6174

Table 22: Data of the human and the automatic evaluation

First of all, it is important to point out that the human evaluation averages range from 1 to 4, being 1 the worst possible result and 4 the best possible. With regards to the interpretation of the automatic metrics, we already saw in 4.8.2.5, that BLUE and NIST are difficult to interpret.

BLUE results can range from 0 to 1. The closer to 1, the more overlap with human references, that is, a better quality. Lavie (2010b) suggests following interpretation scale for BLEU scores:

Figure 34: BLEU Interpretation according to Lavie (2010b)

With regards to NIST, as Culy & Riehemann (2003), "It is less clear what the range of the NIST metric is. A text compared with itself among the reference translations gets a BLEU score of 1, while the NIST scores for our texts compared to themselves ranged from 12.8744 to 14.5006". According to these values, a first glimpse to the data tells us that the absolute values of the human tests and the automatic evaluation correlate in their positions (rank 1, 2 and 3) for comprehensibility in System B, (all rank in the first position). In Fidelity, human evaluation doesn't correlate that well with automatic metrics since automatic evaluation places System B always in the first place, whereas human evaluation places System B in the second place. Again, post-editability shows a good correlation with a rank 1 for System B in all cases.

		WHOLE CORPUS (3,262 segments)				REDUCED CORPUS (228 segments)			
Comprehensibility	Comprehensibility	BLEU-Score	BLEU-Score (Case sensitive)	NIST-Score	NIST-Score (Case sensitive)	BLEU-Score	BLEU-Score (Case sensitive)	NIST-Score	NIST-Score (Case sensitive)
System A	3	2	2	3	3	3	2	3	2
System B	1	1	1	1	1	1	1	1	1
System C	2	3	3	2	2	2	3	2	3

Table 23: Ranks in comprehensibility and automatic evaluation

		WHOLE CORPUS (3,262 segments)				REDUCED CORPUS (228 segments)			
Fidelity	Fidelity	BLEU-Score	BLEU-Score (Case sensitive)	NIST-Score	NIST-Score (Case sensitive)	BLEU-Score	BLEU-Score (Case sensitive)	NIST-Score	NIST-Score (Case sensitive)
System A	3	2	2	3	3	3	2	3	2
System B	2	1	1	1	1	1	1	1	1
System C	1	3	3	2	2	2	3	2	3

Table 24: Ranks in Fidelity and automatic evaluation

		WHOLE CORPUS (3,262 segments)				REDUCED CORPUS (228 segments)			
Post-editability	Post-editability	BLEU-Score	BLEU-Score (Case sensitive)	NIST-Score	NIST-Score (Case sensitive)	BLEU-Score	BLEU-Score (Case sensitive)	NIST-Score	NIST-Score (Case sensitive)
System A	2	2	2	3	3	3	2	3	2
System B	1	1	1	1	1	1	1	1	1
System C	3	3	3	2	2	2	3	2	3

Table 25: Ranks in human and automatic evaluation

6.2.7 Conclusions of Phase 1

After analysing and summarising all these data, the following conclusions can be made:

System A does not offer the desired output quality and falls behind systems B and C. This can be clearly seen both in the human evaluation and in the automatic evaluation.

Systems C offers middle results, and sometimes even better results than the other two systems. This is especially important in post-editability, where results of B and C are very close together.

System B offers the best overall results, both in the human evaluation and in the automatic evaluation, which reflects the results of the human assessments. Since System B offers the best comprehensibility results, this system would be good for its deployment as a system for information gisting or rapid translation of e-mails, company reports etc. The good post-editability results also show that this system is the most appropriate for translation[113]. Therefore, I decided to use System B for the second phase of the empirical part of this research work.

6.3 Phase 2

The test suite of Phase 2 was assessed by 6 evaluators for German and 3 evaluators for English. There were some inconsistencies in the data since some evaluators forgot to evaluate some of the sentences. Therefore, the data represented in Annex IX contains both the absolute frequencies and the relative frequencies.

To evaluate each of the sentences, the scale proposed in 5.6 was converted to numeric values:

- Improvement: 4 points
- No effect +: 3 points
- No effect -: 2 points
- Worsening: 1 point

The result was an average of 3.61 for German sentences and 2.95 for English sentences, indicating that German sentences were between a positive non-effect and improvement, whereas the sentences of the English test oscillated between a negative and a positive non-effect.

A closer look in the results, which can be seen in Annex IX, indicates that an average of 67.80% of the German sentences showed an improvement, followed by a 16.33% with a positive non-effect, 8.50% with a negative non-effect and 7.37% with a worsening effect. These figures indicate the average number of sentences that were assessed by each of the evaluators. On the other side, the English test resulted in 37.44% of sentences with an improvement, 30.14% with a positive non-effect, 22.83% with a negative non-effect and 9.59% with a worsening effect. These results are the average percentages, though if we look at the results of each evaluator, we can find segments

that have been evaluated differently by each of them. However, the agreement among evaluators, as we will see in next section, was fairly high.

There is a clear tendency to assess that the texts rewritten in German following the rules of a controlled language show a general improvement in terms of a better intelligibility or comprehensibility. These results are also reflected in the translations in English, though to a lesser extent, with a bigger amount of sentences being rated in the middle ranges.

It is remarkable that in both languages some segments are marked to have suffered a worsening effect, which is a counter effect of what CL should achieve. In the next section we will analyse what could be the reason for this.

6.3.1 Interannotation agreement: the Kappa coefficient

With regards to the agreement among evaluators, the free-marginal kappa coefficient was also applied in this case. The values can be seen in Table 40 in Annex IX0. We find a moderate agreement both among the evaluators in German (with an average of 0.55 for the sentences rated by 5 and 6 evaluators) and among evaluators in English with a free-marginal kappa value of 0.47.

The number of sentences that obtained a perfect agreement in German were 55 out of 146 sentences (37%), whereas in English there were 64 sentences with perfect agreement (43%).

It is remarkable that, despite the scarce number of evaluators, the agreement figures are pretty high. This might be due, as it was mentioned before, to the high specialization and preparation of evaluators, who belong to the automotive world.

6.3.2 Controls

In order to determine which rules most affected the quality both of the original text and the translations, the rules that applied to every segment of the evaluation were annotated next to it. This gave us an overview of how the different controls affected the quality of the source and target language and, within those controls, which rules were more prone to affect the quality of the segments. However, it was not always straightforward to determine which rule had an effect on the quality of the segment, since many segments were affected by more than one rule, and even within the same category (for instance by two grammar rules). Besides, sometimes the rules gave advice or recommendations that were not followed by the authors and in some cases the rule did not apply to the sentence (the rule was wrongly applied).

In general, however, we can observe a majority of rules causing improvement in the grammar and the orthography controls, both in German and English, whereas the effect of the Terminology and Style controls does not seem to be as positive as it might be expected.

The average rating for each sentence and the percentages with respect to all evaluated sentences were calculated, as it can be seen in this figure[114]:

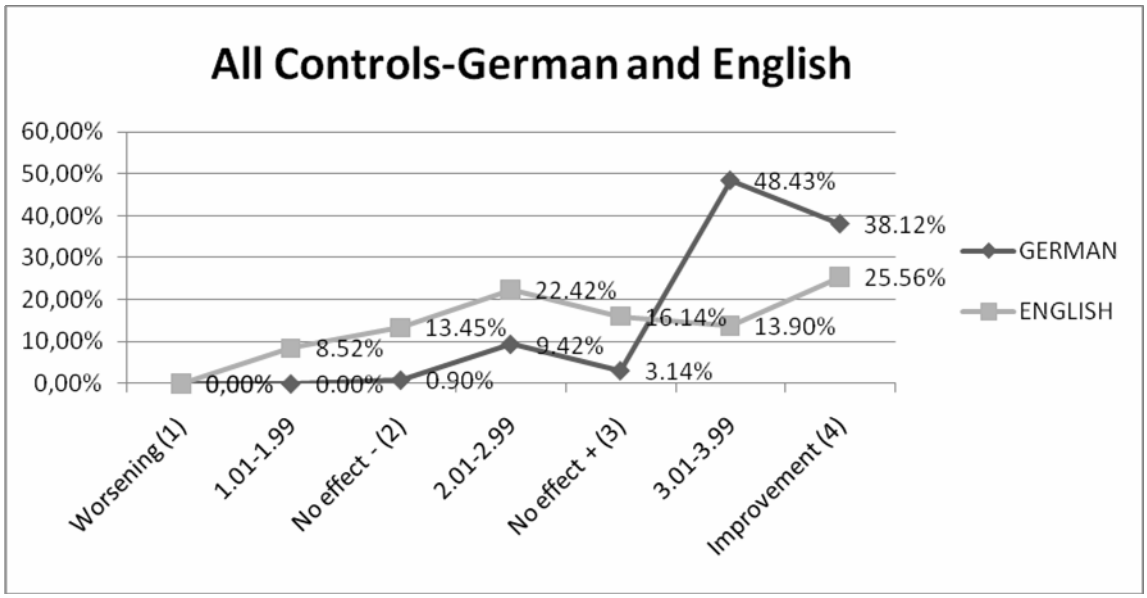

Figure 35: All Controls. Phase 2 Evaluation

In average, 38.12% in German and 25.56% of the sentences in English show an improvement. A total of 3.14% of sentences were evaluated in German as having a positive non-effect, whereas this category is attained by 16.14% of the English sentences. Finally, rules were considered to have a negative non-effect in 0.90% of the sentences in German, whereas in English this category obtained 13.45% of the sentences. There were no sentences that were considered to have worsened by all evaluators. These figures contrast with the data exposed above (in 6.3), since they represent the average values for each sentence, and not the total number of sentences that were evaluated within each category. This is the reason why there are "exact" categories, that is, sentences that obtained in average a 4, 3, 2 or 1 rating and "inexact" categories, that is, the average values among the exact categories.

With regards to the "inexact" categories, that is, sentences that obtained decimal averages, it is remarkable that in German most sentences were evaluated as being between improvement and a positive non-effect (50%). In English, this category was attained by 16.44% of the sentences. With regards to the next "inexact" category, in English there is an important amount of sentences that were considered as having no effect on the resulting translation, being these sentences as correct as before or as incomprehensible and wrong as before (21.92%). In German this category was attained by 8.22% of the sentences. Finally, the last "inexact" category was that of sentences considered as having worsened or as remaining as incomprehensible and wrong as before: in English it was 9.59% of the sentences, whereas in German no sentences were evaluated in this category.

Therefore, in general we can conclude that evaluators perceived a greater improvement in German than in English. Though the inter-annotator agreement in German was complete for 36.63% of the sentences, there was a 43.53% of the sentences where evaluators considered that there had been either an improvement or no effect with a good quality, summing up 80.17% of all sentences. In English, however, 27.40% of the

sentences were considered to have improved, whereas only 11.54% had been considered to have improved or to be as correct as before, summing up 38.94% of all sentences.

In the next sections I will try to clarify which controls had a bigger impact both on the original German text and the translated version.

6.3.2.1 Grammar

The grammar control contains rules related to morphology and sentence structure. Sometimes it includes rules related to punctuation and orthography if they are related to structure of the sentence, such as for instance if a bracket or a hyphen between two words is missing. There were a total of 40 sentences affected by grammar rules.

The rules that affected most sentences were related to inflection, hyphenation, the concordance of word endings (plurals, German datives) and punctuation (commas between main and subordinate clauses) (see Table 56 in Annex X). Of all the sentences affected by the grammar control, only one evaluator for German rated two sentences as having worsened after applying the correction suggested by MULTILINT/CLAT. In one case, the sentence was affected by three controls (terminology, style and grammar) and the worsening was probably due to the Terminology control, since the right term could not be found in the database. In the second case the author of the original text did not follow the recommendation given by MULTILINT/CLAT to improve the sentence and therefore one evaluator considered that the sentence had worsened. The original sentence was “Hinweis: Nach Austausch oder Programmierung des DME-Steuergeräts” and the recommendation was directed to write "nach" in lowercase after the colon. However, the writer decided to remove it, thus changing the meaning of the sentence: “Hinweis: Austausch oder Programmierung des DME-Steuergeräts.”

With regards to English, only one evaluator considered there was worsening in three sentences when implementing Grammar rules. Of all the three sentences, two sentences were affected also by the orthography and the style control, while only one was only affected by the grammar control. As we can see in the following figure, in average,

none of the sentences affected by the Grammar control were rated as having worsened, neither in German nor in English. We find a big amount of sentences between the positive non-effect and the improvement in Germany, whereas in English there are two peaks: 27.50% of the sentences being rated on average as having improved and 22.50% of the sentences being between the negative and the positive non-effect.

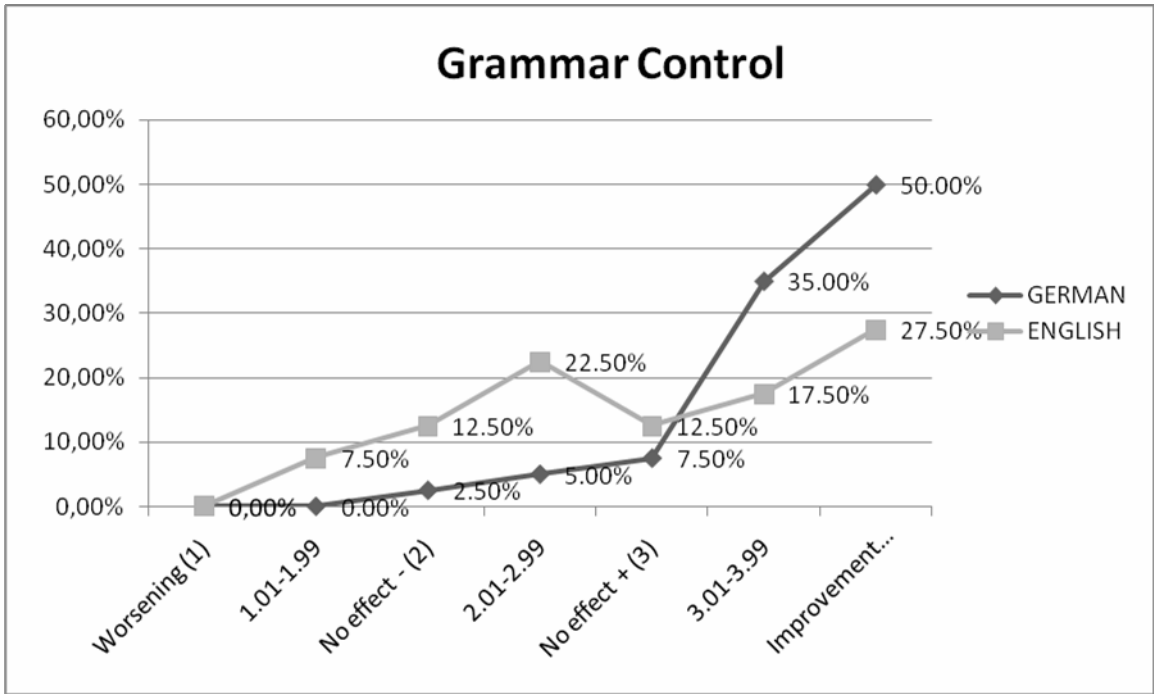

Figure 36: Grammar Control-Phase 2 Evaluation

Table 56 in Annex X shows that, of all the sentences, nine were affected by a rule related to hyphenation; seven recommended to check the inflection; six were affected by a rule concerning word endings (referred to the inflection); five related to punctuation (the comma should be written between two subordinate clauses); three were related to the fixed space between number and measure; two by a rule advising that two words should be written together; two by a rule indicating that subject and predicate should have the same person and numerus; two by a rule related to punctuation (the comma should be removed) and two by the confusion between *dass* (conjunction) and *das* (article, pronoun) in German. The rest of the rules only affected one sentence each. In general, the most common rules had a positive (improvement) or neutral effect (positive non-effect) both in German and in English.

6.3.2.2 Orthography

The orthography control comprehends rules related to misspelling, use of capitals or lower case, spaces and the new German orthography rules. There were a total of 45 sentences affected by orthographic rules.

The rules that affected most sentences were related to misspelling or unknown words (see Table 57 in Annex X). As we can see in the following figure, in average, none of the sentences affected by the Orthography control were rated as having worsened, neither in German nor in English. We find a big amount of sentences between the positive non-effect and the improvement in Germany, whereas in English there is a majority of sentences having improved and a high percentage (summing up 41.10%) of the sentences being rated as being between the negative and the positive non-effect.

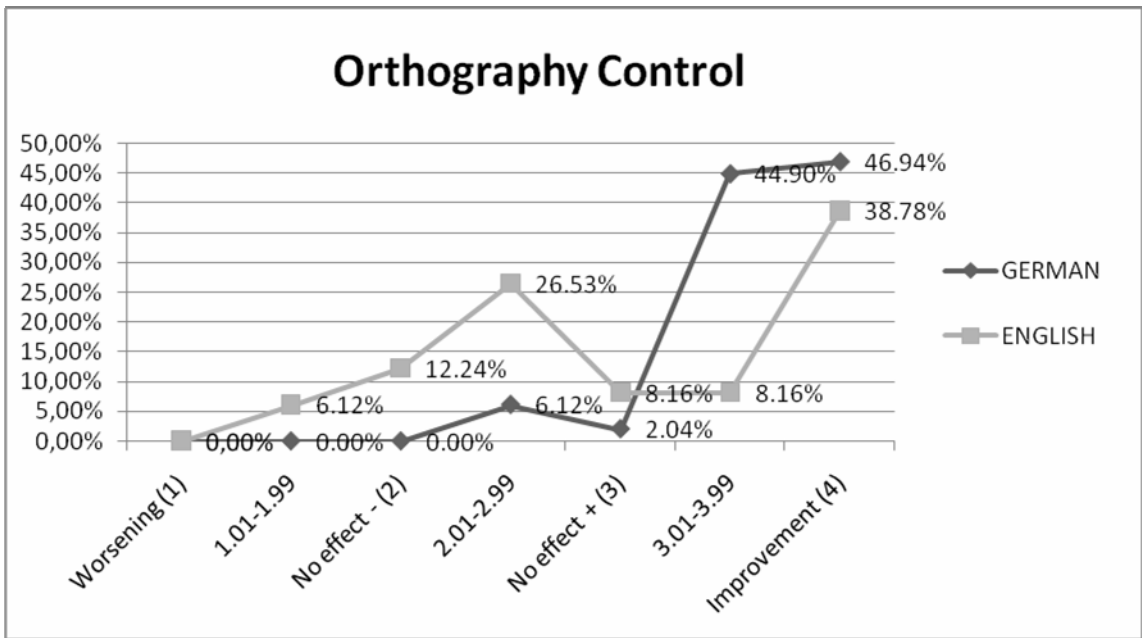

Figure 37: Orthography Control-Phase 2 evaluation

As we can see in Table 57 in Annex X, of all sentences, a total of 29 sentences (60%) were affected by a rule concerning misspelling or unknown words. This is without doubt one of the rules that can most affect the results of automatic translation, since if the word is not recognized or not contained in the dictionary, it will be rendered incorrectly or it will not be translated. Out of the 29 sentences affected by this rule, 16

and 14 were rated as having improved in German and English respectively and 13 and 3 were between improvement and a positive non-effect. The rest of the sentences in English were scattered among the resting values, with 2 having a positive non-effect, 6 being between the positive and negative non-effect, 3 being rated as having a negative non-effect and 1 being rated between a negative non-effect and worsening.

Eight sentences were affected by a rule that states that a fixed space should be placed among multiword acronyms or abbreviations. This rule affected sentences containing the abbreviation *z.B.* Although in general it had a positive impact in German, when including the space the translation into English, the translation was not rendered properly (the abbreviation was not recognized and thus not translated). This could be solved with a special rule in order to recognize *z. B.* (with space) as *zum Beispiel* (for instance).

The rest of the rules concerned the use of capitals or lowercase, incorrect spelling (especially regarding German compounds containing a number and a noun), the misapplication of the new German orthographic rules and the use of wrong compounds.

6.3.2.3 Terminology

The Terminology control contains rules related to the use of the terms stored in a terminology management system, where terms can have various status: preferred, deprecated, variants etc. A total of 79 sentences were affected by terminology rules that conducted to changes in 62 sentences. The resting sentences were marked with the rule POSNEG, which indicates that a term can be preferred or deprecated depending on the context. The author must then evaluate if the use of the term in the given context is correct or not. Only in two cases the authors did not follow the recommendations given by CLAT/MUTILINT. For instance, in the sentence "Beim **der 7er Baureihe** erfolgt die Spannungsversorgung vom **Power Modul**, bei **der 5er** und 6er Baureihe vom Stromverteiler im Kofferraum", it was suggested to substitute *Kofferraum* by *Gepäckraum*. However, the author did not follow this direction and the term was translastaed as "boot". This would have rendered a better translation and, thus, a better

result, since the latter term was included in the MT system dictionary (as "luggage compartment").

As we can see in the following figure, evaluators perceived that 26.58% and 24.05% in German and English of the sentences had improved with the application of the terminology control, a very similar average in both languages. For German, most sentences were rated between improvement and a positive non-effect, whereas in English the effect was more scattered, with a peak in sentences being rated as being between a positive and a negative non-effect.

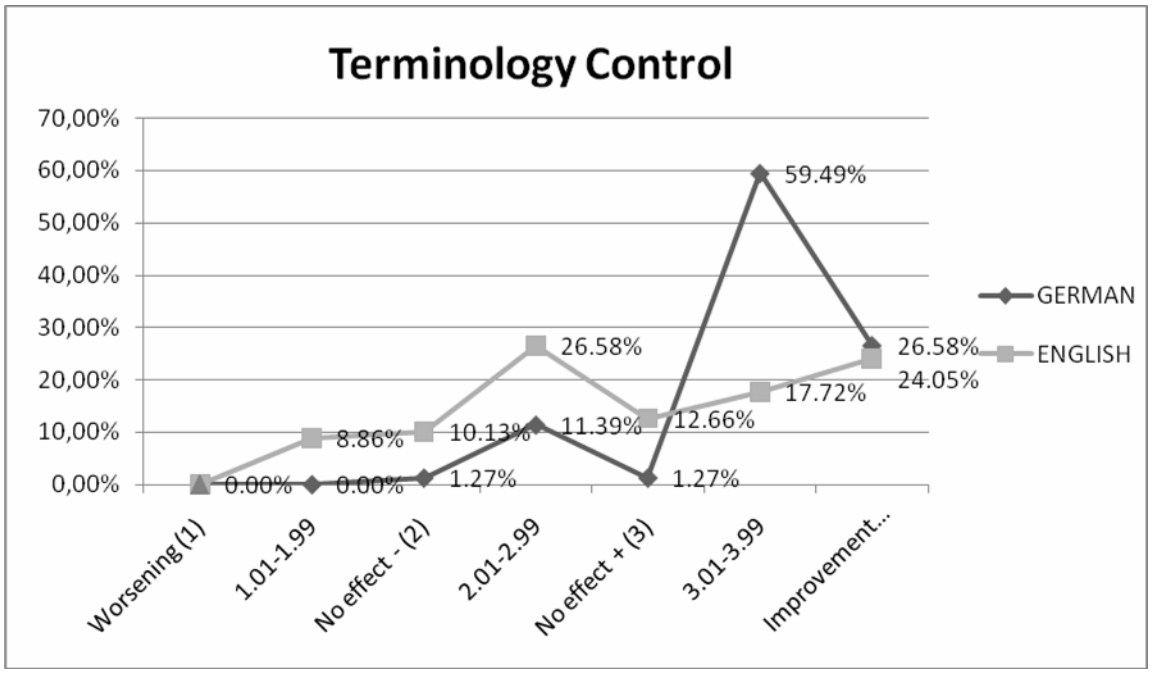

Figure 38: Terminology Control-Phase 2 evaluation

Most sentences were affected by the use of deprecated terms, as we can see in Table 58, followed by cases where the term could be deprecated or not depending on the context and that in most cases did not imply any change in the corrected sentence. Another group of sentences were affected by a rule indicating that the term as such was not stored in the database and suggested a variant that was the preferred one. Finally, only two sentences were affected by a rule indicating what the abbreviation stood for. This

rule is only intended to clarify what the abbreviation means so that the author knows if he is using it correctly.

6.3.2.4 Style

The Style control contains rules related to sentence structure, the use of pronouns and demonstratives, the number of words in the sentence etc. There were a total of 55 sentences affected by style rules.

Most sentences in German experienced an improvement (38.18%) or were between the improvement and the positive non-effect (45.45%) and a negative non-effect (3.64%). 10.91% of the sentences were rated between a negative and a positive non-effect, whereas only 1.82% was rated as being between worsening and a negative non-effect. With regards to English, though there are 14.55% of sentences with an improvement and 10.91% between the improvement and the positive non-effect, most sentences were not affected by the rules according to the evaluators, with 30.91% of positive non-effect, 20% of negative non-effect and 12.73% between the positive and the negative non-effect. 10.91% of the sentences were rated between the worsening and a negative non-effect. The following figure illustrates these data:

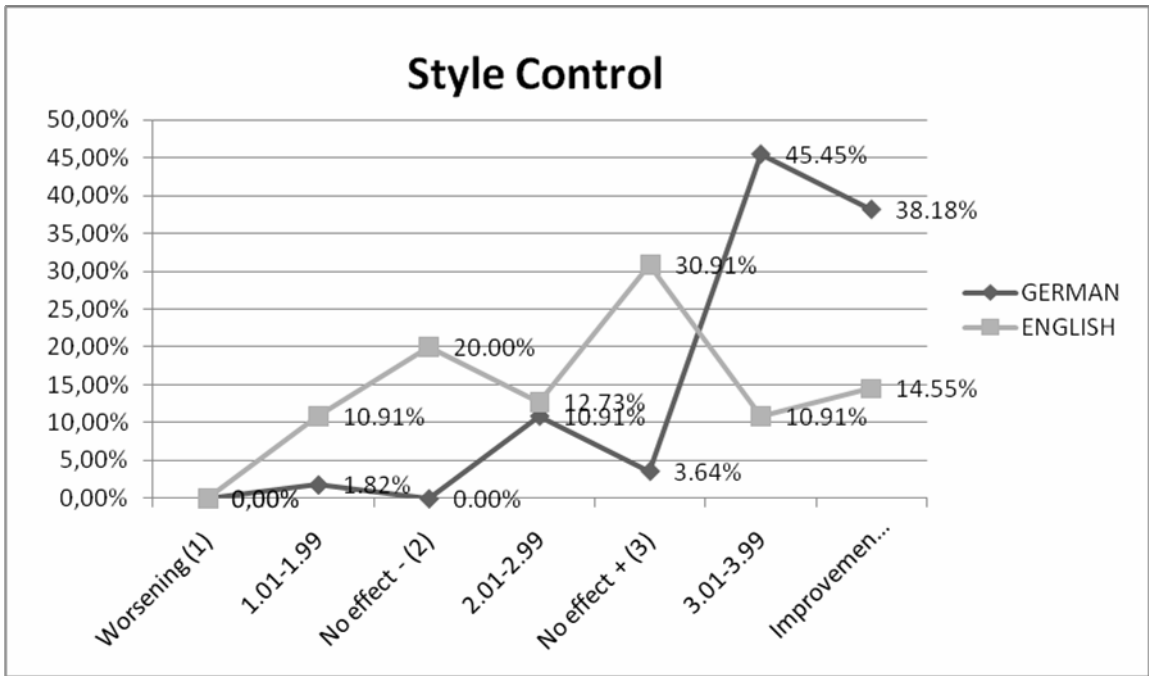

Figure 39: Style Control-Phase 2 evaluation

As we can see in Table 59 in Annex X, the rule that affected most sentences (21.05%) addressed structures with too many nouns or "meaningful units" and recommended its reformulation. Out of the 12 sentences affected by this rule, 4 and 3 were rated as having improved in German and English respectively and 6 and 1 were between improvement and a positive non-effect. There was a sentence valued as having a positive non-effect in German and a sentence valued as being between the positive and the negative non-effect. It is remarkable that in English there were 5 sentences being rated as having a negative non-effect and 1 being rated between a negative non-effect and worsening.

The next rule affecting most sentences was related to their length: "Reduce or split the sentence in two" affected 9 segments and the derived rule "Split the sentence in two if possible" affected 6 sentences, summing up a total of 26.32% of all sentences. Out of all the sentences, 6 were considered to have improved in German and 2 in English; 7 to be between improvement and a positive non-effect in German and 1 in English; 6 to have a positive non-effect in English and 2 to be between the positive and the negative non-effect in German and 3 in English. Finally, one sentence in English was considered to have worsened and 1 was between worsening and a negative non-effect (for the first rule). The final most relevant rule was related to the use of pronouns: here, the results in German were unanimous. Out of 5 sentences affected by this rule, 5 were considered to have improved in German. In English, however, the evaluators considered that only one sentence had improved, whereas 2 of them had a positive non-effect, 1 was between the positive and the negative non-effect and 1 had a negative non-effect.

The rest of the sentence were scattered among 14 other rules pertaining to the Style Control.

6.3.2.5 Term candidates

There were only five sentences where new term candidates were detected, being one of them repeated. In all cases the sentences were affected by more than one control:

terminology, orthography, grammar and style. Therefore, it is difficult to state how they affected the quality of the translations. The terms were translated in all cases with the standard terminology stored in the MT system, though it was not always the correct equivalent.

6.3.3 Conclusions of Phase 2

As a conclusion of this phase we can state that there was a greater improvement in the German sentences than in the English sentences. However, this should be not a big surprise, since the English sentences were translated automatically and their quality was expected to be worse than the sentences written originally by German native speakers. Nevertheless, the evaluators considered that 37.44% of all the English sentences showed an improvement after applying controlled language rules and 30.14% were as good as they were before, summing up 67.58% of the sentences. It is remarkable, however, that some sentences were considered to have worsened after applying the controlled language rules, both in German (7.37%) and in English (9,59%), attaining thus the counter effect.

The agreement among evaluators was 0.55 for German and 0.47 for English, showing a moderate agreement, which is a pretty high figure despite the scarce number of evaluators.

With regards to the different controls, we can observe a majority of rules causing improvement in the grammar and the orthography controls, both in German and English, whereas the effect of the terminology and style controls does not seem to be as positive as it might be expected. There were a total of 40 segments affected by grammar rules, 45 segments affected by orthography rules, 79 sentences affected by the terminology control and 55 sentences affected by the style control. Finally, there were five new terms that were proposed as candidates for the terminology management system.

In the grammar control, rules related to orthography such as the confusion between *dass* and *das* in German, inflections (case and numerus) and the use of hyphens to separate multiword expressions formed by a number and an abbreviation obtained the best general impact. With regards to orthography, the rule that affected most sentences and obtained the best impact was related to unknown or misspelled words. The terminology control contained a rule related to the use of deprecated terms that obtained the best results. Finally, the style control obtained the best results with rules affecting segments with too many nouns or too long sentences, advising the author to paraphrase or split the sentence into two. The rule related to the substitution of pronouns to avoid ambiguities was valued very positively in German, whereas in English had not the same desired effect.

As saw in 1, Reuther (2003) classified the rules according to their priority for translatability. In her study, rules were divided in seven categories:

- Typographic rules,
- Avoidance of ambiguous structures
- Lexical rules
- Avoidance of elliptical structures
- Avoidance of complex structures
- Rules regarding word-order and sequence of sentence chunks
- Stylistic rules
- Company-specific rules

The category “avoidance of complex structures” had the bigger number of rules with priority 1. This category was followed by “stylistic rules” and “lexical rules”. However, in our study, the rule that obtained the highest number of improved sentences was a rule related to misspelling or unknown words. In our test, 16 sentences in German and 14 in English improved after applying the suggested correction.

Our study seems to confirm that sentences affected by terminology rules obtain good results, especially those affected by the rule related to deprecated terms, with 10 sentences in German and 11 in English showing improvement. The avoidance of

complex structures is also represented in our corpus with rules related to sentences with too many nouns or too long, where the author is advised to split the sentence in two. It is difficult to state, however, that it is a top priority rule for translatability, both due to the limited number of sentences affected by it in our corpus and also due to the scattered results between all categories (between improvement and worsening), especially in English.

6.4 Summary and final remarks

In this chapter I have presented the results of the empirical part of this research work, the methodology of which was settled down in Chapter 5. In the first phase, first I selected three MT systems and the best text type for my purposes. I subsequently constructed two corpora, one for the human evaluation and one for the automatic evaluation, in order to elucidate which was the best MT system for my evaluation. This corresponded to the phase related to the selection of results as outlined in the methodology.

In the second part of the chapter I present the results of the second phase of the study, where the CL rule suite and the tool MULTILINT/CLAT were evaluated. I present both the general results of the evaluation and then I analyse the different rules and their effect on the quality of the original text and the translations, comparing the results with the theoretical approaches and previous studies tackled in the first part of this work.

7 WORKFLOW AND FEASIBILITY CONSIDERATIONS

A penny saved is a penny earned.
Benjamin Franklin (1706-1790)

7.1 Introduction

The goals of this dissertation are to evaluate the effectiveness of MULTILINT/CLAT with respect to the translatability of the documents checked using this tool and, at the same time, to study the requirements and consequences of implementing Machine Translation (MT) technology in the translation processes at an automotive company.

In this chapter I present two aspects of this study: on one hand, I discuss the role of MT in the translation process and how this technology can be integrated within the translation workflow. On the other hand, I carry out an economic analysis in the form of an ROI, to determine the investment needed and the savings obtained in such a new scenario.

Further, I study the different scenarios in which MT can be applied, to concentrate afterwards on the most suitable scenario for an automotive company. I specify the characteristics of this scenario, such as the types of text, the language pairs, or the MT system chosen. Finally, I suggest a possible workflow to be used for MT.

The chapter finally contains an economic analysis intended to determine the return on investment of implementing MT technology. This is part of an Operational evaluation as it was indicated in Chapter 4. As White (2000: 105) describes, "Operational evaluations generally address the question of whether an MT system will actually serve its purpose in the context of its operational use. The primary factors include the cost-

benefit of bringing the system into the overall process". Later on, J. S. White (2003: 221) also points out that "a meaningful measure in operational evaluation is return on investment, which implies comparison of the measurement of the real costs of an MT application, and the real benefit (revenue, cost savings, etc.)."

For this purpose, I first consider theoretically the cost factors in a traditional translation process, so that they then can be compared to the potential savings. Then I calculate the costs involved in the various stages of the implementation of MT: the analysis and refinement of the translation processes, the installation of the system, as well as the customization and maintenance costs. I quantify the cost savings by calculating the translation costs, first for a standard translation process and then for a process with MT and post-editing, extracting the difference between both and determining the reduced translation costs. The time saved by the user is also calculated from the data available, allowing us to determine whether there is a gain in productivity. I use all these data to determine the Return on Investment, which is in the form of a percentage.

Finally I summarise the results obtained, draw a conclusion and give a recommendation for the implementation of MT.

7.2 Translation and Authoring Processes in Industrial Environments

Nowadays, companies face a great pressure due hard competitiveness of the global market. Expanding model and product series, coupled with shorter product development cycles and the growing complexity of products, have seen a sharp rise in demand for technical information on the wholesale and retail level. Not only does this imply an increase in source language texts, but also an exploding number of documents in a number of languages when the company has an international presence and the documentation has to be translated.

It is a fact that the amount of documentation produced increases year to year due to the reasons mentioned above. The need to maintain a high quality of language, both in the

source and in the target texts, without increasing authoring and translation costs, is thus absolute and pressing. Many companies have long recognised all these hurdles and have been working in the past years on the creation and maintenance of linguistic resources such as terminology databases and translation memories. Although these efforts are valuable and contribute to gaining in quality and reducing costs in the processes of content creation, further options have to be considered and evaluated in order to face the imminent increase in content and costs. Therefore it has become necessary to adjust the information flow within these companies and consider other options.

7.3 Automating the Process: Reasons to use MT

One of the main ways to optimise the processes is to automate as much as possible of the manual work that is involved in creating and managing multilingual content. In this way, "human resources are freed from repetitive, non-productive labour and can be redeployed to more productive and strategic tasks" (Lawlor, 2005: 2). Tools such as Terminology Management Systems, Translation Memories (TM) and Translation Management Systems help in automating tasks such as: the detection of changes in content, the extraction and packaging of content, the preparation and conversion of files, the customization of workflows, the leveraging of content already translated and the use of consistent terminology.

A further step in the automation of translation tasks is fully automated translation, which consists of the fully automated translation of new content. The quality of MT output has not dramatically improved in recent years (Hutchins, 2003) and, despite the recent advances in data-driven machine translation[115], this also seems to be the case for the near future. However, the commercial interest in MT has been experiencing a significant rebound since the beginning of this century. Successful case studies by companies such as Daimler Chrysler (Flanagan, 2002), Ford (Rychtyckyi, 2000) or Caterpillar (Nyberg, 1997) confirm this trend.

There are three main reasons for a company to decide to use MT in their translation processes: saving costs, saving time and offering a better service. I will now analyze these three reasons in detail.

7.3.1 Saving Costs

A company always seeks to reduce costs while maintaining high quality in its product. The internationalization and localization of a product for different markets results in high translation costs that in certain cases have to be reduced. MT can be an option to help reduce these translation costs.

Nevertheless, at this point it is necessary to point out that the cost savings are closely related to the expected quality. Depending on the level of quality required, cost savings can vary. For example, when post-editing is used to bring MT quality on a par with human translation, the costs can increase to the point of equivalence, due to higher fixed costs associated with implementation and maintenance.

A white paper by Lionbridge (2001) examines three levels of cost savings:

- MT with minimal customization: this is the case of MT with some amount of customization done before translation, where MT is used by the client "out of the box". This variant provides the largest cost savings. Customization might range from the import of company-specific terminology to its linguistic coding and the writing of scripts to clean up both the input and the output. However, minimal customization implies that the output is not reliable, producing results that might vary from perfect to unintelligible. Therefore, this approach is best used in internal communications such as e-mail, or for getting a gist of the content.

- MT with customization and ongoing maintenance: within a specific subject domain, MT can become moderately reliable if the right terminology is used and the linguistic resources are maintained over a time period. Depending on the volume, the savings can be significant. However, customization and maintenance do not

necessarily avoid all mistakes, and human post-editing is necessary if a publishable quality is to be obtained.

- MT with customization, ongoing maintenance, and post-editing: this means automating the process to obtain results comparable to those obtained by human translation. This implies, however, that the likelihood of having savings in cost decreases. This can depend, on one hand, on the volume of translation, and, on the other, on the degree of quality required. These grades of quality can range from correcting only the terminology and smoothing grammatical irregularities to make the text readable, to correcting style and polishing the text to achieve the quality of a professional translation. At this level the cost of human translation might be equated or even exceeded.

The three scenarios are up for consideration, depending on the type of text and the process in which MT is embedded. However, in this document I only analyze the third case, i.e. the deployment of MT with customization, ongoing maintenance and post-editing. Although this alternative might not deliver the largest savings in cost, it was chosen due to the characteristics of this research.

7.3.2 Saving Time

Saving time can be another interesting goal for the company. In cases where documentation needs immediate translation due to its ephemeral nature, MT without post-editing might be appropriate. This would include real-time or near real-time communication (e-mails, chats), technical data or news. In other occasions, translation processes have to meet the time-to-market requirements of shortened product life cycles.

As an example, in an automotive company such as BMW, the translation from German into English in a reduced time span could contribute to reducing the time-to-market of certain information and products for the Asian markets, since these languages are directly translated not from German, but from English.

In any of the arrangements detailed above, the time required to produce results is less than for human translation.

7.3.3 Improving Service

Companies might want to offer a better service to their international customers or to improve the communication among workers from different nationalities within the company. Very often, human translation is not possible, since neither the budget nor the resources are available. MT with customization and ongoing maintenance can deliver the message, even if it is not perfect. I can distinguish between two different scenarios:

- Batch mode MT: where all the content is translated and made available. Since no post-editing is applied in this process, it is recommendable to inform potential users about the origin of the translation. This will set appropriate expectations and avert damage to the corporate brand.

- On-demand MT: allowing users to request a translation online, and get an immediate result. When users need to access content, but find that it is in a language they do not understand, they can choose to use MT service.

7.4 MT in the translation process

7.4.1 Three scenarios in which to use MT

First of all, it is necessary to consider the different scenarios in which MT can be implemented within an industrial context. Basically, three scenarios can be identified within a company:

- Urgent translation of relevant content in e-mails, texts or chats that need to be translated as quickly as possible.
- Translation of knowledge databases and technical support documentation.

- Full translation of documents, where the viability of machine translation with human post-editing needs to be calculated.

The first scenario would correspond to the translation of e-mails within the company. In my study, for example, translation is done from German into English and vice versa. This could be useful to communicate with other partners from abroad. The translation has to be fast and efficient enough so that the reader of the translation can get the basic gist of the information contained in the e-mail or text.

The second scenario would be a knowledge database which contains, on one hand, the customer inquiries (Customer Original Inquiries or Kunden Originalton), and, on the other, the solution to a certain problem provided by the technical support. In this case, translation has to be fast and the accuracy thereof medium to high, so that the technical content is trustworthy.

The third scenario is the use of MT to produce translation drafts which, after some amount of post-editing (perhaps considerable), are ready for dissemination, i.e. for publication. In this case it is necessary to calculate the viability of such an arrangement. MT can speed up the translation process and guarantee the consistency of terminology, but the quality of the translation has to be good enough to justify its use.

The first two scenarios could be interesting to consider. Nevertheless, since the main goal of this research was to investigate the effectiveness of MULTILINT/CLAT, both scenarios were disregarded, as the tool had not been applied in any of them.

The third scenario, i.e. translation drafts to calculate the viability of machine translation with human post-edition, is the scenario that has been considered for this work, and therefore for the present document.

7.4.2 The Translation Workflow

Depending on the translation processes and the systems involved in these processes, the translation workflow within a company may vary substantially. Besides, depending on the player, vendor or client, the workflow might cover only one part of the process or be part of another more complex workflow.

In this section I cover standard translation workflows as seen by both the client and the vendor, and give an overview of the future translation processes within an automotive company, as well as the possible changes resulting from the introduction of MT technology.

7.4.2.1 Standard Translation Workflow with MT

In this section, I summarise four case studies where MT was applied in an industrial translation workflow. This analysis will be useful to define and suggest a workflow for an automotive company such as BMW in the next section.

7.4.2.1.1 First Case Study: Baan Development B.V.

Baan was a vendor of popular enterprise resource planning software. Carmen Lange presented in 1999 a paper that contained a case study dealing with how to combine MT with a TM (Lange & Benett, 1999).

In this case, online help texts were translated by integrating the Translation Memory System Transit (version 2.7) with the machine translation system Logos (version 7.8.2.). The integrated workflow consisted of these main components:

- Source texts are adapted to the rules of the MT system with the help of macros
- Texts are imported in Transit.
- The segments that are not in the Translation Memory are sent to Logos in an extract file.
- Logos sends back the extract file to Transit.

- The file is revised.
- The MT output is optionally improved with entries in the terminology database or in the lexical components of the MT system. In this case, the text is imported again into Transit.
- Texts are post-edited.

Figure 40: Translation Workflow at Baan shows a flow diagram representing the translation workflow. The authors of the study also mention the importance of well-defined roles and propose the role of the translator or reviewer, who will review the text and detect those terms or expressions that must be coded again or newly in the lexical component of the MT system, and the role of the super-user, who has access to the LogoServer, and who can write or rewrite the rules. In this way, only an experienced user has access to the sensitive information contained in lexical resources. However, I have not included these roles in my diagram since all other tasks remain undefined in terms of who should carry them out.

Finally, the authors end up concluding that the implementation of such a workflow has reduced the translation time by up to 50%.

7.4.2.1.1 Second Case Study: CNH

CNH is the largest manufacturer of agricultural tractors and combine harvesters in the world, and one of the largest producers of construction equipment. It also has one of the industry's largest equipment finance operations.

In contrast to the previous case, the solution applied by CNH is much more sophisticated, due to the integration of a Translation Management System and the technology developed in the past few years, which implies much more automation in the process.

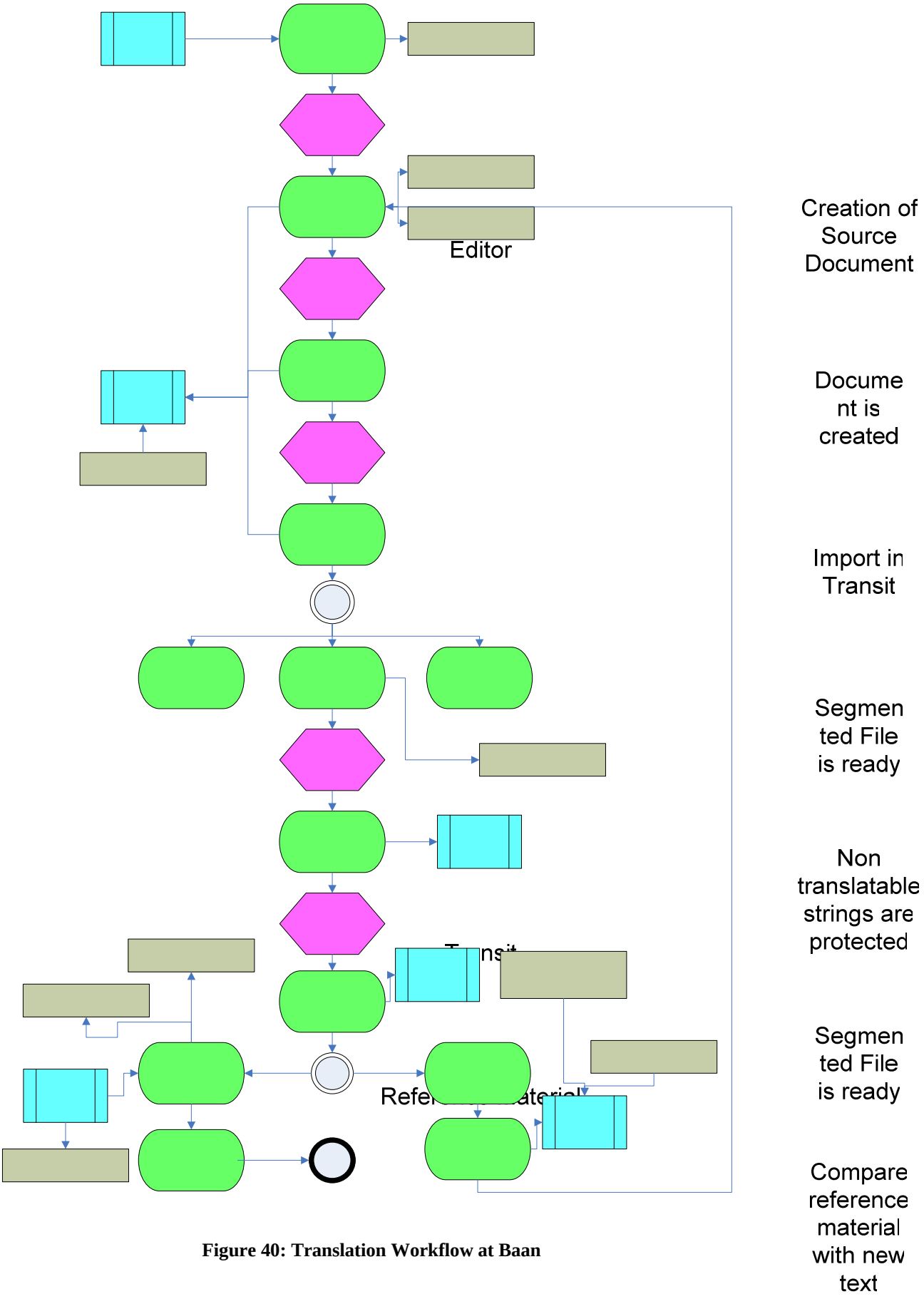

Figure 40: Translation Workflow at Baan

Machine Translation at CNH was implemented within the ASIST project, which resulted in the development of a tool called ASIST to "improve service quality, reduce the problem-experience gap and increase customer loyalty" (Healy, 2004). ASIST is a tool where customers and dealers can post their technical problems through the system and, if there is no "packaged" solution yet, a new one will be developed by the technical support at CNH.

Although the documentation available does not provide any details about the workflow, it makes it clear that the solution deployed by CNH also combines MT (KbTS by SDL), SDLX translation memory and TermBase terminology management with human post-editing elements.

The implementation of this integrated solution has resulted in cost savings of up to 50% as well as productivity gains of 60% for CNH.

7.4.2.1.2 Third Case Study: Volkswagen

The case of Volkswagen is explained in a paper authored by Ulrike Bernardi, Andras Bocsak and Jörg Porsiel (2005) and presented at the Annual Conference of the European Association for Machine Translation in Budapest.

Volkswagen tested 6 commercial systems and chose Comprendium's Traslator Server (Braintribe Group) as well as Comprendium's monolingual and bilingual terminology extraction tools to build a terminology base and import it into the dictionaries of the system.

The implementation of MT at Volkswagen was realised in two scenarios:

Scenario 1: Translation portal that can be accessed by all company employees over an Intranet, providing them with raw translations to get a gist of a document in a split second, for sources such as e-mails, reports, websites etc. In order to maintain terminological consistency within the company, integration of terminological tools and

large-scale terminology imports into the machine translation system proved to be indispensable. The language pairs offered in this portal are: German ↔ English, German ↔ Spanish. The portal is widely accepted by employees and the demand for fast raw translations within the company is striking high.

Scenario 2: Client application (written in Java) for localizing Assembly Instructions at Volkswagen. This application can be downloaded from the Intranet, and the clients can access a central server from every assembly and production site. Assembly Instructions consist of short sentences with simple grammatical structures and with a specific, but restricted, vocabulary. For this type of text, Translation Memories can only be used restrictively since, although the texts contain similar structures, very often small differences require complete rewriting of the sentence. The language pairs included are German ↔ English, Spanish and French and, according to requirements, either the translation is executed interactively field by field or the whole document is translated by the MT system and post-edited by the translators afterwards.

With regards to details of the technology and workflow applied by Volkswagen, the article explains that, for the successful implementation of MT, it is most important to have a high performance server that can cope with all translation requests, as well as an easy integration with other applications, including the import and export of terminological resources. Comprendium's Translator Server works with a Task Scheduler that manages the different tasks, reducing the workload of the pool and increasing the overall performance. It also includes APIs for JAVA, CORBA, COM, SOAP and HTTP), which allows for integration with other applications. Before the implementation of the system, a terminology migration from the Terminology Management System of Volkswagen to the MT system was made. Terms were first imported into Comprendium's professional dictionary administration tool LexShop, which includes an automatic input parser and defaulter to create the MT system values.

Although the authors of this article do not give any data regarding gains in productivity or cost savings, they emphasise the high user acceptance and positive feedback, which has encouraged Volkswagen to undertake new projects in this area, such as the

inclusion of new language pairs, or the offering of MT translation services to other departments as part of their authoring processes.

7.4.2.1.3 Fourth Case Study: SAP

SAP is currently applying MT to different translation scenarios. Bernardi, Bocsak & Porsiel (2005) report about four systems being deployed for the translation of offline texts, i.e. texts extracted from SAP systems, converted into an "MT-suitable" format before machine translation and re-imported into the systems after the translation process has been completed.

The MT systems deployed are:

- LOGOS (used for English–French and English–Spanish)
- PROMT (used for English–Russian and English–Portuguese)
- METAL (used for German–English)
- LOGOVISTA (used for English–Japanese)

The translation workflow varies from system to system, the LOGOS and the PROMT processes being very similar to each other and differing greatly from the METAL process.

LOGOS and PROMT are used to translate SAP documentation material and training courses, whereas METAL and LOGOVISTA are used exclusively for the translation of "SAP notes" (standardised documents for troubleshooting and customer support).

Schaefer (2003) illustrates the workflows with the systems PROMPT and METAL, and makes references as necessary to the major differences between the workflows connected to the four systems.

As for PROMPT, this system has been productively deployed in the translation of SAP documentation and training courses since August 2000. The software is locally installed at SAP, but MT output is sent to external agencies for post-editing.

The translation process with PROMPT is outlined in further detail in Boehme & Svetova (2001). Here, after a new document is created, it is first pre-processed with the tool PROPMT TerM by the PROPMT dictionary developer responsible for SAP TM and MT dictionaries, in order to extract terminology candidates, and compare them with existing terms in the PROMPT and TRADOS TWB dictionaries. The dictionary developer and the translator define the translations of the new terms, update the dictionaries before translation and report, any terminological problems to SAP.

The second step of the process is the translation in TRADOS TWB and PROMPT, which are integrated via the module P4T developed by PROMPT. Here, the text is first sent to the TM and the non-matches to the MT system. The translator then gets the text to be post-edited in the TRADOS TWB. The translator is also expected to report to PROMPT about possible dictionary entry improvements.

The METAL technology has been used in the translation of SAP notes since 1993/94.

Although initially Machine Translation and post-editing were done internally at SAP, in 1996 the translation of notes was outsourced to an external translation agency, which also meant a change in the translation workflow itself. In contrast to the other workflows, the software is installed externally, which means that the whole translation workflow takes place externally.

The following figures (the second one extracted directly from the article by Schaefer, 2003), illustrate the workflows with PROMPT and METAL:

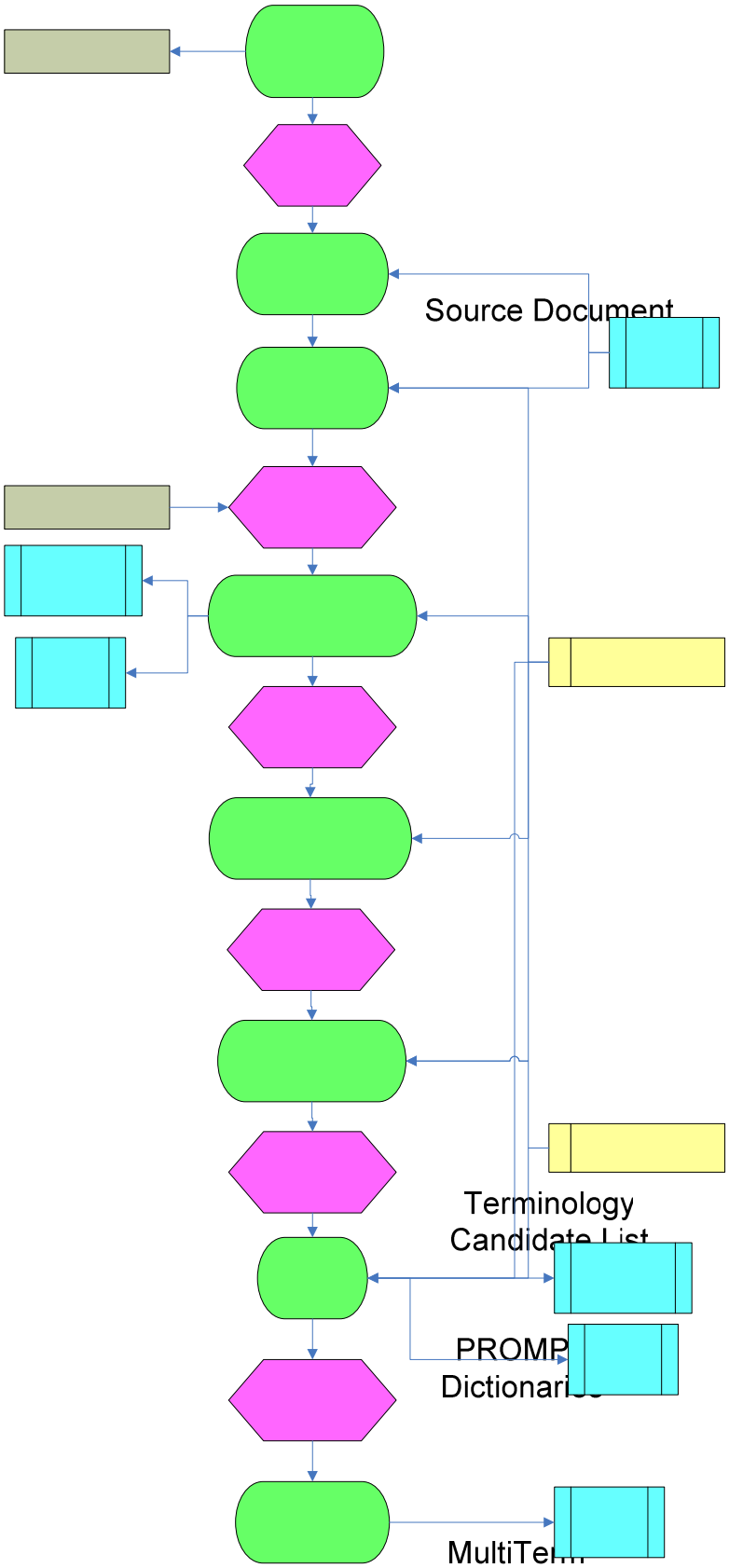

Figure 41: Translation Workflow at SAP with PROMPT

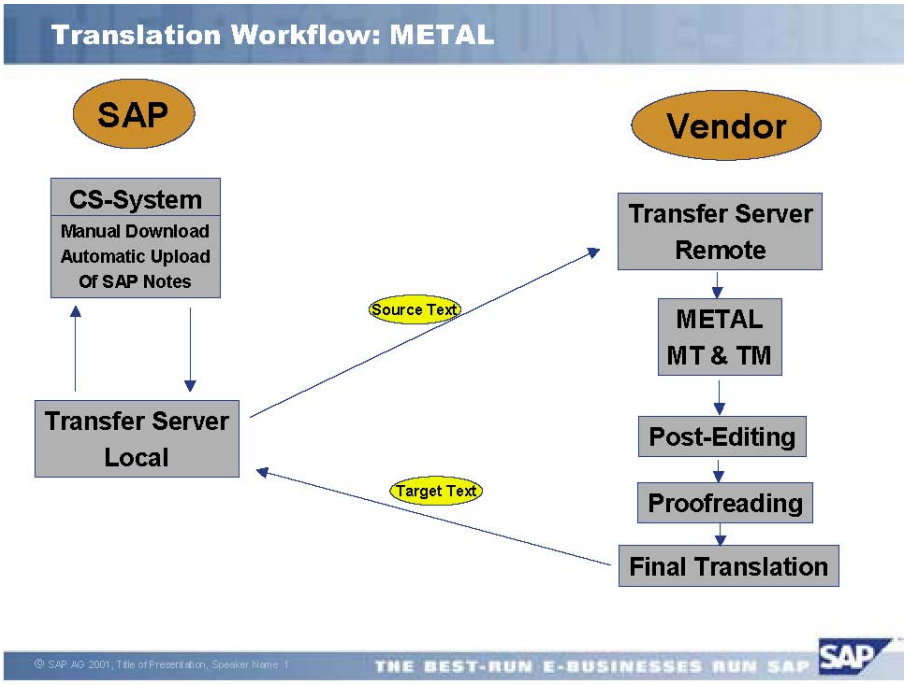

Figure 42: Translation Workflow at SAP with METAL

The Annex XI summarizes the most important data from all four case studies.

There are other case studies, such as those reported by Rychtyckyi (2000) at Ford or Routurier (2004) at Symantec. Furthermore, a considerable amount of other companies are implementing this kind of technology and workflows, and are not releasing their data due to sensitive information or marketing strategies.

7.4.3 Workflow Proposal for an automotive company: the case of BMW

We propose two different processes for MT: MT translation pre-processing and the MT-process, which also includes post-editing. In the following sections I explain the details of the process in three steps: MT Translation Pre-Processing, MT Translation and MT Translation Post-Processing.

7.4.3.1 MT Translation Pre-Processing

The processes Translation Pre-Processing and MT Translation Pre-Processing are two processes that share the same steps up to the point when it is decided whether the translation should be carried out by a human or by the MT system.

First, the project manager is informed about the existence of new material that needs to be translated. A project is created in the Translation Management System and in the TMS. Subsequently, the new material is imported in the TMS and compared to the reference material, the statistics are analysed and a package with the new segments is created.

At this point, the common steps end. The translation method has to be determined. In case the decision[116] for a translation with the MT system is made, a package with the remaining segments (fuzzy matches and no matches) is created and sent to the MT system, and the MT translation process begins.

It is necessary to point out two aspects that might result in slight changes of the process if not applied as suggested:

First of all, the implementation of MT in the translation process might have a previous implication in the process of creation of multilingual content. Special controlled language rules adapted to the MT system might be applied during the creation of the content in the source language in order to obtain better output quality results and smooth the post-editing process.

When the package for the MT system is created, two options are possible: either all segments that have not been previously translated are included in this package (fuzzy matches and no matches), or only the non-matches are included. It might also be necessary to protect those segments that are not to be translated, so that the MT does not process them (for better understanding, see Figure 43: MT Translation Pre-Processing) As I will explain later, I opt for the first option.

Finally, it is important to define well the integration between the TMS and the MT system. It is possible to work "manually" with both systems, generating and extracting the delta file from the TMS, importing it into the MT system and starting the translation once the text is prepared. However, I advocate the full integration of the systems, so that the generation of the delta file, the export from the TMS, the import into the MT system, the translation itself and the import back into the TMS take place automatically. The forecasts realised for the ROI are based on this level of integration.

7.4.3.2 MT Process

The MT Process is the process after the MT Translation Pre-Processing. The file with the segments to be translated is in the MT system and the automatic translation can start.

The translation can be triggered off in different ways. I presuppose that, in a workflow where big text quantities have to be translated, a fully automated batch mode might be recommendable. For instance, all translation tasks could be collected in a task pool and translated overnight, so that the MT manager could start doing his or her work of analysis or workflow management on the next day. Besides, there has to be the possibility of starting a translation manually, in case there is an urgent project or something goes wrong with the workflow (technical failure).

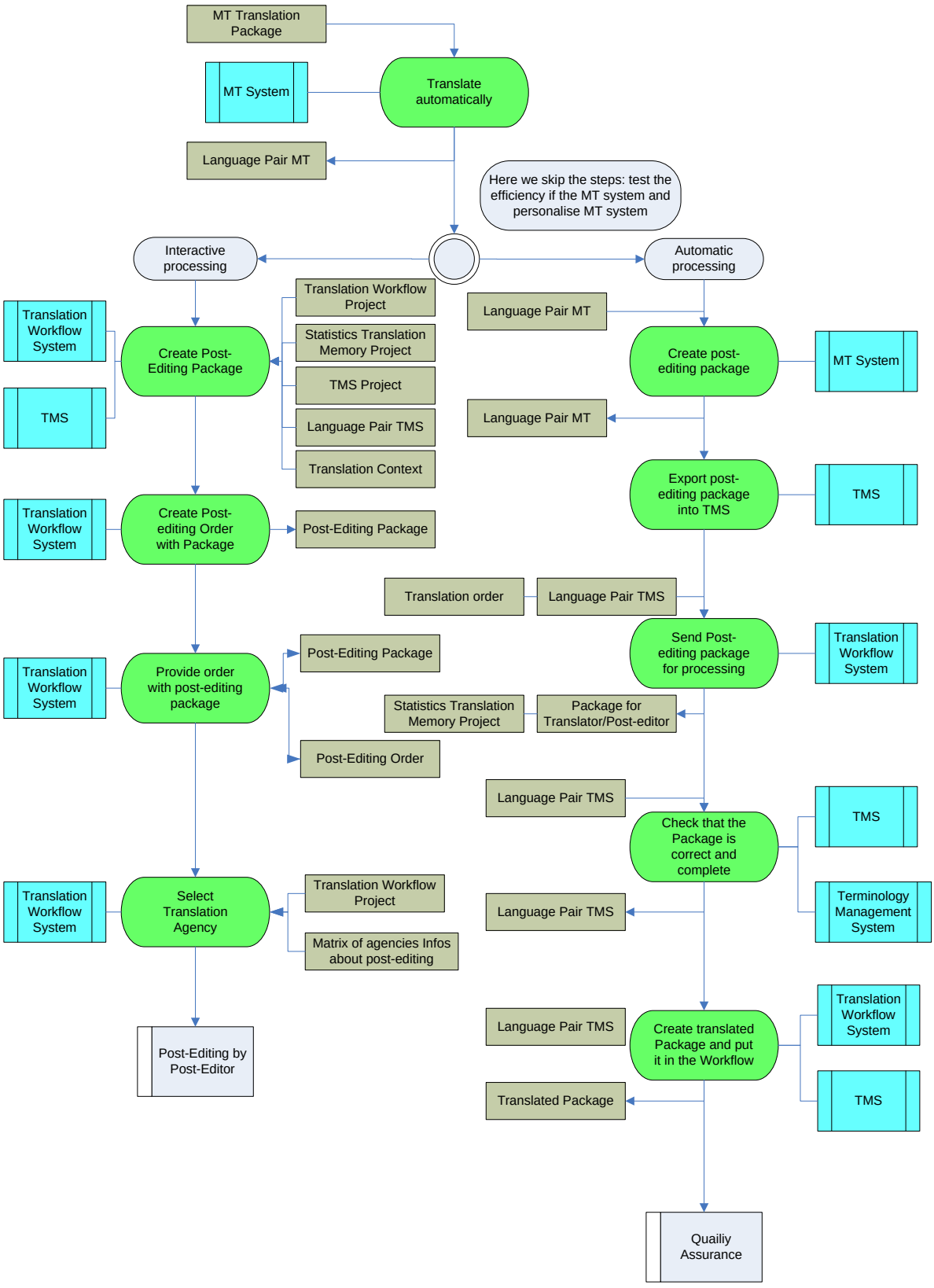

Figure 43: MT Translation Pre-Processing

After translation, the text is checked by the MT manager, who analyses the text in order to assess whether the system can be improved in order to obtain better output quality results. The optimisation measures include modifying or creating new analysis, transfer or generation rules for the MT system, add unknown terms to the dictionary, add semantic or grammatical information to the terms for a better analysis by the MT system, or adapt any of the customisation possibilities the system offers.

Depending on the depth of the analysis and the nature of the modifications, the MT manager might need to be proficient in the language, or have to work in close collaboration with a native speaker (either a translator or an specialist in the target market). Furthermore, the MT manager has to receive the appropriate training and have access to the components of the system in order to carry out this work. Some changes might also have to be undertaken in close collaboration with the MT system manufacturer, who will deliver the new rules in the form of Service Packages or new releases.

Once the system has been optimised, the translation is repeated[117] and, if no opportunities for further optimisation can be identified, the translation is finished and a decision as to how the text will be further processed has to be made.

At this point, there are two possibilities:

- The translation can be edited interactively.
- The translation can be edited automatically.

7.4.3.3 MT Post-Processing

The first option is discussed further in the process “MT Translation Pre-Processing” (see 7.4.3.1). However, I would recommend the creation of a new process called MT Post-Processing for a better division of the different steps in the translation workflow. Besides, though the steps are the same, the contents of the files in both cases are different: in the case of real MT Pre-Processing, the translation package created and sent

to the agency has not been translated yet (is sent for translation), while the steps coming from the MT process itself imply that the package includes the finished translation, and this is sent to the agency for post-editing.

If the translation is going to be processed interactively, the package with the finished translation is created, an order with all the necessary details is prepared and a translation agency is chosen, and the next process can take place: Translation by the translator, which in this case would be Post-Editing by the post-editor. At this point it is necessary to add the following remarks:

a) The selection of the agency for post-editing implies that information about the services offered (translation, interpreting, desktop publishing etc.) is listed in the agency profile as well as the prices (in this case for post-editing).

b) As for the post-editing itself (in the current process, translation), the translator must know that, as explained in section 7.5.3.2, fuzzy matches will have two translations, depending on the system they come from (TMS or MT). It is also necessary to train translators or post-editors so that texts are post-edited according to certain requirements in order to avoid the complete rewriting of the translation.

c) Once the post-editing is finished, the process of quality assurance can start.

d) The second option, the automatic post-editing, is explained in the process “MT translation”.

e) In this process, a post-editing package is created and the MT translated version is sent for post-editing first to the Translation Workflow System and then to the TMS. Here, a person in charge of the quality assurance of the target language checks the text to guarantee that it is formally correct and complete. Finally, the package is sent back to the Workflow System that, in turn, sends it for the next

process: Quality Assurance. In contrast to the interactive processing, here no post-editing by a professional post-editor or translator takes place, but the document is sent to the Quality Assurance after a formal check.

In the following figures summarize the data flow during the whole process:

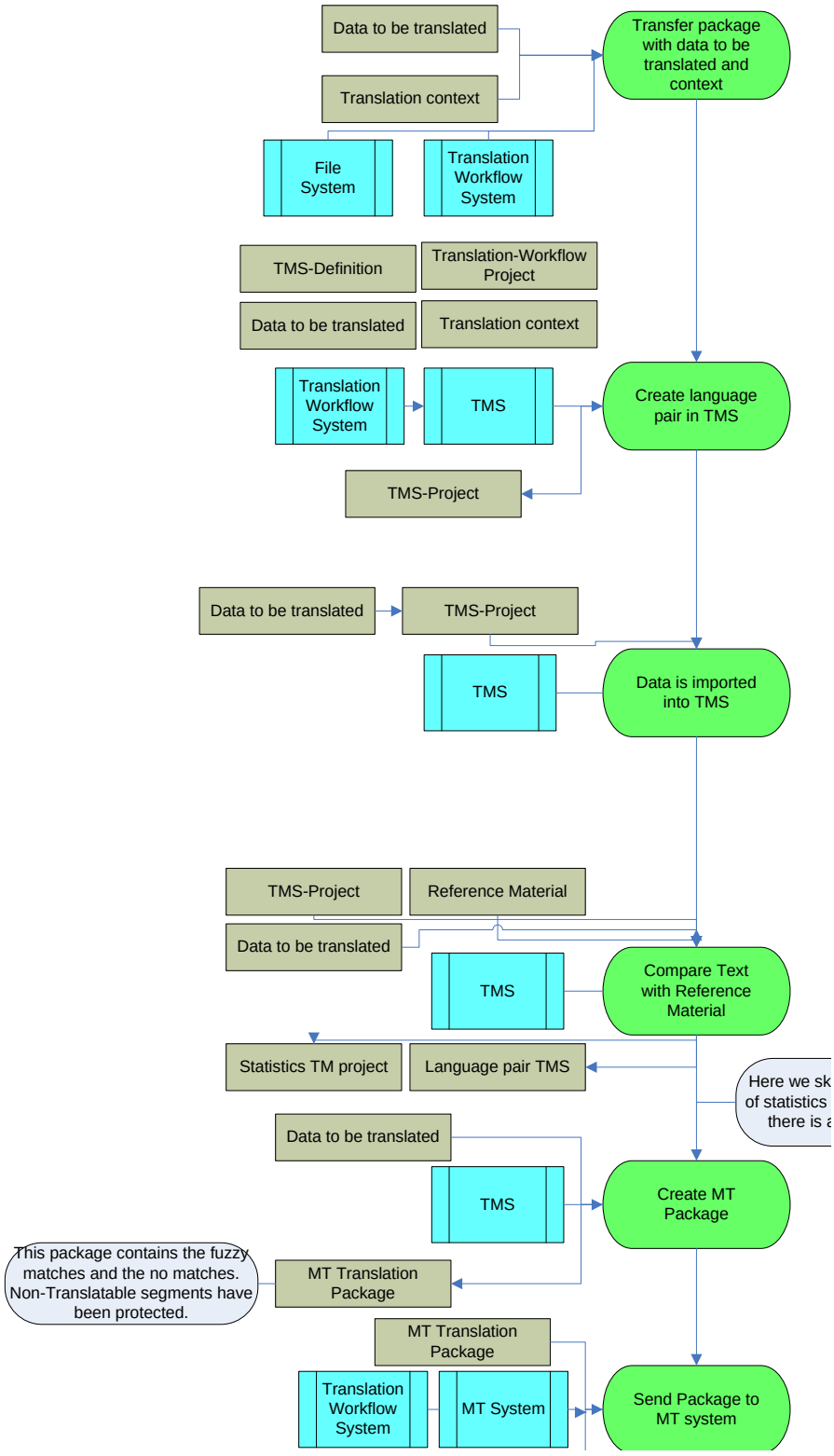

Figure 44: Data Flow during MT Pre-Processing

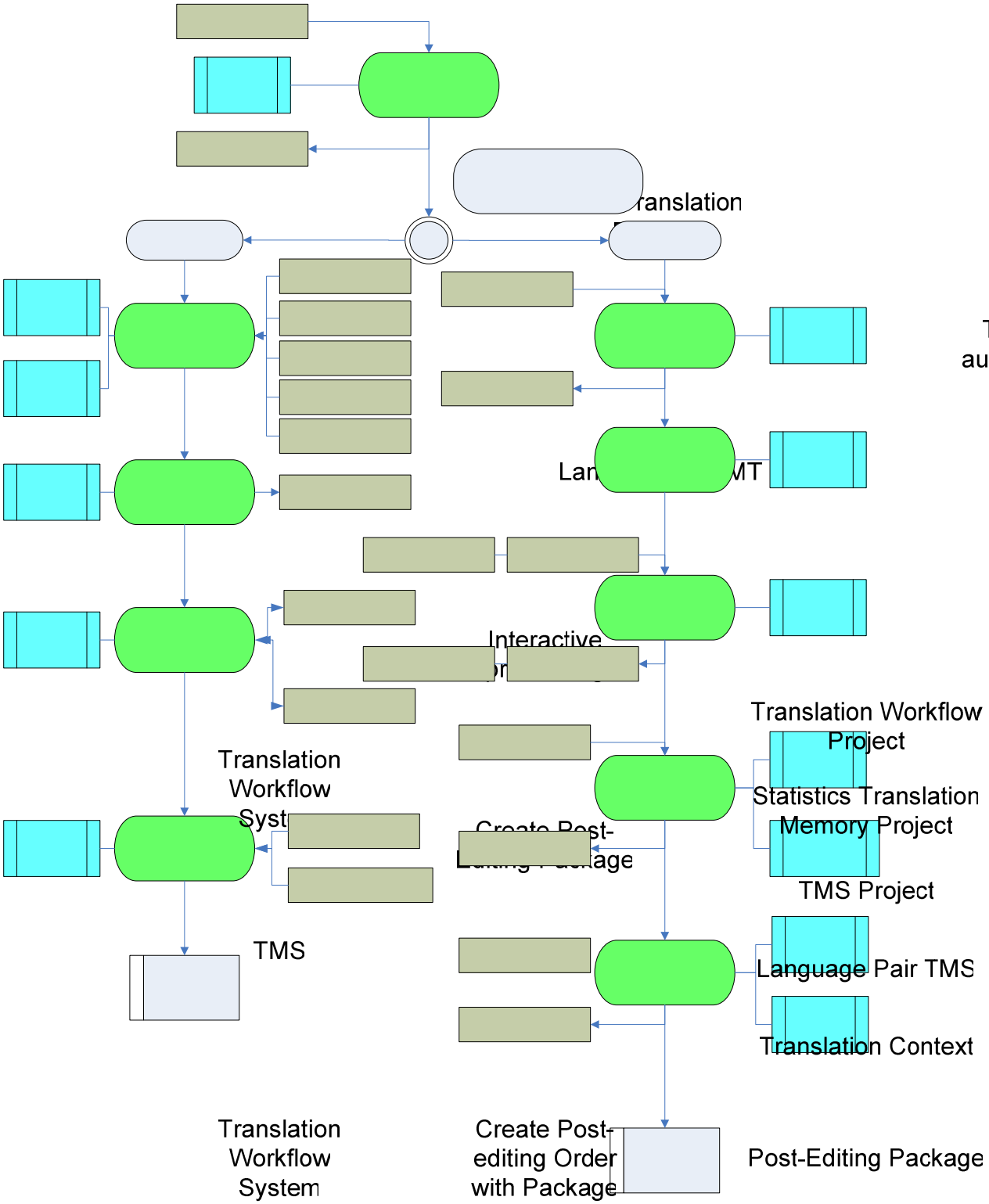

Figure 45: Data Flow during MT Process

7.5 Economic Analysis

7.5.1 Return on Investment (ROI)

ROI (Return on Investment) is a concept deriving from the field of Capital Investment and is a measure used to evaluate the efficiency of an investment. It can be defined as "A general concept referring to Earnings from the Investment of Capital, where the earnings are expressed as a proportion of the outlay."[118]

The ROI is calculated as follows:

$$\text{ROI} = \frac{\text{Profits} - \text{Invested Capital}}{\text{Invested Capital}} * 100$$

The result is a percentage. If the ROI is less than 100%, a project may not be undertaken. For example, a 300% ROI over five years indicates a return of 3 times the original investment over a five-year period. Breakeven analysis is used to indicate after how many months the investment is recuperated.

This definition of ROI only includes hard benefits, i.e. financial benefits that can be easily quantified. However, and especially when calculating ROI for translation processes, it is necessary to take into account other kinds of benefits, such as corporate image and better time-to-market, which are difficult to define and quantify. These are the so-called "soft benefits", and it is important to take them into account when analyzing the advantages of MT.

In the coming sections, I will give an overview of the costs involved in the translation processes of an automotive company, taking the data of at BMW, and estimate the costs of implementing Machine Translation within the translation workflow described in the previous section. Subsequently, after quantifying the potential cost savings obtained by deploying this new technology, I will determine the Return on Investment.

7.5.2 Cost Factors in the Translation Process

According to Lawlor (2005), when we talk about translation costs in general, it is necessary to consider them as falling into one of three categories:

- Translation activities which include all linguistic tasks related to the actual rendering of source words into target languages: translating, editing, proof-reading and review.
- Transaction activities which include project management and coordination for source content and its manipulation throughout the entire translation process.
- Other costs, which include avoidable costs and opportunity costs associated with reduced efficiency, such as lost revenue.

7.5.2.1 Translation (Linguistic) Costs

Despite the fact that these costs are not directly derived from the implementation of MT, it is necessary to take them into account in order to be able to quantify cost savings as the difference between the costs of a traditional translation process and the costs involved in a machine translation process.

Translation costs are a result of one of these main activities: translation, editing, proof-reading and review. All of these tasks are performed by highly-qualified translators. Translation and editing fees are traditionally calculated on a per-word or per-line basis, whereas proof-reading and review tasks (quality assurance) can be calculated either on a per-word basis or on an hourly basis. In the case of BMW Service literature, the main vendor calculates the pricing on a per-line basis and includes both translation and quality assurance costs.

It is obvious that, since translation is paid for on a per-word or per-line basis, more words result in a higher cost. In order to reduce the costs it is therefore necessary to reduce the number of words which need to be translated. Four main strategies help to achieve this goal:

- Translation Memory (TM) Tools: these tools support the re-use of previously translated segments, thus reducing the number of new words that need to be translated.
- Terminology Management Tools: these tools support the management and definition of unambiguous terminology. This speeds up the translation process and improves the quality and consistency of translations. If consistent terminology is always used for the same concept, the number of words that need to be translated diminishes.
- Controlled Authoring. The use of controlled language when producing technical documentation improves the quality and consistency of source texts. This results in content that is more clear and predictable, which not only helps final readers and human translators in their work, but can be also processed more efficiently by automation technologies.
- These three strategies are already implemented in the production of technical documentation and in the translation processes at BMW Service. As seen in section 7.3, another strategy, translation automation, can further help to reduce costs.

7.5.2.2 Transaction Costs

In any translation project there are a number of steps around the translation task itself that consume time and contribute to additional costs both for the client and for the localization vendor. These include data handling, clearing up of reference material, project management and other administrative tasks.

7.5.2.3 Other Costs

Other costs that might arise during the translation process can be classified as avoidable costs, or the cost of lost or delayed revenue.

Other additional costs might result from delays in delivering translated content, especially to Asian markets, since these have to wait for the English version in order to translate into their respective languages. Besides the additional costs incurred by not

being able to deliver translated support content on time, there could be also negative impacts on customer service, damage to global brand and reputation, and even reduced future sales in that market.

7.5.2.4 Costs Factors More Favourably Affected by Automation

Lawlor (2005) presents eleven source material and project factors that can determine how much can be saved through automation. The more these factors come into play, the more can be saved. Those factors closely related to MT technology will now be examined and related to the situation at an automotive company such as BMW.

7.5.2.4.1 Source Material Characteristics

Updated or existing material has more corresponding segments in TM than new material. Although this is not directly related with MT, the more text can be leveraged from the TM, the less text has to be sent to MT and, thus, to post-editing.

Within an automotive company, the quantity of new text and pre-translated text usually varies from year to year, depending on whether new models have been produced that year or special technical problems have arisen. In the case of BMW, the quantity of new text and fuzzy matches for Reparaturanleitungen (repair instructions) has swung over the last years, whereas the quantity of pre-translated text has grown steadily, which indicates that the use of TM and controlled language have been effective.

For Service Information, the quantity of pre-translated text has grown steadily over the past 4 years. However, the amount of new text has also increased progressively, due to the overall increase in text volume. Although this means more text volume for human translation, or for MT and post-editing, it allows us to obtain representative results when calculating the productivity gains resulting from the implementation of MT technology.

Another aspect that might favour automation is the nature of the material. Technical material is more appropriate for translation automation. This kind of text uses more repetitive elements and univocal terminology, which can be enforced by the use of

controlled language. The types of text I have chosen for my study are service documents intended for service technicians or mechanics. The content is thus technical and there should be no room for ambiguities in the language. The authors use MULTILINT/CLAT to produce the texts, applying the rules of controlled language. In this way, a univocal terminology as well as a standard, neutral technical style is attained.

Besides, technical documentation in big automotive companies is usually produced in an authoring system with fixed structures and modules. Texts are written in structured formats such as SGML or XML, which means that tags are used to mark the different elements of the text. Although not yet the case, in the future this might help MT to resolve ambiguities or to translate identical text chunks in different ways depending on the context. For instance, a title and a sentence inside a paragraph, even if they have the same form, may be translated differently, according to the meta information contained in the tags.

7.5.2.4.2 Project Characteristics

As I stated before, the key to reducing costs using translation automation is by reducing the number of words to translate. Thus, the more words there are to manipulate, the more potential there is for cost savings, since usually an exponential model is given when automating translation processes. This implies that the larger the quantity gets, the faster it grows. The experiments I have carried out with the text types RA (Reparaturenleitung or repair instructions) and SI (Service Information) amount to an average of 80,000 lines per year, which mean an MT solution delivers productivity in a world-class workflow context for projects with over 500,000 words a year into a target language set". According to him, these text types have a potential for productivity when translated into different target languages. However, for my study, I only consider the language pair German-English, reducing thus the potential for cost savings in this regard.

Other aspects, such as changes made during the project, the number of files involved, the definition of the process, or whether there are custom processing requirements, may also affect savings through automation. However, these aspects should be solved by a Translation Management System rather than by MT.

7.5.3 Estimating Implementation Costs

I will now give an overview of the different costs possibly incurred by the introduction of MT technology into a standard translation workflow. I will start by estimating the implementation costs, including installation and customization of the system, training of users and testing.

Lawlor (2005) defines three phases in the implementation process that have to be considered:

- Define, asses and refine the current process
- Install the Machine Translation System
- Train the users; customise, maintain and manage the System

7.5.3.1 Define, Assess and Refine Current Process

The first phase consists in analyzing the current processes and studying possible solutions to optimise these processes in order to meet business needs. This implies analyzing content, developing workflows and establishing standard processes. Furthermore, configuration information must be defined: users, roles, languages, content types, views etc.

For this purpose, either an internal workforce or an external consultant will have to be hired, in order to present a business proposal. In this case, most of the work has already been done within the scope of this work. Therefore, I will not include these costs in the ROI.

7.5.3.2 Install the Machine Translation System

When the empirical part of this work was carried out, Linguatec provided a single demo version for 50 € for language pair. Linguatec sent a version of the Personal Translator 2004 Office Plus for free (German ↔ English). Besides, Linguatec offered a network solution for 250 €. The current version is Personal Translator 14 and offers the hybrid technology, based on neuronal networks, thanks to which general-common knowledge is applied to the translation and common names are detected automatically. They offer a stand-alone version for 49 €, a server-based solution for company Intranets for 799 € (5 licenses) and an Intranet version for 4,975 € for language pair.

Systran offers a free SYSTRAN Premium version (expiration 30 days) and a corporate network solution: evaluation license for the client-server solutions SYSTRAN Enterprise or SYSTRAN WebServer (full functionality) priced pro rata temporis. The new version also offers hybrid technology and statistical rules.

Comprendium offers a demo version with dictionary manager for 650 € for language pair (German-> English). Each further language pair costs 500 €.

The second phase involves the installation itself and all the steps required to fully integrate the system into the workflow:

1. Set new workflow rules to adapt the workflow to the characteristics of MT: in case MT is implemented afterwards, either new workflow rules or a completely new workflow will have to be defined within the Translation Management System, in order to cope with the needs of such a technology. I assume, however, that the decision to implement MT within the translation processes in the company would be made beforehand. Therefore, I do not consider these costs as relevant to my ROI.

2. Cost of integration with existing systems: Integration with the TM system will be necessary in order to establish a workflow in which content is processed by the TM Server and subsequently by the MT server for automatic translation. Therefore, it is

important to assess at this point which APIs are offered by the MT vendor and whether these are compatible with the applications that have to be integrated. For the interaction between the translation memory (TM) and the MT system, there are three approaches:

- Approach 1: Traditional Localization Process enhanced with MT. In this scenario, the TM client is used as a front end tool for the translator, while the MT server is used as a back end tool. This means that the translator only works with the TM tool, while the MT server is managed centrally by a MT manager. Texts are first pre-translated (100% matches) and the rest of the segments are sent to the MT system. The translated sentences are then sent back to the Translation Memory and marked with a special status (threshold value) or a fuzzy match penalty (e.g. 15%). When the translator works in the TM, for some sentences he gets a single fuzzy match of 85% similarity, indicating that they come from MT. For some other sentences, the translator gets a fuzzy match of 85% and another rate above 85%, which tells him that the latter comes from the TM. At this point, there are two possibilities: either the translator adapts the real fuzzy matches (above 85%), while the post-editor adapts the segments coming from the MT engine (85%), or the post-editor adapts all fuzzy segments. Once post-edited, the sentences are included in the memory as reference material.

- Approach 2: Pure MT output. Texts are first pre-translated (use 100% matches from existing TMs) and the rest is pushed through the MT engine. The content is then combined, that is, TM and MT output are merged, and published. This is the low cost alternative, but also the low quality one. No translation costs are involved in this scenario, however it is important to flag the text as MT output "to set appropriate expectations for users and avert damage to the corporate brand" (Lionbridge, 2001).

- Approach 3: A combination of scenarios 1 and 2. Depending on the importance or purpose of the text, scenario 1 or 2 is chosen. The importance of the text might be defined beforehand or deduced from user statistics or feedback. The

quality standards have to be defined by the company, so that the decision whether to adopt approach 1 or 2 can be made automatically.

These two degrees of integration require a similar investment in resources, but differ slightly in their respective processes. Since this is only an approach to what language technology systems can be integrated in the workflow, only an approximate value can be given. For further details, section 7.4.2 gives an overview of the different workflow possibilities.

3. License costs and technical support: the type of license will depend on the process in which MT is implemented. Different scenarios, such as in-house implementation, outsourcing or translation of e-mails, will imply different requirements. For this study, the following scenario will be considered: as mentioned before, I am interested in using MT for translation drafts that will subsequently be post-edited for publication. The translation will be carried out in-house, using both the existing translation memories and MT technology for the new fragments. The text will then be sent to an agency for post-edition. Thanks to the fuzzy match percentage, the post-editor can see which segments come from the memory and which from the machine. The purpose of this mark-up is so that the proof-reader can concentrate on the segments which were translated completely automatically. Once post-edited, the segments will be added to the memory.

7.5.3.3 Customization and Maintenance, Training and Management of the System

For the maintenance and operation of this system, a new role is necessary. The role of the MT manager comprises the following tasks:

- Update terminology: even if the import of terminology from the Terminology Management System takes place automatically, the MT system needs more information than the one contained in this system. This information can be of a semantic nature, collocations, syntactic patterns etc. There are direct implications

of this for the workflow, since it means more workload for the Machine Translation Manager.

- Monitoring the translation process: creating job queues (batch translation), starting the translation, attending to any prompts from the system, solving any terminology problems or conflicts that might arise during translation, analyzing or collecting examples to make proposals for new rules to improve the quality of the translation. Part of these tasks (creating job queues, sending them to the Machine Translation System etc.) can be automated through the introduction of a Translation Management System. Therefore, only part of this work load will be considered in terms of costs.

- Any other management task that might arise from the daily use of MT.

The role of the MT manager could be filled by one person for all language pairs and text types. This could be an in-house employee who is in charge of the translation process. The specialist nature of some of the tasks, such as the specification of the terminology, makes it necessary for the person in charge of the management of the MT system to receive the appropriate training required in order to know how to code the information for the system.

7.5.3.4 Costs for the company

7.5.3.4.1 Installation of the MT System

I take as a model an enterprise license with 5 clients (usually the minimum number of clients for this type of license). This implies that different employees within the same department or different departments can request translations. This type of license would allow sharing resources such as memories, configuration and terminology.

By maintenance I refer to the percentage or amount of money required for technical support and software updates and upgrades. Usually an upgrade for this kind of product is launched into the market every 2 to 3 years.

Technical support implies on-line support and the resolving of technical problems in the system. Linguatec annually charges 20% of the product net price.

In my example, the basic version of PT Network costs 799 € and includes five clients. Furthermore, I add the price of a specialised automotive dictionary. It is foreseeable that an update comes to the market approximately every 3 years. This would cost 625 € for the basic version of the dictionary. I then add the technical support costs. All this adds up to 2,809.50 € in license costs for a period of 5 years.

7.5.3.4.2 Customization, Maintenance and Management

For the first year, the customization of the system and the maintenance will be more costly, since different tests are needed to state how the rules of the MT system work and if these can be initially tuned to obtain better results. Besides, tests to assess the terminology and dictionary coverage are needed in order to estimate the effort needed to update the dictionaries of the MT system. For this initial work, and for the text types Repair Instructions and Service Information, I calculate that an internal or an external workforce working part-time during the first year would be needed, i.e. that is, half a man-year. After this initial effort, a quarter man-year would be needed for the management and maintenance of the system for the language pair German-English. This would involve adding new dictionary entries or optimizing new rules to improve the system performance. I base my calculation on 200 man-days a year and a 500 € fee per day. These costs would be incurred on a yearly basis.

In the following table I can see a summary of the costs estimated for the implementation of MT:

	2005	2006	2007	2008	2009	2010	TOTAL accumulated
License							
Clients	799.00 €			625.00 €			1,424.00 €
Extra dictionaries	426.70 €						426.70 €
Technical Support	159.80 €	159.80 €	159.80 €	159.80 €	159.80 €	159.80 €	958.80 €
Integration Costs with TM							
Realization Study (in-house) 30 man-days	15,000.00 €						15,000.00 €
Realization (external) 30 man-days	15,000.00 €						15,000.00 €
Human Resources							
MT manager (in-house)	50,000.00 €	25,000.00 €	25,000.00 €	25,000.00 €	25,000.00 €	25,000.00 €	175,000.00 €
TOTAL	81,385.50 €	25,159.80 €	25,159.80 €	25,784.80 €	25,159.80 €	25,159.80 €	207,809.50 €

Table 26: Implementation costs for MT

7.5.4 Quantifying Cost Savings

As mentioned in section 7.5.2.4, the potential for ROI increases depending on a number of factors, such as: the volume of translation; the number of target languages; the frequency of updates; and the quality and structure of the documents. In the case of MT, the primary contributors to ROI come from saved user time and, depending on the quality of the MT output, reduced amount of words to translate.

7.5.4.1 Quantifying Translation Costs

Although there can be different ways of measuring translation costs, these are usually calculated on a per-line basis. This makes it possible to divide the lines into different groups depending on the degree of coincidence with the reference material from the TM:

Pre-Translation Grade	% of line price
Previously Translated Text that has been fully pre-translated (100% similarity with a segment in the memory or reference material)	0 of line price
Fuzzy Match Text for which a similar translation exists (85-99% of similarity)	80 of line price
New Text completely new text that has not been previously translated Fuzzy index less than 85%)	100 of line price

Table 27: Line Classification depending on Pre-Translation Grade

I proceed now to calculate the translation costs based on the data delivered by one of the main translation vendors in Germany. The net standard prices for Service Literature from German into English are 0.95 € per line. For fuzzy matches, the translation vendor charges 80% of the line price. The number of lines is multiplied by 0.97, which is the weighting factor established for English to help in determining the estimated effort required for the translation of a certain language pair.

In Table 28 and Table 29, we can see that there has been an overall increase in repair instructions text volume of 14.87%, whereas the translation costs have sunk by a 16.75%. This is especially due to the reduction in the amount of new text and the notable increase in the amount of pre-translated text. On the other hand, although the overall increase in Service Information text volume amounts to only 8.18%, the dramatic increase in the amount of new text, and especially of fuzzy matches, shoots the translation costs up to a 157% increase from 2004 to 2005.

	Total	New Text	Fuzzy Match	Pre-translated
RA 2002	19,136.05 €	16,589.76 €	2,546.29 €	0.00 €
%	-2.88	-54.99	336.65	
RA 2003	18,585.36 €	7,466.91 €	11,118.45 €	0.00 €
%	34.95	119.82	-22.05	
RA 2004	25,081.02 €	16,413.76 €	8,667.26 €	0.00 €
%	-14.21	-15.00	-12.72	
RA 2005	21,515.92 €	13,951.51 €	7,564.41 €	0.00 €

Table 28: Translation Costs for RA

	Total	New Text	Fuzzy Match	Pre-translated
SI 2002	10,254.27 €	9,542.13 €	712.14 €	0.00 €
%	180.27	142.39%	687.78	
SI 2003	28,739.74 €	23,129.65 €	5,610.09 €	0.00 €
%	23.80	32.41%	-11.68	
SI 2004	35,580.77 €	30,626.05 €	4,954.72 €	0.00 €
%	157.01	143.67	239.47	
SI 2005	91,445.79 €	74,625.83 €	16,819.96 €	0.00 €

Table 29: Translation Costs for SI

7.5.4.2 Post-Editing Costs

If MT is to be implemented without sacrificing quality, the translation process has to be complemented with a new task: post-editing. This corresponds to the first scenario outlined before and means that either skilled post-editors or translators[119] have to revise and, if necessary, correct the sentences that have been automatically translated. It is necessary to point out that not all sentences will need post-editing, though this can only be stated by the post-editor. This means that all sentences have to be revised, but only a percentage will have to be rewritten. Although this allows us to calculate how productivity can increase, it is not easy to convert these data into cost reductions. This is due to the impossibility of knowing beforehand how many sentences have to be post-edited, which in turn makes it difficult to know how much time is necessary or how many lines need to be post-edited.

I calculate the post-edition costs based on the information delivered by the same translation vendor that provides translation services. This vendor prices new lines at 0.95 €, whereas lines for post-editing are priced at 80% of the line price, that is, at 0.88 €, which is the same price as that of a fuzzy match.

Other translation vendors set hourly rates for post-editing. However, the prices can vary considerably, depending on a number of reasons, such as the degree of post-editing (full or light), whether controlled language has been previously applied etc. Prices can range from 25 to 60 € per hour (Van der Meer, 2006: personal communication).

Table 30 and Table 31 show post-editing costs for the text types RA and SI, based on the data from 2002 to 2005.

	Total	New Text	Fuzzy Match	Pre-translated
RA 2002	15,818.10 €	13,271.81 €	2,546.29 €	0.00 €
RA 2003	17,091.98 €	5,973.53 €	11,118.45 €	0.00 €
RA 2004	18,085.73 €	13,131.01 €	4,954.72 €	0.00 €
RA 2005	18,725.62 €	11,161.21 €	7,564.41 €	0.00 €

Table 30: Post-Editing Costs for RA

	Total	New Text	Fuzzy Match	Pre-translated
SI 2002	8,345.84 €	7,633.71 €	712.14 €	0.00 €
SI 2003	24,113.81 €	18,503.72 €	5,610.09 €	0.00 €
SI 2004	29,455.56 €	24,500.84 €	4,954.72 €	0.00 €
SI 2005	76,520.62 €	59,700.67 €	16,819.96 €	0.00 €

Table 31: Post-Editing Costs for SI

7.5.4.3 Quantifying Reduced Translation Costs with MT

As mentioned in section 7.5.3.2, different approaches can be outlined when implementing MT technology. These scenarios lead to different reductions in translation costs, depending on the degree of post-editing applied.

Table 32 shows an overview of the costs of the different scenarios. For RA, average savings of up to 29.83% can be attained, whereas for SBT the average savings amount to 20.67%. It is necessary to point out, however, that the volume increase and thus, the saving potential, can vary considerably from year to year, depending on the launching of new models, or the emergence of special technical incidences. We see, for instance, that the saving potential for RAs in 2004 is considerable bigger than that for other years. This is due to the larger amount of new text compared to 2003, caused by the introduction of a new model. We can observe the same phenomenon for the SIs in 2005. Until that point, the amount of text had increased steadily, but in 2005 the difference shoots up to 143.67%. This may be due to various reasons, such as technical incidences that required special service information.

From the data of 2002 to 2005, we see that an average of 20.26% was saved in translation costs, thanks to a combination of MT and Post-Editing.

7.5.5 Quantifying User Time Saved by Translation Automation

Another of the goals of applying MT technology is saving time to increase productivity. For this purpose, it is necessary to calculate how much time can be saved by introducing the new process with MT and post-editing. In order to do this, I need to compare how much time a translator needs in a standard process to translate a certain amount of text, and how much time is needed with MT and post-editing.

As for the standard times, it is difficult to calculate how many lines a translator can translate in one hour. Discussions in forums and some references (Allen, 2003) set the average translation speed at 2,000-3,000 words per day, which results in an average of 250-375 words per hour. If a line contains 8 to 9 words, this would mean that a translator could translate between 27 and 46 lines per hour, which results in 220 to 330 lines per day, that is, between 6 and 9 standard pages a day.

	Total price with MT and post-editing	Total Price without MT	Saving potential	Saving potential %
TIS RA 2002	15,818.10 €	19,136.05 €	3,317.95 €	20.98
TIS RA 2003	17,091.98 €	18,585.36 €	1,493.38 €	8.74
TIS RA 2004	18,085.73 €	25,081.02 €	6,995.29 €	38.68
TIS RA 2005	18,725.62 €	21,515.92 €	2,790.30 €	14.90
Average RA	17,430.36 €	21,079.59 €	3,649.23 €	20.94
SI 2002	8,345.84 €	10,254 €	1,908.43 €	22.87
SI 2003	24,113.81 €	28,740 €	4,625.93 €	19.18
SI 2004	29,455.56 €	35,581 €	6,125.21 €	20.79
SI 2005	76,520.62 €	91,446 €	14,925.17 €	19.50
Average SI	34,608.96 €	41,505.14 €	6,896.18 €	19.93
Average 2002	12,081.97 €	14,695.16 €	2,613.19 €	21.63
Average 2003	20,602.90 €	23,662.55 €	3,059.66 €	14.85
Average 2004	23,770.65 €	30,330.90 €	6,560.25 €	27.60
Average 2005	47,623.12 €	56,480.85 €	8,857.73 €	18.60
Average per year	26,019.66 €	31,292.37 €	5,272.71 €	20.26

Table 32: Overview of Costs

Regarding the new process, the following table outlines the different productivity values that have been derived from the experiments carried out for this research work.

Number of sentences for post-editing	125
Number of characters without blank spaces	9612
Number of words (approximated)	1842.50
Number of standard lines	160.20
Average time needed for revising and post-editing[120]	1.68 hours
Average number of norm lines revised and post-edited in 1 hour	95.35
Average number of norm pages revised and post-edited in 1 hour	2.72
Average number of norm pages revised and post-edited in 1 day	21.79

Table 33: Overview of Productivity

For the calculations I count the number of characters that the translator had to post-edit, and divide them by 55 (standard number of characters in a line including spaces), obtaining thus the standard number of lines. Besides, I calculate the average time needed to post-edit these lines, which allows us to calculate how many lines can be translated in one hour or in a day. Post-editors were instructed to do light post-editing, where intelligibility and accuracy were the main goals, and where no kinds of mistakes were permitted.

If I compare the 6 to 9 pages translated per a day in a traditional workflow with the 21.79 machine translated and post-edited pages, there is a clear gain in productivity and time saving. However, due to the pricing by line schema, these productivity gains cannot be directly converted into cost savings, since the number of lines to be revised remains the same.

The productivity increase can nevertheless be a great advantage for time-to-market requirements.

7.5.6 Determining Return on Investment (ROI)

According to Lawlor (2005), "the successful case for implementation must demonstrate that a positive ROI will be reached within a reasonable amount of time". In other words, the cumulative savings resulting from implementation must exceed the total of the upfront costs plus the ongoing running costs. The definition of a reasonable timeframe may differ from company to company. It is necessary to take into account that successful implementation requires a significant commitment of money and time.

The data used to calculate my ROI can be seen in Table 61 in Annex XII. The forecasted figures are marked in light yellow for a better overview. In order to better understand cash flow statements, I follow a simple plus/minus convention: all cash inflows are positive numbers, whereas all cash outflows are negative numbers.

First, I proceed to make a forecast of the translation costs for the next years in order to evaluate how much could be saved with the introduction of MT technology. Since the translation volume and, therefore the costs, vary from year to year due to new product releases, I calculate a geometrical increase in translation volume every two years, taking as a base the years where significant increases happened, i.e. new products were released. These data allow me to calculate the average of the two groups (years without and years with new product releases), which result in the maximal translation costs. The following charts show this prognosis:

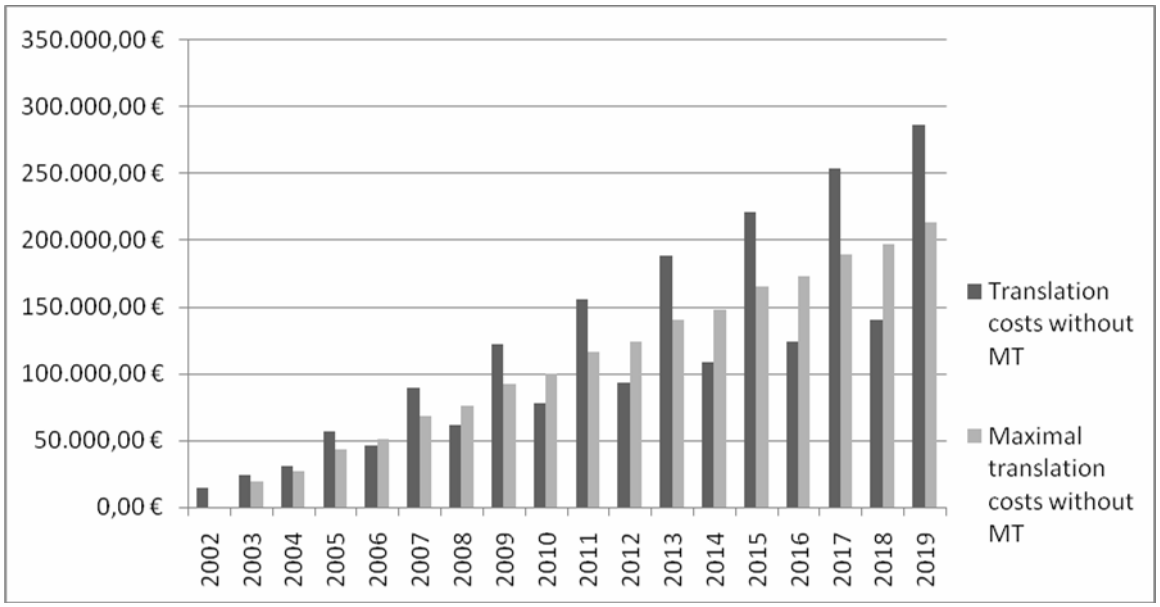

Figure 46: Maximal Translation Costs without MT and with/without product release

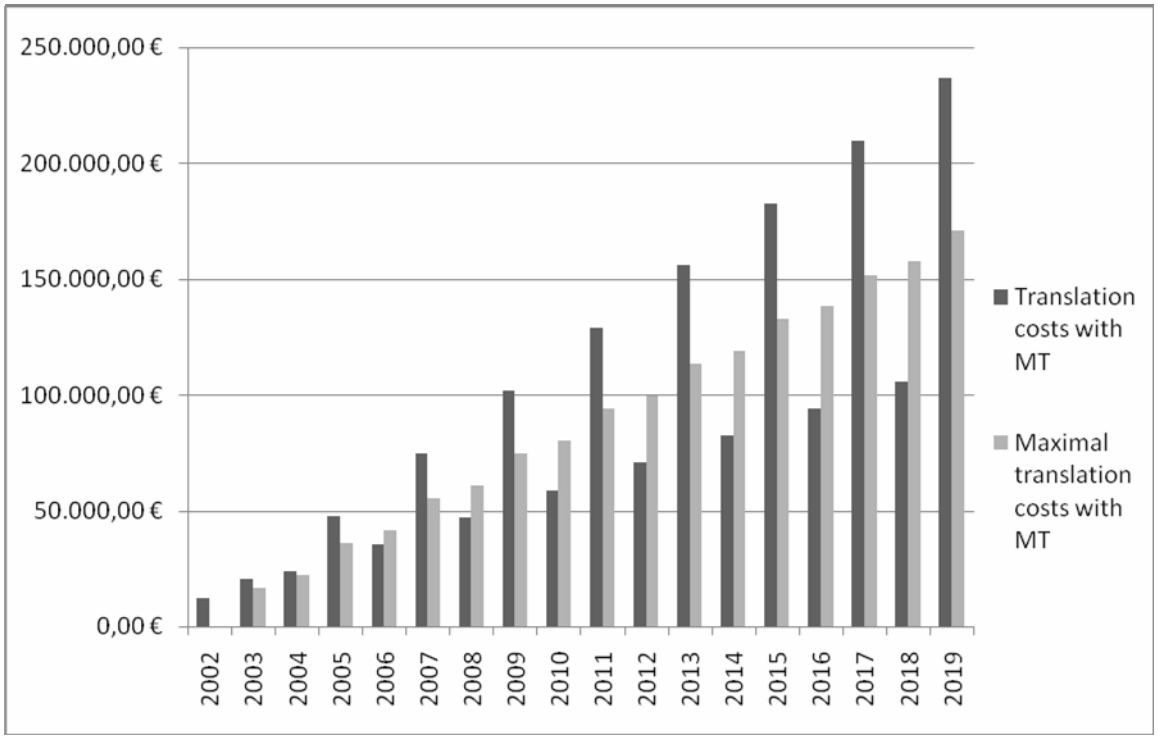

Figure 47: Maximal Translation Costs with MT and with/without product release

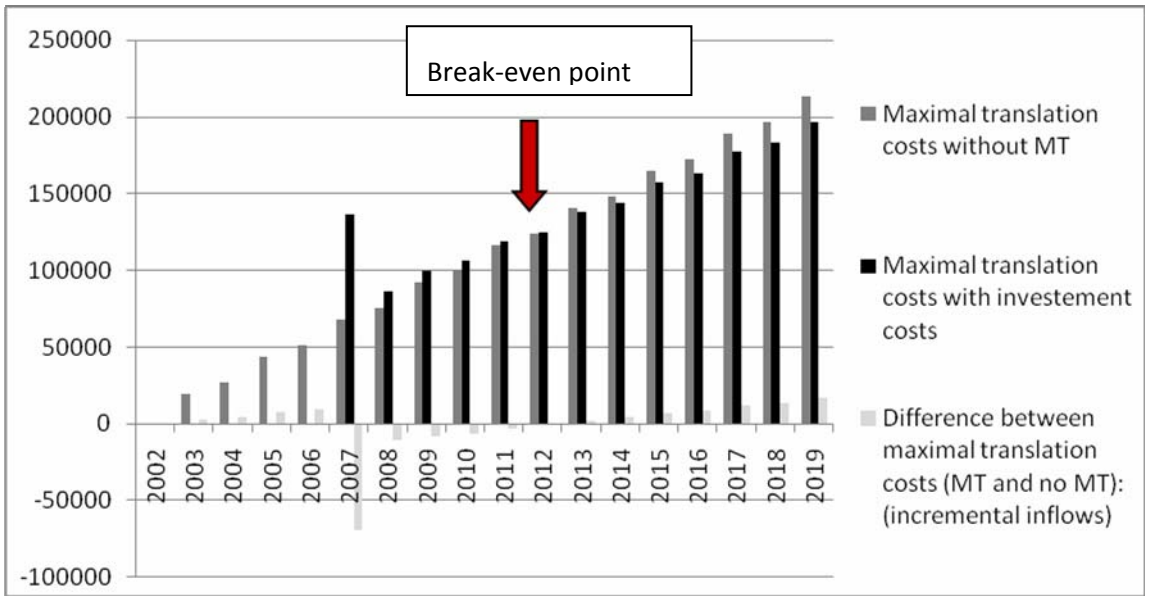

Figure 48: Break-even point

An ROI for MT is difficult to calculate, since there are no cash inflows as such, that is, no benefits or gains, but cumulative savings as a result of implementing the technology. Therefore, in order to calculate the ROI, I only need to take into account the costs and expenses incurred by the introduction of MT and the net cash flows, that is, the savings in the translation process.

I compare two scenarios: business as usual and my proposal (MT). In both scenarios there are outflows (costs), whereas there are inflows (savings) only in the MT scenario. The difference between both is the net cash flow. Furthermore, in order to be able to evaluate the investment of MT technology, an incremental cash flow statement is included. Each value in the incremental statement is the difference (delta) between a value in the proposed scenario and the corresponding "Business as Usual" value. A positive incremental cash flow means that the company's cash flow will increase with the acceptance of the project and is a good indication that an organization should spend some time and money investing in the project. In my case, a positive incremental cash flow is reached already after the second year of implementation. A graphical illustration of these data can be seen in the following charts:

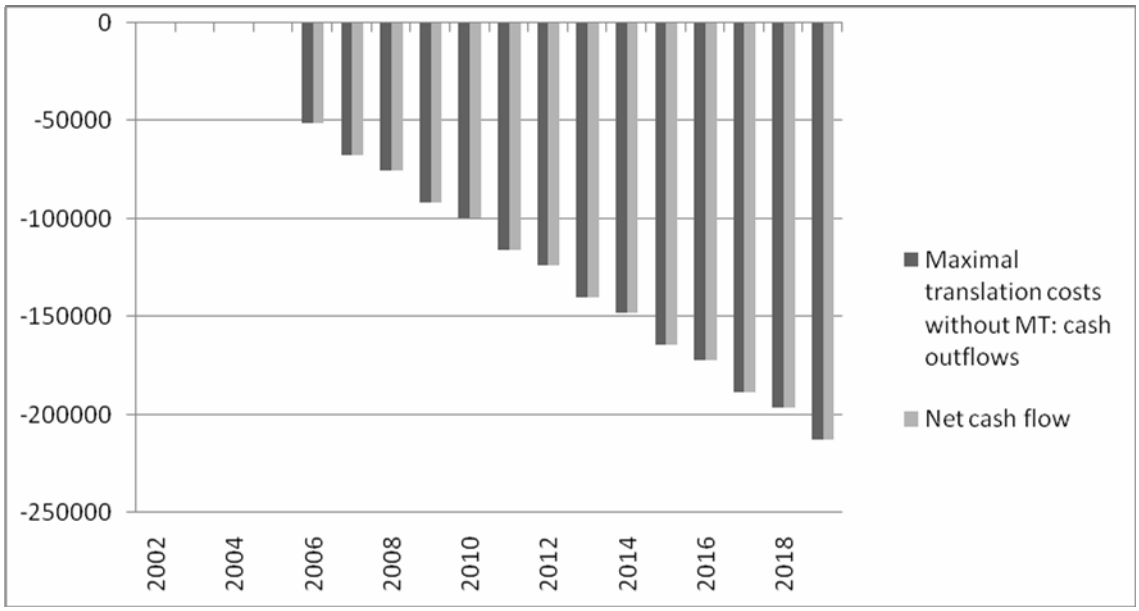

Figure 49: Business as Usual

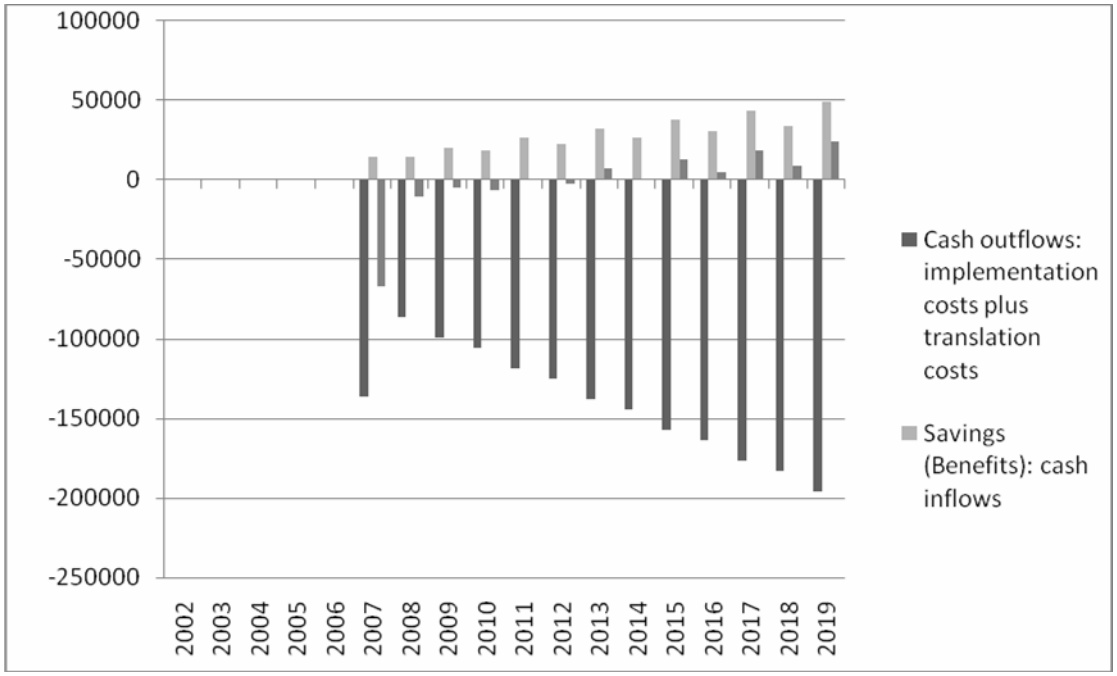

Figure 50: Proposal

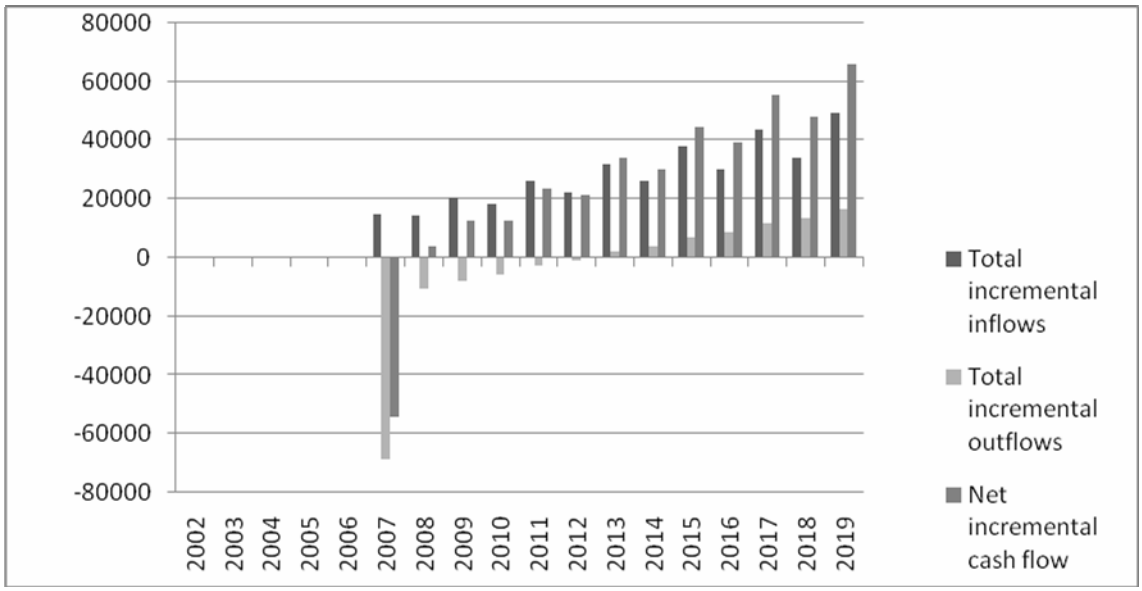

Figure 51: Incremental Cash Flows

I also calculate the cumulative cash flow. Cumulative cash flow is simply the sum of all cash flows up to one point in time. The cumulative figures show the total net flow up to the end of each year. The cumulative cash flow numbers and graphs show roughly when "payback" occurs, i.e. when the cumulative inflows are exactly equal to the cumulative outflows (Figure 52). In my case, this happens after 5 years, with cumulative inflows over 20,000 €. The payback period metric estimates that point in time precisely at 5.1 years.

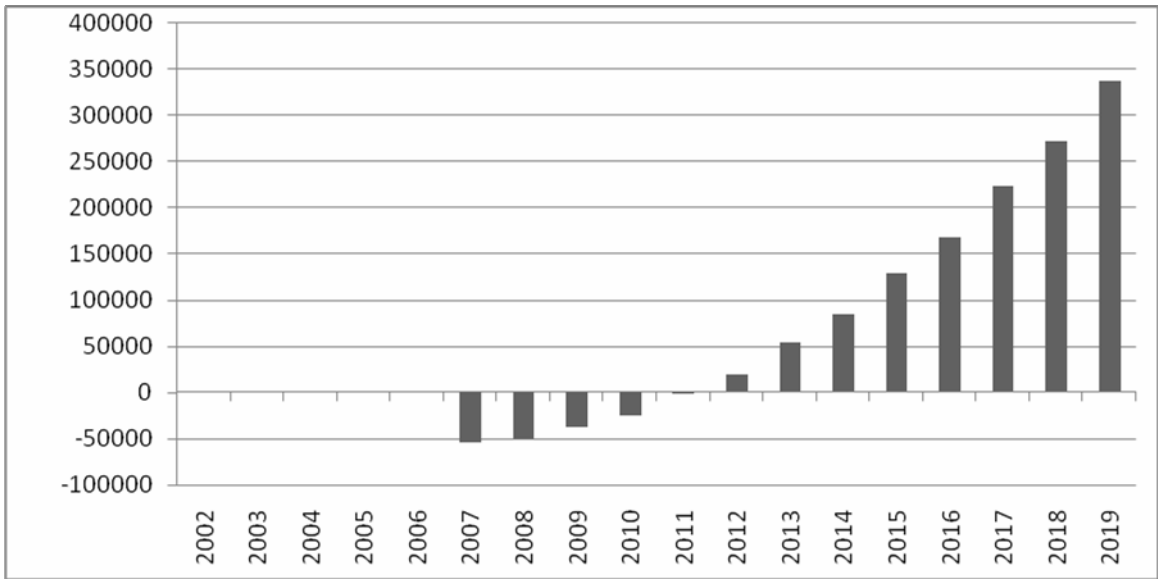

Figure 52: Cumulative Incremental Cashflows

Finally, I calculate the ROI ratio, which can be seen in Table 61 and Table 62 in Annex XII. Here I observe that, since payback occurs 5.1 years after initial outlay, the ROI ratio begins to be positive at that point, with a value of 20.76%. This means that every invested Euro has a payback of 1 Euro and 20.76 cents from this point on. An ROI of 100% will be reached after approximately 7.5 years, i.e. we will be saving exactly as much as it was invested.

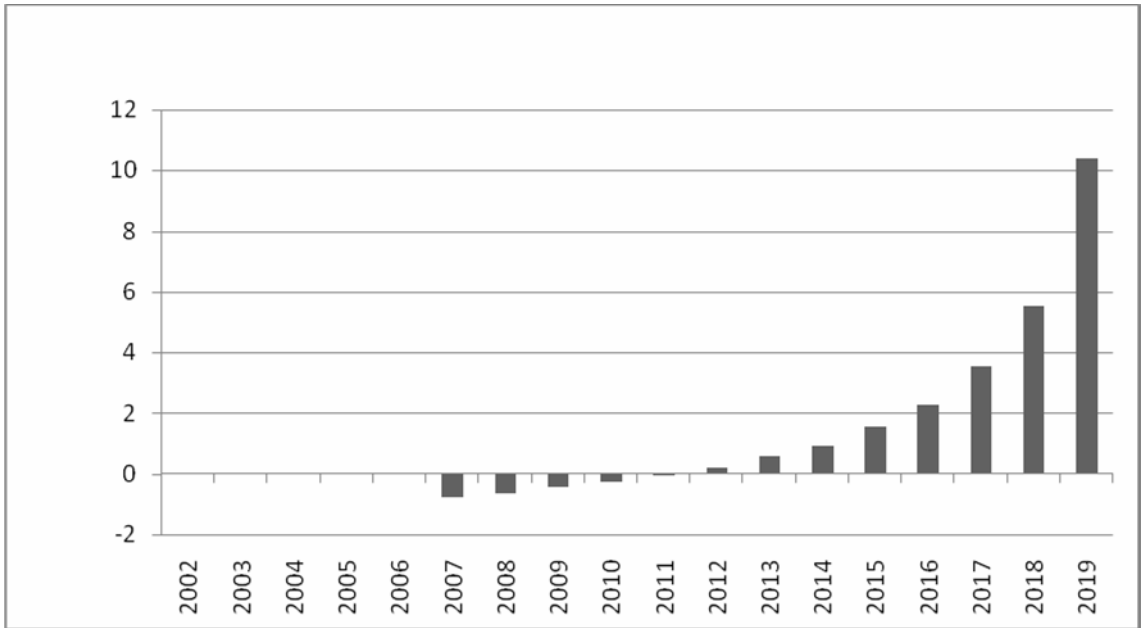

Figure 53: Return on Investment

7.6 Summary and final remarks

In this chapter I have analyzed the different factors involved in the calculation of an ROI for the implementation of MT technology in a translation process. After analyzing a credible scenario, and defining different workflow possibilities and scenarios where MT could be applied, I quantified the costs and calculated how much could be saved by deploying the new technology. With these data I was able to calculate an ROI ratio as well as a payback period.

With quality requirements close to those of human translation, an ROI of 20.76% is achieved after a payback period of 5.1 years. This means that the implementation of MT in this setting is a feasible alternative to human translation, though the goal of saving

costs is only achieved with a narrow margin. This is due, as I said before, to strict requirements on translation quality. An ROI without post-editing would reduce the payback period to 1 year. However, with both solutions, it is possible to attain significant productivity gains that would, in turn, improve the translation processes, especially for the Asian languages that translate directly from English.

Besides, this ROI is only based on the language pair German-English, and for the text types RA and SI. Additional language pairs and text types could see exponential increases in the ROI ratio.

All in all, the decision whether to deploy MT technology in translation processes or not will depend on the business goals that are intended to be attained by implementing MT. Although cost savings are not as significant as expected (yet still positive), productivity gains can lead to better processes, avoiding those costs incurred by delays in delivering translated content. Other scenarios, such as offering on-line MT to employees, can also improve communication within the company and guarantee that sensitive information is not sent to a public on-line translation service, but remains within the company.

8 CONCLUSIONS AND FUTURE WORK

I am turned into a sort of machine for observing facts and grinding out conclusions.
Charles Darwin

I started this research work due to the need to establish a methodology to assess the effectiveness and impact of controlled languages in the production of technical documentation and industrial contexts and, more specifically, in the creation of technical documentation for vehicles. In particular, I was interested in studying if automatic translatability and the quality of the target texts written using controlled language rules was improved or not with regards to texts not following these rules.

This main goal was further split into three hypotheses:

- First: texts written in accordance with the rules of a controlled language and assistance of a tool for applying it show improved intelligibility, comprehensibility and translatability.

- Second: Machine Translation (MT) is a technology that can represent an "objective" evaluator with regards to "translatability", since it is free of the variability of human translation.

- Finally, and as a collateral effect, MT represents a technology that can deal with the growing translation volumes of technical documentation. Using well-defined processes, this technology can bring about considerable savings both in time and in translation costs without neglecting quality.

In order to achieve this goal and confirm or reject the hypotheses posed before, I defined a set of specific goals:

- The development of a theoretical framework to define, describe and analyze the concept of controlled language, delimiting it from other similar concepts such as natural language or sublanguage. Furthermore, to study the application of these languages in industrial contexts and the tools used to automate their implementation, specifically MULTILINT / CLAT, which was evaluated in the empirical part of this work. In addition, to carry out a descriptive study of the problems and peculiarities of technical documentation translation as well as an analysis of different methods for the evaluation of language technology, in particular controlled language rules sets and MT.

- The design of a theoretically well-founded methodology in order to discern whether texts written and edited with the rules of a controlled language are more (automatically) translatable than others. This approach is innovative since, so far, most studies use human translators for assessment, without establishing clear differences between the rules that can improve human and automatic translatability. Furthermore, there are no studies with real texts used in the automotive field, and only a few studies in other areas of the industry that deal with this issue. This methodology was divided into three stages or phases, namely:

    1. Phase 1. In this phase, a microevaluation was performed to determine what resources were best suited to carry out the evaluation of phase 2. A text type as well as the most suitable MT system for our purposes were selected. For this evaluation I applied human and automatic evaluation methods in order to prove their reliability.

    2. Phase 2. In a second stage, I performed a macroevaluation with a corpus of texts in the source language (German) written without following the guidelines of a controlled language, and with the same texts rewritten in

accordance with the rules of MULTITERM/CLAT. These texts were then translated using the MT system selected in phase 1. The evaluation of the quality of both corpora allowed me to draw conclusions about the impact of the implementation of a controlled language in the source and the target text.

3. Phase 3. Finally I conducted an economic study to analyse the feasibility and the return on investment of implementing a process with a controlled language and MT in an industrial context, taking into account the adaptation of the processes and the characteristics of each technology.

The application of this methodology in three stages has revealed which resources are best suited to carry out this research, as well as what effect controlled language rules implemented by the tool MULTILINT / CLAT have. Specifically, I was able to figure out what kind of rules have a greater effect in the target text, though results are not conclusive due to the subjectivity of the human evaluation in the second phase and the differences between raters, an aspect that should be considered in further studies. Furthermore, the economic and process analysis has revealed that, in order to apply this type of technology, a detailed study of all factors and the definition of an optimal process is required. Furthermore, this does not necessarily imply a reduction in costs and, in any case, not in the short term, since the implementation of this technology is coupled with expenses and the restructuring of processes. Moreover, results highly depend on the characteristics of the text, the volume, quality requirements and the target languages. However, other soft benefits can be obtained, such as a reduction in complexity of the translation process, shortening of the process or an improvement in the consistency of documents with more control of terminology and language resources.

In general, it may be concluded that the implementation of a controlled language is perceived as positive, especially for the source language, as shown by the data presented in chapter 6. However, it is not infallible, since some rules may not cause improvements and may even lead to a deterioration in the quality of text. This becomes even more clear in the case of translatability, although this cannot be attributed solely to the effect

of the controlled language. MT itself, an imperfect technology, clearly contributes to the poor quality. Therefore, the improvements brought about with regards to machine translatability cannot be proven conclusively. An alternative to try to solve this difficulty would be to implement a statistical MT engine trained and adapted exclusively to automotive texts. In addition, human translation could also be implemented, although in this case the subjective factor of the evaluation would increase and the evaluation would not be effective as recommended by White & Taylor (1998), who argue that an ideal evaluation method for MT "should be readily reusable, with a minimum of preparation and participation of raters or subjects".

Among the rules that have a positive impact on both the source language and target language, there are rules regarding spelling and unfamiliar words, as well as rules concerning the use of approved terminology and the avoidance of complex sentence structures. This confirms the approach defended by Reuther (2003) and other authors who have studied various aspects of translatability (see Annex 2). Unfortunately we were unable to draw conclusive results regarding the rules that produce a deterioration in quality, since the results are highly segmented and cannot be assigned to a single rule or a set of rules.

With respect to the feasibility and economic analysis, in order to apply these technologies effectively, two premises must be made:

- Optimal processes are designed to ensure quality as well as time and cost savings;

- Volumes of translation are large, and translation is done into several languages, if possible. In this way, the investment in resources will be recovered more rapidly.

The setting out of optimal processes with quality assurance will require specialized reviewers to correct the output of MT. These must be trained first and must be offered decent rates to perform quality work. Therefore, a positive return on investment will only occur in the medium to long term. In our case, there is a return of 20.76% after five

years. This is a fairly long term due to the high level of quality required. A process not including post-editing would get a positive return in less than a year, but this was not the aim of this study.

The evaluation of language technologies is a complex issue that requires a thorough analysis of the context to apply the most appropriate methodology, including the time constraints and the economic conditions of a project. Therefore, when an evaluation scenario is proposed, it is necessary to define with precision what the goals are and what the context is in order to make the evaluation as optimal and reusable as possible with regards to the selected resources and the results obtained.

Our study has followed these guidelines to establish limits for the evaluation and define in detail the context in which it took place. A more comprehensive analysis would include more target languages and text types, as well as statistical MT systems trained specifically in order to obtain more insightful results. In this way we could get a better return on investment and would be able to compare whether controlled language rules have the same effects in different target languages. Furthermore, the inclusion of new text types would allow us to know whether the rules of a controlled language have the same effects in texts from contexts other than that to technical documentation.

Further research is also needed with regards to new standards and metrics that allow for objective and efficient evaluations, optimizing resources and allowing for correlation with other metrics and measurements, as well as the establishment of relationships between different aspects of an evaluation, such as for example, in our case, between the comprehensibility and the translatability of texts or the text quality of the source and the target text. In this sense, it would be beneficial and necessary to develop applications for the evaluation that facilitated easy corpus processing, the selection of metrics, and the collection and analysis of results, in the wake of the tool developed by Nießen et al. (2000) or Language Studio, a translation tool developed by the company Asia Online[121]. In this way, the evaluation of language technologies would become accessible to a larger number of potential users and would allow for improvement of the evaluation process. There are, however, not such tools for the evaluation of controlled

languages and there is a need for standard metrics, largely due to the particularities of each language and the difficulty of accessing information and rule sets. Therefore, a goal for the future would be the development of such tools to facilitate evaluation, allowing for example the compilation of corpora or the creation of tests to evaluate the effects and alleged improvements that this technology brings.

9 REFERENCES

Adams, A. H., Austin, G. W., & Taylor, M. (1999). Developing a Resource for Multinational Writing at Xerox Corporation. *Technical Communication*, *46*(2), 249-254.

Adriaens, G. (1994). Simplified English Grammar and Style Correction in an MT Framework: The LRC SECC Project (pp. 78-88). London.

Adriaens, G. (1996). SECC: Using Text Structure Information to Improve Checker Quality and Coverage (pp. 226-232). Leuven, Belgium: Katholieke Universiteit Leuven Centre for Computational Linguistics.

Adriaens, G., & Macken, L. (1995). Technological evaluation of a controlled language application: precision, recall and convergence tests for SECC. Leuven.

Adriaens, G., & Scheurs, D. (1992). From COGRAM to ALCOGRAM: Toward a Controlled English Grammar Checker (pp. 595-600). Nantes.

AECMA. (2004). AECMA Simplified English: A GUIDE FOR THE PREPARATION OF AIRCRAFT MAINTENANCE DOCUMENTATION IN THE INTERNATIONAL AEROSPACE MAINTENANCE LANGUAGE (No. AECMA Document: PSC-85-16598, Issue 2). Austria.

Agarwal, A., & Lavie, A. (2008). Meteor, M-BLEU and M-TER: Evaluation metrics for high-correlation with human rankings of machine translation output (pp. 115-118.). Ohio, USA.

Ahrenberg, L., & Merkel, M. (2000). Correspondence Measures for MT Evaluation (pp. 41-46). Athens.

Aikawa, T., Schwartz, L., Corston-Oliver, M., King, R., & Lozano, C. (2007). Impact of controlled language on translation quality and post-editing in a statistical machine translation environment (pp. 1-7). Copenhagen.

Akiba, Y., Imamura, K., & Sumita, E. (2001). Using Multiple Edit Distances to Automatically Rank Machine Translation Output (pp. 15-20). Presented at the Proceedings of MT Summit VIII, Santiago de Compostela.

Akiba, Y., Nakaiwa, H., Sumita, E., Yamamoto, S., & Okuno, H. G. (2003). Experimental Comparison of MT Evaluation Methods: RED vs. BLEU (pp. 1-8). Presented at the Proceedings of Machine Translation Summit IX, New Orleans, USA.

Albrecht, J. S., & Hwa, R. (2007a). Regression for Sentence-Level MT Evaluation with Pseudo References (pp. 296–303). Prague, Czech Republic.

Albrecht, J. S., & Hwa, R. (2007b). A Re-examination of Machine Learning Approaches for Sentence-Level MT Evaluation (pp. 880–887). Prague, Czech Republic.

Allen, J. (1999a). Different Types of Controlled Languages. *Technical Communicators'(TC-) Forum*, *1*.

Allen, J. (1999b). Comment on a Message by Amo Fuchs to the tcf-gen Mailing-List: CL 18). *Technical Communicators'(TC-) Forum*, *3*.

Allen, J. (1999c). Adapting the concept of "Translation Memory" to "Authoring memory" for a Controlled Language writing environment. London.

Allen, J. (2000). Taking on the Critics: Giving the Machine Equal Time - An MT expert takes on one of machine translation's most vocal critics. *Special issue on Machine Translation in Language International magazine*, *12*(3), 23-25,44-45.

Allen, J. (2003). Controlled Langauge for Authoring and Translation. In H. Somers (Ed.), *Computers and Translation: A Handbook*, Benjamins Translation Library (Vol. 35, pp. 125-146). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Almqvist, I., & Hein, A. S. (1996). Defining ScaniaSwedish. A Controlled Language for Truck Maintenance (pp. 159-165). Leuven, Belgium: Katholieke Universiteit Leuven Centre for Computational Linguistics.

Alonso Cortés, Á. (1994). Sublenguajes: notas sobre el lenguaje de la física. *DICENDA. Cuadernos de Filología Hispánica*, *12*, 243-253. 4

Alonso, J. A. (2005). Machine translation for Catalan <->Spanish: the real case for productive MT (pp. 23-26). Budapest.

ALPAC. (1966). *Language and Machines — Computers in Translation and Linguistics* (A report by the Automatic Language Processing Advisory Committee, Division of Behavioral Sciences). Washington, DC: National Academy of Sciences, National Research Council.

Altwarg, R. (2000, November 20). Controlled Languages: An Introduction. Retrieved from <http://www.shlrc.mq.edu.au/masters/students/raltwarg/clindex.htm#About%20this>

Amigó, E., Giménez Linares, J., Gonzalo, J., & Márquez, L. (2006). MT Evaluation: Human-like vs. Human Acceptable. Sydney, Australia.

AMTA. (1994, January 7). Nes from Smart Communications, Inc. MT News International: Newsletter of the International Association for Machine Translation, 7, 11.

Andersen, P. (1994). ClearCheck demonstration. Columbia, Maryland, USA.

Aranberri-Monasterio, N., & O'Brien, S. (2009). Evaluating RBMT output for -ing forms: A study of four target languages. (W. Daelemans & V. Hoste, Eds.)*Linguistica Antverpiensia New Series*, Themes in Transalation Studies, *8*, 105-122.

Arnold, D., Balkan, L., Meijer, S., Humphreys, R. L., & Sadler, L. (1994). *Machine Translation. An Introductory Guide*. Cambridge, Massachussets: Blackwell Publishers. Retrieved from <http://clwww.essex.ac.uk/~doug/book/book.html>

Arnold, D., Sadler, L., & Humphreys, R. L. (1993). Evaluation: an assessment. *Machine Translation*, *8*(12), 1-24.

Aymerich, J. (2005). Using Machine Translation for fast, inexpensive, and accurate health information assimilation and dissemination: Experiences at the Pan American Health Organization. Presented at the 9th World Congress on Health Information and Libraries, Salvador de Bahia.

Babych, B., Elliott, D., & Hartley, A. (2004). Extending MT evaluation tools with translation complexity metrics. University of Geneva, Switzerland.

Babych, B., & Hartley, A. (2003). Improving machine translation quality with automatic named entity recognition (pp. 1-8). Budapest, Hungary.

Babych, B., & Hartley, A. (2004a). Weighted N-gram model for evaluating Machine Translation output (pp. 15-22). Unviersity of Birmingham.

Babych, B., & Hartley, A. (2004b). Modelling legitimate translation variation for automatic evaluation of MT quality (pp. 833-836). Barcelona, Spain.

Babych, B., & Hartley, A. (2004c). Extending the BLEU MT evaluation method with frequency weightings (pp. 621-628). Barcelona, Spain.

Babych, B., & Hartley, A. (2008). Sensitivity of automated MT evaluation metrics on higher quality MT output: BLEU vs task-based evaluation methods. Marrakech, Morocco.

Babych, B., & Hartley, A. (2009). Automated error analysis for multiword expressions: Using BLEU-type scores for automatic discovery of potential translation errors. (W. Daelemans & V. Hoste, Eds.)*Linguistica Antverpiensia New Series*, Themes in Transalation Studies, *8*, 81-104.

Babych, B., Hartley, A., & Atwell, E. (2003). Statistical modelling of MT output corpora for Information Extraction (pp. 62-70). Lancaster, UK.

Babych, B., Hartley, A., & Elliott, D. (2004). Calibrating resource-light automatic MT evaluation: a cheap approach to ranking MT systems by the usability of their output (pp. 2031-2034). Lisbon, Portugal.

Baker, K., Franz, A. M., Jordan, P. W., Mitamura, T., & Nyberg, E. (1994). Coping with ambiguity in a large-scale machine translation system (pp. 90-94). Kyoto, Japan.

Balkan, L. (1994). Test suites: some issues in their use and design. Cranfield University, England: Cranfield University Press.

Balkan, L., Arnold, D., & Fouvry, F. (1995). Test Suites for NLP. Dublin City University.

Balkan, L., Arnold, D., & Meijer, S. (1994). Test Suites for Natural Language Processing. London.

Balkan, L., & Netter, K. (1994). Test Suites for NLP. Groningen.

Balkan, L., Netter, K., Arnold, D., & Meijer, S. (1994). TSNLP. Test Suites for Natural Language Processing. Paris.

Banerjee, S., & Lavie, A. (2005). METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgements (pp. 65-72). Ann Arbor.

Barrett, L. (2003). Considerations of methodology and human factors in rating a suite of translated sentences (pp. 13-19). New Orleans, USA.

Barthe, Kathleen. (1998). GIFAS Rationalised French. Designing one Controlled Language to Match Another (pp. 98-101). Pittsburgh (Pennsylvania): Language Technologies Institute, Carnegie Mellon University.

Barthe, Kathleen, Bès, G. ., Escande, J., Pinna, D., & Rodier, E. (1998). Issues related to realistic evaluation of controlled language checkers (pp. 134-144). Pittsburgh (Pennsylvania): Language Technologies Institute, Carnegie Mellon University.

Barthe, Kathy, Juaneda, C., Leseigneur, D., Loquet, J.-C., Morin, C., Escande, J., & Vayrette, A. (1999). GIFAS Rationalized French: A Controlled Language for Aerospace Documentation in French. *Technical Communication*, *46*(2), 220-229.

Bernardi, U., Bocsak, A., & Porsiel, J. (2005). Are We Making Ourselves Clear? Terminology Management and Machine Translation at Volkswagen. Budapest, Hungary.

Berns, K., & Ramírez Polo, L. (2008). Machine Translation: is it worth the trouble? *MultiLingual*, *19*(3), 44-46.

Bernth, A. (1997). EasyEnglish: A Tool for Improving Document Quality. Washington, DC: Association for Computational Linguistics.

Bernth, A. (1998). EasyEnglish: Preprocessing for MT (pp. 30-41). Pittsburg, Pennsylvania: Language Technologies Institute, Carnegie Mellon University.

Bernth, A. (1999). A Confidence Index for Machine Translation (pp. 120-127). Chester College, England.

Bernth, A. (2006). EasyEnglishAnalyzer: Taking Controlled Language from Sentence to Discourse Level. Cambridge, Massachussets.

Bernth, A., & Gdaniec, C. (2001). MTranslatability. *Machine Translation*, *16*(3), 175-218.

Bernth, A., & MCCord, M. (2000). The Effect of Source Analysis on Translation Confidence (pp. 89-99). Presented at the Proceedings of the 4th Conference of the Association for Machine Transation in the Americas-AMTA: Envisioning Machine Translation in the Information Future, Cuernavaca: Springer.

Betts, R. (2003, May 15). *Challenges in Cross-Cultural Communication*. PPT Presentation presented at the Joint Conference of the 8th International Workshop of the European Association for Machine Translation and the 4th Controlled Language Applications Workshop, Dublin.

Birch, A., Osborne, M., & Blunsom, P. (2010). Metrics for MT evaluation: evaluating reordering. *Machine Translation*, *24*(1), 15-26.

Blatz, J., Fitzgerald, E., Foster, G., Gandrabur, S., Goutte, C., Kulesza, A., Sanchis, A., et al. (2004). Cofidence Estimation for Machine Translation (pp. 315-321). Geneva, Switzerland.

Bock, G. (1993). Ansätze zur Verbesserung von Technikdokumentation. Eine Analyse von Hilfsmitteln für Technikautoren in der Bundesrepublik Deutschland. Technical Writing. Frankfurt am Main: Peter Lang Verlagsgruppe.

Boehme, D. U., & Svetova, S. (2001). An Integrated Solution: Applying PROMT Machine Translation Technology, Terminology Mining, And TRADOS's TWB Translation Memory to SAP Content Translation. Santiago de Compostela, Spain.

Bohan, N., Breidt, E., & Volk, M. (2000). Evaluating Translation Quality as Input to Product Development. Athens.

Bond, F. (2002). Toward a Science of Machine Translation. Keihanna, Japan.

Bouillon, P., Rayner, M., Chatzichrisafis, N., Hockey, B. A., Santaholma, M., Starlander, M., Nakao, Y., et al. (2005). A generic multi-lingual open source platform for limited-domain medical speech translation. (pp. 50-58). Budapest.

Bowker, L. (2002). *Computer-Aided Translation Technology: A Practical Introduction.* Didactics of Translation (Vols. 1-5, Vol. 5). Otawa: University of Otawa Press.

Bowker, L. (2009). Can Machine Translation meet the needs of the official language minority communities in Canada? A recipient evaluation. (V. Hoste & W. Daelemans, Eds.)*Linguistica Antverpiensia New Series*, Themes in Transalation Studies, *8*, 123-158.

Bowne Global Solutions (último). (2002). *Kontrollierte Sprache - Nutzen und Vorteile* (p. 6).

Brockmann, D. (1993). Was kann LOGOS? Linguistische Bewertung eines kommerziellen MÜ-Systems unter Berücksichtigung weiterer markt- und forschungsorientierter Systeme. Untersuchte Sprachrichtung: Deutsch-Englisch (Diplomarbeit (Master Thesis)). Universität des Saarlandes.

Brockmann, D. (1997, December). Controlled Language and Translation Memory Technology: a Perfect Match to Save Translation Cost. *Technical Communicators'(TC-) Forum*, *4*, 10-11.

Bruckner, C., & Plitt, M. (2001). Evaluating the Operational Benefit of Using Machine Translation Output as Translation Memory Input (pp. 18-22). Santiago de Compostela, Spain.

Brundage, J. A. (2001). Machine Translation. Evolution not Revolution. Santiago de Compostela, Spain.

Buchmann, B., Warwick, S., & Shane, P. (1984). Design of a Machine Translation System for a Sublangauge (pp. 334-337). Presented at the 10th International Conference on Computational Linguistics and 22nd Annual Meeting of the Association for Computational Linguistics, California: Stanford University.

Bühler, K. (1969). *Die Axiomatik der Sprachwissenschaften.* (E. Ströker, Ed.). Frankfurt am Main: Vittorio Klostermann.

Cadwell, P. (2008). Readahility: c. Localisation Focus: The International Journal of Localisation, 7(1), 34-45.

Caeyers, H. (1997a). Presentation of LANT technology (pp. 253-254). San Diego (California).

Caeyers, H. (1997b, May 21). *Machine Translation and controlled English*. Copenhagen. Retrieved from <http://www.mt-archive.info/MTS-1997-Caeyers.pdf>

Callison-Burch, C., & Flournoy, R. S. (2001). A Program for Automatically Selecting the Best Output from Multiple Machine Translation Engines (pp. 63-66). Santiago de Compostela, Spain.

Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., & Schroeder, J. (2007). (Meta-) Evaluation of Machine Translation. Prague, Czech Republic. Retrieved from <http://www.statmt.org/wmt07/>

Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., & Schroeder, J. (2008). Further Meta-Evaluation of Machine Translation. Prague, Czech Republic. Retrieved from <http://www.statmt.org/wmt07/>

Callison-Burch, C., Osborne, M., & Koehn, P. (2006). Re-evaluating the Role of BLEU in Machine Translation Research. Trento, Italy.

Canadian International Trade Tribunal. (2002). *Procurement John Chandioux Experts-Conseils Inc.* (Vol. File Nos. PR–2001-029 and PR-2001-032). Retrieved from <http://www.citt.gc.ca/procure/determin/pr2b029_e.asp#P90_10741>

Carbonell, J., & Wilks, Y. (1991). Machine Translation: An In-Depth Tutorial. *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*. Berkley, California: University of California.

Carl, M. (1999). Inducing Translation Templates for Exampled-Based Machine Translation (pp. 250-258). Presented at the Proceedings of the Machine Translation Summit VII, Singapore: Asia-Pacific Association for Machine Translation (AAMT), National University of Singapore.

Carl, M. (2003). Data-Assisted Controlled Translation (pp. 16-24).

Carl, M., Haller, J., Horschmann, C., Maas, D., & Schütz, J. (2002). The TETRIS Terminology Tool. *Traitement Automatique des Langues*, *43*(1), 73-102.

Carl, M., Hernandez, M., Preuß, S., & Enguehard, C. (2004). English Terminology in CLAT. Lisbonne.

Carl, M., Iomdin, L. L., Pease, C., & Streiter, O. (2000). Towards a dynamic linkage of example-based and rule-based machine translation. *Machine Translation*, *15*(3), 223-257.

Carl, M., & Langlais, P. (2003). Tuning General Translation Knowledge to a Sublanguage (pp. 25-34). Dublin City University.

Carl, M., Schmidt-Wigger, A., & Hong, M. (1997). KURD: A Formalism for Shallow Post Morphological Processing. Phuket, Thailand.

Carl, M., Way, A., & Schäler, R. (2002). Toward a Hybrid Integrated Translation Environment (pp. 11-21). London, UK: Springer-Verlag.

Castilla, A., Babic, A., & Furuie, S. (2005). Machine Translation on the Medical Domain: The Role of BLEU/NIST and METEOR in a Controlled Vocabulary Setting (pp. 47-54). Phuket.

Chan, Y. S., & Tou Ng, H. (2008). MAXSIM: a maximum similarity metric for machine translation evaluation (pp. 55-62). Ohio, USA.

Chandioux, J., & Grimaila, A. (1996). Specialized Machine Translation. Montreal, Quebec.

Chatzichrisafis, N., Bouillon, P., Rayner, M., Santaholma, M., Starlander, M., & Hockey, B. A. (2006). Evaluating task performance for a unidirectional controlled language medical speech translation system (pp. 9-16). New York.

Chervak, S., Drury, C. G., & Ouellette, J. (1996). Field Evaluation of Simplified English for Aircraft Workcards. Alexandria (Virginia).

Chevalier, M., Dansereau, J., & Poulin, G. (1978). *TAUM-METEO: description du système* (Technical report) (p. 113). Université de Montreal.

Chevrak, S. G., & Drury, C. G. (2003). Effects of Job Instruction on Maintenance Task Performance. *Occupational Ergonomics*, *3*(2), 121-131.

Chung-ling, S. (2007). Teaching Translation of Text Types with MT Error Analysis and Post-MT Editing. *Translation Journal*. Retrieved from <http://accurapid.com/journal/>

Church, K. W., & Hovy, E. H. (1993). Good Applications for Crummy Machine Translation. *Machine Translation*, *8*, 239-258.

Ciarlone, L., Kadie, K., & Laplante, M. (2008). Multilingual Communications as a Business Imperative: Why Organizations Need to Optimize the Global Content Value Chain. The Gilbane Group.

Ciravegna, F. (1995). Understanding messages in a diagnostic domain. *Information Processing and Management*, *31*(5), 687-701.

Civil Aviation Authority. (2006). *Radiotelephony Manual* (Manual No. CAP 413 (16th edition)). West Sussex.

Clark, P., Harrison, P., Jenkins, T., Thompson, J., & Wojcik, R. (2005). Acquiring and Using World Knowledge using a Restricted Subset of English.

Clark, P., Harrison, P., Thompson, J., Wojcik, R., Jenkins, T., & Israel, D. (2006). *Reading to Learn: Final Report*. Retrieved from <http://www.cs.utexas.edu/users/pclark/rtol/final-report.doc>

Climent, S., Moré, J., & Oliver, A. (2003a). Building an Environment for unsupervised automatic e-mail translation (pp. 45-53). Dublin.

Climent, S., Moré, J., & Oliver, A. (2003b, May 15). *Building an Environment for unsupervised automatic e-mail translation*. PPT Presentation presented at the Joint Conference of the 8th International Workshop of the European Association

for Machine Translation and the 4th Controlled Language Applications Workshop, Dublin.

Correa, N. (2003). A fine-grained evaluation Framework for machine translation system development. New Orleans, USA.

Corston-Oliver, S., Gamon, M., & Brockett, C. (2001). A machine learning approach to the automatic evaluation of machine translation (pp. 140-147). Toulouse, France.

Coughlin, D. (2003). Correlating Automated and Human Assessments of Machine Translation Quality (pp. 63-70). New Orleans, USA.

Coulombe, C., Doll, F., & Drouin, P. (2005). Intégration d'un analyseur syntaxique à large couverture dans un outil de langage contrôlé en français. *Lingvisticæ Investigationes*, *28*(1), 19–36.

Cregan, A., Schwitter, R., & Meyer, T. (2007). Sydney OWL Syntax - towards a Controlled Natural Language Syntax for OWL 1.1. Innsbruck.

Cremers, L. (2003, May 15). *Controlled Language in an automated localisation environment*. PPT Presentation presented at the Joint Conference of the 8th International Workshop of the European Association for Machine Translation and the 4th Controlled Language Applications Workshop, Dublin.

Cremers, L. (2008). Putting MT to work. *MultiLingual*, *19*(3), 38-40.

Crystal, D. (1987). *The Cambridge Encyclopedia of Language*. Cambridge: Cambridge University Press.

Cucchiarini, C. (2002). Euromap HLT Case Study: How HLT Applications Can Lead to Higher Quality Translations at Lower Costs: The Experience of Océ Technologies.

Culy, C., & Riehemann, S. (2003). The limits of N-Gram Translation Evaluation Metrics. New Orleans, USA.

Dabbadie, M., Hartley, A., King, M., Miller, K. J., Mustafa El Hadi, W., Popescu-Belis, A., Reeder, F., et al. (2002). A Hands-On Study of the Reliability and Coherence of Evaluation Metrics (pp. 8-16). Las Palmas de Gran Canaria, Spain.

Dabbadie, M., Mustafa El Hadi, W., & Timimi, I. (2004). CESTA: The European MT Evaluation Campaign. *Multilingual Computing & Technology*, *15*(5), 10-12.

Daelemans, W., & Hoste, V. (Eds.). (2009). *Evaluation of Translation Technology*. Linguistica Antverpiensia. New Series - Themes in Translation Studies (Vol. 8). Artensis Hogeschool Antwerpen.

Danlos, L., Lamapalme, G., & Lux, V. (2002). Generating a Controlled Language. Retrieved from <http://en.scientificcommons.org/574368>

Darwin, M. (1999). Trial and Error: An Evaluation Project with Japanese (pp. 77-82). Singapore: Asia-Pacific Association for Machine Translation (AAMT), National University of Singapore.

Dauphin, E., & Lux, V. (1996). Corpus-based annotated test set for Machine Translation evaluation by a Industrial User (Vol. 1).

Davis, B., Handschuh, S., Cunningham, H., & Tablan, V. (2006). Further Use of Controlled Natural Language for Semantic Annotation of Wikis. Athens, Georgia, USA.

de Gispert Ramis, A. (2006, October). *Introducing Linguistic Knowledge into Statistical Machine Translation* (PhD Thesis). Universitat Politècnica de Catalunya.

de Koning, M. (1996). Bringing Controlled Language Support to the Desktop (pp. 11-20). Vienna.

de Pedro, R. (1999). The Translatability of Texts: A Historial Overview. *Meta*, *44*(4).

de Pedro, R. (2001). Translatability and the Limits of Communication. *Critical Studies*, *VII*(16: Language - Meaning – Social Construction), 107-122.

Decrozant, L., & Voss, C. R. (1999). Dual Use of Linguistic Resources: Evaluation of MT Systems and Language Learners (pp. 32-38). Maryland.

Denkowski, M., & Lavie, A. (2010). Choosing the Right Evaluation for Machine Translation: an Examination of Annotator and Automatic Metric Performance on Human Judgment Tasks. *Proceedings of the Ninth Conference of the Association for Machine Translation in the Americas*. Denver, Colorado.

Dervišević, D., & Steensland, H. (2005, September 23). *Controlled Languages in Software User Documentation* (Master Thesis). Linköpings Universitet. Retrieved from <http://www.ep.liu.se/undergraduate/abstract.xsql?dbid=4637>

Dillinger, M. (2001). Dictionary Development Workflow for MT: Design and Management (pp. 83-87). Santiago de Compostela, Spain.

Doddington, G. (2002). Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. In M. Marcus (Ed.), (pp. 138-145). San Diego, California.

Doheny-Farina, S. (1992). *Rethoric, Innovation, Technology. Case Studies of Technical Communication in Technology Transfers*. MIT Press Series in Technical Communication. Cambridge, Massachussets: MIT Press. Retrieved from <http://books.google.es/books?id=N1g00v_Zld8C&pg=PP1&dq=Rhetoric,+Innovation,+Technology+Doheny-Farina&sig=ACfU3U2fU1LXlS5DKKpsrsNR_7SdX3nUPA>

Doherty, S., O'Brien, S., & Carl, M. (2010). Eye tracking as an MT evaluation technique. *Machine Translation*, *24*(1), 1-13.

Douglas, J., & Rusk, G. M. (2000). Toward a scoring function for quality-driven machine translation (pp. 376-382). Saarbrücken, Luxemburg, Nancy.

Douglas, S., & Hurst, M. (1996). Controlled Language Support for Perkins Approved Clear English (PACE). *Proceedings of the First International Workshop on Controlled Language Applications* (pp. 93-105). Presented at the Proceedings of the First International Workshop on Controlled Language Applications, Leuven, Belgium.

Drury, C. G., & Ma, J. (2004). Experiments on Language Errors in Aviation Maintenance. *Human Factors and Ergonomics Society Annual Meeting Proceedings*, *48*, 118-122.

DuBay, W. (2004). The Principles of Readability. Impact Information. Retrieved from <http://www.eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/1b/bf/46.pdf>

EAGLES. (1996). *EAGLES. Editor's Introduction*. Retrieved from <http://www.ilc.cnr.it/EAGLES96/edintro/edintro.html>

Eck, M., Vogel, S., & Waibel, A. (2006). A Flexible Online Server for Machine Translation Evaluation (pp. 89-94). Oslo, Norway.

Elliott, D., Hartley, A., & Atwell, E. (2003). Rationale for a multilingual corpus for machine translation evaluation. In D. Archer, P. Rayson, A. Wilson, & T. McEnery (Eds.), (pp. 191-200). Lancaster, UK.

English, P. M., & Tenneti, R. (1994). Interleaf active documents. *Electronic Publishing*, *7*(2), 75-87.

Estrella, P., Hamon, O., & Popescu-Belis, A. (2007a). How Much Data is Needed for Reliable MT Evaluation? Using Bootstrapping to Study Human and Automatic Metrics. Copenhagen.

Estrella, P., Hamon, O., & Popescu-Belis, A. (2007b). A new method for the study of correlations between MT evaluation metrics (pp. 55-64). Skövde, Sweden.

Estrella, P., Popescu-Belis, A., & King, M. (2009). The FEMTI guidelines for contextual MT evaluation: Principles and resources. (W. Daelemans & V. Hoste, Eds.)*Linguistica Antverpiensia New Series*, Themes in Transalation Studies, *8*, 43-64.

Evaluation Working Group (ISLE). (2003). *International Standards for Language Engineering* (Final Report No. ISLE – IST-1999-10647) (p. 24).

Fais, L. (2004). Inferable Centers, Centering Transitions, and the Notion of Coherence. *Computational Linguistics*, *30*(2), 119-150.

Falkedal, K. (1991). Evaluation methods of Machine Translation Systems: An historial overview and a critical account. Repport to Suissetra. ISSCO - University of Geneva.

Farwell, D., & Helmreich, S. (2003). Pragmatics-based Translation and MT Evaluation (pp. 21-27). New Orleans, USA.

Feely, A. J., & Harzing, A.-W. (2003). Language management in Multinational Companies. *Cross Cultural Management: An International Journal*, *10*(2), 37-52.

Fiederer, R., & O'Brian, S. (2009). Quality and machine translation: A realistic objective? *JoSTrans: The Journal of Specialised Translation*, *11*. Retrieved from <http://www.jostrans.org/issue11/art_fiederer_obrien.pdf>

Flanagan, M. A. (1994). Error classification for MT Evaluation (pp. 65-72). Columbia, Maryland, USA.

Flanagan, M. A. (2002). *SYSTRAN: DaimlerChrysler's Language Engine* (p. 5). Soisy-sous-Montmorency: SYSTRAN.

Flanagan, M. A. (2009). *Recycling Texts: Human Evaluation of Example-Based Machine Translation Subtitles for DVD* (Doctor of Philosophy). Dublin City University, Dublin, Ireland.

Flanagan, M. A., & McClure, S. (2002). *SYSTRAN and the Reinvention of MT* (IDCBulletin No. #26459). Retrieved from <http://wwwv5.systransoft.com/IDC/26459.html>

Flint, P., Lord van Sylke, M., Stärke-Meyerring, D., & Thomspon, A. (1999). Going online: Helping Technical Communicators Help Translators. *Technical Communication*, *46*(2), 238-248.

Forcada, M. L., Canals-Marote, R., Esteve-Guillén, A., Garrido-Alenda, A., Guardiola-Savall, M. I., Iturraspe−Bellver, A., Montserrat−Buendia, S., et al. (2001). The Spanish<->Catalan Machine Translation System interNOSTRUM. Retrieved from <http://www.internostrum.com/docum/iN-MTS.pdf>

Forsbom, E. (2003). Training a Super Model Look-Alike: Featuring Edit Distance, N-Gram Occurrence, and One Reference Translation (pp. 29-36). New Orleans, USA.

Fouvry, F., & Balkan, L. (1996a). Test Suites for Controlled Language Checkers (pp. 179-136).

Fouvry, F., & Balkan, L. (1996b). Test Suites for Quality Evaluation of NLP Products. Moncton, N.-B., Canada.

Fouvry, F., Balkan, L., & Arnold, D. (1995). Test Suites for Quality Evaluation of NLP Products. London.

Franco Sabarís, M., Rojas Alonso, J. L., Dafonte, C., & Arcay, B. (2001). Multilingual Authoring through an Artificial Language",. Santiago de Compostela, Spain.

Frederking, R., Nirenburg, S., Farwell, D., Helmreich, S., Domashnev, C., Attardo, D., Grannes, D., et al. (1994). Integrating translations from multiple sources within the PANGLOSS Mark III machine translation system (pp. 73-80). Columbia, Maryland, USA.

Friedman, C., Kra, P., & Rzhetsk, A. (2002). Two biomedical sublanguages: a description based. *Journal of Biomedical Informatics*, *35*, 222–235.

Friske, H.-J. (1996). *Technische Dokumentation*. Kommunikation über Kommunikation. Münster: Lit Verlag.

Fritz, M. (2003). Die tekom wird 25. *technische Kommunikation*, *25*(5/2003), 14-20.

FSF Editorial Staff. (2006, February). High Stakes in Language Proficiency. *Flight Safety Digest*, *25*, 1-13.

Fuchs, N., Kaljurand, K., & Kuhn, T. (2008). Attempto Controlled English for Knowledge Representation. (C. Baroglio, P. Bonatti, J. Maluszynski, M. Marchiori, A. Polleres, & S. Schaffert, Eds.)*Lecture Notes in Computer Science*, Reasoning Web, Fourth International Summer School 2008, (5224), 104–124.

Fuchs, N., Schwertel, U., & Schwitter, R. (1999). Attempto Controlled English (ACE): Language Manual, Version 3.0. Zürich.

Fuchs, N., & Schwitter, R. (1996). Attempto Controlled English (ACE) (pp. 124-136). Leuven, Belgium.

Fuji, M., Hatanaka, N., Ito, E., Kamei, S., Kumai, H., Sukehiro, T., Yoshimi, T., et al. (1999). Evaluation Method for Determining Groups of Users Who Find MT "Useful." Singapore: Asia-Pacific Association for Machine Translation (AAMT), National University of Singapore.

Funk, A., Tablan, V., Bontcheva, K., Cunningham, H., Davis, B., & Handschuh, S. (2007). CLOnE: Controlled Language for Ontology Editing (p. (no page numbers)). Busan, Korea.

Gamero Pérez, S. (2001). La traducción de textos técnicos. Descripción y análisis de textos (alemán-español). Ariel Lenguas Modernas. Barcelona: Editorial Ariel.

Gamon, M., Aue, A., & Smets, M. (2005). Sentence-level MT evaluation without reference translations: beyond language modeling (pp. 103-111). Budapest, Hungary.

Gamon, M., Suzuki, H., & Corston-Oliver, S. (1999). Using Machine Learning for System-Internal Evaluation of Transferred Linguistic Representations. Singapore: Asia-Pacific Association for Machine Translation (AAMT), National University of Singapore.

Gaser, R., Guirado, C., & Rey, J. (Eds.). (2004). *Insights into Scientific and Technical Translation*. Barcelona: PPU. Promociones y Publicaciones Universitarias.

Gaspari, F. (2004). Online MT services and real users' needs: An empirical usability evaluation. In R. Frederking & K. B. Taylor (Eds.), (pp. 74-85). Washington DC, USA: Berlin: Springer Verlag.

Gavieiro-Villatte, E., & Spaggiari, L. (1999). Open-ended overview of controlled languages. *BULAG: Bulletin de Linguistique Appliquée et Générale*, Presses universitaires de Franche-Comté, *24*(Génie Linguistique et Génie Logiciel), 89-100.

Gdaniec, C. (1994). The Logos Translatability Index (pp. 97-105). Columbia, Maryland.

Gendner, V., Illouz, G., Jardino, M., Monceaux, L., Paroubek, P., Robba, I., & Vilnat, A. (2002). A Protocol for Evaluating Analyzers of Syntax (PEAS) (pp. 590-596). Las Palmas de Gran Canaria, Spain.

Gerber, L., & Hovy, E. (1998). Improving Translation Quality by Manipulating Sentence Length. In D. Farwell, L. Gerber, & E. H. Hovy (Eds.), (pp. 448-460). Langhorne, PA, USA.

Gerzymisch-Arbogast, H., & Mundersbach, K. (1998). *Methoden des wissenschaftlichen Übersetzens*. UTB für Wissenschaft. A. Francke Verlag.

Giménez, J., & Márquez, L. (2007). Linguistic featuresfor automatic evaluation of heterogenous mt systems (pp. 256–264).

Giménez, J., & Márquez, L. (2008). Heterogeneous Automatic MT Evaluation Through Non-Parametric Metric Combinations (pp. 319–326).

Giménez Linares, J. (2008, July). *Empirical Machine Translation and its Evaluation* (Ph.D. Thesis on Artificial Intelligence). Universitat Politècnica de Catalunya. Retrieved from <http://www.lsi.upc.edu/~jgimenez/thesis/thesis.pdf>

Godden, K. (1998). Controlling the Business Environment for Controlled Language (pp. 185-190). Pittsburg, Pennsylvania: Language Technologies Institute, Carnegie Mellon University.

Göpferich, S. (1998). *Interkulturelles Technical Writing: Fachliches adressatengerecht vermitteln; ein Lehr- und Arbeitsbuch*. Forum für Fachsprachen-Forschung. Tübingen: Gunter Narr Verlag. Retrieved from <http://books.google.es/books?id=4Zjl5RZw7ukC&dq=Die+Verst%C3%A4ndlichkeit+von+Texten&source=gbs_summary_s&cad=0>

Göpferich, S. (2003). Thesaurus verknüpft Informationsarten: Terminologie-Management – DaimlerChrysler und sein Projekt 'DAiSY. *technische Kommunikation – Fachzeitschrift für technische Dokumentation und Informationsmanagement*, *5*, 43-47.

Göpferich, S. (2004, March 18). *Übersetzungsgerechte Dokumentationserstellung*. Presented at the Treffen der tekom-Regionalgruppe Rhein-Main, Bonn.

Gorm Hansen, I., & Selsøe Sørensen, H. (2002). LinguaNet: Embedded MT in a Cross-Border Messaging System for European Law Enforcement. *Machine Translation*, *17*, 139-163.

Gough, N., & Way, A. (2004). Example-Based Controlled Translation (pp. 73-81). Malta.

Gow, F. (2003). *Metrics for Evaluating Translation Memory Software* (Master Thesis). University of Ottawa.

Grasse, N. (2001, October). Qualitätskontrolle des MÜ-Systems DCINTRANS in der Anwendung des Sprachendienstes der DaimlerChrysler AG (Diplomarbeit (Master Thesis)). Universität des Saarlandes.

Grishman, R., & Kittredge, R. (Eds.). (1986). *Analyzing Language in Restricted Domains: Sublanguage Description and Processing*. New Jersey: Hillsdale: Lawrence Erlbaum Associates.

Grover, C., Holt, A., Klein, E., & Moens, M. (1999). *Description of restricted natural language*. ESPRIT LTR Project PROSPER (26241), part of deliverable D5.1b. Edinburgh: University of Edinburgh.

Guerberof Arenas, A. (2008). Productivity and quality in the post-editing of outputs from translation emmories and machine translation. *Localisation Focus: The International Journal of Localisation*, *7*(1), 11-21.

Guerra Martínez, L. (2003, August). Human Translation versus Machine Translation and Full Post-Editing of Raw Machine Translation Output (Master Thesis). Dublin City University.

Guessoum, A., & Zantout, R. (2001). Semi-Automatic Evaluation of the Grammatical Coverage of Machine Translation Systems (pp. 133-138). Santiago de Compostela, Spain.

Guyon, A. (2003). *Machine Translation and the Virtual Museum of Canada (VMC)*. Retrieved from <http://www.chin.gc.ca/English/Pdf/Digital_Content/Machine_Translation/Machine_Translation.pdf>

Hahn, M. (1992). *The Key to Technical Translation* (Vol. Volume 1: Concept Specification). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Hajič, J., Hric, J., & Kubon, V. (2000). Machine Translation of Very Close Languages (pp. 7-12). Seattle (Washington).

Haller, J. (1996). MULTILINT- Multilingual Intelligence for Technical Documentation. Presented at the Aslib, London: The Association for Information Management, Information House.

Haller, J. (2001). MULTIDOC: Authoring Aids for Multilingual Technical Documentation. In J. Chabás (Ed.), (pp. 143-147). Barcelona: Universidad Pompeu Fabra.

Haller, J., & Fottner-Top, C. (2001). Multilint - eine toolgestützte Lösung für die Kontrolle von Textqualität. Presented at the tekom, Frühjahrstagung. Retrieved from <http://www.tekom.de/index_neu.jsp?url=/servlet/ControllerGUI?action=voll&id=394>

Haller, J., & Ramírez Polo, L. (2005). Controlled Language and the Implementation of Machine Translation for Technical Documentation. London.

Hamon, O., Hartley, A., Popescu-Belis, A., & Choukri, K. (2007). Assessing Human and Automated Quality Judgments in the French MT Evaluation Campaign CESTA (pp. 231-238). Copenhagen, Denmark.

Hamon, O., Mostefa, D., & Arranz, V. (2008). Diagnosing Human Judgments in MT Evaluation: an Example based on the Spanish Language. Donostia-San Sebastian, Spain.

Hamon, O., Popescu-Belis, A., Choukri, K., Dabbadie, M., Hartley, A., Mustafa El Hadi, W., Rajman, M., et al. (2006). CESTA: First Conclusions of the Technolangue MT Evaluation Campaign (pp. 179-184). Genoa, Italy.

Hamon, O., & Rajman, M. (2006). X-Score: Automatic Evaluation of Machine Translation Grammaticality (pp. 22-28). Genoa, Italy.

Harkus, S. (2000, April). *Translation and Localisation. Can writing style reduce localisation time?* PPT Presentation presented at the Third Australasian Online Documentation Conference, Brisbane.

Harris, Z. (1968). *Mathematical Structures of Language*. Interscience Tracts in Pure and Applied Mathematics. New York: John Wiley & Sons.

Harris, Z. (1988). *Language and Information*. Bampton Lectures in America. New York: Columbia University Press.

Harris, Z. (1991). A theory of language and information: a mathematical approach. Oxford: Clarendon Press.

Hart, G., Dolbear, C., & Goodwin, J. (2007). Lege Feliciter: Using Structured English to represent a Topographic Hydrology Ontology. Innsbruck.

Hart-Davidson, W. (2001). The Core Competencies of Technical Communication. *Technical Communication*, *48*(2), 145-155.

Hartley, A., & Paris, C. (2001). Translation, controlled languages, generation. In E. Steiner & C. Yallop (Eds.), *Exploring Translation and Multilingual Text Production: Beyond Content* (pp. 307-325). Berlin: Walter De Gruyter Inc.

Hartley, A., Scott, D., Bateman, J., & Dochev, D. (2001). AGILE - A system for multilingual Generation of Technical Instructions (pp. 145–150).

Hayes, P., Maxwel, S., & Schmandt, L. (1996). Controlled English Advantages for Translated and Original English Documents (pp. 84-92). Leuven.

He, Y., & Way, A. (2010). Metric and reference factors in minimum error rate training. *Machine Translation*, *24*(1), 27-38.

Heald, I. (2001). Elimination of ambiguity in technical translation. In C. Valero Garcés & I. de la Cruz Cabanillas (Eds.), *Traducción y Nuevas Tecnologías. Herramientas Auxiliares del Traductor* (pp. 287-294). Alcalá de Henares: Servicio de Publicaciones Universidad de Alcalá.

Healey, J. (n.d.). Case New Holland and Translation Technology. An interview with JeanPierre Oorlynck on the development of the ASIST project. *Multilingual Computing & Technology*, *15*(18).

Hebling, U. (2002). *Controlled Language am Beispiel des Controlled English*. Trier, Germany: Wissenschaftlicher Verlag Trier.

Henning, J., & Tjarks-Sobhani, M. (2005). Technische Dokumentation in Deutschland. In J. Henning & M. Tjarks-Sobhani (Eds.), *Stand und Perspektiven der Technischen Dokumentation – International*, tekom-Schriften zur Technischen Kommunikation. Lübeck: Schmidt-Römhild.

Hernandez, M., & Rascu, E. (2004). Checking and Correcting Technical Documents. (S. Vienney & M. Bioud, Eds.)*BULAG: Bulletin de Linguistique Appliquée et Générale*, Presses universitaires de Franche-Comté, *29*(Correction automatique: bilan et perspectives), 69-84.

Hirschman, L., & Sager, N. (1982). Automatic Information Formatting of a Medical Sublanguage. *Sublanguage: Studies of Language in Restricted Semantic Domains*, Library Edition/Foundation of Commnication (pp. 27-80). Berlin: de Gruyter.

Hoard, J., Wojcik, R., & Holzhauser, K. (1992). An Automated Grammar and Style Checker for Writers of Simplified English. In P. Holt & W. Nole (Eds.), *Computers and Writing: State of the Art* (pp. 278-296). Dordrecht: Kluwer Academic Publishers.

Hoenig, H. G., & Kussmaul, P. (1996). *Strategie der Übersetzung : Ein Lehr-und Arbeitsbuch*. Tübinger Beiträge zur Linguistik. Tübingen: Gunter Narr.

Hoft, N. L. (1995). *International Technical Communication. How to export information about high technology*. Wiley Technical Communication Library. New York: John Wiley & Sons, Inc.

Hoft, N. L. (1999). Global Issues, Local Concerns. *Technical Communication*, *Second Quarter*, 145-148.

Höge, M. (2002). *Towards a Framework for the Evaluation of Translators' Aids' Systems* (Doktorarbeit (PhD)). University of Helsinki. Department of Translation Studies.

Holmback, H., Shubert, S., & Spyridakis, J. H. (1996). Issues in Conducting Empirical Evaluations of Controlled Languages (pp. 166–177).

Holt, P., & Nole, W. (Eds.). (1992). *Computers and Writing: State of the Art*. Dordrecht: Kluwer Academic Publishers.

Horn-Heft, B. (1999). *Technisches Übersetzen in Theorie und Praxis*. UTB für Wissenschaft. Tübingen und Basel: A. Francke Verlag.

Houlihan, D. (2009). *Translating Product Documentation: The Right Balance between Cost and Quality in the Localization Chain* (No. Research report) (p. 25). Aberdeen Group. Retrieved from <http://www.aberdeen.com/launch/report/benchmark/6230-RA-translation-product-documentation.asp>

Hovy, E., King, M., & Popescu-Belis, A. (2002a). Principles of Context-Based Machine Translation Evaluation. *Machine Translation*, *17*(1), 43-75.

Hovy, E., King, M., & Popescu-Belis, A. (2002b). An Introduction to MT Evaluation (pp. 1-7). Las Palmas de Gran Canaria, Spain.

Hovy, E., Vassar College, N. I., Frederking, R., Mariani, J., & Zampolli, A. (1999). *Multilingual Information Management: Current Levels and Future Abilities*. A report Commissioned by the US National Science Foundation and also delivered to the European Commission's Language Engineering Office and the US Defense Advanced Research Projects Agency. Retrieved from <http://www-2.cs.cmu.edu/~ref/mlim/index.html>

Hui, B. (2002). Measuring user acceptability of machine translations to diagnose system errors: an experience report (pp. 1-7). Taipei (Taiwan).

Hujisen, W.-O. (1998a). Controlled Language: An introduction (pp. 1-15). Pittsburg, Pennsylvania: Language Technologies Institute, Carnegie Mellon University.

Hujisen, W.-O. (1998b, September 25). *Completeness of Compositional Translation* (PhD Thesis). Universiteit Utrecht.

Hurst, S. (2007). Web Content within a Global Information Management Strategy. SDL White Paper. Retrieved from <http://www.sdl.com/en/globalization-knowledge-centre/whitepapers/Web_Content_within_a_Global_Information_Management_Strategy.asp>

Hutchins, W. J. (1986). *Machine Translation: past, present, future*. Chichester (UK): Ellis Horwood. Retrieved from <http://www.hutchinsweb.me.uk/PPF-TOC.htm>

Hutchins, W. J. (1997a). From first conception to first demonstration: the nascent years of machine translation, 1947-1954. A chronology. *Machine Translation*, *12*(3), 195-252.

Hutchins, W. J. (1997b). Evaluation of Machine Translation and Translation Tools. In G. B. Varile & A. Zampolli (Eds.), *Survey of the state of the art in human language technology* (pp. 418-419). Pisa: Giardini. Retrieved from <http://cslu.cse.ogi.edu/HLTsurvey/ch13node5.html>

Hutchins, W. J. (2003). Has machine translation improved? Some historical comparisons (pp. 181-188). New Orleans, USA.

Hutchins, W. J., Hartmann, W., & Ito, E. (2004). Compendium of Translation Software: directory of commercial machine translation systems and computer-aided translation support tools (p. 127). The European Association for Machine Translation.

Hutchins, W. J., & Somers, H. (1992). *An Introduction to Machine Translation*. London: Academic Press. Retrieved from <http://www.hutchinsweb.me.uk/IntroMT-TOC.htm>

Hyland, C. (2003). Testing "Prompt": the development of a rapid post-editing service at CLS Corporate Language Services AG, Switzerland (pp. 189-193). New Orleans, USA.

Isabelle, P., & Bourbeau, L. (1985). TAUM-AVIATION: Its technical Features and Some Experimental Results. *Computational Linguistics*, *11*(1), 18-27.

Isseroff, A. (1999). Comment on Technical Writers Gain Control. *Technical Communicators'(TC-) Forum*, *3*.

Janowski, W. (1998). Controlled Language - Risks and Side Effects. *Technical Communicators'(TC-) Forum*, *2*.

Johnson, E. (2000). Talking across Frontiers (pp. 1-23). Belfast: Queen's University.

Jurafsky, D., & Martin, J. H. (2000). Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. (S. Russel & P. Norving, Eds.)Prentice Hall Series in Artificial Intelligence. New Jersey: Prentice-Hall, Inc.

Kaji, H. (1999). Controlled Languages for Machine Translation: State of the Art (pp. 37-39). Singapore: Asia-Pacific Association for Machine Translation (AAMT), National University of Singapore.

Kamprath, C., Adolphson, E., Mitamura, T., & Nyberg, E. (1998). Controlled Language for Multilingual Document Production; Experience with Caterpillar Technical English (pp. 51-61). Pittsburg, Pennsylvania: Language Technologies Institute, Carnegie Mellon University.

Karjalainen, M.-L., & Nordlund, J. (1997). The Influence of Language and Culture on Written Communication. *Technical Communicators'(TC-) Forum*, *4*.

Ketzan, E. (2007). Rebuilding Babel: Copyright and the future of machine translation online. *Tulane Journal of Technology & Intellectual Property*, 1-26.

Kimbrough, S. O., Lee, T. Y., Padmanabhan, B., & Yang, Y. (2004). On Original Generation of Structure in Legal Documents. Retrieved from <http://citeseer.ist.psu.edu/kimbrough04original.html>

King, M. (2001). *Standards work related to evaluation* (Handout). MTEval Workshop "An Invitation to Get Your Hands Dirty!" Geneva: ISSCO - University of Geneva.

King, M., & Falkedal, K. (1990). Using Test Suites in the Evaluation of Machine Translation Systems. *Proceedings of the 13th International Conference on Computational Linguistics (COLING)* (pp. 211-219). Helsinki.

King, M., Hovy, E., White, J., T'sou, B. K., & Zaharin, Y. (1999). MT Evaluation. *Proceedings of the Machine Translation Summit VII*. Singapore: Asia-Pacific Association for Machine Translation (AAMT), National University of Singapore.

King, M., Popescu-Belis, A., & Hovy, E. H. (2003). FEMTI: creating and using a framework for MT evaluation (pp. 224-231). New Orleans, USA.

King, M., Wilks, Y., Allen, S., Heid, U., & Albisser, D. (1993). Forum: Evaluation of MT Systems - Setting the Stage". In S. Nirenburg (Ed.), *Progress in Machine Translation* (pp. 267-271). Amsterdan, Oxford, Washington D.C.: IOS Press.

Kit, C., & Tak, M. W. (2008). Comparative Evaluation of Online Machine Translation Systems with Legal Texts. *Law Library Journal*, *100*(2), 299-322.

Kittredge, R. (1982). Sublanguages. *American Journal of Computational Linguistics*, *8*(2), 79-84.

Kittredge, R. (1985). The Significante of Sublanguage for Automatic Translation (pp. 154-166). Hamilton, New York: Colgate University.

Kittredge, R. (2003). Sublanguages and Controlled Languages. In R. Mitkow (Ed.), *The Oxford Handbook of Computational Linguistics* (pp. 430-447). New York: Oxford University Press.

Kittredge, R., & Lehrberger, J. (Eds.). (1982). *Sublanguage: Studies of Language in Restricted Semantic Domains*. Berlin: Walter De Gruyter Inc.

Klein, J., Lehmann, S., Netter, K., & Tillmann, W. (1998). DiET in the context of MT evaluation. In R. Nübel & R. Seewakd-Heeg (Eds.), Sprachwissenschaft, Computerlinguistik, Neue Medien (Vol. 2, pp. 107-126). Bonn: Gardez! Verlag.

Knops, U. (1999). Controlled Language - Issues in Checkers' Design (pp. 40-45). Presented at the Proceedings of the Machine Translation Summit VII, Singapore: Asia-Pacific Association for Machine Translation (AAMT), National University of Singapore.

Koehn, P. (2004). Statistical significance tests for machine translation evaluation (p. 8pp.). Barcelona, Spain.

Koehn, P., & Monz, C. (2006). Manual and automatic evaluation of machine translation between European languages. *Proceedings of NAACL 2006 Workshop on Statistical Machine Translation*.

Koh, S., Maeng, J., Lee, J.-Y., Chae, Y.-S., & Choi, K.-S. (2001). A Test Suite for Evaluation of English-to-Korean Machine Translation Systems (pp. 191-195). Santiago de Compostela, Spain.

Kohl, J. R. (1999). Improving Translatability and Readability with Syntactic Cues. *Technical Communication*, *46*(2), 149-166.

Kontrollierte natürliche Sprache: Attempto. (2000). Institu für Informatik, Universität Zürich.

Kornái, A., & Stone, L. (2004). Automatic Translation to Controlled Medical Vocabularies. In A. Machintosh, R. Ellis, & T. Allen (Eds.), (pp. 413-434). Berlin: Springer Verlag.

Korpela, J. (2006). Translation-friendly authoring, especially in HTML for the WWW. Retrieved from <http://www.cs.tut.fi/ ~jkorpela/transl/master.html>

Kothes Technische Kommunikation GmbH & Co. KG. (2004). The risks of poor documentation. Retrieved from <http://www.kothes.de/public_de/Infoservice_pag_20070615T143404606.aspx>

Kuhn, T. (2010). An Evaluation Framework for Controlled Natural Languages. In N. Fuchs (Ed.), *Proceedings of the Workshop on Controlled Natural Language (CNL 2009)*. Presented at the Lecture Notes in Computer Science 5972, Springer.

Kulesza, A., & Shieber, S. M. (2004). A Learning Approach to Improving Sentence-Level MT Evaluation (pp. 75-84). Baltimore, Maryland, USA.

Kürten, A. (2003). *Lesekompetenz und Textverständlichkeitsmodelle: Überlegungen zum verständlicheren Schreiben* (Zwischenprüfungsarbeit). Erziehungswissenschaftliches Institut der RWTH Aachen.

Langer, I., Schulz von Thun, F., & Tausch, R. (1974). *Verständlichkeit in Schule, Verwaltung, Politik und Wissenschaft*. München: Ernst Reinhard Verlag. Retrieved from <http://books.google.es/books?id=ZXzbAAAACAAJ&output=html>

Langer, I., Schulz von Thun, F., & Tausch, R. (2003). Sich verständlich ausdrücken: Anleitungstexte, Unterrichtstexte, Vertragstexte, Gesetzestexte, Versicherungstexte, Wissenschaftstexte, weitere Textarten. München: Ernst Reinhard Verlag.

Langlais, P., Gandrabur, S., Leplus, T., & Lapalme, G. (2005). The Long-Term Forecast for Weather Bulletin Translation. *Machine Translation*, *19*(1), 83-112.

Langlais, P., Leplus, T., Lapalme, G., & Gandrabur, S. (2005). Approche en corpus pour la traduction: le cas METEO (pp. 463-474). Dourdan, France.

Langlais, P., Leplus, T., Lapalme, G., & Gandrabur, S. (2005). From the Real World to Real Words: The METEO case (pp. 166-175). Budapest, Hungary.

Language Industry Monitor. (1993). Boeing's Simplified English Checker. *Language Industry Monitor*, *Jan/Feb*. Retrieved from <http://www.lim.nl/monitor/boeing.html>

Language Industry Monitor. (1995). Carnegie Group's ClearCheck. *Language Industry Monitor*, *Jan/Feb*. Retrieved from <http://www.lim.nl/monitor/carnegie.html>

Laoudi, J., Tate, C., & Voss, C. R. (2004). Towards an Automated Evaluation of an Embedded MT System (pp. 106-115). Malta.

Lavie, A. (2009, July 3). *Evaluation of Machine Translation Systems: Metrics and Methodology*. Presented at the Óbidos Workshop, Óbidos, Portugal. Retrieved from <http://www.cs.cmu.edu/afs/cs/user/alavie/www/Presentations/MT-Evaluation-Portugal-Jul09.ppt>

Lavie, A. (2010a). Essentials of machine translation evaluation. *TAUS Blog*. Retrieved from <http://www.translationautomation.com/best-practices/essentials-of-machine-translation-evaluation.html>

Lavie, A. (2010b, October). Essentials of machine translation evaluation. *TAUS*. Retrieved from <http://www.translationautomation.com/best-practices/essentials-of-machine-translation-evaluation.html>

Lavie, A. (2010c, November 31). *Evaluating the Output of Machine Translation Systems*. Presented at the The Ninth Conference of the Association for Machine Translation in the Americas, Denver, Colorado.

Lavie, A., & Denkowski, Michael. (2009). The METEOR Metric for Automatic Evaluation of Machine Translation. *Machine Translation*, *Special Issue*.

Lavie, A., Sagae, K., & Jayaraman, S. (2004). The significance of recall in automatic metrics for MT evaluation (pp. 134-143). Washington DC, USA.

Lawlor, T. (2005). Globalization Management Systems Building a Better Business Case. How to Deliver Rapid ROI from Translation Management (p. 19). SDL International.

LDC. (2002). Linguistic Data Annotation Specification: Assessment of Fluency and Adequacy in Arabic-English and Chinese-English Translations. Retrieved from <http://projects.ldc.upenn.edu/TIDES/translation/TransAssess02.pdf>

Lehmann, S., & Oepen, S. (1996). TSNLP – test suites for natural language processing (pp. 711-716). Copenhagen: Center for Sprogteknologi.

Lehrberger, J. (1982). Automatic Translation and the Concept of Sublanguage. In R. Kittredge & J. Lehrberger (Eds.), *Sublanguage: Studies of Language in Restricted Semantic Domains*, Library Edition/Foundation of Communication (pp. 81-106). Berlin: de Gruyter.

Lehrberger, J., & Bourbeau, L. (1988). Machine translation: linguistic characteristics of MT Systems and general methodology of evaluation. Amsterdam [etc.]: John Benjamins Publishing Company.

Lehrndorfer, A. (1996). Kontrolliertes Deutsch: Linguistische und sprachpsychologische Leitlinien für eine (maschinell) kontrollierte Sprache in der Technischen Dokumentation. Tübingen: Gunther Narr Verlag.

Lehrndorfer, A., & Beceiro Mangold, R. (1997). How to save money in translation cost. *Technical Communicators'(TC-) Forum*, *2*, 12-14.

Leplus, T. (2004, November). *Étude de la traduction automatique des bulletins météorologiques* (Memoire (M.SC. Thesis)). Université de Montréal,.

Leplus, T., Langlais, P., & Lapalme, G. (2004a). A corpus-based Approach to Weather Report Translation. Montréal.

Leplus, T., Langlais, P., & Lapalme, G. (2004b). Weather Report Translation using a Translation Memory. Lecture Notes in AI 3265 (pp. 154-163). Washington DC, USA: Springer.

Lethola, A., Tenni, J., & Bounsaythip, C. (1998). Controlled Language Technology in Multilingual User Interfaces. Stockholm.

Lethola, A., Tenni, J., Bounsaythip, C., & Jaaranen, K. (1999). Controlled Languages as the Basis for Multilingual Catalogues on the WWW. *Business and Work in the Information Society: New Technologies and Applications.* (Roger et al. (Eds.)., pp. 207-213). Amsterdam: IOS-Press.

Leusch, G., Ueffing, N., & Ney, H. (2003). A Novel String-to-String Distance MeasureWith Applications to Machine Translation Evaluation (pp. 240-247). New Orleans, USA.

Leusch, G., Ueffing, N., & Ney, H. (2006). CDER: efficient MT evaluation using block movements. (pp. 241-248). Trento, Italy.

Leusch, G., Ueffing, N., Vilar, D., & Ney, H. (2005). Preprocessing and Normalization for Automatic Evaluation of Machine Translation (pp. 17-24). Ann Arbor, Michigan.

Levin, L., Bartlog, B., Font Llitjos, A., Gates, D. M., Lavie, A., Wallace, D., Watanabe, T., et al. (2000). Lessons Learned from a Task-Based Evaluation of Speech-to-Speech Machine Translation (pp. 721-724). Athens.

Lewis, T., & Meier, R. M. (2005). The MT developer/provider and the global corporation (pp. 176-180). Budapest, Hungary.

Ley, M. (2005). Kontrollierte Textstrukturen. Ein (linguistisches) Informationsmodell für die Technische Kommunikation (Inauguraldissertation (PhD Thesis)). Justus-Liebig-Universität Gießen.

Liénard, F. (2005). Langage texto et langage contrôlé. *Lingvisticæ Investigationes*, *28*(1), 49-60.

Lin, C.-Y., & Hovy, E. H. (2002). *Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics* (p. 8). Gaithersburg, MD: National Institute of Standards and Technology.

Lin, C.-Y., & Och, F. J. (2004a). ORANGE: a method for evaluating automatic evaluation metrics for machine translation. Geneva, Switzerland.

Lin, C.-Y., & Och, F.-J. (2004b). Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigrams Statistics (pp. 605-612). Barcelona, Spain.

Lionbridge. (2001). When to Use Machine Translation. And what business goals can be achieved. (p. 19). Lionbridge Technologies, Inc.

Lita, L. V., Rogati, M., & Lavie, A. (2005). BLANC: Learning Evaluation Metrics for MT. Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP) (pp. 740–747). Vancouver, Canada.

Liu, D., & Gildea, D. (2005). Syntactic Features for Evaluation of Machine Translation. In J. Goldstein, A. Lavie, L. Chin-Yew, & C. Voss (Eds.), (pp. 25-32). Ann Arbor, Michigan.

Liu, D., & Gildea, D. (2006). Stochastic Iterative Alignment for Machine Translation Evaluation. Sidney.

Liu, D., & Gildea, D. (2007). Stochastic Iterative Alignment for Machine Translation Evaluation.

Lockwood, R. (2000). Machine Translation and Controlled Authoring at Caterpillar. In R. C. Sprung (Ed.), *Translating into Success* (pp. 187-202). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Lonsdale, D. W., Franz, A. M. ., & Leavitt, J. R. R. (1994). Large-Scale Machine Translation: An Interlingua Approach (pp. 525-530). Town Lake, Austin, Texas.

Luckhardt, H.-D. (1991). Sublanguages in Machine Translation (pp. 306-308). Presented at the Fifth Conference of the European Chapter of the Association for Computational Linguistics, Berlin.

Macklovitch, E. (2004). The Contribution of End-Users to the TransType2 Project (pp. 197-207). Washington DC, USA.

Mador-Haim, S., Winter, Y., & Braun, A. (2006). Controlled Language for Geographical Information System Queries. Buxton (UK).

Maier, E., & Clarke, A. (2003). Scalability of MT Systems (pp. 248-253). New Orleans, USA.

Maier, E., Clarke, A., & Stadler, H.-U. (1999). Evaluation of machine translation systems at CLS Corporate Language Services AG (pp. 223-228). Singapore: Asia-Pacific Association for Machine Translation (AAMT), National University of Singapore.

Markantonatou, S., Vangelis, K., & Maistros, Y. (2002). An Authoring Tool for Controlled Modern Greek (pp. 165-175). Thessaloniki.

Marrafa, P., & Ribeiro, A. (2001). Quantitative Evaluation of Machine Translation Systems: Sentence Level (pp. 39-43). Santiago de Compostela, Spain.

Martin, A. F., Garofolo, J. S., Fiscus, J. C., Le Audrey, N., Pallet, D. S., Przybocki, M. A., & Sanders, G. A. (2004). NIST Language Technology Evaluation Cookbook. Lisabon.

Matthews, C. (2000). *A guide to Presenting Technical Information. Effective Graphic Communication*. London: Professional Engineering Publishing Limited.

Means, L., & Godden, K. (1996). The Controlled Automotive Service Language (CASL) Project (pp. 106-114). Leuven.

Melamed, D., Green, R., & Turian, J. P. (2003). Precision and recall of machine translation (pp. 61–63).

Miller, K., Gates, D. M., Underwood, N. L., & Magdalen, J. (2001). Evaluation of Machine Translation Output for an Unknown Source Language: Report of an ISLE-Based Investigation. Santiago de Compostela, Spain.

Miller, K. J., & Vanni, M. (2001). Scaling the ISLE Taxonomy: Development of Metrics for the Multi-Dimensional Characterisation of Machine Translation Quality. Presented at the Proceedings of MT Summit VIII, Santiago de Compostela.

Mitamura, T. (1999). Controlled Language for Multilingual Machine Translation (pp. 13-17). Singapore: Asia-Pacific Association for Machine Translation (AAMT), National University of Singapore.

Mitamura, T. (2007). *Controlled Language Input/Output*. Presentation for MT class, Carniege Mellon University. Retrieved from <http://www.cs.cmu.edu/afs/cs/project/cmt-55/lti/Courses/731/www/CL-MTclass-07.pdf>

Mitamura, T., Baker, K., Nyberg, E., & Svoboda, D. (2003). Diagnostics for Interactive Controlled Language Checking. Dublin.

Mitamura, T., Baker, K., Nyberg, E., & Svoboda, D. (2003, May 15). *Diagnostics for Interactive Controlled Language Checking*. PPT Presentation presented at the Joint Conference of the 8th International Workshop of the European Association for Machine Translation and the 4th Controlled Language Applications Workshop, Dublin.

Mitamura, T., Baker, K., Svoboda, D., & Nyberg, E. (2003). Source Language Diagnostics for MT. New Orleans.

Mitamura, T., & Nyberg, E. (1995). Controlled English for Knowledge-Based MT: Experience with the KANT System (pp. 158-172). Leuven.

Mitamura, T., & Nyberg, E. (2001). Automatic Rewriting for Controlled Language Translation (pp. 1-12). Tokyo.

Mitamura, T., Nyberg, E., Baker, K., Svoboda, D., & Torrejón Díaz, E. (2001). Pronominal Anaphora Resolution in KANTOO English-to-Spanish Machine Translation System.

Mitamura, T., Nyberg, E., Baker, K., Svoboda, D., Torrejón Díaz, E., & Duggan, M. (2001). The Kantoo MT System: Controlled Language Checker and Knowledge Maintenance Tool.

Mitamura, T., Nyberg, E., & Carbonell, J. G. (1991). An Efficient Interlingua Translation System for Multilingual Document Production. Presented at the Third Machine Translation Summit.

Møller, M. H. (2003a). Grammatical Metaphor, Controlled Language and Machine Translation (pp. 95-104). Dublin.

Møller, M. H. (2003b, May 15). *Grammatical Metaphor, Controlled Language and Machine Translation*. PPT Presentation presented at the Joint Conference of the

8th International Workshop of the European Association for Machine Translation and the 4th Controlled Language Applications Workshop, Dublin.

Montalt i Resurrecció, V. (2005). *Manual de traducció cientificotècnica*. Biblioteca de traducció i interpretació. Vic: Eumo Editorial.

Moré-López, J., & Climent-Roca, S. (2006a). A cheap MT-evaluation method based on Internet searches (pp. 19-26). Oslo, Norway.

Moré-López, J., & Climent-Roca, S. (2006b). La tradautomaticidad: un concepto aplicado a la evaluación de sistemas de traducción automática. *Revista de Procesamiento del Lenguaje Natural*, *37*, 233-240.

Moré-López, J., & Climent-Roca, S. (2007). A cheap MT evaluation method based on the notion of machine translationness. Leuven, Belgium.

Moré-López, J., & Climent-Roca, S. (2008). A Machine Translationness Typology for MT Evaluations. Hamburg, Germany.

Morland, D. V. (2002). Nutzlos, Bien Pratique, or Muy Util? Business Users Speak Out on the Value of Pure Machine Translation. London.

Muegge, U. (2007). Controlled language: the next big thing in translation? *ClientSide News Magazine*, *7*(7), 21-24.

Muldoon, D. (1999). A Writer's View of Using a Controlled Language. *Technical Communicators'(TC-) Forum*, *3*.

Mustafa El Haidi, W., Dabbadie, M., Timimi, I., Rajman, M., Langlais, P., Hartley, A., & Popescu-Belis, A. (2004). Work-in-Progress project report: CESTA - Machine Translation Evaluation Campaign (pp. 16-26). Geneva, Switzerland.

Mustafa El Haidi, W., Timimi, I., & Dabbadie, M. (2001). Setting a Methodology for Machine Translation Evaluation (pp. 49-54). Santiago de Compostela, Spain.

Namgoong, H., & Kim, H.-G. (2007). Ontology-based Controlled Natural Language Editor Using CFG with Lexical Dependency (p. (no page numbers)). Busan, Korea.

Näsström, J. (1997). A Note on Controlled Language. *Technical Communicators'(TC-) Forum*, *4*.

Neal, J. G., & Walter, S. M. (1991). *Natural language processing systems evaluation workshop* (Technical report No. RL-TR-92-308). Rome Laboratory. Retrieved from <http://www.dtic.mil/cgi-bin/GetTRDoc?Location=U2&doc=GetTRDoc.pdf&AD=ADA247286>

Netter, K., Amstrong, S., Kiss, T., Lehman, S., Milward, D., Regnier-Prost, S., Schäler, R., et al. (1998). DiET - Diagnostic and Evaluation Tools for Natural Language Application. *Proceedings of the 1st International Conference on Language Resources and Evaluatio* (pp. 573-579). Granada, Spain.

Neufeld, J. K. (1987). *A Handbook for Technical Communication*. New Jersey: Prentice-Hall, Inc.

Newmark, P. (1988). Technical Translation. *A textbook of translation* (pp. 151-161). Hertfordshire: Prentice-Hall International.

Newton, J. (Ed.). (1992). *Computers in Translation: A Practical Appraisal*. London: Routledge.

Nickels Shirk, H. (2000). Researching the History of Technical Communication: Accessing and Analyzing Corporate Archives.

Nießen, S., Och, F. J., Leusch, G., & Ney, H. (2000). An evalution tool for machine translation: fast evaluation for MT Research (pp. 39-45). Athens.

Niremburg, S., Somers, H., & Wilks, Y. (Eds.). (2003). *Readings in Machine Translation*. Cambridge, Massachussets: MIT Press.

Nirenburg, S. (Ed.). (1987). *Machine Translation. Theoretical and Methodological Issues*. Studies in Natural Language Processing. Cambridge, UK: Cambridge University Press.

NIST. (2005). *The 2005 NIST Machine Translation Evaluation Plan (MT-05)* (p. 6). Gaithersburg, MD: National Institute of Standards and Technology.

Nübel, R. (1998). MT Evaluation in Research and Industry: Two Case Studies. In N. Weber (Ed.), *Machine Translation: Theory, Applications, and Evaluation. An assessment of the state-of-the-art*, Sprachwissenschaft, Computerlinguistik und neue Medien (pp. 85-118). St. Augustin: Gardez!-Verl.

Nübel, R., & Schütz, J. (2000). Evaluation as a Language Technology Deployment Trigger (pp. 69-75). Ljubljana, Slovenia.

Nyberg, E., Kamprath, C., & Mitamura, T. (1998, March). The KANT Translation System. From R&D to Large-Scale Deployment. *LISA Newsletter*, *2*(1).

Nyberg, E., & Mitamura, T. (1996). Controlled Language and Knowledge-Based Machine Translation: Principles and Practice (pp. 74-83). Leuven, Belgium: Katholieke Universiteit Leuven Centre for Computational Linguistics.

Nyberg, E., Mitamura, T., & Carbonell, J. G. (1994). Evaluation Metrics for Knowledge-Based Machine Translation (pp. 95-99). Kyoto, Japan.

Nyberg, E., Mitamura, T., & Carbonell, J. G. (1997). The KANT Machine Translation System: From R&D to Initial Deployment. Washington D.C.

Nyberg, E., Mitamura, T., & Hujisen, W.-O. (2003). Controlled Langauge for Authoring and Translation. In H. Somers (Ed.), *Computers and Translation: A Handbook*, Benjamins Translation Library (Vol. 35, pp. 71-110). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Nyberg, E., Mitamura, T., Svoboda, D., Brunner, A., & Baker, K. (2003). An integrated system for source language checking, analysis and term management (pp. 115-124). New Orleans, USA.

O'Brien, S. (2003a). Controlling Controlled English: An Analysis of Several Controlled Language Rule Sets (pp. 105-114). Dublin: Dublin City University.

O'Brien, S. (2003b, May 15). *An Analysis of Several Controlled English Rule Sets*. PPT Presentation presented at the Joint Conference of the 8th International Workshop

of the European Association for Machine Translation and the 4th Controlled Language Applications Workshop, Dublin.

O'Brien, S. (2005). Methodologies for measuring the correlations between post-editing effort and machine translatability. *Machine Translation*, *19*(1), 37-58.

O'Brien, S. (2006a). Pauses as indicators of cognitive effort in post-editing machine translation output. *Across Languages and Cultures*, *7*(1), 1-21.

O'Brien, S. (2006b, May). Machine-Translatability and Post-Editing Effort: An Empirical Study using Translog and Choice Network Analysys (PhD Thesis). Dublin City University.

O'Brien, S., & Roturier, J. (2007). How Portable are Controlled Languages Rules? A Comparison of Two Empirical MT Studies (pp. 345-352). Copenhagen.

O'Hara, F. M. (2001). A Brief History of Technical Communication. Retrieved from from <http://www.stc.org/confproceed/2001/PDFs/STC48-000052.pdf>

Oard, D., & Gonzalo, J. (2001). The CLEF 2001 Interactive Track. In C. Peters, M. Braschler, J. Gonzalo, & J. Kluck (Eds.), (pp. 308-319). Darmstadt, Germany: Springer.

Oepen, S., Dyvik, H., Flickinger, D., Lønning, J. T., Meurer, P., & Rosén, V. (2005). Holistic Regression Testing for High-Quality MT. Some Methodological and Technological Reflections.

Offersgaard, L., Povlsen, C., Almsten, L., & Maegaard, B. (2008). Domain specific MT in use. Hamburg, Germany.

Osborne, M. (2010, November 29). Statistical Machine Translation. *Encyclopedia of Machine Learning*. Springer.

Owczarzak, K. (2008, April). *A Novel Dependency-Based Evaluation Metric for Machine Translation* (Doctor of Philosophy (Ph.D.)). Dublin City University. Retrieved from <http://www.nclt.dcu.ie/mt/papers/Owczarzak_thesis_08.pdf>

Owczarzak, K., van Genabith, J., & Way, A. (2007a). Evaluating machine translation with LFG dependencies. *Machine Translation*, *21*(2), 95-119.

Owczarzak, K., van Genabith, J., & Way, A. (2007b). Dependency-based automatic evaluation for machine translation (pp. 80-87). Rochester, NY.

Owczarzak, K., van Genabith, J., & Way, A. (2007c). Labelled dependencies in machine translation evaluation (pp. 104-111). Prague, Czech Republic.

Paepcke, F. (1975). Gemeinsprache, Fachsprachen und Übersetzen. In K. Berger & H.-M. Speier (Eds.), *Im Übersetzen leben*, Tübinger Beiträge zur Linguistik: Übersetzen und Textvergleich (pp. 291-312). Tübingen: Gunter Narr Verlag. Retrieved from <http://books.google.es/books?id=RF7OgGghOPsC&printsec=frontcover#PPR9,M1>

Palmer, M., & Finin, T. (n.d.). Workshop on the evaluation of natural language processing systems. *Computational Linguistics*, *16*(3), 175-181.

Papineni, K., Roukos, S., Ward, T., Henderson, J., & Reeder, F. (2002). Corpus-based Comprehensive and Diagnostic MT Evaluation: Initial Arabic, Chinese, French, and Spanish Results (pp. 132-137). Presented at the Proceedings of the second international conference on human language technology research, San Diego, California.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: a Method for Automatic Evaluation of Machine Translation. In A. H. Adams, G. W. Austin, & M. Taylor (Eds.), (pp. 311-318). Pennsylvania, USA.

Paroubek, P., Chaudiron, S., & Hirschman, L. (2007). Principles of Evaluation in Natural Language Processing. *TAL*, *48*(1), 7-31.

Parr, B., & McManus, M. (2001). Web-Site Globalization: The Next Imperative for the Internet 2.0 Era. Retrieved from <http://www.etranslate.com/en/downloads/IDC_Globalization_report.pdf>

Paul, M., Finch, A., & Sumita, E. (2007). Reducing Human Assessment of Machine Translation Quality to Binary Classfiers (pp. 154-162). Skövde, Sweden.

Pease, A., & Murray, W. (2003). An English to Logic Translator for Ontology-based Knowledge Representation Languages (pp. 777-783). Beijing.

Peterson, K. (2009, February 24). Announcing NIST Open MT09. Retrieved from <http://www.mail-archive.com/mt-list@eamt.org/msg01271.html>

Petris, A. (2001). *EC SYSTRAN: The Comission's Machine Translation System* (p. 32). Brussels: European Comission Translation Service (SDT).

Pfaffin, S. M. (1965). Evaluation of Machine Translations by Reading Comprehension Tests and Subjective Judgments. *Mechanical Translation and Computational Linguistics*, *8*(2), 2-8.

Philippe, M. (2002). Knowledge representation in CGLF, CGIF, KIF, Frame-CG and Formalized-English (pp. 77-91). Borovets (Bulgaria): Springer Verlag.

Polvsen, C., & Bech, A. (2001). Ape: Reducing the Monkey Business in Post-Editing by Automating the Task Intelligently (pp. 283-386). Santiago de Compostela, Spain.

Polvsen, C., Underwood, N. L., Music, B., & Neville, A. (1998). Evaluating Text-type Suitability for Machine Translation a Case Study on an English-Danish MT System. In A. Rubio, N. Gallardo, R. Castro, & A. Tejada (Eds.), (pp. 129-134). Granada, Spain.

Pool, J. (2006a). Can Controlled Languages Scale to the Web? *5th International Workshop on Controlled Language Applications*.

Pool, J. (2006b). Can Controlled Languages Scale to the Web? Cambridge, Massachussets. Retrieved from

<http://www.utilika.org/pubs/etc/ambigcl/clweb.html>

Popescu-Belis, A. (2003). An experiment in comparative evaluation: human vs. computers (pp. 307-314). New Orleans, USA.

Popescu-Belis, A., King, M., & Benatar, H. (2002). Towards a corpus of corrected human translations (pp. 17-21). Las Palmas de Gran Canaria, Spain.

Popescu-Belis, A., Manzi, S., & King, M. (2001). Towards a Two-stage Taxonomy for Machine Translation Evaluation (pp. 1-8). Presented at the MT Summit VIII Workshop, Santiago de Compostela, Spain.

Porsiel, J. (2008a). Sprache als betriebswirtschaftlicher Faktor. *MDÜ*, *5*. Retrieved from <http://www.mt-archive.info/Multilingual-2008-Porsiel.pdf>

Porsiel, J. (2008b, December). Machine translation at Volkswagen: a case study. *Multilingual Computing & Technology*, *100*. Retrieved from <http://www.mt-archive.info/Multilingual-2008-Porsiel.pdf>

Power, R., Hartley, A., & Scott, Donia. (2003). Multilingual generation of controlled languages (pp. 115-123). Dublin.

Pratt-Hartmann, I. (2003). A Two-Variable Fragment of English. *Journal of Logic, Language and Information*, *12*, 13-35.

Przybocki, M. A., Sanders, G. A., & Le Audrey. (2006). Edit distance: a metric for machine translation evaluation (pp. 2038-2043). Genoa, Italy.

Pulman, R. (1996). Controlled Language for Knowledge Representation (pp. 233-242). Leuven: Katholieke Universiteit Leuven Centre for Computational Linguistics.

Quintal, P. (2002, September 26). *AECMA Simplified English*. Presented at the PLAIN Conference, Toronto. Retrieved from <http://www.plainlanguagenetwork.org/conferences/2002/aecma/aecma.pdf>

Quirk, C. B. (2004). Training a Sentence-Level Machine Translation Confidence Measure (pp. 825-828). Presented at the Proceedings of LREC-2004: Fourth International Conference on Language Resources and Evaluation, Lisbon, Portugal.

Rajman, M., & Hartley, A. (2001). Automatically predicting MT systems rankings compatible with Fluency, Adequacy or Informativeness scores (pp. 29-34). Santiago de Compostela, Spain.

Rascu, E. (2006). A Controlled Language Aproach to Text Optimisation in Technical Documentation. In M. Butt (Ed.), (pp. 107-114). Presented at the (Konferenz zur Verarbeitung natürlicher Sprache), Universität Konstanz.

Reeder, F. (2001). In One Hundred Words or Less. Presented at the MT Summit VIII Workshop, Santiago de Compostela, Spain.

Reeder, F., Miller, K., Doyon, J., & White, J. (2001). The Naming of Things and the Confusion of Tongues: an MT metric (pp. 55-59). Santiago de Compostela, Spain.

Reeder, F., Siddharthan, A., Mitamura, T., Miller, K., Dorr, B., Farwell, D., Habash, N., et al. (2004). Semantic Annotation for Interlingual Representation of Multilingual Texts. Presented at the International Conference on Language Resources and Evaluation, Lisbon.

Reeder, F., & White, J. (2003). Granularity in MT Evaluation (pp. 37-42). New Orleans, USA.

Reiss, K. (1983). Texttyp und Ubersetzungsmethode□: der operative Test. Heidelberg: Julius Groos.

Reiter, E., Mellish, C., & Levine, J. (1995). Automatic Generation of Technical Documentation. *Applied Artificial Intelligence*, *9*(3), 259-287.

Reskin, P. (1997). Evaluating Multilingual Gisting of Web Pages (pp. 129-135). Stanford, CA.

Reuther, U. (1998). Controlling Language in an Industrial Application. Pittsburg, Pennsylvania: Language Technologies Institute, Carnegie Mellon University.

Reuther, U. (1999). Technical Writers gain Control. *Technical Communicators'(TC-) Forum*, *2*.

Reuther, U. (2003). Two in one: Can it work? Readability and Translatability by means of Controlled Language (pp. 124-132). Dublin.

Reuther, U. (2007, May). *Controlled Language! Controlled Translation?* PPT Presentation presented at the CAT Workshop "The Next Chapter," Jülich (Germany).

Reuther, U., & Schmidt-Wigger, A. (2000). Designing a Multi-Purpose CL Application (pp. 120-133). Seattle (Washington).

Reuther, U., Schmidt-Wigger, A., & Fottner-Top, C. (1998). *MULTILINT: Abschlussbericht* (p. 17). Retrieved from <http://www.iai.uni-sb.de/docs/bericht5.pdf>

Richardson, S. D., & Braden-Harder, L. C. (1988). The Experience of Developing a Large-Scale Natural Language Processing System: Critique (pp. 195-202). Presented at the Second Conference on Applied Natural Language Processing, Austin, Texas.

Rico, C., & Torrejón, E. (2004). Controlled Translation as a New Translation Scenario: Training the Future User. London: ASLIB.

Riezler, S., & Maxwell, J. T. (2005). On Some Pitfalls in Automatic Evaluation and Significance Testing for {MT} (pp. 57-64). Ann Arbor, Michigan.

Rinsche, A. (1991). Comparative MT performance evaluation: an empirical study. In K. Falkedal (Ed.), (pp. 169-180.). Les Rasses, Vaud, Switzerland: Geneva: ISSCO.

Rinsche, A. (1993a). Evaluationsverfahren für maschinelle Übersetzungssysteme□: zur Methodik und experimentellen Praxis / Zugl.: Bonn, Univ., Diss., 1992. Universität Bonn.

Rinsche, A. (1993b). Towards a MT evaluation methodology (pp. 266-275). Kyoto, Japan.

Rintanen, K., & Zetzsche, J. (2002). Integrating Translation Tools in Document Creation. Retrieved from <http://www.internationalwriters.com/dejavu/Integrating_tools.html>

Rojas, D. M., & Aikawa, T. (2006). Predicting MT quality as a function of the source language (pp. 2534-2537). Genoa (Italia).

Roturier, J. (2004). Assessing a set of Controlled Language rules: Can they improve the performance of commercial Machine Translation systems? London.

Roturier, J. (2006). An Investigation into the Impact of Controlled English Rules on the Comprehensibility, Usefulness, and Acceptability of Machine-Translated Technical Documentation for French and German Users (PhD Thesis). Dublin City University.

Ruffino, R. (1982). Coping with Machine Translation. *Proveedings of Translating and the Computer 5: Tools for the trade:* (pp. 57-60). London: North-Holland Publishing Company.

Ruiz Casales, R., & Sutcliffe, R. F. E. (2003a). A specification and validating parser for simplified technical Spanish (pp. 35-44). Dublin.

Ruiz Casales, R., & Sutcliffe, R. F. E. (2003b, May 15). *A specification and validating parser for simplified technical Spanish*. PPT Presentation presented at the Joint Conference of the 8th International Workshop of the European Association for Machine Translation and the 4th Controlled Language Applications Workshop, Dublin.

Rychtyckyj, N. (2002). An assessment of Machine Translation for Vehicle Assembly Process Planning at Ford Motor Company (pp. 207-215). Presented at the Machine Translation: From Research to Real Users, 5th Conference of the Association for Machine Translation in the Americas, AMTA, Tiburon: Springer Verlag.

Rychtyckyj, N. (2006). Standard Language at Ford Motor Company: A Case Study in Controlled Language Development and Deployment. Cambridge, Massachussets.

Sager, N., & Nhàn, N. T. (2002). The computability of strings, transformations, and sublanguage. In B. E. Nevin & S. M. Johnson (Eds.), *The Legacy of Zellig Harris*, Chapter 4 (Vol. 2, pp. 79–120). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Sågvall Hein, A. (1997). Language control and machine translation. Santa Fe (New Mexico).

Sammer, M., Reiter, K., Soderland, S., & Etzioni, O. (August 8-12). Ambiguity reduction for machine translation: human-computer collaborationç (pp. 193-202). Cambridge, Massachussets.

Sargent, B. B. (2002, July). Calculating ROI in Software Localization. *Software Business Magazine*.

Schachtl, S. (1996). Requirements for Controlled German in Industrial Applications (pp. 143-149). Leuven.

Schäfer, F. (2002). Die maschinelle Übersetzung von Wirtschaftsfachtexten: Eine Evaluierung anhand des MÜ-Systems der EU-Kommission, SYSTRAN, im Sprachenpaar Französisch-Deutsch (PhD Thesis). Universität des Saarlandes.

Schäfer, F. (2003). MT post-editing: How to shed light on the "unknown task" Experices made at SAP (pp. 115-123). Dublin.

Scheurs, D., & Adriaens, G. (1992). Controlled English (CE): From COGRAM to ALCOGRAM. In P. Holt & W. Nole (Eds.), *Computers and Writing: State of the Art* (pp. 206-221). Dordrecht: Kluwer Academic Publishers.

Schmidt-Wigger, A. (1998). Grammar and Style Checking for German (pp. 76-86). Pttsburgh, PA.

Schmitt, P. A. (1994). Die Eindeutigkeit von Fachtexten: Bemerkungen zu einer Fiktion. In M. Snell-Horby (Ed.), *Übersetzungswissenschaft. Eine Neuorientierung*, UTB wür Wissenschaft (2nd ed., pp. 252-282). Túbingen und Basel.

Schnitzlein, M. (2003, August). Zum Aussagewert von Qualitätsnormen und Qualitätssicherungssystemen für die Translationsqualität - eine exemplarische Analyse (Diplomarbeit (Master Thesis)). Universität des Saarlandes.

Schütz, J. (1994). Towards Text-Based Machine Translation (pp. 185-192). Columbia, Maryland, USA.

Schütz, J. (1996). Combing language technolgy and web technology to streamline an automotive hotline support service (pp. 180-189). Montreal, Quebec, Canada.

Schütz, J. (1999). Deploying the SAE J2450 Translation Quality Metric in MT Projects (pp. 278-284). Presented at the Proceedings of the Machine Translation Summit VII, Singapore: Asia-Pacific Association for Machine Translation (AAMT), National University of Singapore.

Schütz, J. (2001). Blueprint for MT Evolution. Reflections on "Elements of Style" (pp. 9-13). Santiago de Compostela, Spain.

Schwertel, U. (2000). Controlling Plural Ambiguities in Attempto Controlled English (pp. 105-119). Seattle (Washington).

Schwitter, R. (1998). *Kontrolliertes English für Anforderungsspezifikationen* (PhD Thesis). Fakultät I der Universität Zürich. Retrieved from www.ics.mq.edu.au/~rolfs/papers/DissBook.pdf

Schwitter, R., Scott, D., Ljungberg, A., & Hood, D. (2003). ECOLE: A Look-ahead Editor for a Controlled Language. Dublin.

Schwitter, R., & Tilbrook, M. (2004). Controlled Natural Languages meets the Semantic Web. Retrieved from <http://www.ics.mq.edu.au/~rolfs/controlled-natural-languages/>

Schwitter, R., & Tilbrook, M. (2006). Writing RSS Feeds in a Machine-Processable Controlled Natural Language. Cambridge, Massachussets.

Searle, J. R. (1970). *Speech Acts: An Essay in the Philosophy of Language*. Cambridge: Cambridge University Press.

Seewald-Heeg, U. (1998). Textsortenspezifische Evaluation Maschineller Übersetzungssysteme am Beispiel von Instruktionstexten. In R. Nübel & U. Seewald-Heeg (Eds.), Sprachwissenschaft, Computerlinguistik, Neue Medien Band 2 (pp. 19-38). Universität Bonn: Gardez! Verlag St. Augustin.

Seewald-Heeg, U. M. (1998). Linguistic Features of Instructional Texts and their Treatment by Machine Translation Systems. In N. Weber (Ed.), *Machine*

Translation: Theory, Applications, and Evaluation. An assessment of the state-of-the-art, Sprachwissenschaft, Computerlinguistik und neue Medien (pp. 137-165). St. Augustin: Gardez!-Verl.

Senfle, M. (2004). *Writing-for-Translation* (Research Paper). University of Wisconsin.

Shaw, D. (2006). Simplified Technical English in the 21st century. Cambridge, Massachussets.

Sheremetyeva, S. (2006). Integration of Correction Modules in a Controlled Language. Cambridge, Massachussets.

Skuce, D. (2003). *A Controlled Language for Knowledge Formulation on the Semantic Web*. Retrieved from <http://www.site.uottawa.ca:4321/factguru2.pdf>

Slocum, J., & Bennet, W. S. (1985). An Evaluation of METAL: the LRC Machine Translation System (pp. 62-69). Geneva, Switzerland.

Smith, J. (2001, June). How to Save Money on Translation By Editing the Source Text. *STC International Technical Communication SIG Translation Kit*.

Snell-Horby, M. (Ed.). (1994). *Übersetzungswissenschaft. Eine Neuorientierung*. UTB für Wissenschaft (2nd ed.). Tübingen und Basel: A. Francke Verlag.

Snover, Mathew, Dorr, B., Schwartz, R., Micciulla, L., & Makhoul, J. (2006). A Study of Translation Edit Rate with Targeted Human Annotation.

Snover, Matthew, Dorr, B., Schwartz, R., Makhoul, J., Micciulla, L., & Weischedel, R. (2005). *A Study of Translation Error Rate with Targeted Human Annotation* (No. LAMP-TR-126, CS-TR-4755, UMIACS-TR-2005-58). College Park. MD: University of Maryland.

Softky, B. (2007, May 15). How Google translates without understanding. Most of the right words, in mostly the right order. *The Register*. Portal, . Retrieved February 24, 2010, from <http://www.theregister.co.uk/2007/05/15/google_translation/>

Somers, H. (Ed.). (2003). *Computers and translation. A translator's guide*. Benjamins Translation Library. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Somers, H., & Sugita, Y. (2003). Evaluating Commercial Spoken Language Translation Software. New Orleans, USA.

Sowa, J. F. (2004). Common Logic Controlled English. Retrieved from <http://www.jfsowa.com/clce/specs.htm>

Spaggiari, L. (2003). A controlled language at Airbus (pp. 151-159). Dublin.

Sparck Jones, K., & Galliers, J. (1995). *Evaluating natural language processing systems: an analysis and review*. Lecture notes in computer science. Berlin: Springer. Retrieved from <http://books.google.es/books?id=8xPCC7H9c9oC&lpg=PA157&ots=o4j7p2r47X&vq=Adequacy%20Evaluation%20EAGLES&dq=Adequacy%20Evaluation%20EAGLES&hl=es&pg=PP1#v=onepage&q&f=true >

Specia, L., & Farzindar, A. (2010). Estimating Machine Translation Post-Editing Effort with HTER. AMTA 2010- workshop, Bringing MT to the User: MT Research

and the Translation Industry, at The Ninth Conference of the Association for Machine Translation in the Americas,. Denver, Colorado. Retrieved from <http://rali.iro.umontreal.ca/Publications/files/Specia-Farzindar_AMTA_workshop.pdf>

Specia, L., Raj, D., & Turchi, M. (2010). Machine translation evaluation versus quality estimation. *Machine Translation*, *24*(1), 39-50.

Spyridakis, J. H., Holmback, H., & Schubert, S. K. (1997). Measuring the Translatability of Simplified English in Procedural Documents. *IEEE Transactions on Professional Communication*, *40*(1).

Streiff, A. A. (1985). New developments in TITUS 4. Proveedings of Translating and the Computer 5: Tools for the trade: London.

Sukkarieh, J. Z., Hartley, A., & Scott, D. (2003). Mind your Language! Controlled Language for Inference Purposes (pp. 160-169). Dublin.

Surcin, S., Hamon, O., Hartley, A., Rajman, M., Popescu-Belis, A., Mustafa El Hadi, W., Timimi, I., et al. (2005). Evaluation of Machine Translation with Predictive Metrics beyond BLEU/NIST: CESTA Evaluation Campaign#1 (pp. 117-124). Phuket, Thailand.

Symonenko, S., Rowe, S., & Liddy, E. (2006). Illuminating Trouble Tickets with Sublanguage Theor. *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL* (pp. 169–172). New York: Association for Computational Linguistics.

Tablan, V., Polajnar, T., Cunningham, H., & Bontcheva, K. (2006). User-friendly ontology authoring using a controlled language. Genova, Italy.

Tak Ming, W. (2008, February). *Machine Translation and Evauation: Online Systems* (Thesis of Master).

Tate, C., Lee, S., & Voss, C. R. (2003). Task-based MT Evaluation: Tackling Software, Experimental Design, & Statistical Models (pp. 43-50). New Orleans, USA.

TCEurope. (n.d.). *Usable and safe operating manuals for consumer goods - guideline*. Retrieved from <http://www.tceurope.org/pdf/securedoc1_04.pdf>

Temnikova, I. (2010). A Cognitive Evaluation Approach for a Controlled Language Post-Editing Experiment. *Proceedings of the International Conference "Language Resources and Evaluation" (LREC2010),*. Valletta, Malta.

The EAGLES MT Evaluation Working Group. (1996). EAGLES Evaluation of Natural Language Processing Systems. Final Report. EAGLES Document EAG-EWG-PR.2 (p. 271). Copenhagen: Center for Sprogteknologi.

The EAGLES MT EValuation Working Group. (1995). *EAGLES Evaluation of Language Processing Systems. EAGLES Document EAG-EWG-PR.2* (p. 271). Copenhagen: Center for Sprogteknologi.

The EAGLES MT EValuation Working Group. (1999). *EAGLES Evaluation Working Group Final Report. EAG-II-EWG-PR.1* (p. 164). Copenhagen: Center for Sprogteknologi.

Thompson, H. S. (1991). Automatic evaluation of translation quality: outline of methodology and report on pilot experiment. In K. Falkedal (Ed.), (pp. 215-223). Les Rasses, Vaud, Switzerland.

Thompson, H. S. (1992). Automatic evaluation of translation quality: outline of methodology and report on pilot experiment. (pp. 24-26). San Diego (California).

Thrust, E. A. (2001). Plain English? A Study of Plain English Vocabulry and International Audiences. *Technical Communication*, *48*(3), 289-296.

Tillmann, C., Vogel,, S., Ney, H., Zubiaga, A., & Sawaf, H. (1997). Accelerated DP based Search for Statistical Translation. *Proceedings of European Conference on Speech Communication and Technology.*

Tomás, J., Angel Mas, J., & Casacuberta, F. (2003). A quantitative method for machine translation evaluation (pp. 12-17). Budapest, Hungary.

Torrejón Díaz, E., & Rico Pérez, C. (2002). Controlled Translation: A New Teaching Scenario Tailor-made for the Translation Industry (pp. 107-116). European Association for Machine Translation.

Trujillo, A. (1999). Translation engines: Techniques for Machine Translation. London: Springer.

Tuggy, D. (2006). Schematic network: Ambiguity, polysemy, and vagueness. In D. Geeraerts (Ed.), *Cognitive Linguistics: Basic Readings* (pp. 167-184). Berlin: Mouton de Gruyter. Retrieved from

<http://books.google.es/books?id=canMZSZ32ZgC&printsec=frontcover&dq=Cognitive+linguistics:+Basic+Readings&sig=ACfU3U30jzxMhIXYF1fzeT67elJW19dKNQ#PPA167,M1>

Turian, J. P., Shen, L., & Melamed, D. I. (2003). Evaluation of machine translation and its evaluation (pp. 23-28). New Orleans, USA.

Tylor, K., & White, J. S. (1998). Predicting What MT is Good for: User Judgments and Task Performance. In D. Farwell, L. Gerber, & E. H. Hovy (Eds.), (pp. 364-373). Langhorne, PA, USA.

Ueffing, N., Macherey, K., & Hermann, N. (2003). Confidence Measures for Statistical Machine Translation (pp. 394-401). New Orleans, USA.

Ueffing, N., & Ney, H. (2005). Word-Level Confidence Estimation for Machine Translation using Phrase-Based Translation Models (pp. 763-770). Vancouver, Canada.

Underwood, N. L., & Jongejan, B. (2001). Translatability Checker: A Tool to Help Decide Whether to Use MT (pp. 363-368). Santiago de Compostela, Spain.

Unwalla, M. (2004). AECMA Simplified English. *Communicator*, *Winter*.

Uszkoreit, H., & Koehn, P. (2008). Statistical and hybrid machine translation between all European languages. Publishable executive summary (p. 6). The Euromatrix consortium.

van der Eick, P., de Koning, M., & van der Steen, G. (1996). Controlled language correction and translation (pp. 64-73). Leuven.

van der Meer, J. (2003, June 18). The Business Case for MT: The Breakthrough Is for Real. *Globalization Insider: The LISA Newsletter*, *XII*(2.6).

van der Steen, G., & Dijenborgh, B. J. (1992). Online correction and translation of industrial texts. *Translating and the Computer 14* (pp. 135-164). London: Aslib.

van Slype, G. (1979). *Critical study of methods for evaluating the quality of machine translation. Final Report* (p. 187). Brussels: Ingenieurs-Conseils en methodes de direction.

Vandeghinste, V. (2009). Scaling up a hybrid MT System: From low to full resources. (W. Daelemans & V. Hoste, Eds.)*Linguistica Antverpiensia New Series*, Themes in Transalation Studies, *8*, 65-80.

Vanni, M., & Miller, K. (2001). Scaling the ISLE Framework: Validating Tests of Machine Translation Quality for Multi-Dimensional Measurement. Santiago de Compostela, Spain.

Vanni, M., & Reeder, F. (2000). How are you doing? A look at MT Evaluation. In J. S. White (Ed.), (pp. 109-116). Cuernavaca, Mexico: Springer.

Vasconcellos, M. (1992a). Panel: Apples, oranges, or kiwis? Criteria for the comparison of MT systems. *MT Evaluation: Basis for Future Directions. Proceedings of a workshop sponsored by the National Science Foundation* (pp. 37-50). San Diego, California.

Vasconcellos, M. (Ed.). (1992b). MT Evaluation: Basis for Future Directions. Proceedings of a workshop sponsored by the National Science Foundation. Association for Machine Translation.

Vashee, K. (2009, July 29). The importance of Measuring Translation Quality – BLEU. *Tomedes Blog: Smart Human Translation Services*. blog, . Retrieved February 23, 2010, from <http://blog.tomedes.com/measuring-machine-translation-quality/>

Vassiliou, M., Markantonatou, S., Maistros, Y., & Karkaletsis, V. (2003). Evaluating Specifications for Controlled Greek. Dublin.

Vertan, C., & von Hahn, W. (2003). Menu choice translation: a flexible menu-based controlled natural language system (pp. 194-199). Dublin. Retrieved from <http://www.mt-archive.info/CLT-2003-Vertan.pdf>

Viren, J. (2003). *Improved metrics for machine translation evaluation*. Department of Computer & Information Science University of Pennsylvania.

Voss, C. R., & Van Ess-Dykema, C. (2000). When is an Embedded MT System "Good Enough" for Filtering? (pp. 9-16). Seattle, Washington.

Walker, D. J., Clements, D. E., Darwin, M., & Amtrup, J. W. (2001). Sentence Boundary Detection: A Comparison of Paradigms for Improving MT Quality. Santiago de Compostela, Spain.

Walmer, D. (1999). One Company's Efforts to Improve Translation and Localization. *Technical Communication*, *46*(2), 230-237.

Way, Andrew, & Gough, N. (2005). Controlled translation in an example-based environment: what do automatic evaluation metrics tell us? *Machine Translation*, *19*(1), 1-36.

Way, Andy. (2009). A critique of Statistical Machine Translation. (W. Daelemans & V. Hoste, Eds.)*Linguistica Antverpiensia New Series*, Themes in Transalation Studies, *8*, 17-42.

Way, Andy (Ed.). (2010). *Machine Translation.* Special Issue: Topics in Machine Translation Evaluation/Guest Edited by Alon Lavie and Mark Przybocki (Vol. 24/1).

Weber, N. (1998). Machine Translation, Evaluation, and Translation Quality Assessment. In N. Weber (Ed.), *Machine Translation: Theory, Applications, and Evaluation. An assessment of the state-of-the-art*, Sprachwissenschaft, Computerlinguistik und neue Medien (pp. 47-84). St. Augustin: Gardez!-Verl.

Wells-Akis, J., & Sisson, W. R. (2002). Improving Translatability - A Case Study at Sun Microsystems Inc. *The LISA Newsletter: Globalization Insider*, *4.5*.

Werfelman, L. (2007, August). Simplifying the Technicalities. *AerosafetyWorld*, 16-21.

Westendrop, P. (2003). Das Gesetz von Moore für Anwendungsunterstützung Produktkomplexität contra Anwenderunterstützung. *technische Kommunikation*, *25*(5), 33-37.

White, J. (1992). MT evaluation: basis for future directions. Proceedings of a workshop. San Diego (California).

White, J., O'Conell, T., & Carlson, L. M. (1993). Evaluation of Machine Translation. Princeton, New Jersey.

White, J. S. (1995). Approaches to Black Box MT Evaluation. Luxembourg.

White, J. S. (2000). Contemplating automatic MT evaluation. In: White, John S. (ed.). Envisioning Machine Translation in the Information Future. Berlin/Heidelberg/NewYork/Barcelona/HongKong/London/Milan/Paris/Singapore/Tokyo: Springer. pp. 100-108.

White, J. (2000). Toward an Automated, Task-Based MT Evaluation Strategy. In Maegaard, B., ed., Proceedings of the Workshop on Machine Translation Evaluation at LREC-2000. Athens, Greece.

White, J. S. (2001). Predicting Intelligibility from Fidelity in MT Evaluation (pp. 35-37). Presented at the MT Summit VIII Workshop, Santiago de Compostela, Spain.

White, J. S. (2003). How to evaluate machine translation. In: Somers, H. (ed.). Computers and Translation: A Translator's Guide. Amsterdam/Philadelphia: John Benjamins Publishing. pp. 211-244.

White, J. S., Doyon, J. B., & Talbott, S. W. (2000). Task Tolerance of MT Output in Integrated Text Processes (pp. 9-16). Seattle, Washington.

White, J. S., & Forner, M. (2001). Predicting MT fidelity from noun-compound handling (pp. 45-48). Santiago de Compostela, Spain.

White, J. S., O'Conell, T., & O'Mara, F. (1994). The ARPA MT Evaluation Methodologies: Evolution, Lessons/ and Future Approaches (pp. 193-205). Columbia, Maryland, USA.

White, J. S., & O'Connell, Theresa. (1994). Evaluation in the ARPA machine translation program: 1993 methodology (pp. 135 - 140). Plainsboro, NJ: Association for Computational Linguistics.

White, J. S., & Taylor, K. B. (1998). A Task-Oriented Evaluation Metric for Machine Translation (pp. 21-25). Granada.

Wilks, Y. (2009). *Machine Translation: Its Scope and Limits*. New York: Springer Verlag.

Wojcik, Richard, Harrison, P., & Bremer, J. (1993). Using Bracketted Parses to Evaluate a Grammar Checking Application (pp. 38-49). Columbus, Ohio.

Wojcik, Richard, & Holmback, H. (1996). Getting a controlled language off the ground at Boeing (pp. 22-31). Leuven.

Wojcik, Richard, Holmback, H., & Hoard, J. (1998). Boeing Technical English: An Extension of AECMA SE beyond the Aircraft Maintenance Domain (pp. 114-123). Pittsburgh (Pennsylvania): Language Technologies Institute, Carnegie Mellon University. Retrieved from <http://64.233.179.104/scholar?q=cache:eFl91MyxwpQJ:www4.ncsu.edu>

Wojcik, Rick, & Hoard, J. E. (1996). Controlled Languages in Industry. In R. A. Cole, J. Mariani, H. Uszkoreit, A. Zaenen, & V. Zue (Eds.), *Survey of the State of the Art in HUman Language Technology*. Retrieved from <http://cslu.cse.ogi.edu/HLTsurvey/HLTsurvey.html>

Wright, S. E., & Wright, L. D. (1993). *Scientific and Technical Translation*. American Translators Association Series. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Xiaohong, W. (2005). Controlled Language: A useful technique to facilitate machine translation of technical documents. *Lingvisticæ Investigationes*, John Benjamins Publishing Company, *28*(1), 123–131.

Yamada, S., Sumita, E., & Kashioka, H. (2000). Translation using information on dialogue participants (pp. 37–43). Seattle, WA.

Yasuda, K., Sugaya, F., Takezawa, T., Yamamoto, S., & Yanagida, M. (2001). An Automatic Evaluation Method of Translation Quality Using Translation Answer Candidates Queried from a Parallel Corpus (pp. 373-378). Santiago de Compostela, Spain.

Yasuda, K., Sugaya, F., Takezawa, T., Yamamoto, S., & Yanagida, M. (2003). Applications of Automatic Evaluation Methods to Measuring a Capability of Speech Translation System (pp. 371-378). Budapest, Hungary.

Ye, Y., Zhou, M., & Chin-Yew, L. (2007). Sentence Level Machine Translation Evaluation as a Ranking Problem: one step aside from BLEU (pp. 240–247). Prague, Czech Republic.

Ying Zang, & Stephan Vogel. (2004). Measuring Confidence Intervals for the Machine Translation Evaluation Metrics. *TMI 04*, 85-94.

Yokoyama, S., Kashioka, H., Kumano, A., Matsudaira, M., Shirokizawa, Y., Kodama, S., Ehara, T., et al. (2001). An Automatic Evaluation Method for Machine Translation using Two-way MT (pp. 379-384). Santiago de Compostela, Spain.

Yuste, E. (2004). Corporate Language Resources in Multilingual Content Creation, Maintenance and Leverage (pp. 9-15). Geneva, Switzerland.

Yuste, E., & Braun-Chen, F. (2001). Comparative Evaluation of the Linguistic Output of MT Systems for Translation and Information Purposes. Santiago de Compostela, Spain.

Zhang, W., Zhou, X., & Yu, S. (1998). Construction of Controlled Chinese Lexicon (pp. 159-173). Pittsburgh (Pennsylvania): Language Technologies Institute, Carnegie Mellon University. Retrieved from <http://64.233.179.104/scholar?q=cache:eFl91MyxwpQJ:www4.ncsu.edu>

Zhang, Y., & Vogel, S. (2004). Measuring Confidence Intervals for the Machine Translation Evaluation Metrics (pp. 85-94). Baltimore, Maryland, USA.

Zhang, Y., & Vogel, S. (2010). Significance tests of automatic machine translation evaluation metrics. *Machine Translation*, *24*(1), 51-65.

Zwicky, A. M., & Zwicky, A. D. (1982). Register as a Dimension of Linguistic Variation. In R. Kittredge & J. Lehrberger (Eds.), *Sublanguage: Studies of Language in Restricted Semantic Domains*, Library Edition/Foundation of Communication (pp. 213-218). Berlin: de Gruyter.

ANNEX I: OVERVIEW OF CLs

ACE (Avaya Controlled English)	
Based on: **Language(s):** English **Organization:** Avaya **Year of development: (**2004) **Domain:** Telecommunications **CL checker(s):** Avaya Controlled English Checker	**Restrains:** Avaya Controlled English provides a set of principles to control vocabulary, sentence construction, and sentence length. **Classification:** Multilingual, HOCL **Classification (Pool, 2006):** Formalistic/Restricted **Relevant literature:** Avaya Style Guide (2004: 30) **Comments:**

AECMA Simplfied English (SE)	
Based on: ILSAM **Language(s):** English **Organization:** AECMA **Year of development**: In development since 1979, it was made mandatory in 1987 (Quintal, 2002) **Domain:** Aircraft	**CL checker(s):** **Restrains:** American Spelling **Classification:** Monolingual, HOCL **Classification (Pool, 2006):** Naturalistic/Restricted **Relevant literature:** Quintal (2002); Unwalla (2004) **Comments:**

Airbus Warning Language	
Based on:	**Restrains:** Short industrial warnings
Language(s): English	**Classification:** Monolingual, HOCL
Organization: Airbus	**Classification (Pool, 2006):** Naturalistic/Restricted
Year of development: 1998	**Relevant literature:** Spaggiari (2003)
Domain: Aircraft	**Comments:**
CL checker(s):	

ALCOGRAM	
Based on: COGRAM	**Restrains:**
Language(s): English	**Classification:** Multilingual, MOCL
Language orientation:	**Classification (Pool, 2006):** Naturalistic/Restricted
Organization: Alcatel-Bell Company	**Relevant literature:** Adriaens & Scheurs (1992); Scheurs & Adriaens (1992)
Year of development:	
Domain: Telecommunications	**Comments:** Algorithmic representation of COGRAM
CL checker(s): ALCOGRAM	

ARREX Controlled Language	
Based on: **Language(s):** Italian **Organization:** ARREX Le Cucine **Year of development:** **Domain:** Kitchen Furniture **CL checker(s):**	**Restrains:** **Classification:** Monolingual, MOCL **Classification (Pool, 2006):** **Relevant literature:** Fellet (2011) **Comments:**

ASD STE (Simplified Technical English)	
Based on: AECMA SE **Language(s):** English **Organization:** AECMA **Year of development:** **Domain:** Aircraft **CL checker(s):**	**Restrains:** • Use the active voice • Use articles wherever possible • Use simple verb tenses • Use language and terminology consistently • Avoid lengthy compound words • Use relatively short sentences **Classification:** Monolingual, HOCL **Classification (Pool, 2006):** Naturalistic/Restricted **Relevant literature:** Shaw (2006) **Comments:** Further development of AECMA Simplified English.

BASiC (Basic American Scientific International Commercial)	
Based on: **Language(s):** English **Language orientation:** Monolingual **Organization:** **Year of development:** 1930 (by Odgen) **Domain:**	**CL checker(s):** **Restrains:** 850-word simplified vocabulary **Classification:** **Classification (Pool, 2006):** **Relevant literature:** <http://ogden.basic-english.org/> **Comments**: This can be considered the first attempt in creating a controlled language.

BTE (Boeing Technical English)	
Based on: AECMA SE **Language(s):** English **Organization:** Boeing **Year of development: 1990** **Domain:** Technical documentation **CL checker(s):** Boeing Simplified English Checker (BSEC)	**Restrains:** **Classification:** Monolingual, HOCL **Classification (Pool, 2006):** Naturalistic/Generalistic **Relevant literature:** Wojcik, J. E. Hoard, & Holzhauser (1990); Wojcik & Holmback (1996); Wojcik, Holmback, & J. Hoard (1998) **Comments:**

BULL Controlled English	
Based on: **Language(s):** English **Language orientation:** Monolingual **Organization:** Groupe Bull (France) **Year of development:** **Domain:** Aircraft	**CL checker(s):** MaxIt Grammar Checker Spelling Checker **Restrains:** 10 generic rules **Classification:** Monolingual, MOCL **Classification (Pool, 2006):** **Relevant literature:** Lee (1993) **Comments:**

Cap Volmac Lingware Services	
Based on: **Language(s):** English/Dutch **Organization:** Cap Volmac Lingware Services **Year of development:** **Domain:** textile and insurance companies **CL checker(s):**	**Restrains:** **Classification:** **Classification (Pool, 2006):** **Relevant literature:** de Koning (1996); van der Steen & Dijenborgh (1992); Van der Eick, de Koning, & van der Steen (1996) **Comments:** This company has developed a series of controlled languages for textile and insurance companies

CASL (General Motor's Controlled Automottive Service Language)	
Based on: **Language** **Organization:** General Motors **Year of development:** **Domain:** Automotive **CL checker(s):**	**Restrains:** 62 rules **Classification:** Multilingual, HOCL **Classification (Pool, 2006):** Naturalistic/Restricted **Relevant literature:** Godden (1998); Means & Godden (1996) **Comments:**

CFE (Caterpillar Fundamental English)	
Based on: BASIC **Language(s):** English **Organization:** Caterpillar **Year of development:** 1972 (Kamprathetak98) **Domain:** Heavy Machinery **CL checker(s):**	**Restrains:** CFE provides a restricted vocabulary of around 850 words (initially) 1000 specialised vocabulary 1200 word vocabulary based on BASIC Intended as a Form of English as a Second Language for non-English speakers, who would be able to read the service manuals **Classification:** Monolingual, HOCL **Classification (Pool, 2006):** Naturalistic/Restricted **Relevant literature:** Kamprath, Adolphson, Mitamura, & Nyberg (1998) **Comments:**

COGRAM (Controlled English Grammar)	
Based on: AECMA SE, IBM Easy English, Ericsson **Language(s):** English **Organization:** Alcatel-Bell Company **Year of development:** **Domain:** Telecommunications **CL checker(s):**	**Restrains:** Grammar with 150 rules **Classification:** Multilingual, HOCL **Classification (Pool, 2006):** Naturalistic/Restricted **Relevant literature:** Adriaens & Scheurs (1992); Scheurs & Adriaens (1992) **Comments:** Paper grammar

Controlled Chinese	
Based on: **Language(s):** Chinese **Organization:** **Year of development:** **Domain:** General **CL checker(s):**	**Restrains:** **Classification:** Monolingual, MOCL **Classification (Pool, 2006):** Naturalistic/Generalistic **Relevant literature:** Zhang, Zhou, & Yu (1998) **Comments:**

Controlled Modern Greek	
Based on: **Language(s):** Greek **Organization:** **Year of development:** **Domain:** General **CL checker(s):**	**Restrains:** **Classification:** Monolingual, MOCL **Classification (Pool, 2006):** Naturalistic/Generalistic **Relevant literature:** Vassiliou, Markantonatou, Maistros, & Karkaletsis (2003) **Comments:**

CTE (Caterpillar Technical English)	
Based on: CFE **Language(s):** English **Organization:** Caterpillar **Year of development:** 1991-1997 **Domain:** Heavy Machinery **CL checker(s):**	**Restrains:** 65,000 CTE terms 130 rules **Classification:** Multilingual, HOCL **Classification (Pool, 2006):** **Relevant literature:** Kamprath, Adolphson, Mitamura, & Nyberg (1998) **Comments:** Related to Carniege Group, for KANT MT system

DCE (Diebold Controlled English)	
Based on: **Language(s):** English **Organization:** Diebold **Year of development:** **Domain:** Security **CL checker(s):** Diebold's Controlled Language Checker	**Restrains:** **Classification:** Multilingual **Classification (Pool, 2006):** **Relevant literature:** Moore (2000) **Comments:** Related to Carniege Group

Douglas Aircraft Simplified English	
Based on: **Language(s):** English **Organization:** McDonnel Douglas Corp. **Year of development:** 1979 **Domain:** Aircraft **CL checker(s):**	**Restrains:** 2000 words taken from the list of the preferred verbs used in the Navy, in the Air Force, and in McDonnell 50's technical manuals. This technical vocabulary was one of the sources studied for the creation of the AECMA SE lexicon. **Classification:** Monolingual, HOCL **Classification (Pool, 2006):** Naturalistic/Restricted **Relevant literature:** Gingras (1987); Hujisen (1998a; 1998b); Stewart (1998) **Comments:**

Ericsson English	
Based on: ILSAM **Language(s):** English **Language orientation:** **Organization:** Ericsson **Year of development:** **Domain:** Telecommunications	**CL checker(s):** COGRAM **Restrains:** 2-level lexicon: Level 1 documents might only contain those lexical terms that are marked 1, whereas Level 2 documents can be edited using a more extended vocabulary **Classification:** **Classification (Pool, 2006):** Naturalistic/Restricted **Relevant literature:** Scheurs & Adriaens (1992) **Comments:**

GIFAS Français Rationalisé	
Based on: **Language(s):** French **Organization:** Dassault Aerospace **Year of development:** 1985 **Domain:** Aircraft **CL checker(s):**	**Restrains:** **Classification:** Monolingual, HOCL **Classification (Pool, 2006):** Naturalistic/Restricted **Relevant literature:** Barthe et al., (1999) **Comments:**

GE (General Motors Global English)	
Based on: **Language(s):** English **Language orientation:** Monolingual **Organization:** General Motors **Year of development:** **Domain:** Automotive **CL checker(s):**	**Restrains:** 12 general rules **Classification:** **Classification (Pool, 2006):** **Relevant literature:** Allen (2004) **Comments:**

HELP (Hyster's Easy Language Program)	
Based on: PEP **Language(s):** English **Language orientation:** **Organization:** Rockwell International **Year of development:** **Domain:** Electric-powered lift trucks **CL checker(s):**	**Restrains**: 2,500-word Controlled English vocabulary **Classification:** **Classification (Pool, 2006):** **Relevant literature:** Smart (2003) **Comments:**

IBM EasyEnglish Language	
Based on: ILSAM **Language(s):** English **Organization:** IBM **Year of development:** **Domain:** Software **CL checker(s):** EasyEnglish Analizer	**Restrains:** Marginal ! Symbol indicates that the word has some restriction X symbol indicates a word to be avoided British Spelling **Classification:** Multilingual, MOCL **Classification (Pool, 2006):** Naturalistic/Generalistic (checker) **Relevant literature:** Bernth (1997; 1998; 2000; 2006) **Comments:**

ILSAM: White's International Language of Service and Maintenance	
Based on: CFE **Language(s):** English **Organization:** **Year of development:** **Domain:** **CL checker(s):**	**Restrains:** Around 800 words + specific terms **Classification:** Monolingual **Classification (Pool, 2006):** **Relevant literature:** Reference in Kaji (1999) **Comments:**

KANT Controlled English	
Based on: **Language(s):** English **Language orientation:** **Organization:** Carnegie Mellon University **Year of development:** 1989 **Domain:** Heavy equipment **CL checker(s):**	**Restrains:** **Classification:** Monolingual, HOCL **Classification (Pool, 2006):** Naturalistic/Generalistic **Relevant literature:** Mitamura & Nyberg (1995); Nyberg et al. (1998); Nyberg & Mitamura (1996) **Comments:**

KISL (Kodak International Service Language)	
Based on: **Language(s):** English **Language orientation:** Multilingual **Organization:** Kodak **Year of development:** **Domain:** **CL checker(s):**	**Restrains:** Fewer than 1100 words **Classification:** Monolingual, MOCL **Classification (Pool, 2006):** **Relevant literature:** Muldoon (1999) **Comments:**

Langage Documentaire Canonique	
Based on: **Language(s):** English, German, Spanish French **Language orientation:** **Organization:** **Year of development:** **Domain:** Textile **CL checker(s):**	**Restrains:** **Classification:** Multilingual, MOCL **Classification (Pool, 2006):** **Relevant literature:** Ducrot (1984) in Adriaens & Scheurs (1992) **Comments:** Implemented together with MT system TITUS

LinguaNet	
Based on: Airspeak, Seespeak, Policespeak, INTACOM **Language(s):** English/French **Organization:** Prolingua /Channel Police **Year of development:** 1994 to 1998 **Domain:** A specially designed, messaging system for cross border, mission critical operational communication by police, fire, ambulance, medical, coastguard, disaster response coordinators. **CL checker(s):**	**Restrains:** **Classification:** Multilingual, MOCL **Classification (Pool, 2006):** Naturalistic/restrictive **Relevant literature:** Johnson (2000) <http://www.prolingua.co.uk/Linguanet/index.html> **Comments:**

MCE (Multinational Customized English)	
Based on: ILSAM **Language(s):** English **Language orientation:** **Organization:** Xerox Corporation **Year of development:** 1978 **Domain:** **CL checker(s):**	**Restrains:** A vocabulary customized for the company's technology and products Clear, simple, logical writing style Rules of grammar aligned to the translation software rules Avoidance of ambiguous words, expressions, and sentence structure **Classification:** Multilingual, MOCL **Classification (Pool, 2006):** **Relevant literature:** Elliston (1979); Ruffino (1982)**;** Adams, Austin, & M. Taylor (1999) **Comments:** Uses Systran and ALPS in conjunction with a Controlled Language Input

OCÉ Controlled English	
Based on: **Language(s): English** **Organization:** Océ **Year of development:** 1999 **Domain:** Printers	**CL checker(s):** **Restrains:** **Classification:** Multilingual **Classification (Pool, 2006):** **Relevant literature:** Cremers (2003); Cucchiarini (2002) **Comments:**

NSE Nortel Standard English	
Based on: advanced version of ASD-STE100 Simplified Technical English (STE) **Language(s):** English **Language orientation:** Multilingual **Organization:** Nortel **Year of development:** 1995 **Domain:** Telecommunications equipment **CL checker(s):** MaxIt Checker	**Restrains:** Little over a dozen rules **Classification:** Monolingual **Classification (Pool, 2006):** **Relevant literature:** J.M. Smart (2006) **Comments:**

ORACAL (Oracle Controlled Language)	
Based on: **Language(s):** English **Organization:** Oracle **Year of development:** **Domain:** **CL checker(s):**	**Restrains:** **Classification:** Multilingual **Classification (Pool, 2006):** **Relevant literature:** J. Allen (2004) **Comments:**

PACE (Perkins Approved Clear English)	
Based on: ILSAM CTE (see Newton 92: 47) **Language(s):** English **Organization:** Perkins International Limited **Year of development:** 1980 (Huijsen 98: 45) **Domain:** Heavy Machinery (manufacturer of Diesel engines) **CL checker(s):** **Restrains:** PACE guidelines: 1. Keep sentences short 2. Omit redundant words 3. Order the parts of the sentence logically 4. Don't change constructions mid-sentence 5. Take care with logic of 'and' and 'or' 6. Avoid elliptical constructions	**Classification:** Monolingual **Classification (Pool, 2006):** Naturalistic/Generalistic **Relevant literature:** Douglas & Hurst (1996); Newton (1992)**;** (Hujisen, 1998a; Schwarze, 2001) **Comments:** Consists primarily of a single wordlist (2500 words in 1990, 10% verbs) plus ten very general rules of writing MT system used: Weidner's MicroCat

PEP (Smart's Plain English Program)	
Based on: CFE **Language(s):** English **Organization:** **Year of development:** **Domain:** **CL checker(s):**	**Restrains:** **Classification:** **Classification (Pool, 2006):** **Relevant literature:** Reference in Kaji (1999) **Comments:**

Plain Japanese	
Based on: **Language(s):** Japanese **Organization:** **Year of development:** **Domain:**	**CL checker(s):** **Restrains:** **Classification:** **Classification (Pool, 2006):** Naturalistic/Generalistic **Relevant literature:** Sato, Utusro, Tsuchinya, Asaoka, & Matsuyoshi (2004) **Comments:**

Rockwell International	
Based on: PEP **Language(s):** English **Organization:** Rockwell International's Automotive Division **Year of development:** **Domain:** **CL checker(s):**	**Restrains:** **Classification:** **Classification (Pool, 2006):** **Relevant literature:** Scheurs & Adriaens (1992) **Comments:** Uses MaxTrans (Smart Communications, AMTA 1994)

Scania Swedish	
Based on: Scania **Language(s):** Swedish **Language orientation:** **Organization:** **Year of development:** **Domain:** **CL checker(s):**	**Restrains:** **Classification:** Multilingual **Classification (Pool, 2006):** **Relevant literature:** Almqvist & Sångall Hein (1996) **Comments:**

SDD (Siemens Dokumentationsdeutsch) CSDG (Controlled Siemens Documentary German)	
Based on: Siemens **Language(s):** German **Language orientation:** **Organization:** **Year of development:** **Domain:** **CL checker(s):**	**Restrains:** **Classification:** **Classification (Pool, 2006):** Naturalistic/Generalistic **Relevant literature:** Lehrndorfer (1996) **Comments:**

SeaSpeak	
Based on: **Language(s):** English **Organization:** **Year of development:** **Domain:** **CL checker(s):**	**Restrains:** **Classification:** **Classification (Pool, 2006):** Naturalistic/Restricted **Relevant literature:** Kimbrough, T. Y. Lee, Padmanabhan, & Yang (2004) **Comments:**

Simplified Technical Spanish	
Based on: **Language(s):** Spanish **Organization:** **Year of development:** **Domain:** **CL checker(s):**	**Restrains:** **Classification:** Monolingual **Classification (Pool, 2006):** Naturalistic/Restricted **Relevant literature:** Ruiz Cascales & Sutcliffe (2003) **Comments:**

Smart Controlled English	
Based on: **Language(s):** English **Organization:** Smart NY **Year of development:** **Domain:** Financial systems and banking, automotive and capital equipment, medican and measuring equipment etc. **CL checker(s):**	**Restrains:** SMART Controlled English (CE) is a technical vocabulary of approximately 1,200 basic words, plus product terminology. **Classification:** Monolingual **Classification (Pool, 2006):** **Relevant literature:** J.M. Smart (2006) **Comments:**

Standard Language	
Based on: **Language(s):** English **Organization:** Ford **Year of development:** **Domain:** **CL checker(s):**	**Restrains:** **Classification:** **Classification (Pool, 2006):** **Relevant literature:** Rychtyckyj (2002; 2006b) **Comments:**

STE (Simplified Technical English) ASD-STE100	
Based on: AECMA STE **Language(s):** English **Organization:** European Aeronautic Defence and Space Company (EADS) Rolls-Royce Saab Systems Boeing **Year of development:** **Domain:** **CL checker(s):**	**Restrains:** **Classification:** **Classification (Pool, 2006):** **Relevant literature:** Macdonald (2008) **Comments:**

Sun Controlled English	
Based on: Sun **Language(s):** English **Language orientation:** Multilingual **Organization:** Sun Microsystems **Year of development:** **Domain:** **CL checker(s):**	**Restrains:** **Classification:** **Classification (Pool, 2006):** Naturalistic/Restricted **Relevant literature:** O'Brien (2003b); Wells-Akis & Sisson (2002) **Comments:**

WebTran	
Based on: **Language(s):** Swedish **Organization:** **Year of development:** **Domain:** **CL checker(s):**	**Restrains:** **Classification:** MOCL **Classification (Pool, 2006):** Naturalistic/Realistic **Relevant literature:** Lethola, Tenni, & Bounsaythip (1998); Lethola, Tenni, Bounsaythip, & Jaaranen (1999) **Comments:**

ANNEX II: CL COMPLIANCE AND LINGUISTIC ANALYSIS

Between the end of April 2005 and the beginning of May 2005, a batch analysis of following document packages was carried out:

Year	SBT	SI
2002	112	158
2003	483	74
2004	269	68
2005 (until 20.04.05)	81	9
TOTAL	945	309
Model	**RA**	**TNU**
E65 MÜ	114	
E83	527	
E87	618	
E90	320	32
E91 (until 20.04.05)	50	133
TOTAL	1629	165

Table 34: Document types (figures)

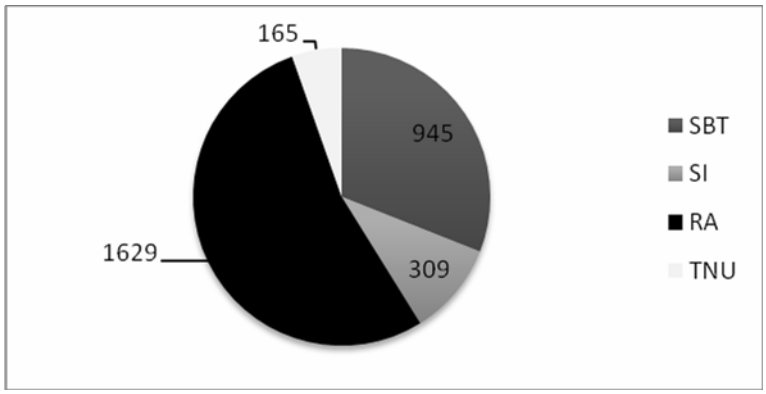

Figure 54: Document type distribution for the analysis

These documents were analysed by the linguistic engine of MULTILINT/CLAT with respect to orthography, terminology, abbreviations, term candidates, style and grammar.

It is important to point out, before going on with the analysis of the data, that the figures of this study have been directly extracted from the batch analysis, without evaluating error messages. Thus it must be taken into account that the data are not 100% reliable since precision and recall do not match. This is due to wrong error messages or errors that have not been flagged.

Besides, the figures delivered by the batch analysis are absolute, that is, they do not take into account repetitions. If a terminology error occurs twice, it will be showed and counted twice.

However, as we consider that these "emissions" or "over generations" have been produced in all documents, we assume that the results compensate and represent a true image of reality. Indeed, the results obtained and analysed taking into account extrinsic and intrinsic factors of the text production seem to confirm this thesis.

Analysis

After extracting the number of errors for each category for each document type, the relative frequencies of the packages have been calculated, taking as a basis a norm line

(55 characters). By this method, we assume that a line can contain at the most one type of error (e.g. per line only one terminology error, one grammar error, etc.). We are aware that this is not always the case. However, we think that figures compensate since there are error types which strike much less frequently than others.

Document types

As it can be observed in the following chart, the document type that presents the highest frequency of errors is the Teilnehmerunterlagen (TNU), with a frequency of errors of 80,62% in 6,125 lines of text. Indeed, this is not a surprising result, since TNUs and in general training documentation are not checked consistently with MULTILINT/CLAT. Only some authors do it from time to time voluntarily and the terminology is not as controlled as in other document production areas, such as RAs.

	SBTs		SIs		RAs		TNUs		
Lines	66,314.78		16,489.50		24,534.65		6,125.89		
	Absolute	Relative	Absolute	Relative	Absolute	Relative	Absolute	Relative	TOTAL
Orthography	2,226	3.36%	1,386	8.41%	530	2.16%	454	7.41%	21.33%
Terminology	6,792	10.24%	2,397	14.54%	2,599	10.59%	1,141	18.63%	54.00%
Abbreviations	9,543	14.39%	1,686	10.22%	490	2.00%	754	12.31%	30.48%
Term candidates	937	1.41%	913	5.54%	440	1.79%	237	3.87%	12.61%
Style	2,135	3.22%	1,479	8.97%	426	1.74%	456	7.44%	21.37%
Grammar	932	1.41%	525	3.18%	114	0.46%	1,905	31.10%	36.15%
		34.03%		50.86%		18.74%		80.76%	

Table 35: Relative Frequencies per Document Package and per Control

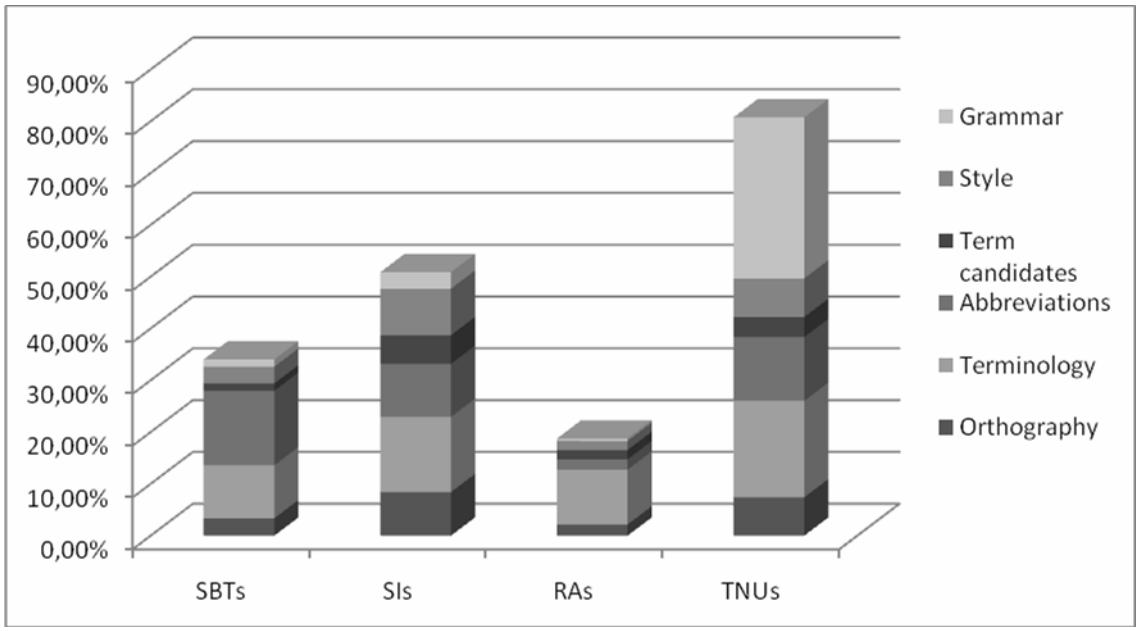

Figure 55: Relative Frequencies per Document Type and per Control

Therefore, authors do not experience a learning process and do not adapt their writing to CLAT's rules. Contrarily, RAs, where authors use MULTILINT/CLAT regularly as part of the authoring process, present an insignificant frequency of errors. SBTs also present low figures.

A striking difference of the TNUs with respect to the other types of documents is the great number of grammar errors. However, after revising the results of the analysis, we have concluded that many of these mistakes are false messages due to fail parses, or due to inconsistencies in the terminology.

Terminology and Abbreviations represent the most important types of errors in all documents. However, here again we must be careful when interpreting the results. Many messages are not necessarily errors, but indications for the author. For instance, the terminology code POSNEG indicates a term which is preferred in a certain domain, but deprecated in another one. Since the system is not able to distinguish to which domain the document belongs, it is not possible to know if these are real errors or not.

It is also interesting to observe that, despite the fact that RAs have been consistently produced with CLAT for nearly 5 years and the rest of controls are quite humble, the number of terminology errors, especially in the "deprecated"-category, is quite meaningful, even in new produced documents, such as the RAs for the series E91. This might indicate that, although the authors use this tool regularly, terminology is one of the most difficult areas to be controlled by humans. The implementation of CLAT is therefore especially important in this case, since terminology is one of the most problematic issues, both for the comprehension of the text and for the translation (including economic factors).

	SBT 2002	SBT 2003	SBT 2004	SBT 2005	SI 2002	SI 2003	SI 2004	SI 2005	RA E65 MÜ	RA E83	RA E87	RA E90	RA E91	TNU E90	TNU E91
Orthography	10.39%	3.85%	4.20%	3.50%	9.71%	7.00%	6.38%	10.87%	8.72%	1.29%	1.57%	1.77%	3.24%	6.79%	7.56%
Terminology	15.33%	9.35%	10.09%	11.80%	18.49%	11.04%	8.95%	9.91%	16.12%	10.39%	9.39%	10.21%	16.36%	25.37%	17.02%
Abbreviations	12.57%	13.99%	15.38%	13.32%	11.45%	7.83%	10.08%	9.91%	2.83%	1.94%	1.70%	2.21%	2.92%	15.36%	11.58%
Term candidates	5.09%	0.89%	1.50%	0.90%	6.27%	4.21%	4.81%	9.18%	3.65%	1.63%	1.61%	1.55%	2.27%	5.43%	3.50%
Style	5.79%	3.36%	2.92%	1.76%	10.53%	6.62%	7.01%	15.22%	3.65%	1.84%	1.49%	1.29%	1.30%	8.23%	7.26%
Grammar	2.09%	1.12%	1.63%	1.33%	4.14%	2.40%	1.82%	1.45%	0.34%	0.63%	0.43%	0.35%	0.32%	4.24%	2.87%
TOTAL	51.25%	32.57%	33.07%	32.61%	60.60%	39.10%	39.05%	56.53%	35.31%	17.72%	16.19%	17.38%	26.41%	65.43%	49.78%

Table 36: Relative Frequencies per Document Package and per Control

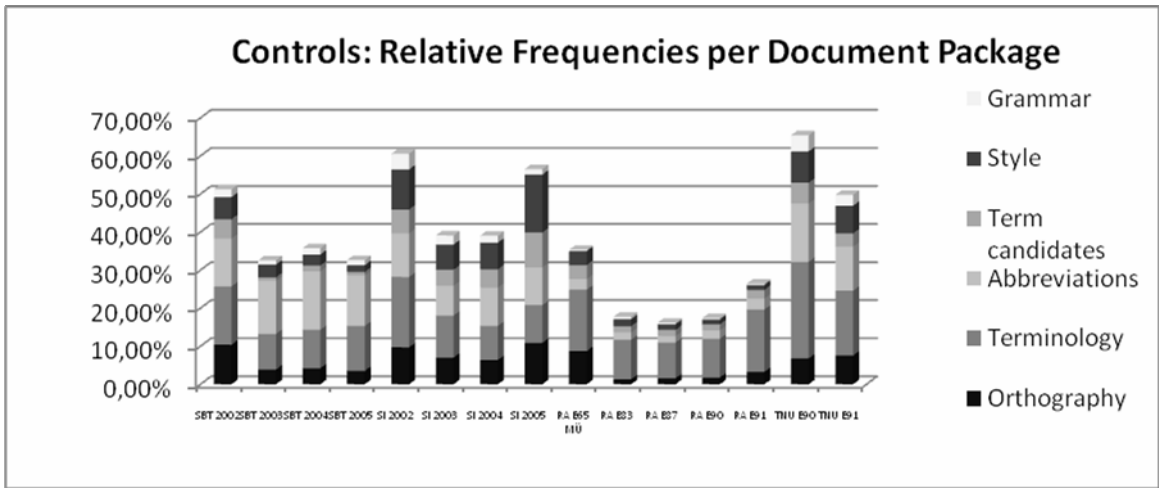

Figure 56: Relative Frequencies per Document Package

Here we can observe how the different document packages behave with respect to the different controls. In the case of the SBTs, a significant reduction of error messages can be observed from 2002 to 2003. Afterwards, the figures seem to stabilize, appearing in the SBTs of 2005 errors with a frequency of around 30%. However, this figure is based on documents produced until the end of April 2005 and it is not certain if the trend has been maintained or if the number of errors has increased or decreased.

In the case of SIs, we observe a similar development. After a reduction of error frequencies from 2002 to 2003, figures seem to be relative constant during until 2004. However, the number of errors in 2005 is strikingly high. We could argue that this is due to new technology, new models and thus, new terminology and abbreviations. Term candidates have indeed grown, but so have style errors and orthography errors. This could be due to new authors (trainees). It is also possible that, as argued in the case of SBTs, this trend changes when taking into account the results of the whole year.

RA is the document type that presents the less number of errors. Again, this result is not surprising since this is the only information type where MULTILINT/CLAT has been consistently used for more than 4 years. There is an obvious learning effect by the

authors, who adapt their writing to the controlled language checked with MUTILINT/CLAT. For this reason, this is the most appropriate information type for the pre-selection phase of our study.

Error types

With regards to the error types, we find that terminology errors, followed by abbreviations, grammar, style, orthography and term candidates.

	SUM
Orthography	21.33%
Terminology	54.00%
Abbreviations	30.48%
Term candidates	12.61%
Style	21.37%
Grammar	36.15%

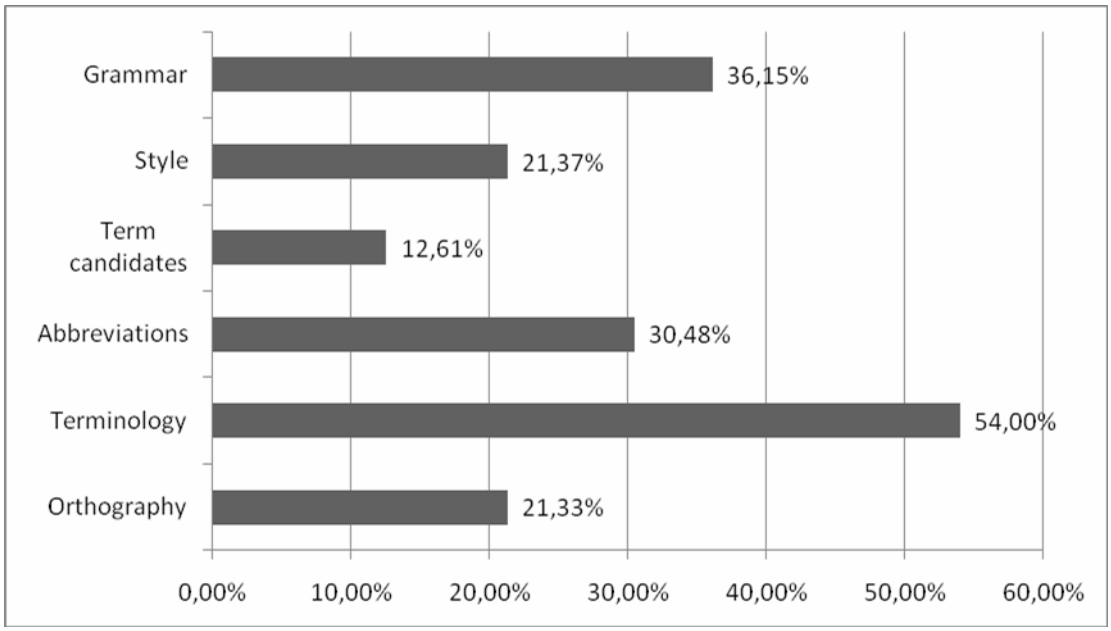

Figure 57: Relative Error Frequencies per Category (all documents)

Terminology errors are especially frequent in the area of deprecated terms, that is, terms that are not allowed within a concept. This error code is closely followed by the code POSNEG. This is the case when terms are preferred in a certain domain, but deprecated in another one. As mentioned before, however, we cannot be sure if all these are real errors or just indications for the authors that are not necessarily applicable. The codes DEFTERM (orthographic variants not present in the terminology data-base) and VARPOSNEG (variants that coincide with a preferred or a deprecated term in the terminology) follow at a certain distance.

Style rules and Translatability

Especially interesting in this context is to study the style rules. In 2003 Julia Reuther published an article in which style rules where measured regarding their degree of human translatability (Reuther, 2003). In this study, Reuther let style rules to be evaluated by professional translators and native speakers with regards to their effect on translation. Evaluators had to give the rule a priority of 1 to 3. Priority 1 meant high importance for translation, 2 medium importance, and 3, less importance. Rules could also be pointed out as "irrelevant" with an x.

Rules in MULTILINT/CLAT are organized in 7 different categories:

- Typographic rules
- Avoidance of ambiguous structures
- Lexical rules
- Avoidance of elliptical structures
- Avoidance of complex structures
- Rules regarding word-order and sequence of sentence chunks
- Stylistic rules

- Company-specific rules

Of a total of 91 rules, these are as distributed as follows:

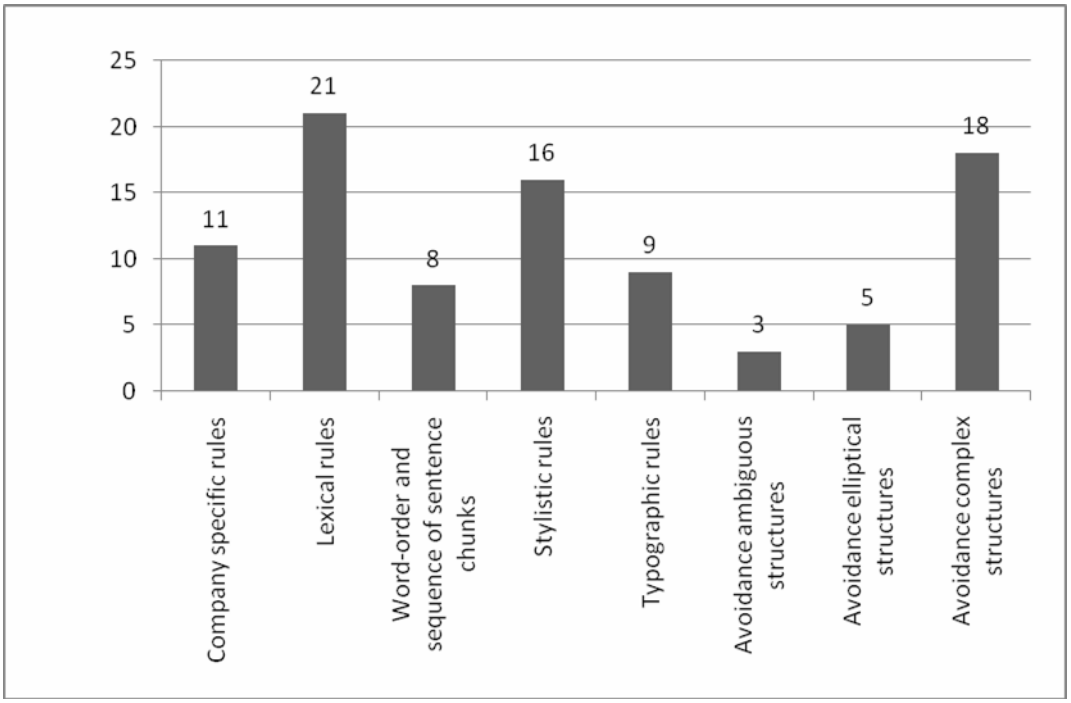

Figure 58: Style Rules in MULTILINT/CLAT

As we can observe, most rules pertain to the category “lexical rules” that is, controlled vocabulary. Another important pillar is the category “avoidance of complex structures”. Indeed, as we will see in the next chart, this category seems to play an important role with respect to translatability:

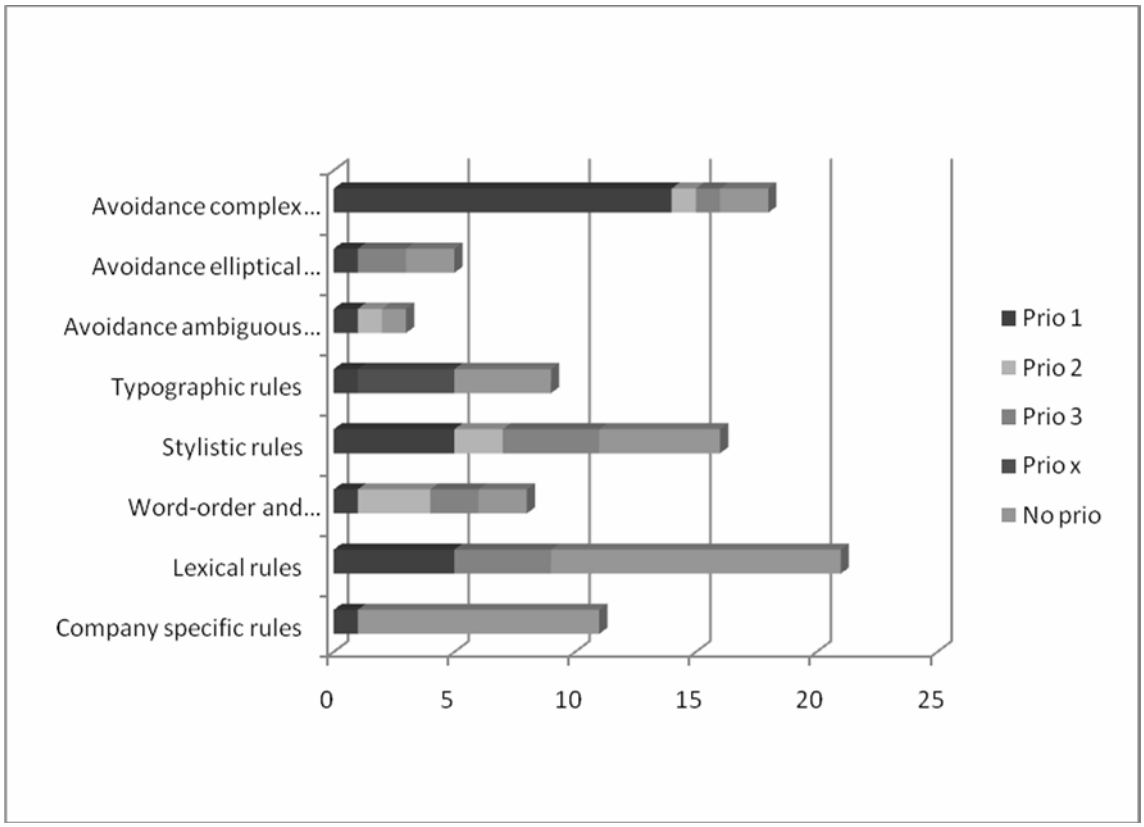

Figure 59: Style Rules in MULTILINT/CLAT

We can observe here that the category "avoidance of complex structures" has the bigger number of rules with prio 1. This category is followed by "stylistic rules" and "lexical rules". In the latter case, however, we have to take into account that many of the rules (57%) have not been prioritised. This will be done in the evaluation phase of our study. Rules will be evaluated with respect to machine translatability. In this respect we expect to confirm the human prioritisation made by Reuther and to find certain differences, especially in certain categories such as typography, where rules seem to have especial important for MT (Bernth & Gdaniec, 2001; Grasse, 2001).

Conclusion

After a detailed analysis of the results of the batch analysis of more than 3000 documents of different information types, we can draw the following conclusions:

- TNUs is the information type which presents the highest frequency of errors. This is due to the absence of a consistent linguistic quality assurance process.

- RAs is the information type which presents the lowest frequency of errors. This information type has been consistently checked with MULTILINT/CLAT for more than 4 years now. Authors have learned how to write in controlled German and to stick to the CLAT rules.

- The most frequent error category is terminology. Many errors are produced in the error code "deprecated term". This could be due to the always-increasing number of new terms and to the reduced human capacity to remind which terms are preferred in a certain context. Usually, humans tend to use synonyms and variants when writing texts to avoid monotony. However, this is not desirable from a language processing point of view. In this sense, the implementation of MULTILINT/CLAT is absolutely useful and necessary in order to maintain a consistent terminology that fosters comprehensibility and translatability.

- Most style errors are produced in the category of "avoidance of complex structures", which have, at the same time, the highest priority for translatability. In this sense, the implementation of MULTILINT/CLAT in this case is also indispensable in order to avoid unclear structures that present a hurdle for readability, comprehensibility and, thus, translatability.

-

ANNEX III: TRANSLATABILITY CRITERIA

Sources:

Bernth, A., & Gdaniec, C. (2001). MTranslatability. *Machine Translation*, *16*(3), 175-218.

Gdaniec, C. (1994). The Logos Translatability Index (págs. 97-105). Columbia, Maryland.

Grasse, N. (2001, Octubre). *Qualitätskontrolle des MÜ-Systems DCINTRANS in der Anwendung des Sprachendienstes der DaimlerChrysler AG* (Diplomarbeit (Master Thesis)). Universität des Saarlandes. (pp. 90-94).

Reuther, U. (2003). Two in one: Can it work? Readability and Translatability by means of Controlled Language (págs. 124-132). Dublin.

Underwood, N. L., & Jongejan, B. (2001). Translatability Checker: A Tool to Help Decide Whether to Use MT (págs. 363-368). Santiago de Compostela, Spain.

Category	Rule	Author	Rule	Author	Rule	Author	Rule	Author
Formal, Formatting, Punctuation, Layout...	Parentheses (short, unmatched)	(Gdaniec, 1994)	Rule 19: Do not include parenthesized expressions in segment unless the segment is still valid syntactically when you remove the parentheses while leaving the parenthesized expressions.	(Bernth & Gdaniec, 2001)	Avoid complete sentences in brackets Avoid parenthesis starting with d.h. (corresponding i.e.)	(Reuther, 2003)		
Formal, Formatting, Punctuation, Layout...	Punctuation marks	(Reuther, 2003)	Rule 20: Use punctuation prudently.	(Bernth & Gdaniec, 2001)	Auf korrekte Zeichensetzung achten: auf eine richtige Interpunktion achten!	(Grasse, 2001)		
Formal, Formatting, Punctuation, Layout...	Spacing	(Reuther, 2003)	Leerzeichen: Überflüssige bzw. fehlende Leerzeichen vermeiden	(Grasse, 2001)	Nicht zulässige Formatierungen: Formatierungen durch manuell eingefügte Returns-, Leerzeichen, Tabulatoren oder Zeilenumbrüche führen zu Ungenauigkeiten in der Segmentierung.			

Category	Rule	Author	Rule	Author	Rule	Author	Rule	Author
Formal, Formatting, Punctuation, Layout...	Typographic elements (e.g. lists)	(Reuther, 2003)	Formale Gestaltung von Aufzählungen	(Grasse, 2001)	Aufbau von Aufzählungen: Keine Trennung des Einleitungstexts der Aufzählung.			
Formal, Formatting, Punctuation, Layout...	Hervorhebung durch Sperren vermeiden	(Grasse, 2001)						
Formal, Formatting, Punctuation, Layout...	Rule 21: Avoid using (s) to indicate plural	(Bernth & Gdaniec, 2001)	Avoid additional plural forms in brackets (Translation Memory)	(Reuther, 2003)				
Formal, Formatting, Punctuation, Layout...	Rule 22: Avoid using / as in and/or and user/system	(Bernth & Gdaniec, 2001)						
Formal, Formatting, Punctuation, Layout...	Rule 23: Check your spelling	(Bernth & Gdaniec, 2001)	Rechtschreibung: Stellen Sie sicher, dass ihr ausgangssprachlicher Text keine Rechtschreibfehler enthhält-	(Grasse, 2001)				

Category	Rule	Author	Rule	Author	Rule	Author	Rule	Author
Formal, Formatting, Punctuation, Layout...	Rule 18: Avoid footnotes in the middle of a segment, and make footnotes independent segments.	(Bernth & Gdaniec, 2001)						

Category	Rule	Author	Rule	Author	Rule	Author	Rule	Author
Lexical Ambiguity Terminology	Spelling variants: Lambdasonde vs. Lambda-Sonde	(Reuther, 2003)	Komposita: Deutsche Komposita werden entweder zusammen oder mit Bindestrich geschrieben.	(Grasse, 2001)				
Lexical Ambiguity Terminology	Morphological variants: Ankühlungsvorgang vs. Abkühlvorgang	(Reuther, 2003)	Auf Groß- und Kleinschreibung bei Anrede achten.	(Grasse, 2001)				
Lexical Ambiguity Terminology	Synonym Variants: Kältetest vs. Käteprüfung	(Reuther, 2003)						
Lexical Ambiguity Terminology	Avoid ambiguous genitive constructions	(Reuther, 2003)						
Lexical Ambiguity Terminology	Einheitliche Terminologie verwenden	(Grasse, 2001)	Standardabkürzungen oder gar keine Abkürzungen verwenden	(Grasse, 2001)				
Lexical Ambiguity Terminology	Großschreibung vermeiden (sie werden als Akronyme erkannt und nicht übersetzt)	(Grasse, 2001)						

Category	Rule	Author	Rule	Author	Rule	Author	Rule	Author
Lexical Ambiguity Terminology	Mischtexte vermeiden (Deutsch-Englisch)	(Grasse, 2001)						

Category	Rule	Author	Rule	Author	Rule	Author	Rule	Author
Syntax: Grammatical Ambiguity Complexity	Ambigous structures	(Reuther, 2003)	Zweideutige Satzkonstruktionen vermeiden	(Grasse, 2001)				
Syntax: Grammatical Ambiguity Complexity	Präpositionen: Präpositionen, die sowohl vor als auch nach dem Wort stehen können, werden von System besser übersetzt, wenn sie vor dem Wort stehen.	(Grasse, 2001)	PPs and/or subclases	(Underwood & Jongejan, 2001)				
Syntax: Grammatical Ambiguity Complexity	Pronouns	(Reuther, 2003)	Rule 8: Minimize use of personal pronouns	(Bernth & Gdaniec, 2001)	Pronomina: Das System kann nur satzweise Zusammenhänge herstellen. Versuchen Sie, eindeutige Bezüge herzustellen.	(Grasse, 2001)		

Category	Rule	Author	Rule	Author	Rule	Author	Rule	Author
Syntax: Grammatical Ambiguity Complexity	Complexity	(Reuther, 2003)	Einfacher Satzbau: klare, einfache Satzstruktur; Vermeidung verschachtelte Konstruktionen und unnötig lange Sätze; teilen Sie Säatzen wenn möglich in mehrere Einzelsätze auf; einfache Subjekt-Objekt-Struktur; vermeiden so weit wie möglich Bandwurm- und Schachtelsätze	(Grasse, 2001)				
Syntax: Grammatical Ambiguity Complexity	Telegrammstil vermeiden: vermeiden Sie unvollständige Sätze	(Grasse, 2001)	Short sentence (< 3 words) Long sentence (> 25 words)	(Underwood & Jongejan, 2001)	Rule 13: Avoid overly long sentences and very short sentences	(Bernth & Gdaniec, 2001)		

Category	Rule	Author	Rule	Author	Rule	Author	Rule	Author
Syntax: Grammatical Ambiguity Complexity	Order of Elements In a condition/Action sentence, the condition part should precede the action part (circunvent grammatical parsing problems)	(Reuther, 2003)	Nomina/Substantive: Mit verschiedenen Konstruktionen bzw. Wörtern, die dem Substantiv vorausgehen, hat das System in bestimmten Kontexten Übersetzungsproblemme. Adverbien: Adverb am Anfang des Satzes, inbesondere wenn es sich um eine negierte Aussage handelt.	(Grasse, 2001)				
Syntax: Grammatical Ambiguity Complexity	Einschübe als eingenständige Sätze formulieren: Formen Sie aus Einschüben eigenständige Sätze.	(Grasse, 2001)						
Syntax: Grammatical Ambiguity Complexity	Relativsätze nicht zu komplex gestalten.	(Grasse, 2001)	Rule 6: Do not omit relative pronouns; write "that" (which, who, etc.) explicitely					
Syntax: Grammatical Ambiguity Complexity	Rule 1: Avoid ungrammatical constructions	(Bernth & Gdaniec, 2001)						

Category	Rule	Author	Rule	Author	Rule	Author	Rule	Author
Syntax: Grammatical Ambiguity Complexity	Coordination: Repeat final words of the left conjunct or initial words of the right conjunct, as necessary, to disambiguate coordination.	(Bernth & Gdaniec, 2001)	Multiple coordination	(Underwood & Jongejan, 2001)				
Syntax: Grammatical Ambiguity Complexity	Rule 7: Avoid post-modifying adjective phrases	(Bernth & Gdaniec, 2001)						
Syntax: Grammatical Ambiguity Complexity	Rule 9: Always write the complementizer that explicitely	(Bernth & Gdaniec, 2001)						
Syntax: Grammatical Ambiguity Complexity	Rule 10: Avoid long noun phrases, if possible	(Bernth & Gdaniec, 2001)	One or more nominal compunds (> 2 nouns)	(Underwood & Jongejan, 2001)				
Syntax: Grammatical Ambiguity Complexity	Rule 12: Use one-word verbs instead of verb-particle whenever possible	(Bernth & Gdaniec, 2001)						

Category	Rule	Author	Rule	Author	Rule	Author	Rule	Author
Syntax: Grammatical Ambiguity Complexity	Verben: Verben, die im Deutschen sowohl mit als auch ohne "es"-Korrelat verwendet werden können, sollten stets in der Form mit "es"-Korrelat geschrieben werden. Verbteile nicht trennen: Versuchen Sie, Ihre Sätze so zu formulieren, dass die Verbteile möglichst nahe beianderstehen. Imperativ: Wenn Sie Befehlsformen verwenden, formulieren Sie den Imperativ bitte als reinen Imperativ und vermeiden den Einsatz des Infinitivs, um eine Anordnung bzw. Aufforderung auszudrücken	(Grasse, 2001)			.			

Category	Rule	Author	Rule	Author	Rule	Author	Rule	Author
Style	Rule 14: Avoid metaphors, idioms, slang, and dialect.	(Bernth & Gdaniec, 2001)	Idiomatische Redewendungen vermeiden; Umgangsprachliche Formulierungen vermeiden	(Grasse, 2001)				
Style	Rule 15: Avoid elipsis	(Bernth & Gdaniec, 2001)	Verbauslassungen in langen Sätzen vermeiden	(Grasse, 2001)	No verb present No finitive verb present	(Underwood & Jongejan, 2001)	Elliptical constructions	(Reuther, 2003)
Style	Rule 16: Avoid passive constructions, if possible.	(Bernth & Gdaniec, 2001)	General stylistic recommendations: passive, future tense, negation etc.	(Reuther, 2003)	Akriv/Passiv: Verwenden Sie wenn möglich Aktivkonstruktione n Fragesätze: Vermeiden Sie umgangsprachliche Formen der Frageststellung	(Grasse, 2001)		
Style	Rule 17: Make sure that each segment can stand alone syntactically.	(Bernth & Gdaniec, 2001)						

Category	Rule	Author	Rule	Author	Rule	Author	Rule	Author
Style	Avoid double negation (circunvent parsing problems)	(Reuther, 2003)	Negation: Konstruktionen zu vermeiden: wenn das negierte Pronomen "keine(r,s)" in substantivischer Objektposition steht oder wenn ein Satz negiert wird und gleichzeitig das Adverb "sehr" verwendet wird.	(Grasse, 2001)				

Table 37: Translatability Criteria by author

Formal Rules			
	Punctuation		
		Parentheses	(Bernth & Gdaniec, 2001; Gdaniec, 1994; Reuther, 2003)
		Punctuation marks	(Bernth & Gdaniec, 2001; Grasse, 2001; Reuther, 2003)
	Formatting		
		Lists	(Grasse, 2001; Reuther, 2003) TETRIS
		Spacing Hervorhebung durch Sperren (space out)	(Grasse, 2001; Reuther, 2003) TETRIS
		Plural forms in brackets	(Bernth & Gdaniec, 2001; Reuther, 2003)
		Use of / as and-or	(Bernth & Gdaniec, 2001)
	Layout		
		Footnotes	(Bernth & Gdaniec, 2001)
	Ortography		
		Spelling	(Bernth & Gdaniec, 2001; Grasse, 2001) TETRIS

Terminology			
	Variants		
		Spelling Variants: Compounds (Schreibvarianten, Bindestrichvarianten, Zahlen); Großschreibung vermeiden wenn kein Akronym	(Grasse, 2001; Reuther, 2003) TETRIS
		Morphological Variants	(Grasse, 2001; Reuther, 2003)
		Synonym Variants	(Reuther, 2003) TETRIS
		Einheitliche Terminologie	(Grasse, 2001)
		Abbreviations and Acronyms	(Grasse, 2001) TETRIS

Grammar			
	Ungrammatical Constructions		(Bernth & Gdaniec, 2001)
	Noun phrases		(Bernth & Gdaniec, 2001)
	Verbs		
		One-word Verbs instead of verb-particles	(Bernth & Gdaniec, 2001)
		Imperative	(Grasse, 2001)
		"es"-Korrelat (German)	(Grasse, 2001)
	Syntax		
		Ambiguous structures	(Grasse, 2001; Reuther, 2003)
		Prepositions	(Grasse, 2001; Underwood & Jongejan, 2001)

Grammar			
			TETRIS
		Pronouns	(Bernth & Gdaniec, 2001; Gdaniec, 1994; Reuther, 2003)
		Articles	TETRIS
		Complexity	(Grasse, 2001; Reuther, 2003)
		Order of Elements	(Grasse, 2001; Reuther, 2003)
		Subordinate and Relativ Clauses	(Bernth & Gdaniec, 2001; Grasse, 2001)1
		Sentence length	TETRIS (Bernth & Gdaniec, 2001; Underwood & Jongejan,

Grammar			
			2001)
		Coordination	(Bernth & Gdaniec, 2001; Underwood & Jongejan, 2001) TETRIS

Style			
	Elliptical constructions		(Bernth & Gdaniec, 2001; Grasse, 2001; Reuther, 2003; Underwood & Jongejan, 2001)
	Passive constructions		(Bernth & Gdaniec, 2001; Gdaniec, 1994; Reuther, 2003)
	Methaphors, idioms, slang, dialect		Bernth & Gdaniec, 2001; Grasse, 2001)
	Negation		(Grasse, 2001; Reuther, 2003)

Table 38: Translatability criteria by type

ANNEX IV: FEMTI EVALUATION PLAN

EVALUATION TYPE

- Declarative evaluation: The purpose of declarative evaluation is to measure the ability of an MT system to handle texts representative of an actual end-user. It is concerned with coverage of linguistic phenomena and handling of samples of real text. Declarative evaluations generally test for the functionality attributes of intelligibility, (how fluent or understandable it appears to be) and fidelity (the accurateness and completeness of the information conveyed).

CONTEXT CHARACTERISTICS

- Machine translation user: This refers to the person who interacts with the machine translation system and with the output produced by it.

- Genre: Genre refers to the characteristic or definitive form and style peculiar to a type of document. Examples of genre are: newspaper articles; scientific and technical articles; recipes and instructions; correspondence; business/commercial reports; marketing texts and advertisements; legal texts; literature: novels, poetry, etc.; and many others.

- Organisational user: An organisational user of MT may be a corporate user, a translation service, a translation agency or other provider of translation.

- Domain or field of application: Domain refers to topic, the field of interest for which the document is relevant, and the potential sublanguage effects germane to MT, for example technical/scientific (specific field being biology, chemistry, automotive mechanics, etc.), social, etc.

- Dissemination: The ultimate purpose of dissemination is to deliver to others a translation of documents produced inside the organization.

- Quantity of translation: This concerns the volume of translation typically dealt with by the organisation.

- Professional training: Related to the experience of the author in producing a particular type of texts, i.e is he familiar with the terminology? how long has he/she been working in similar posts?

- Proficiency in source language: This refers to proficiency in the source language as attested by some recognised measurement, international or regional. Two of the best known language proficiency scales are the ACTFL guidelines (first proposed in 1983 by the American Council for the Teaching of Foreign Languages) and the the ILR (FSI) proficiency scale, a five-level scale originally developed by the Foreign Service Institute (FSI) of the United States government, and later adopted by other services under the name of Interagency Language Roundtable (ILR) scale. The scale proposed for use in FEMTI is based on the ACTFL guidelines.

- Input characteristics (author and text): Input characteristics refer to the stylistic form or

format of the source document, the topic domain, and both the competency and performance qualities of the author.

- **Computer literacy:** This refers to the degree to which the user is at ease in computer use and manipulation.

- **Internal or in-house dissemination:** In the case of internal / in-house dissemination the translations are sent to other people in the same organization, who share aspects of the culture, terminology, and domain knowledge to some extent. The most important feature for this type of task is: speed - how fast is the system, can it keep up with the demand for input.

- **User characteristics:** This covers the characteristics of users in three senses: the end user who will interact with the machine translation system; the end user of the final product of the translation process which may include for example, post-editing; the organisation deploying the machine translation system. Note however that in the case when machine translation is combined with substantial post-editing, the resulting "system" might no longer fall under the scope of FEMTI, hence the end users are no longer users of a machine translation system.

- **Time allowed for translation.:** This concerns the deadlines for translation production typical within the organisation.

- **Superior:** " Superior-level writers are characterized by the ability to: express themselves effectively in most informal and formal writing on practical, social, and professional topics treated both abstractly as well as concretely; present well developed ideas, opinions, arguments, and hypotheses through extended discourse; control structures, both general and specialized/professional vocabulary, spelling or symbol production, punctuation, diacritical marks, cohesive devices, and other aspects of written form and organization with no pattern of error to distract the reader " ACTFL 2001.

- **Evaluation requirements:**

- **Superior:** " Able to read with almost complete comprehension and at normal speed expository prose on unfamiliar subjects and a variety of literary texts. Reading ability is not dependent on subject matter knowledge, although the reader is not expected to comprehend thoroughly texts which are highly dependent on knowledge of the target culture. [...] Occasional misunderstandings may still occur; for example, the reader may experience some difficulty with unusually complex structures and low-frequency idioms. [...] Material at this level will include a variety of literary texts, editorials, correspondence, general reports, and technical material in professional fields. Rereading is rarely necessary, and misreading is rare. " (ACTFL 1983 guidelines for reading proficiency).

- **Distinguished:** " Able to read fluently and accurately most styles and forms of the language pertinent to academic and professional needs. Able to relate inferences in the text to real-world knowledge and understand almost all sociolinguistic and cultural references by processing language from within the cultural framework. Able to understand a writer's use of nuance and subtlety. Can readily follow unpredictable turns of thought and author intent in such materials as sophisticated editorials, specialized journal articles, and literary texts such as novels, plays, poems, as well as in any subject matter area directed to the general reader. " (ACTFL 1983 guidelines for reading proficiency).

- **Distinguished:** " Able to read fluently and accurately most styles and forms of the language pertinent to academic and professional needs. Able to relate inferences in the text to real-world knowledge and understand almost all sociolinguistic and cultural references by processing language from within the cultural framework. Able to understand a writer's use of nuance and subtlety. Can readily follow unpredictable turns of thought and author intent in such materials as sophisticated editorials, specialized journal articles, and literary texts such as novels, plays, poems, as well as in any subject matter area directed to the general reader. " (ACTFL 1983 guidelines for reading proficiency).

- **Distinguished:** " Able to read fluently and accurately most styles and forms of the language pertinent to academic and professional needs. Able to relate inferences in the text to real-world knowledge and understand almost all sociolinguistic and cultural references by processing language from within the cultural framework. Able to understand a writer's use of nuance and subtlety. Can readily follow unpredictable turns of thought and author intent in such materials as sophisticated editorials, specialized journal articles, and literary texts such as novels, plays, poems, as well as in any subject matter area directed to the general reader. " (ACTFL 1983 guidelines for reading proficiency).

- **Author characteristics:** This set of characteristics covers writer attributes that are relevant to the writing task, which influence the unproofed text that is produced.

- **Document type:** The type of the input document can greatly affect the output of an MT system. For example, inputs to the METEO system are specific and very restricted, mainly weather forecast texts, using a limited lexicon and particular syntactic constructions. As a result the system produces accurate output, comparable to human translation. In contrast, MT of arbitrary text invariably produces output of much lesser quality. Both the genre and the application domain determine the quality.

- **Proficiency in target language:** This refers to proficiency in the source language as attested by some recognised measurement. The level of proficiency may be measured, for example by local education tests, internationally recognised examination schemes or organisation internal testing. Two of the best known language proficiency scales are the ACTFL guidelines (first proposed in 1983 by the American Council for the Teaching of Foreign Languages) and the the ILR (FSI) proficiency scale, a five-level scale originally developed by the Foreign Service Institute (FSI) of the United States government, and later adopted by other services under the name of Interagency Language Roundtable (ILR) scale. The scale proposed for use in FEMTI is based on the ACTFL guidelines. Depending on the operations performed on the translation, it is either the reading or the writing proficiency which are more specifically relevant. We propose to use the ACTFL eading proficiency scale (1985) -- note that only the guidelines for writing/speaking have been recently updated.

- **Proficiency in source language:** This refers to proficiency in the source language as attested by some recognised measurement. The level of proficiency may be measured, for example by local education tests, internationally recognised examination schemes or organisation internal testing. Two of the best known language proficiency scales are the ACTFL guidelines (first proposed in 1983 by the American Council for the Teaching of Foreign Languages) and the the ILR (FSI) proficiency scale, a five-level scale originally developed by the Foreign Service Institute (FSI) of the United States government, and later adopted by other services under the name of Interagency Language Roundtable

(ILR) scale. The scale proposed for use in FEMTI is based on the ACTFL guidelines for reading (1985) -- note that only the guidelines for writing/speaking have been recently updated.

- **Number of personnel:** This concerns the number of personnel within the organisation who will be directly concerned with the use of the MT system.

QUALITY CHARACTERISTICS SUGGESTED BY FEMTI

- **Languages:** "The range of languages which the product supports is a vital selection criterion. In machine translation systems, the languages are classified according to source and target language pairs, due to the need for full linguistic processing capability. In translator workbench products, the languages are not necessarily classified by strict language pairs as these products are interactive and therefore require only partial linguistic information. Terminology products have little or no linguistic ability and therefore the information only relates to the character sets which the product supports." (OVUM report)

Normalized weight: 0.6

Metrics:

- *Languages supported*
  Method: For each component tool of the product (MT, terminology management, translation memory etc) run the tool on texts, or other relevant resources in a variety of languages, and record whether it was possible to treat that particular language.

- **Dictionaries:** The kinds and number of dictionaries available. In this context, a dictionary is assumed to be equivalent to the term lexicon. In MT systems, the term tends to be used interchangeably. Another assumption is that the lexicon carries more information than a standard wordlist or glossary, including grammatical information. Finally, machine translation engines tend to have a general dictionary and other specialized vocabulary dictionaries, which are often called customer specific dictionaries.

Normalized weight: 0.5

Metrics:

- *Format of dictionary entries*
  Method: Examine either the documentation accompanying the system or the dictionaries included to ascertain the format used

- *Kinds of dictionaries available*
  Method: Examine the documentation accompanying the system and ascertain the kinds of dictionaries available or potentially available. Note whether there are domain specific dictionaries also available

- Functionality: The capability of the software product to provide functions which meet stated and implied needs when the software is used under specified conditions.

Normalized weight: 0.4

Metrics:
No selected metrics for this quality characteristic

- Ease of dictionary update: Facility of modifying the dictionary used by an MT system, most often regarding the addition of new words, phrases, grammatical roles, or senses.

Normalized weight: 0.30000000000000004

Metrics:

- *Effort necessary to update dictionary*
  Method: Define a list of dictionary update operations, e.g. insertion of various types of words, etc. Use a pool of typical users of the MT system and ask them to perform the task. Using a questionnaire, estimate the cognitive effort of required from the subjects.

- Well-formedness: Degree to which the output respects the reference rules of the target language at the specified linguistic level.

Normalized weight: 0.1

Metrics:

- *Percentage of phenomena correctly treated.*
  Method:

- Comprehensibility: The extent to which the text as a whole is easy to understand. That is, the extent to which valid information and inferences can be drawn from different parts of the same document. Comprehensibility reflects the degree to which a complete translation can be understood (whereas intelligibility is based on the general clarity of the translation, whether this is considered in its entirety or by segments out of context). (Halliday in Van Slype's Critical Report). Subjective evaluation of the degree of comprehensibility and clarity of the translation. (Van Slype in Van Slype's Critical Report).

Normalized weight: 0.1

Metrics:
No selected metrics for this quality characteristic

- Coverage of corpus-specific phenomena: Coverage refers to the ability of the system to deal satisfactorily with linguistic phenomena, both generally addressing known cross-language phenomena and specifically addressing phenomena in a corpus of

interest. Coverage of corpus-based problematic phenomena concerns the ability of the system to deal with the particular challenges presented by a corpus of interest.

Normalized weight: 0.1

Metrics:

- *Errors*
  Method: By constituting a representative corpus and submitting it to the system in order to observe what errors occur.

- Fidelity - precision: Subjective evaluation of the degree to which the information contained in the original text has been reproduced without distortion in the translation (Van Slype). Measurement of the correctness of the information transferred from the source language to the target language (Halliday in Van Slype's Critical Report).

Normalized weight: 0.1

Metrics:

- *Van Slype Rating of sentences read*
  Method: Rating of sentences read on a 4-point scale.

- *Rank-order evaluation*
  Method: Rank-order evaluation of MT system: correlation of automatically computed semantic and syntactic attributes of the MT output with human scores for adequacy and informativeness, and also fluency.

- *BLEU*
  Method: Bleu evaluation tool kit Automatic n-gram comparison of translated sentences with one or more human reference translations.

- Terminology: Correct translation of technical (domain-specific) terms.

Normalized weight: 0.1

Metrics:

- *Percentage of domain terms correctly translated.*
  Method:

- Dictionary updating: Facilities to assist users in researching and entering terminology which the machine does not recognize into the system's dictionary.

Normalized weight: 0.1

Metrics:
No selected metrics for this quality characteristic

- Cost: Cost here covers all of the monetary costs of introducing MT, maintenance costs implied by operational use of the system and the potential costs of not introducing MT.

Normalized weight: 0.1

Metrics:
No selected metrics for this quality characteristic

ADDITIONAL QUALITY CHARACTERISTICS (NOT SUGGESTED BY FEMTI)

- Translation preparation activities: Translation preparation is related to transferring the source text into a form which the translation process can accept or which will facilitate translation. The more the source text can be designed and created with translation in mind, the less work it will require when passing into translation process (OVUM report).

Normalized weight: 0.0

Metrics:

- *Can edit list of terms to be ignored during translation process.*
  Method: 1) Determine if system has feature through reading documentation.
  2) Test the operation of the feature through one or more test cases.

- Suitability: The capability of the software product to provide an appropriate set of functions for specified tasks and user objectives (ISO 9126: 2001, 6.1.1).

Normalized weight: 0.0

Metrics:
No selected metrics for this quality characteristic

- Transfer-based MT: MT systems which analyse the source text into a syntax tree and then convert the tree into the form required by the target syntax (for example, moving the verb complex) or analyse the source text into some formalism that is intended to capture meaning, not just grammatical form.

Normalized weight: 0.0

Metrics:

- *ease of adding or changing the rules*
  Method: Add or change a grammar rule

- Ease of importing data: As part of translation preparation activities, the user may need to import different types of data (for an example from word processors, see OVUM report)

Normalized weight: 0.0

Metrics:

- *Ease of importing data*
  Method: 1) Review system documentation for list of data types accepted by the system, to include file types, code sets, data formats. 2) Import data into the system for each file type, code set and data format advertised.

- System characteristics:

Normalized weight: 0.0

Metrics:
No selected metrics for this quality characteristic

- Interactive translation activities : Interactive MT systems require user guidance at points when the system reaches an impasse during processing. The user's assistance (whether in the form of menu choices, parameter entry) constitutes a form of editing that can be called "inline editing" or "in-editing" (The Pangloss Mark III MT System).

Normalized weight: 0.0

Metrics:

- *Time for interactive translation*
  Method: Measure the amount of time it takes to perform interactive translation on test corpus.

- *Steps for translation*
  Method: Count number of times system requires assistance when translating a test corpus.

- Post-translation activities: Post-translation activities relate to preparing the output texts to meet the requirements for final publication or delivery (OVUM report). Revision of output translation interactively to produce a final version for printing (Trial of the Weidner Computer-Assisted Translation System, p.12, October,1985). Sometimes this is referred to as the camera-ready copy.

Normalized weight: 0.0

Metrics:

- *Availability of editing functions*
  Method: 1) Check the system documentation to check availability and operation of post-edit functions. 2) Test the operation of each function on test documents.

- Rule-based models : rule-based model, also known as "knowledge based", involves rules to analyse and represent the source text in a more abstract form as well as rules to map this abstract representation to the corresponding target text; these rules can be morphological, lexical, etc.

Normalized weight: 0.0

Metrics:

- *If the system uses a grammar, the ease of adding or changing rules.*
  Method: Design and add/change grammatical rule to the system.

- Methodology: The underlying theoretical methodology behind the development of a given system.

Normalized weight: 0.0

Metrics:

- *Description of theory/method of translation*
  Method: Provision of supporting documentation such as white papers.

ANNEX V: SELECTION OF A TEXT TYPE

			AWKat Arbeitswerte Katalog Flat Rates Catalogue		
	Criteria	**Weight**	**Description**	**Points**	**Total**
Integration within Authoring System	100% Authoring system integration 50% Partial authoring system integration 0% No integration within authoring system	1	Not integrated in an authoring system	0	0
CLAT	100% Quality Assurance with CLAT (MULTILINT) for at least 3 years 50% Quality Assurance with CLAT planned 25% Quality assurance with CLAT sporadically 0% No quality assurance with CLAT, neither now nor planned	3	There is no quality assurance process with CLAT for this information type.	0	0
External Characteristics	100% Translation volume for English > 100.000 Lines/Year 50% Translation volume for English < 100.000 & > 50.000	2	Translation languages: En-UK, Fr, It, Es, Ni, Sv, En-US, Ja, Ru, Ch, Ko, Th, Ind, Tür, Gr, Pt, Fi English is financed by BMW. The markets are responsible for the translations in their official	0	0

			AWKat Arbeitswerte Katalog Flat Rates Catalogue		
	Criteria	**Weight**	**Description**	**Points**	**Total**
	0% Translation volume for English < 50.000		languages Translation volume: according to personal communication with [Gehlich 05], about 11.000 €/Year		
Linguistic Characteristics	100% Compound document & written by professional technical writers 50% Collection of sentences & wirtten by professional technical writers 25% Compouund document & written by non-professionals 0% Collection of sentences & written by non-professionals	2	Collection of sentences (sentence chunks) Written by professional technical writers. Light to absent grammatical complexity; translation quality highly depends on terminology coverage and accuracy	2	4
		TOTAL			**4**

Table 39: Evaluation of the AWK text type

			RA (TIS) Reparaturanleitung Repair Instructions		
	Criteria	**Weight**	**Description**	**Points**	**Total**
Integration within Authoring System	100% Authoring system integration 50% Partial authoring system integration 0% No integration within authoring system	1	Integrated in an authoring system	3	3
CLAT	100% Quality Assurance with CLAT (MULTILINT) for at least 3 years 50% Quality Assurance with CLAT planned 25% Quality assurance with CLAT sporadically 0% No quality assurance with CLAT, neither now nor planned	3	This information type has been regularly checked with CLAT since aprox. 2000. Therefore, it can be assumed that, because of this checking and the experience of the writers, this kind of document complies pretty much with the CL rules.	3	9
External Characteristics	100% Translation volume for English > 100.000 Lines/Year 50% Translation volume for English < 100.000 & > 50.000 0% Translation volume for English < 50.000	2	Translation languages: En-UK, Fr, It, Es, Ni, Sv, En-US, Ja, Ru, Ch, Ko, Th, Ind, Tür, Gr, Pt. English is financed by BMW. The markets are responsible for the translations in their official languages. Translation volume: According to data gathered by [Berns & Törl 02], of 2.859.577 characters for translation in 2002, 984.942 were new. This makes a total of 17908 norm lines, which, at a flat rate of 1,25 the line, makes a total price of 22385 €/Year for English. Apart from this, about 31.000 pre-translated and 2000 were partly pre-translated	0	0

			RA (TIS) Reparaturanleitung Repair Instructions		
	Criteria	**Weight**	**Description**	**Points**	**Total**
Linguistic Characteristics	100% Compound document & written by professional technical writers 50% Collection of sentences & wirtten by professional technical writers 25% Compound document & written by non-professionals 0% Collection of sentences & written by non-professionals	2	Compound document Written by professional technical writers. Since these documents are terminologically checked, quality depends highly on terminology coverage and accuracy. These are instructional texts, with a very high presence of imperatives. Middle grammatical complexity Therefoire, it is important that the MT system can correctly translate this structure in English (and other languages). Up to 5 pages	3	6
		TOTAL			**18**

Table 40: Evaluation of the RA text type

			SBT (TIS)		
	Criteria	**Weight**	**Description**	**Points**	**Total**
Integration within Authoring System	100% Authoring system integration 50% Partial authoring system intgration 0% No integration within authoring system	1	Integrated in an authoring system	3	3
CLAT	100% Quality Assurance with CLAT (MULTILINT) for at least 3 years 50% Quality Assurance with CLAT planned 25% Quality assurance with CLAT sporadically 0% No quality assurance with CLAT, neither now nor planned	3	This information type has been regularly checked with CLAT since aprox. 2000. Therefore, it can be assumed that, because of this checking the and experience of the writers, this kind of document complies pretty much with the CL rules.	3	9
External Characteristics	100% Translation volume for English > 100.000 Lines/Year 50% Translation volume for English < 100.000 & > 50.000 0% Translation volume for English < 50.000	2	Translation languages: En-UK, Fr, It, Es, Ni, Sv, En-US, Ja, Ru, Ch, Ko, Th, Ind, Tür, Gr, Pt. Translation volume:	0	0
Linguistic Characteristics	100% Compound document & written by professional technical writers 50% Collection of sentences & wirtten by professional technical writers 25% Compuound document & written by non-professionals 0% Collection of sentences & written by non-professionals	2	Compound document Written by professional technical writers. Medium to high grammatical complexity Long docum ents (30 pages)	2	4
		TOTAL			**16**

Table 41: Evaluation of the SBT text typ

			SI (TIS) (Service Information)		
	Criteria	**Weight**	**Description**	**Points**	**Total**
Integration within Authoring System	100% Authoring system integration 50% Partial authoring system intgration 0% No integration within authoring system	1	Integrated in an authoring system	3	3
CLAT	100% Quality Assurance with CLAT (MULTILINT) for at least 3 years 50% Quality Assurance with CLAT planned 25% Quality assurance with CLAT sporadically 0% No quality assurance with CLAT, neither now nor planned	3	This information type has been regularly checked with CLAT since aprox. 2000. Therefore, it can be assumed that, because of this checking the and experience of the writers, this kind of document complies pretty much with the CL rules.	2	6
External Characteristics	100% Translation volume for English > 100.000 Lines/Year 50% Translation volume for English < 100.000 & > 50.000 0% Translation volume for English < 50.000	2	Translation languages: En-UK, Fr, It, Es, Ni, Sv, En-US, Ja, Ru, Ch, Ko, Th, Ind, Tür, Gr, Pt. Translation volume:	0	0
Linguistic Characteristics	100% Compound document & written by professional technical writers 50% Collection of sentences & wirtten by professional technical writers 25% Compuound document & written by non-professionals 0% Collection of sentences & written by non-professionals	2	Compound document Written by professional technical writers. Middel to high grammatical complexity; short documents (about 5 pages); Terminology coverage highly influences the quality of the translation.	3	6
		TOTAL			**15**

Table 42: Evaluation of the SI text type

			Technical Campaigns (OSCAR)		
	Criteria	**Weight**	**Description**	**Points**	**Total**
Integration within Authoring System	100% Authoring system integration 50% Partial authoring system intgration 0% No integration within authoring system	1	Not integrated in ANTARES.	0	0
CLAT	100% Quality Assurance with CLAT (MULTILINT) for at least 3 years 50% Quality Assurance with CLAT planned 25% Quality assurance with CLAT sporadically 0% No quality assurance with CLAT, neither now nor planned	3	This information type has been sporadically checked with CLAT since 2000. There is no control on which documents have been checked or not and by whom.	1	3
External Characteristics	100% Translation volume for English > 100.000 Lines/Year 50% Translation volume for English < 100.000 & > 50.000 0% Translation volume for English < 50.000	2	Translation languages: En, Fr, Es, It, Nl, Sv	0	0
Linguistic Characteristics	100% Compound document & written by professional technical writers 50% Collection of sentences & wirtten by professional technical writers 25% Compuound document & written by non-professionals 0% Collection of sentences & written by non-professionals	2	Compound document Written by professional technical writers.	3	6
		TOTAL			**9**

Table 43: Evaluation of the OSCAR text type

			PUMA		
	Criteria	**Weight**	**Description**	**Points**	**Total**
Integration within Authoring System	100% Authoring system integration 50% Partial authoring system intgration 0% No integration within authoring system	1	Not integrated in an authoring system	0	0
CLAT	100% Quality Assurance with CLAT (MULTILINT) for at least 3 years 50% Quality Assurance with CLAT planned 25% Quality assurance with CLAT sporadically 0% No quality assurance with CLAT, neither now nor planned	3	There is no quality assurance process with CLAT for this information type.	0	0
External Characteristics	100% Translation volume for English > 100.000 Lines/Year 50% Translation volume for English < 100.000 & > 50.000 0% Translation volume for English < 50.000	2	Translation languages: En, Fr, Es, It, Nl, Ja Translation volume: Translations are needed ad-hoc, as soon as possible (agency is obligued to provide translation within 24 hours)	0	0
Linguistic Characteristics	100% Compound document & written by professional technical writers 50% Collection of sentences & wirtten by professional technical writers 25% Compuound document & written by non-professionals 0% Collection of sentences & written by non-professionals	2	Compound document Written by mechanics and service advisors. Therefore, some colloquial language can be found. Middle to high grammatical complexity	1	2
		TOTAL			**2**

Table 44: Evaluation of the PUMA text type

			Schulungsunterlagen: SU (Training Documentation)		
	Criteria	**Weight**	**Description**	**Points**	**Total**
Integration within Authoring System	100% Authoring system integration 50% Partial authoring system intgration 0% No integration within authoring system	1	Integrated in ANTARES.	3	3
CLAT	100% Quality Assurance with CLAT (MULTILINT) for at least 3 years 50% Quality Assurance with CLAT planned 25% Quality assurance with CLAT sporadically 0% No quality assurance with CLAT, neither now nor planned	3	This information type has been sporadically checked with CLAT since 2000. There is no control on which documents have been checked or not and by whom.	2	6
External Characteristics	100% Translation volume for English > 100.000 Lines/Year 50% Translation volume for English < 100.000 & > 50.000 0% Translation volume for English < 50.000	2	Translation languages: English, French, Italian, Spanish, Dutch, Swedish and Portuguese. TNU (Teilnehmerunterlagen): En, Fr, Es, It, Nl, Sv, Po THG (Trainerhintergrund): En, Fr, Es TLF (Trainerleitfaden): En, Fr, Es Volume: According to the numbers gathered in the project TERMinator, 28.000 new lines were writen in 2002; about 12.000 were pre-translated, and about 1000 were partly pre-translated. There is a potential for these 28.000 lines. This number can vary from year to year 2004: 82.215 € for English; 102.881,62 for French (about 80.000 lines!) 28.000 new lines at a price of 1,25/line makes a price of 35000 EUR The translation of the created documents needs about 2 months (10 weeks for big projects) and takes place before the production starts.	0	0

			Schulungsunterlagen: SU (Training Documentation)		
	Criteria	**Weight**	**Description**	**Points**	**Total**
Linguistic Characteristics	100% Compound document & written by professional technical writers 50% Collection of sentences & wirtten by professional technical writers 25% Compuound document & written by non-professionals 0% Collection of sentences & written by non-professionals	2	Compound document Written by professional technical writers. Two types of training: technical and non-technical. The problem especially with non-technical training is that authors write in a free way giving place to inconsistent texts There is a real translation volume problem with this type of texts!!	2	4
		TOTAL			**13**

Table 45: Evaluation of the SU text type

	AWKat Arbeitswerte Katalog Flat Rates Catalogue	RA (TIS)	SBT (TIS)	SI (TIS)	Technical Campaigns (OSCAR)	PUMA	Schulungsunterlagen: SU (Training Documentation)
Integration within Authoring System	0	3	3	3	0	0	3
CL-Compliance (Translatability)	0	9	9	6	3	0	6
Translation Languages and Volume	0	0	0	0	0	0	0
Text length	4	6	4	6	6	2	4
TOTAL	4	18	16	15	9	2	13

Table 46: Text type evaluation summary

ANNEX VI: PHASE 1-HUMAN EVALUATION

Human evaluation. Average results for RA and SBT

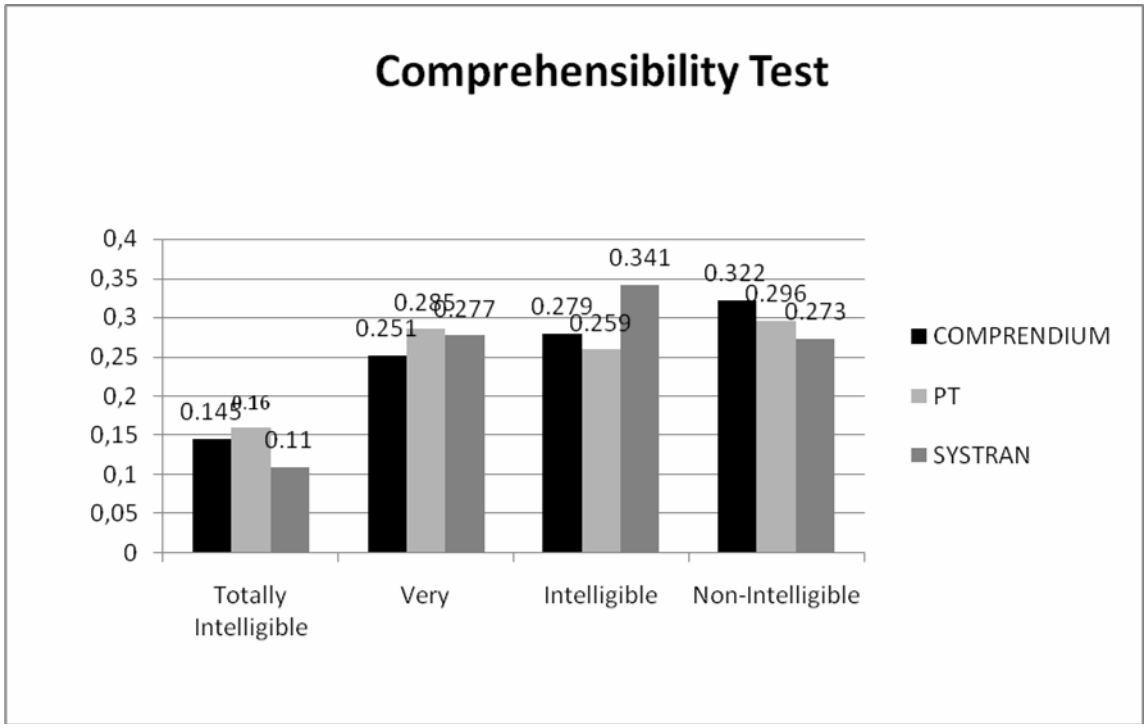

Figure 60: Comprehensibility Test

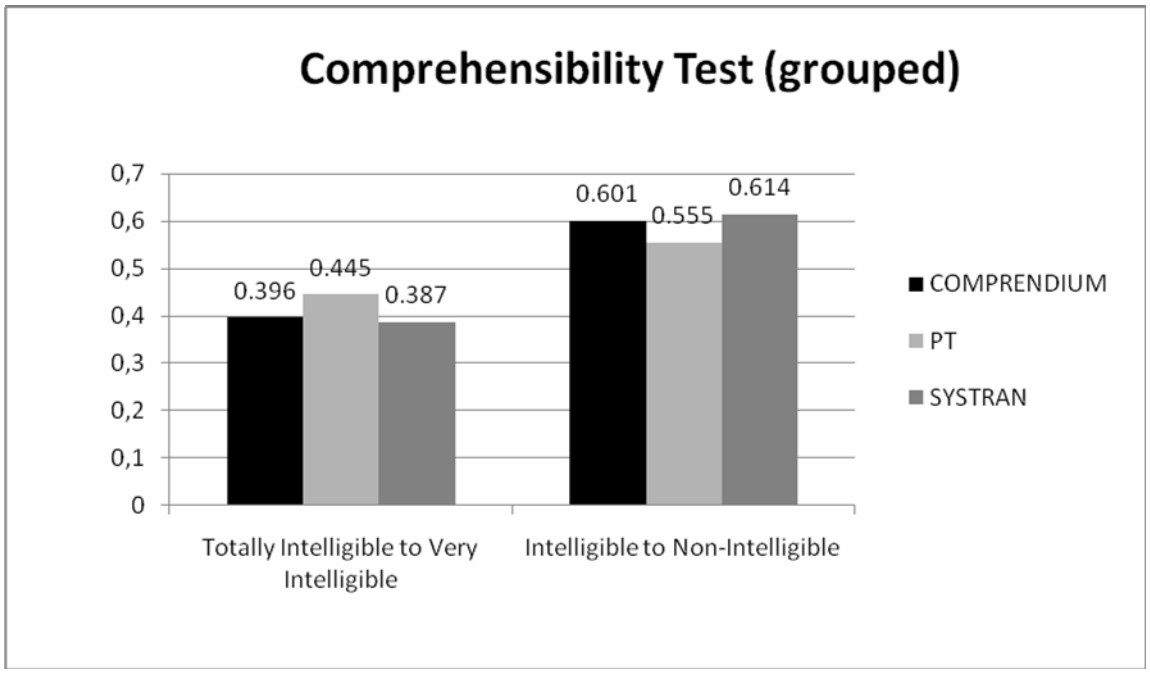

Figure 61: Comprehensibility Test (grouped)

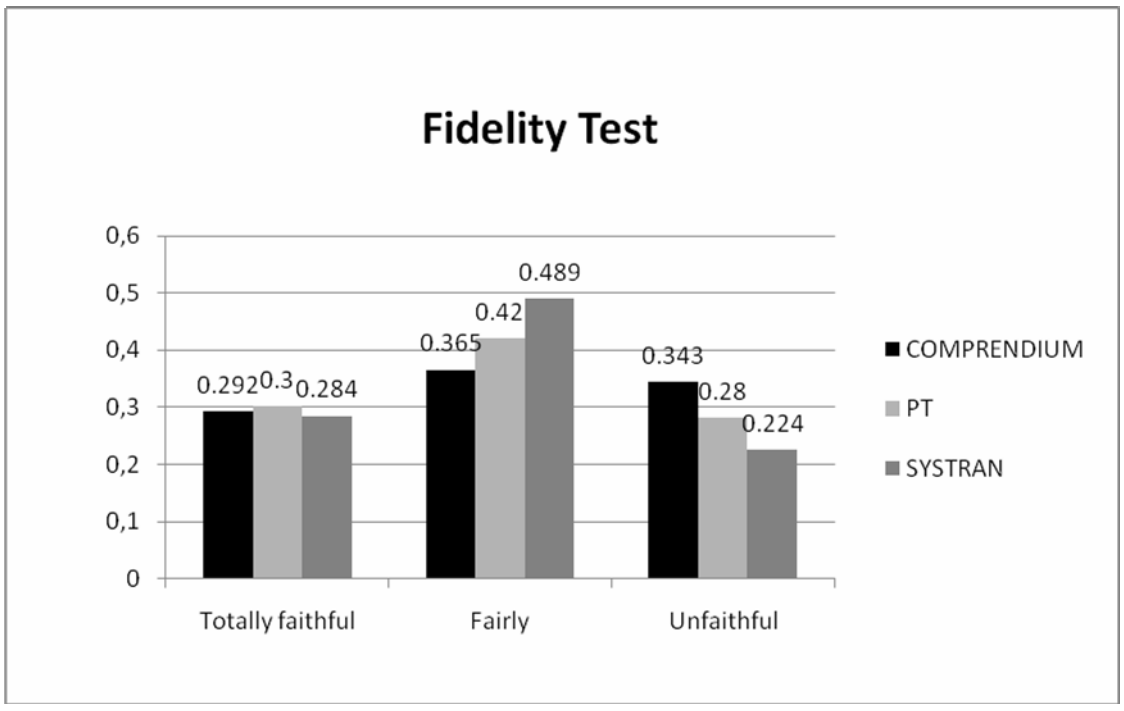

Figure 62: Fidelity Test

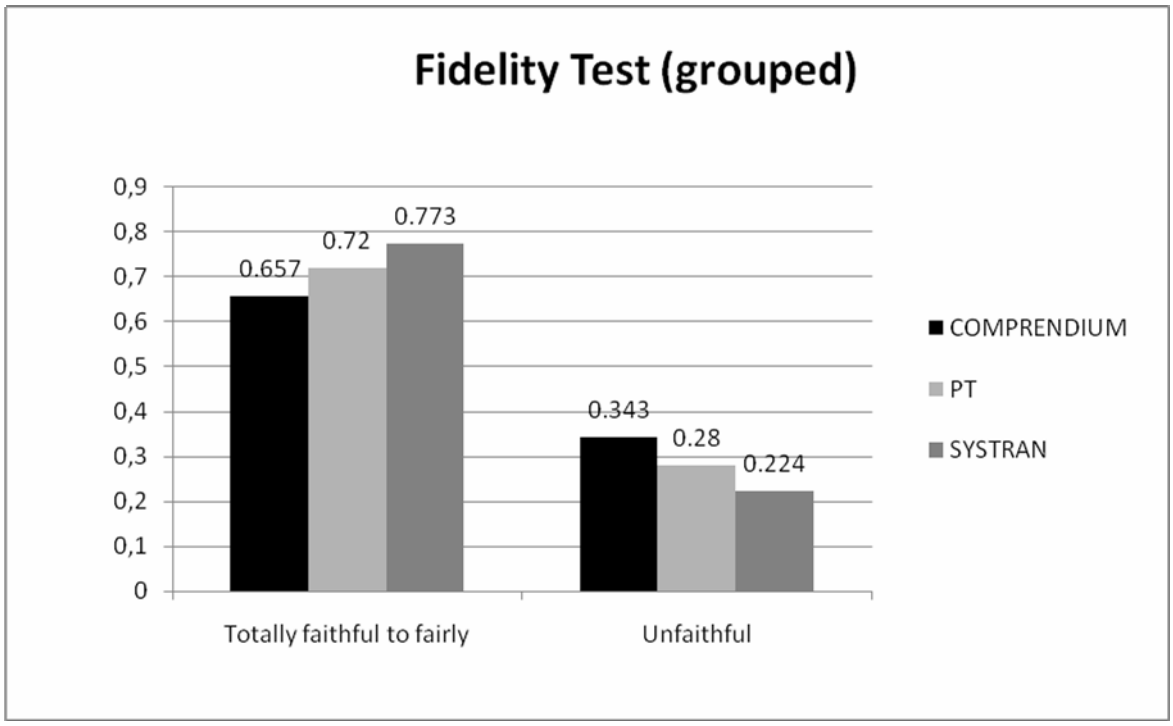

Figure 63: Fidelity Test (grouped)

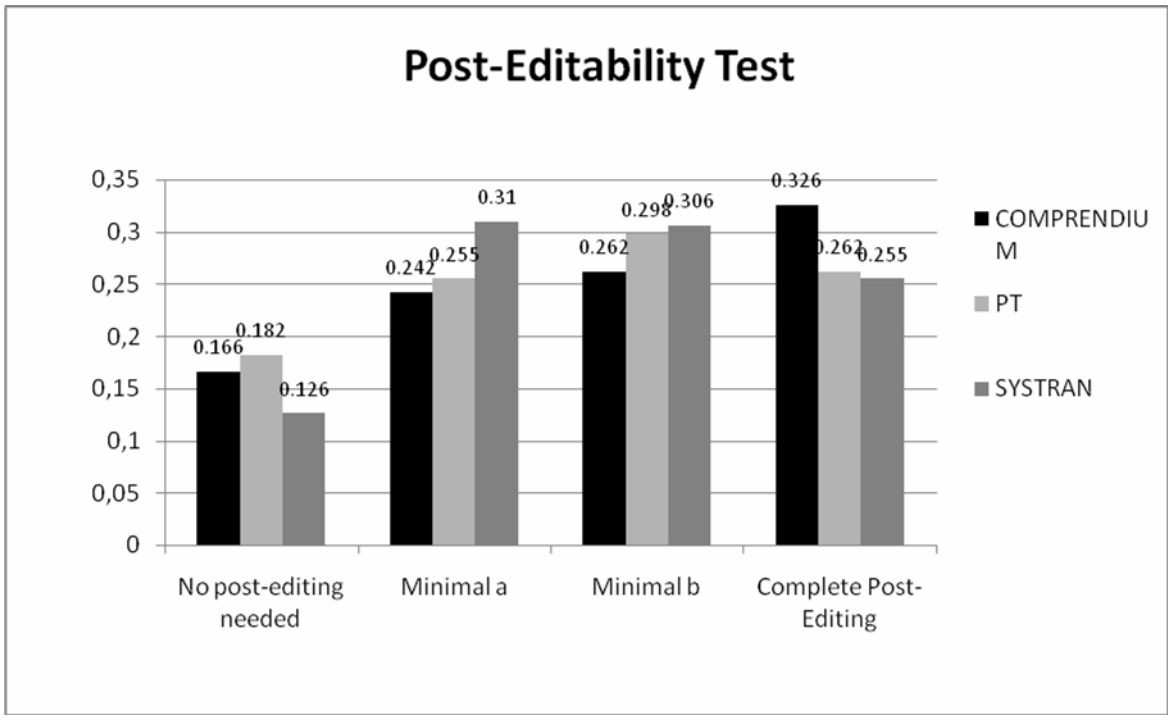

Figure 64: Post-editability Test

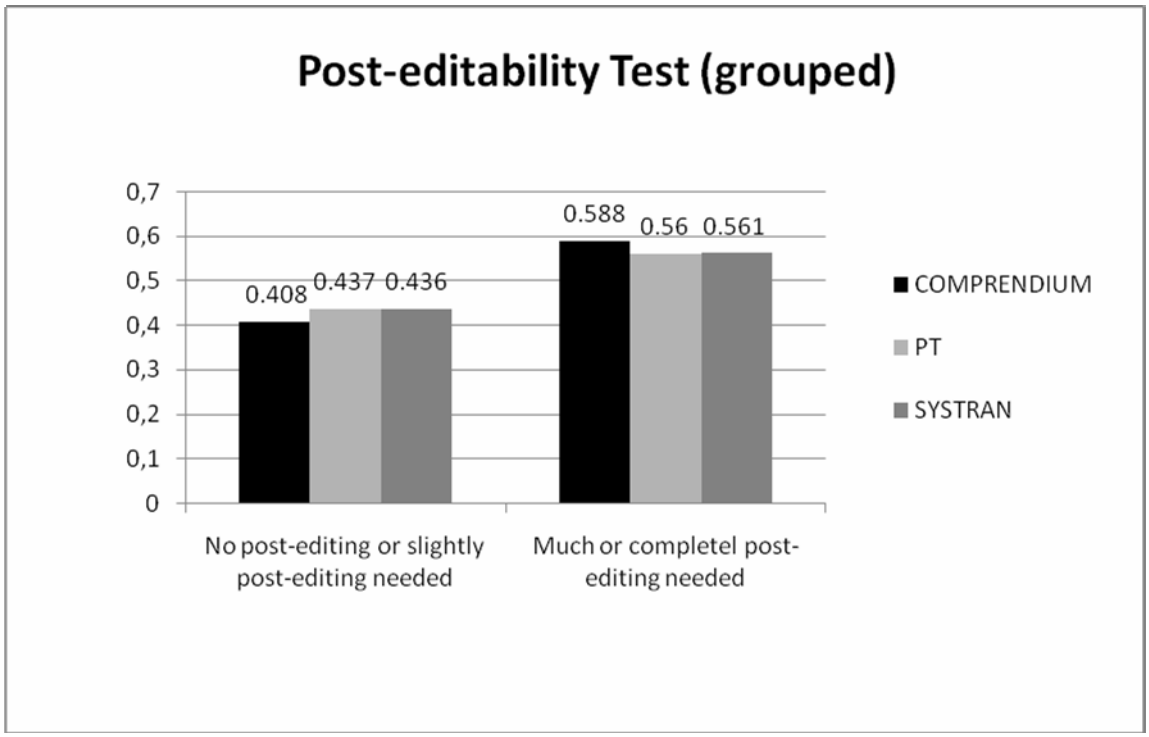

Figure 65: Post-Editability Test (grouped)

TRANSLATOR	1	2	3	4	5	6	7	8	Average
TESTSUITE	II	II	I	I	II	I	I	II	
How many years have you been working as a translator for the language pair German-English?	13	10	8.5	19	8	15	15	7	119.375
How many years of experience do you have with automotive texts?	7	1	4	15	6	6	5	6	6.25
How many years of experience do you have with BMW texts? What kind of texts have you translated?	4	0	1	15	1	0	2	0	2.87
Do you have any experiencie in Evaluating MT-systems?	No	No	Yes	Yes	No	No	No	No	75% No 25% Yes
TRANSLATOR	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Average**
TESTSUITE	**II**	**II**	**I**	**I**	**II**	**I**	**I**	**II**	
How much time did you need to carry out the tests?	13	12	15	16	15.75	11	11	17,5	13.90
Intelligibility (hours)	7	3	5	4	3	2.5	3	4.5	4
Fidelity (hours)	1	4	3	5	3.75	3.5	2	6.5	3.15
Post-Editability (hours)	1	1	3	6	4	3,5	2	6.50	3.37
How much time did you need to correct the sentences? (hours)	4	4	4	3	5	1.5	4	included in hours for post-editability	3.18

TRANSLATOR	1	2	3	4	5	6	7	8	Average
TESTSUITE	II	II	I	I	II	I	I	II	
Would you be ready to work post-translating translations?	Yes	Yes	Yes	Yes	Yes	No	No	No	62.5% Yes 37.5% No
Have you made any experience before in post-editing?	No, but there are similarities to reviewing of non-native speaker translations, in which I have considerable experience.	Yes, but not MT	Yes	No	Yes	No	No	No	62.5% Yes 37.5% No

Table 47: Poll for evaluators

ANNEX VII: PHASE 1-KAPPA VALUES

		System A		
		Percent of overall agreement Po	**Fixed-marginal kappa**	**Free-marginal kappa**
Test 1	**Intelligibility**	0.494667	0.294708	0.326223
	Fidelity	0.626344	0.431352	0.439516
	Post-editability	0.495968	0.305267	0.327957
Test 2	**Intelligibility**	0.406504	0.200063	0.208672
	Fidelity	0.576	0.362912	0.364
	Post-editability	0.45082	0.261591	0.26776

Table 48: Kappa values for System A

		System B		
		Percent of overall agreement Po	Fixed-marginal kappa	Free-marginal kappa
Test 1	Intelligibility	0.465334	0.266399	0.287112
	Fidelity	0.613334	0.399654	0.420001
	Post-editability	0.513334	0.328818	0.351112
Test 2	Intelligibility	0.390667	0.175016	0.187556
	Fidelity	0.541334	0.3013	0.312001
	Post-editability	0.40847	0.209487	0.211293

Table 49: Kappa values for System B

		System C		
		Percent of overall agreement Po	Fixed-marginal kappa	Free-marginal kappa
Test 1	Intelligibility	0.476	0.258437	0.301333
	Fidelity	0.525334	0.18958	0.288001
	Post-editability	0.50542	0.285698	0.34056
Test 2	Intelligibility	0.338667	0.0966976	0.118223
	Fidelity	0.4458	0.145432	0.1687
	Post-editability	0.400273	0.181811	0.200364

Table 50: Kappa values for System C

ANNEX VIII:
PHASE 1-AUTOMATIC EVALUATION

VIII.1 BLUE scores

COMPLETE CORPUS (3,262 segments)		
	BLEU-Score	**BLEU-Score (Case sensitive)**
COMPRENDIUM	**0.3097**	**0.2860**
PT	**0.3225**	**0.3035**
SYSTRAN	**0.3099**	**0.2808**

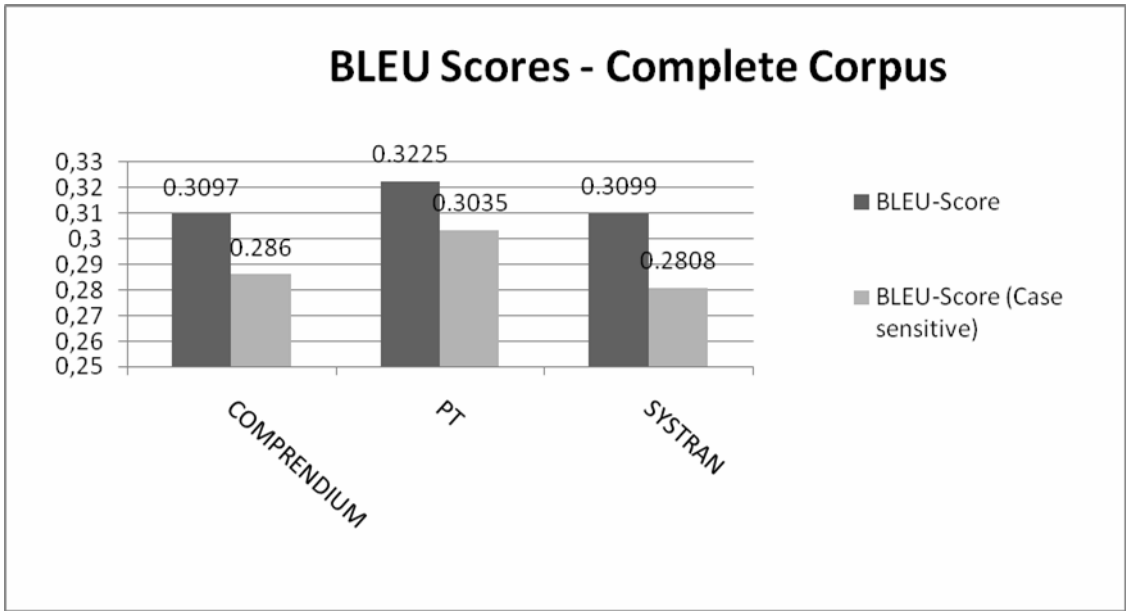

Figure 66: BLUE Scores-Complete Corpus

COMPLETE CORPUS - RA (529 segments)		
	BLEU-Score	**BLEU-Score (Case sensitive)**
COMPRENDIUM	0.3609	0.3502
PT	0.3922	0.3866
SYSTRAN	0.3760	0.3425

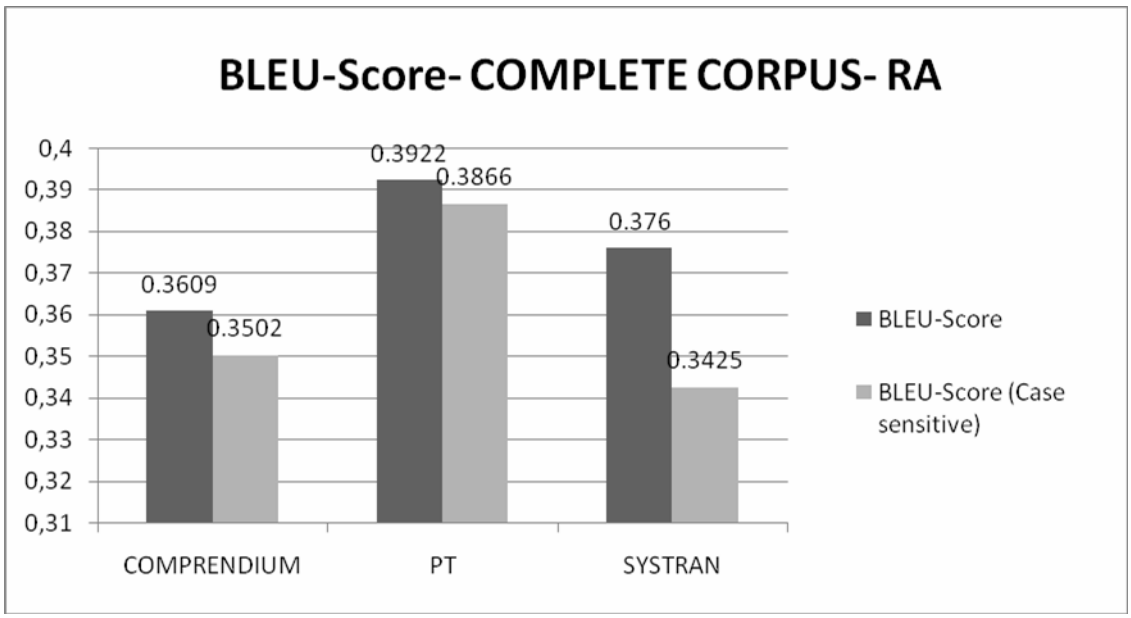

Figure 67: BLUE Scores-Complete Corpus (RA)

COMPLETE CORPUS-SBT (2,733 segments)		
	BLEU-Score	**BLEU-Score (Case sensitive)**
COMPRENDIUM	0.2938	0.2661
PT	0.3023	0.2794

SYSTRAN	0.2883	0.2602

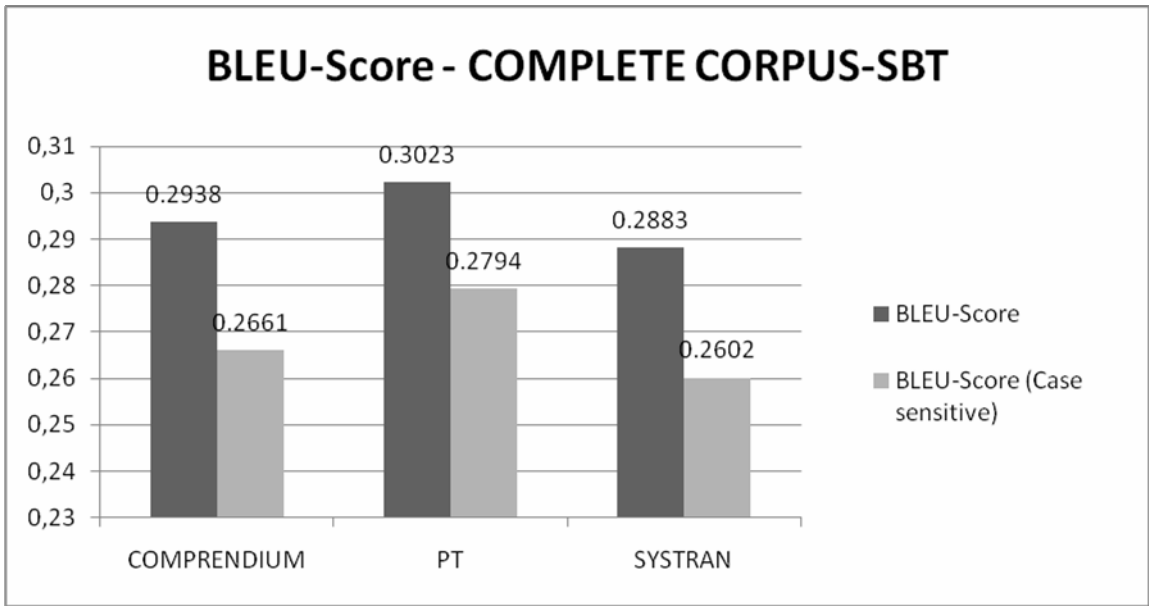

Figure 68: BLEU Scores-Complete Corpus (SBT)

REDUCED CORPUS SBT (228 segments)		
	BLEU-Score	**BLEU-Score (Case sensitive)**
COMPRENDIUM	0.3035	0.2944
PT	0.3311	0.3236
SYSTRAN	0.3083	0.2891

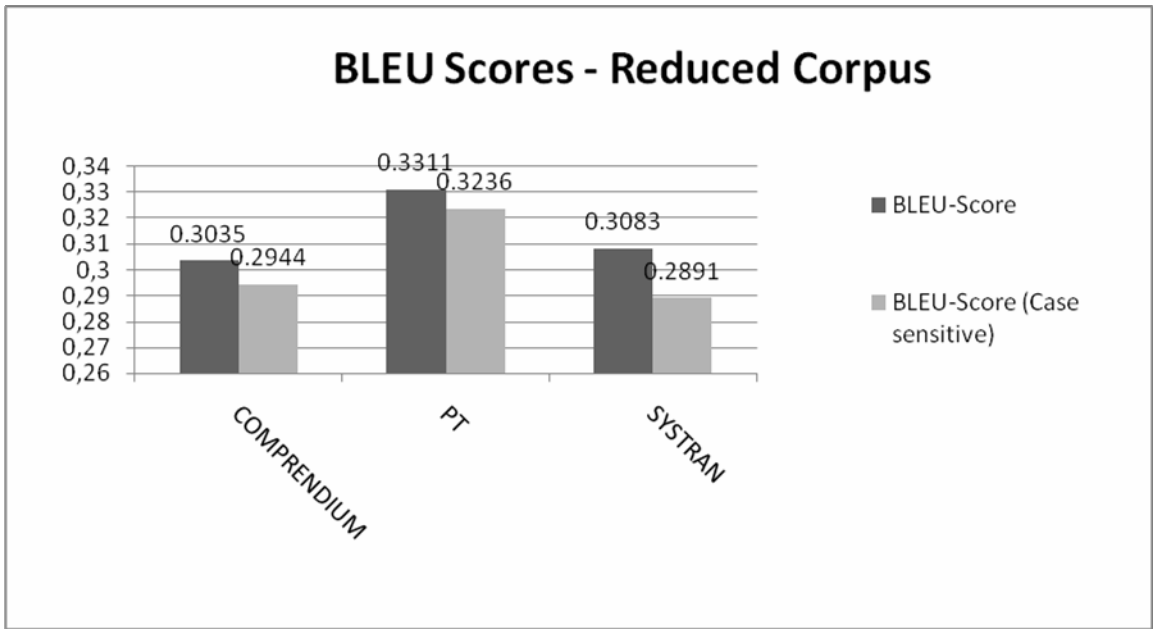

Figure 69: BLUE Scores-Reduced Corpus

COMPLETE CORPUS-RA MONOREF (121 segments)		
	BLEU-Score	**BLEU-Score (Case sensitive)**
COMPRENDIUM	0.3035	0.2944
PT	0.3311	0.3236
SYSTRAN	0.3083	0.2891

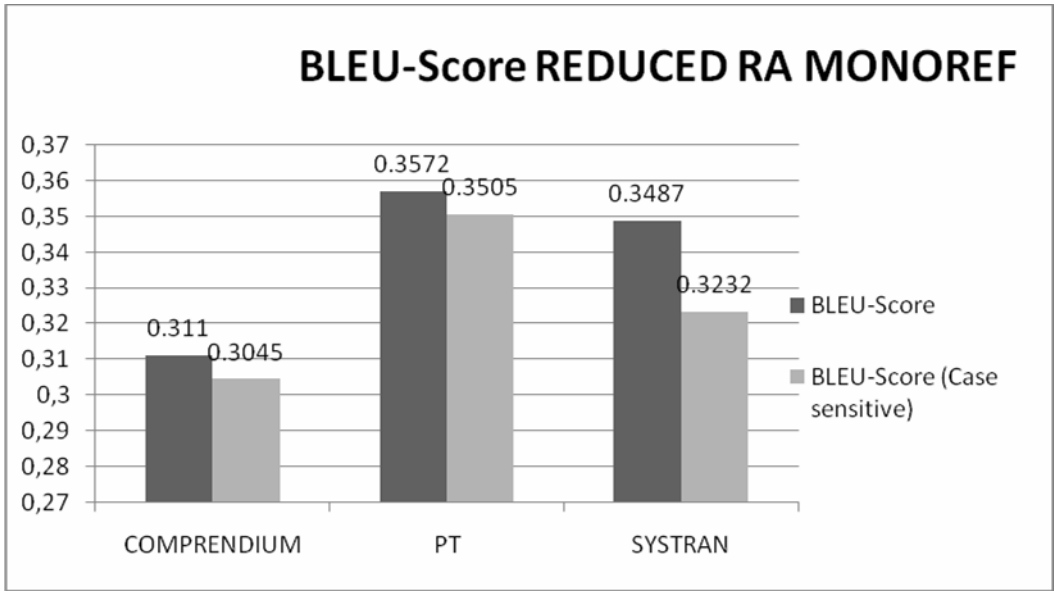

Figure 70: BLEU Scores-Reduced Corpus (RA-Monoreference)

REDUCED CORPUS -SBT MONOREF (107 segments)		
	BLEU-Score	**BLEU-Score (Case sensitive)**
COMPRENDIUM	0.2974	0.2862
PT	0.3106	0.3026
SYSTRAN	0.2756	0.2611

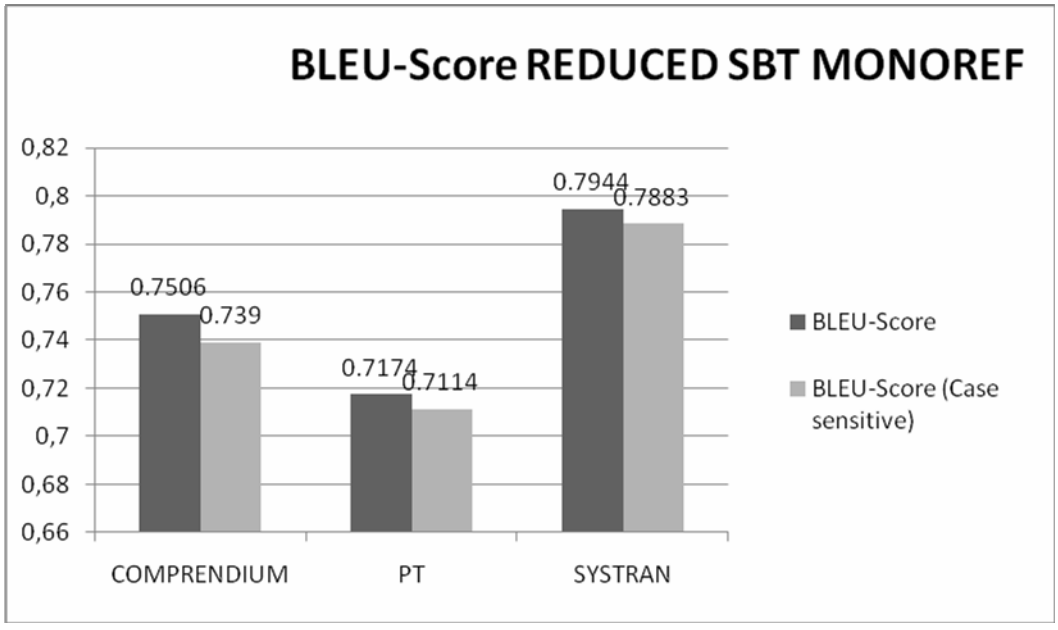

Figure 71: BLEU Scores-Reduced Corpus (SBT-Monoreference)

REDUCED CORPUS - RA MULTIREF (121 segments)		
	BLEU-Score	**BLEU-Score (Case sensitive)**
COMPRENDIUM	0.6541	0.6415
PT	0.7470	0.7404
SYSTRAN	0.6693	0.6105

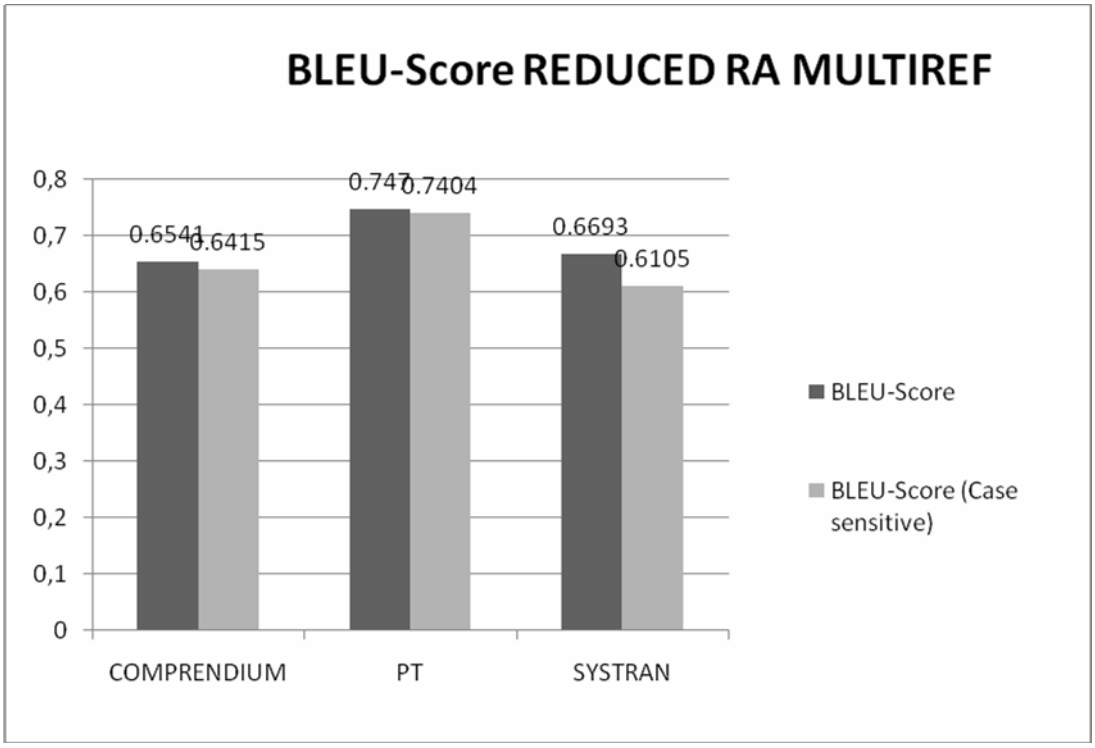

Figure 72: BLEU Scores-Reduced Corpus (RA-Multireference)

REDUCED CORPUS - SBT MULTIREF (107 segments)		
	BLEU-Score	**BLEU-Score (Case sensitive)**
COMPRENDIUM	0.7506	0.7390
PT	0.7174	0.7114
SYSTRAN	0.7944	0.7883

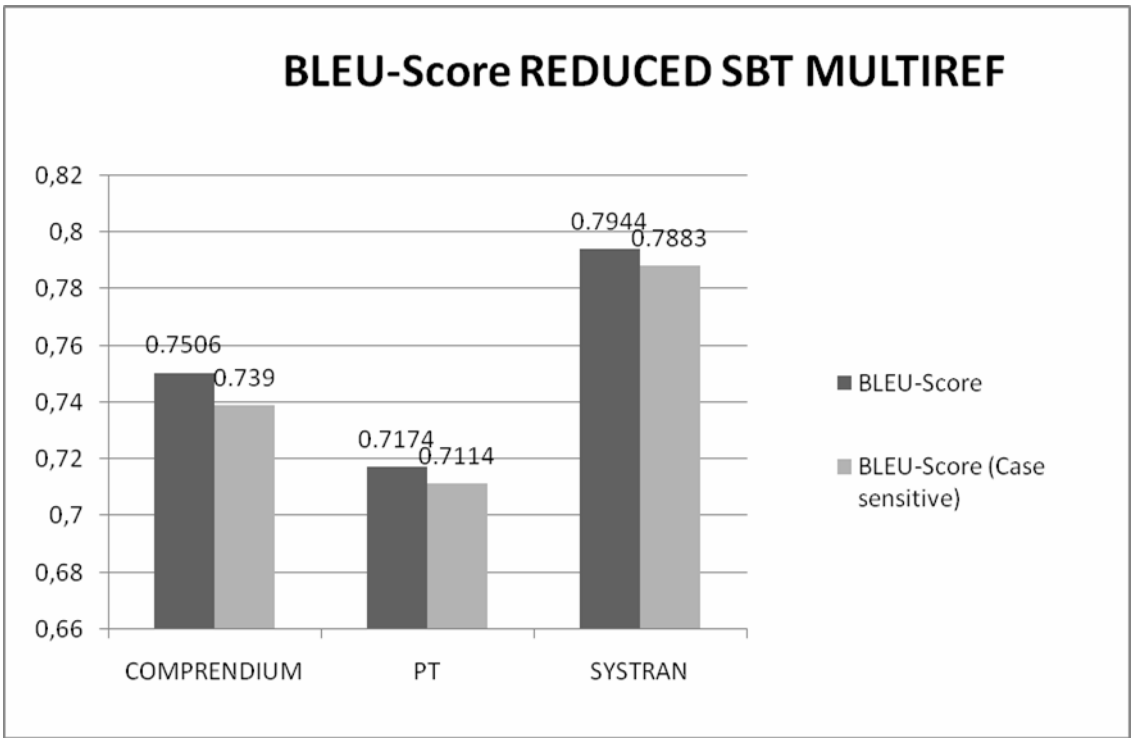

Figure 73: BLEU Scores -Reduced Corpus (SBT-Multireference)

VIII.2 NIST scores

COMPLETE CORPUS (3262 segments)		
	NIST-Score	**NIST-Score (Case sensitive)**
COMPRENDIUM	6.9614	6.6395
PT	7.1161	6.8276
SYSTRAN	7.1137	6.6650

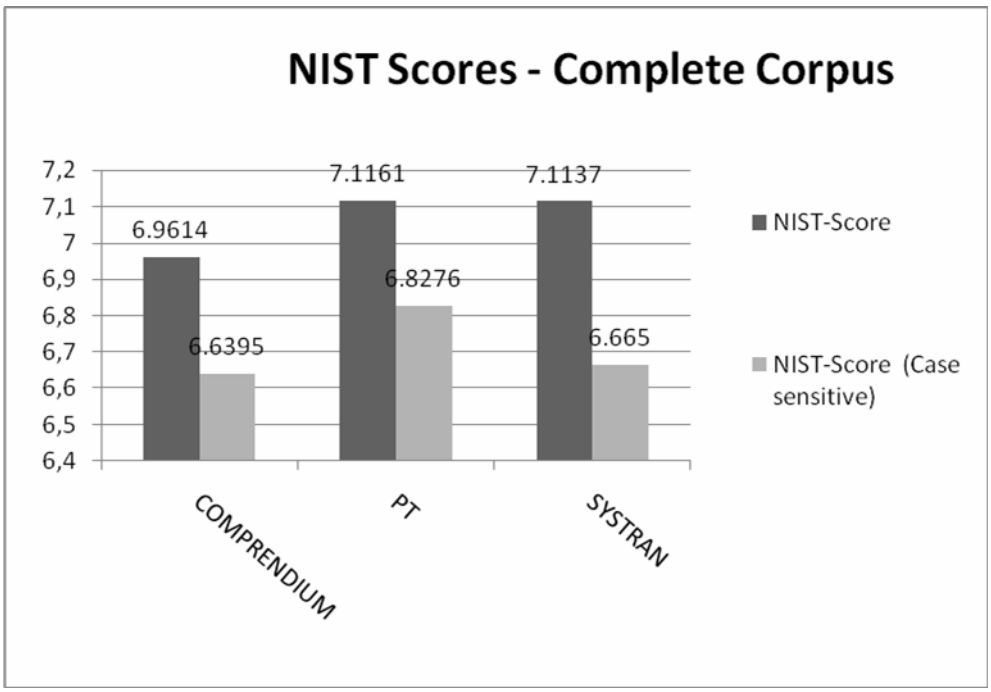

Figure 74: NIST Scores-Complete Corpus

COMPLETE CORPUS - RA (529 segments)		
	NIST-Score	**NIST-Score (Case sensitive)**
COMPRENDIUM	5.6159	5.4405
PT	6.1721	6.0638
SYSTRAN	5.9779	5.3112

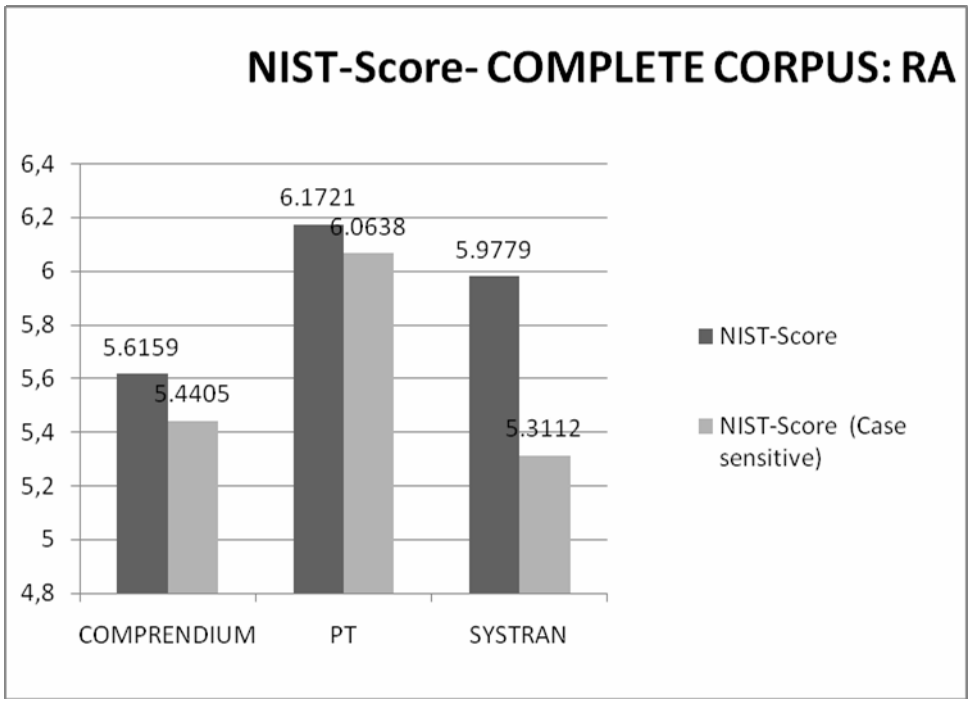

Figure 75: NIST Scores-Complete Corpus (RA)

COMPLETE CORPUS - SBTs (2733 segments)		
	NIST-Score	**NIST-Score (Case sensitive)**
COMPRENDIUM	6.8853	6.5411
PT	6.9010	6.5784
SYSTRAN	6.9509	6.6057

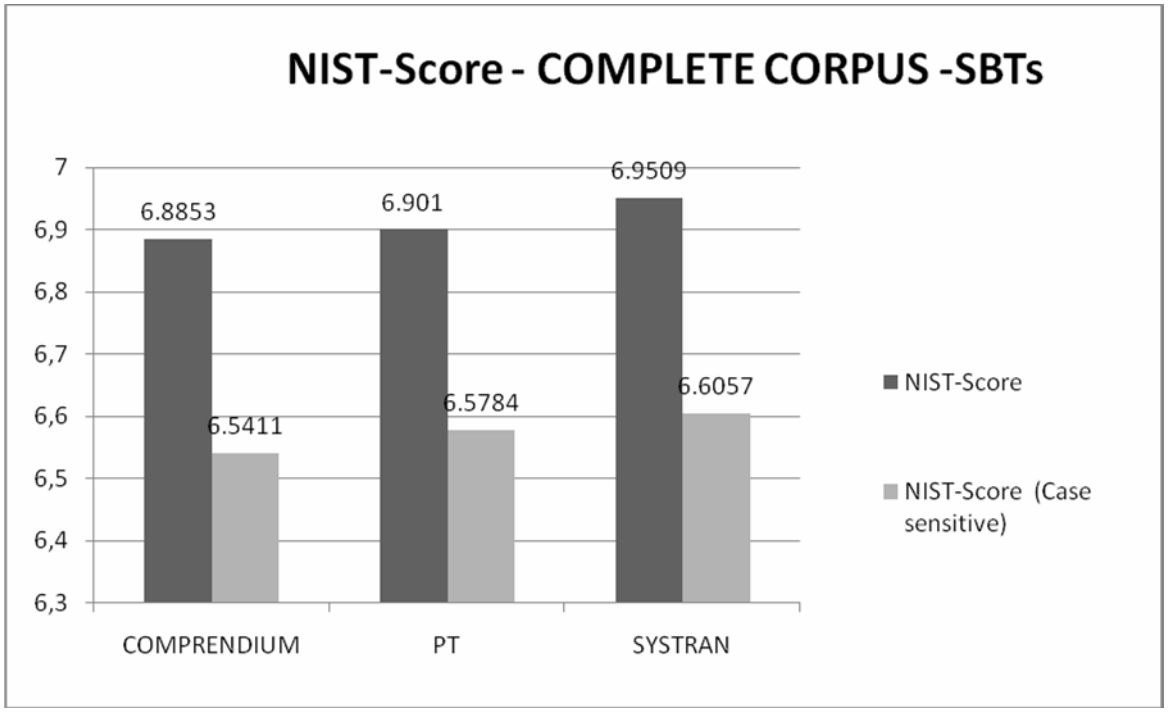

Figure 76: NIST Scores-Complete Corpus (SBT)

REDUCED CORPUS (228 segments)		
	NIST-Score	**NIST-Score (Case sensitive)**
COMPRENDIUM	5.8757	5.7231
PT	6.1528	6.0403
SYSTRAN	5.9743	5.6174

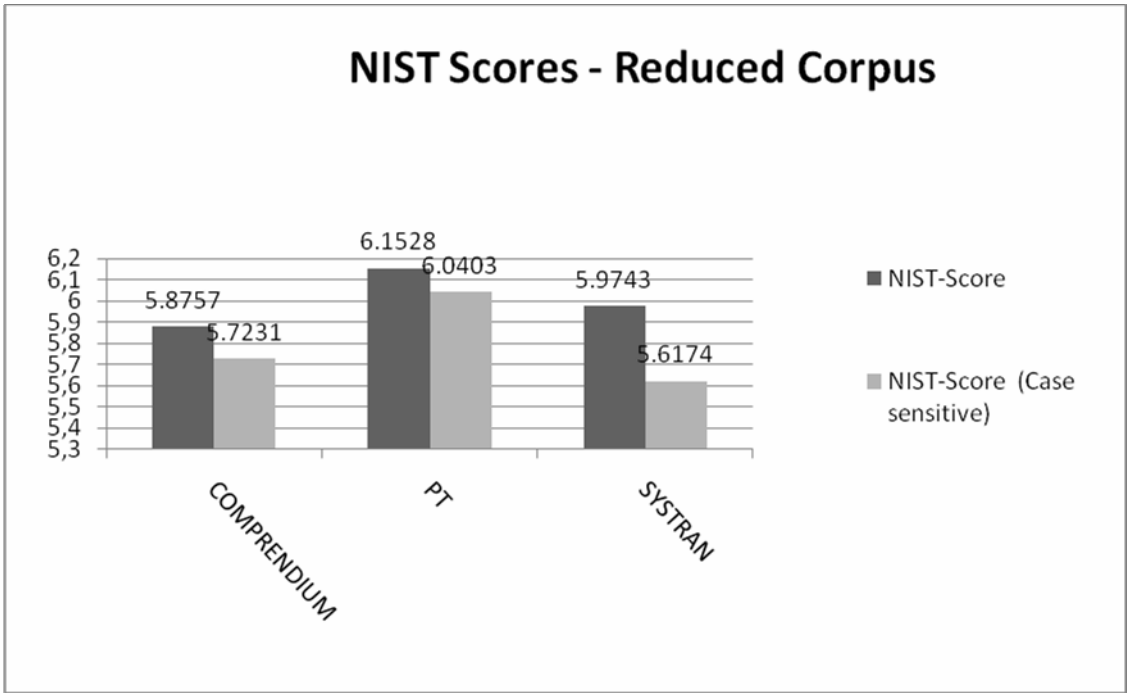

Figure 77: NIST Scores-Reduced Corpus

REDUCED CORPUS- RA MONOREF (121 segments)		
	NIST-Score	**NIST-Score (Case sensitive)**
COMPRENDIUM	5.0049	4.8653
PT	5.5889	5.4842
SYSTRAN	5.4626	4.9338

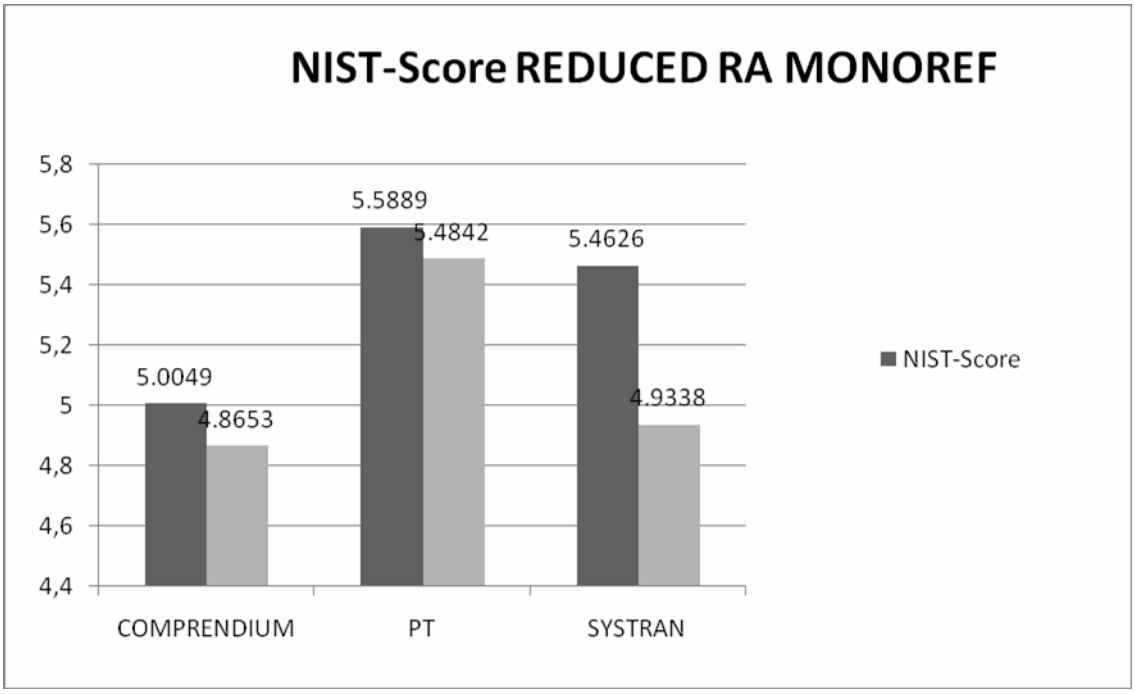

Figure 78: NIST Scores-Reduced Corpus (RA-Monoreference)

REDUCED CORPUS - SBT MONOREF (107 segments)		
	NIST-Score	**NIST-Score (Case sensitive)**
COMPRENDIUM	5.7273	5.5918
PT	5.7063	5.6072
SYSTRAN	5.507	5.3547

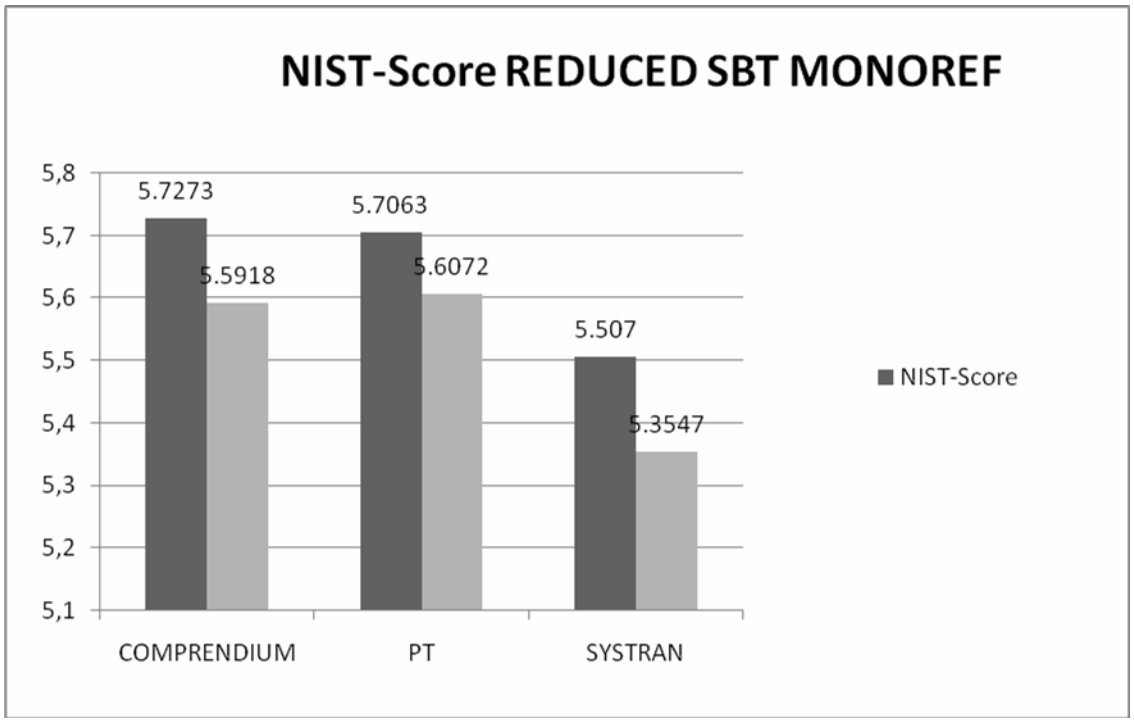

Figure 79: NIST Scores-Reduced Corpus (SBT-Monoreference)

REDUCED CORPUS RA MULTIREF (121 segments)		
	NIST-Score	**NIST-Score (Case sensitive)**
COMPRENDIUM	8.5467	8.3716
PT	9.5913	9.5133
SYSTRAN	9.0667	8.1088

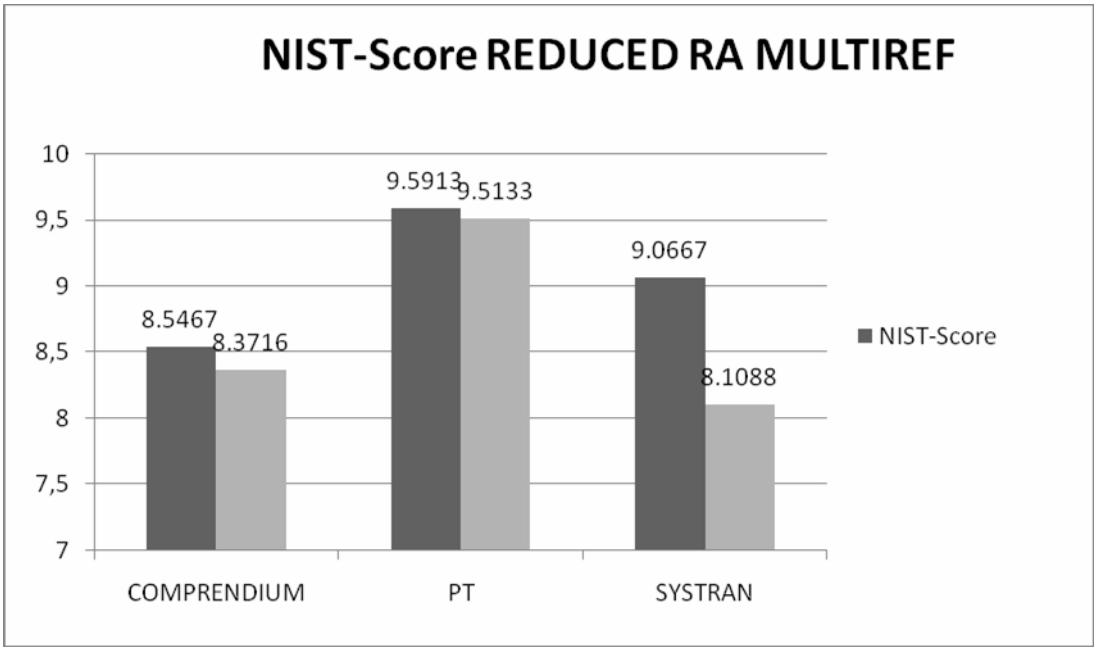

Figure 80: NIST Scores-Reduced Corpus (RA-Multireference)

REDUCED CORPUS SBT MULTIREF (107 segments)		
	NIST-Score	**NIST-Score (Case sensitive)**
COMPRENDIUM	10.1330	10.0447
PT	9.8104	9.771
SYSTRAN	10.5015	10.4597

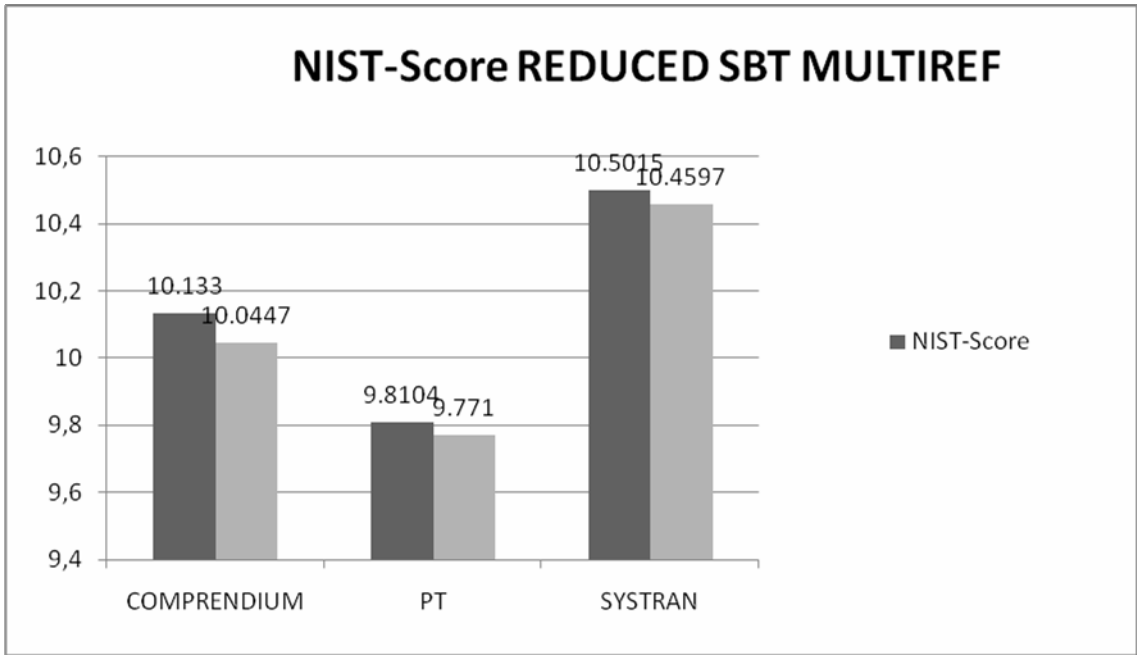

Figure 81: NIST Scores-Reduced Corpus (SBT-Multireference)

ANNEX IX: PHASE 2 EVALUATION -RESULTS BY EVALUATOR

	GERMAN TEST. ABSOLUTE VALUES			
	Improvement	**No effect (+)**	**No effect (-)**	**Worsening**
EVALUATOR 1	111	16	9	11
EVALUATOR 2	95	39	39	39
EVALUATOR 3	123	19	4	1
EVALUATOR 4	107	27	6	5
EVALUATOR 5	49	19	11	5
EVALUATOR 6	113	24	6	4
TOTAL	598	144	75	65

Table 51: German Test. Absolute frequencies.

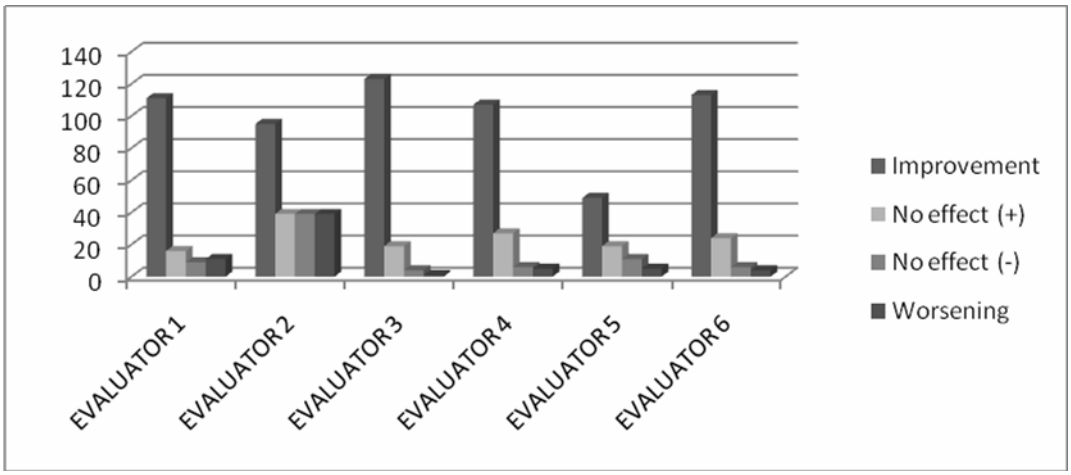

Figure 82: German Test. Absolute frequencies.

	GERMAN TEST: RELATIVE FREQUENCIES			
	Improvement	**No effect (+)**	**No effect (-)**	**Worsening**
EVALUATOR 1	75.51%	10.88%	6.12%	7.48%
EVALUATOR 2	44.81%	18.40%	18.40%	18.40%
EVALUATOR 3	83.67%	12.93%	2.72%	0.68%
EVALUATOR 4	73.79%	18.62%	4.14%	3.45%
EVALUATOR 5	58.33%	22.62%	13.10%	5.95%
EVALUATOR 6	76.87%	16.33%	4.08%	2.72%
TOTAL	67.80%	16.33%	8.50%	7.37%

Table 52: German Test. Relative Frequencies.

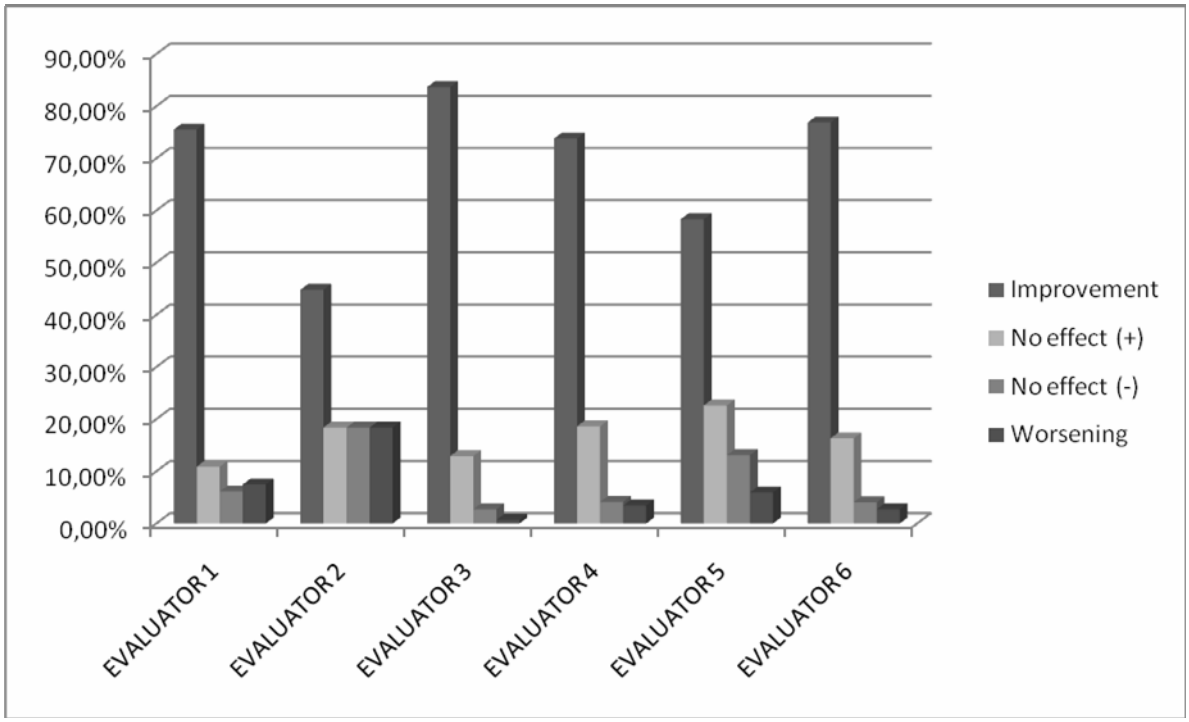

Figure 83: German Test. Relative Frequencies.

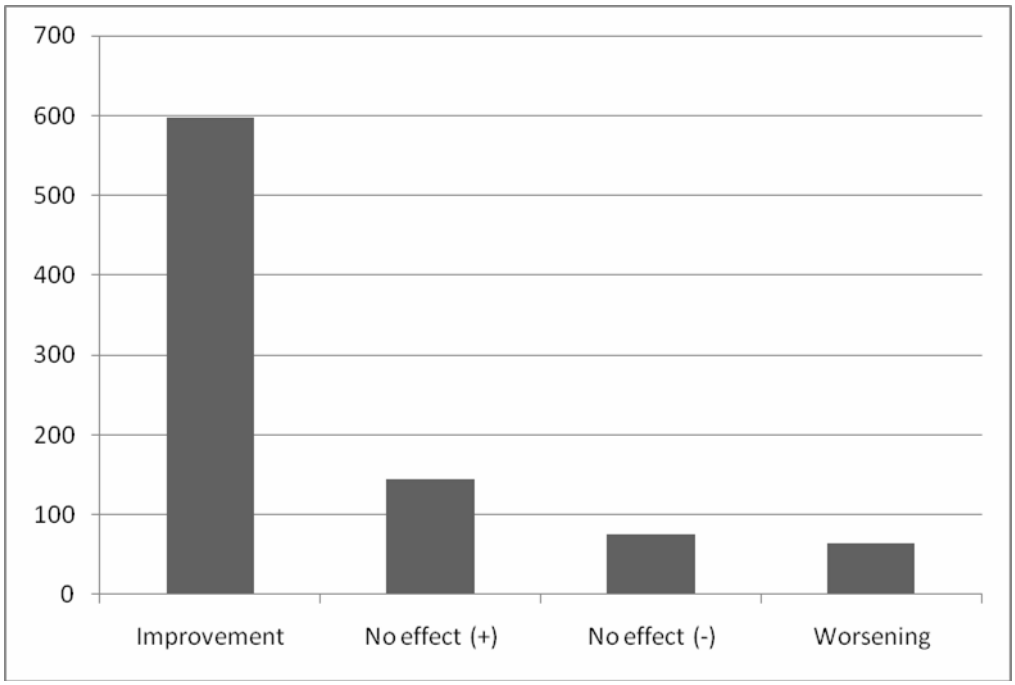

Figure 84: German Test. Total number of sentences. Absolute frequencies.

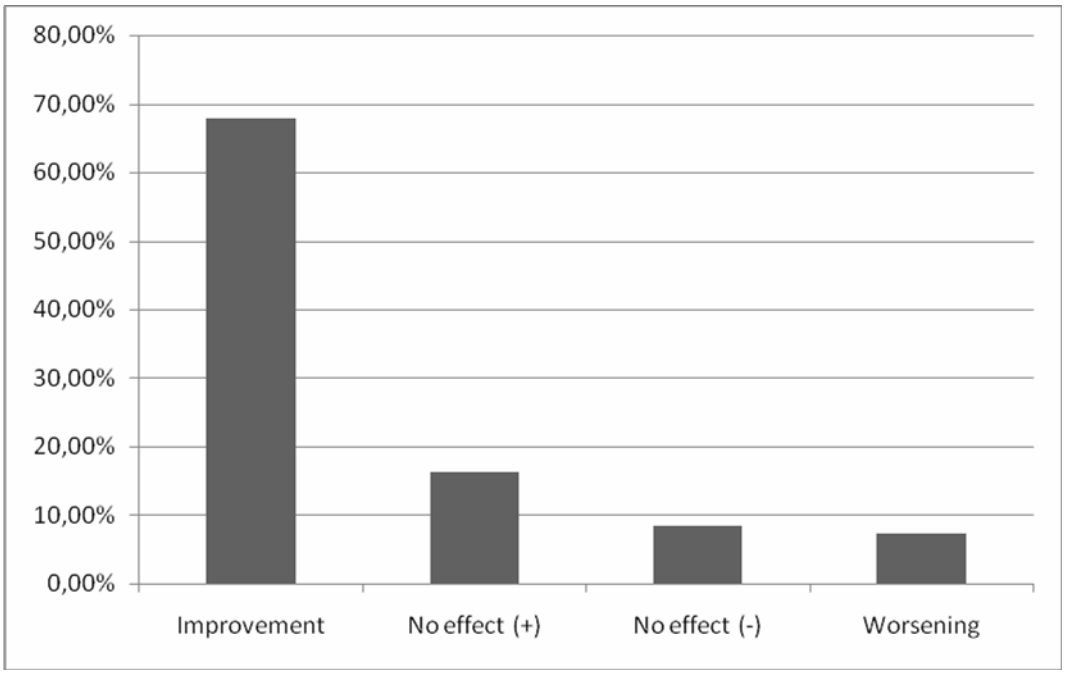

Figure 85: German Test. Total number of sentences. Relative frequencies.

	ENGLISH TEST. ABSOLUTE VALUES			
	Improvement	**No effect (+)**	**No effect (-)**	**Worsening**
EVALUATOR 1	57	43	35	12
EVALUATOR 2	54	58	23	11
EVALUATOR 3	53	31	42	19
TOTAL	164	132	100	42

Table 53: English Test. Absolute frequencies.

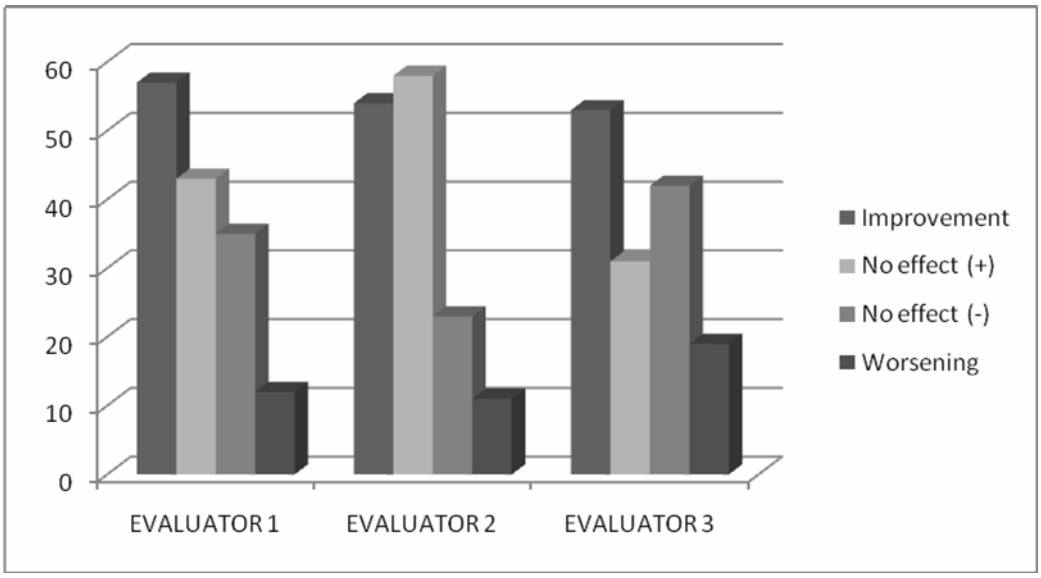

Figure 86: English Test. Absolute frequencies.

	PERCENTAGES			
	Improvement	**No effect (+)**	**No effect (-)**	**Worsening**
EVALUATOR 1	38.78%	29.25%	23.81%	8.16%
EVALUATOR 2	36.99%	39.73%	15.75%	7.53%
EVALUATOR 3	36.55%	21.38%	28.97%	13.10%
TOTAL	37.44%	30.14%	22.83%	9.59%

Table 54: English Test. Relative Frequencies.

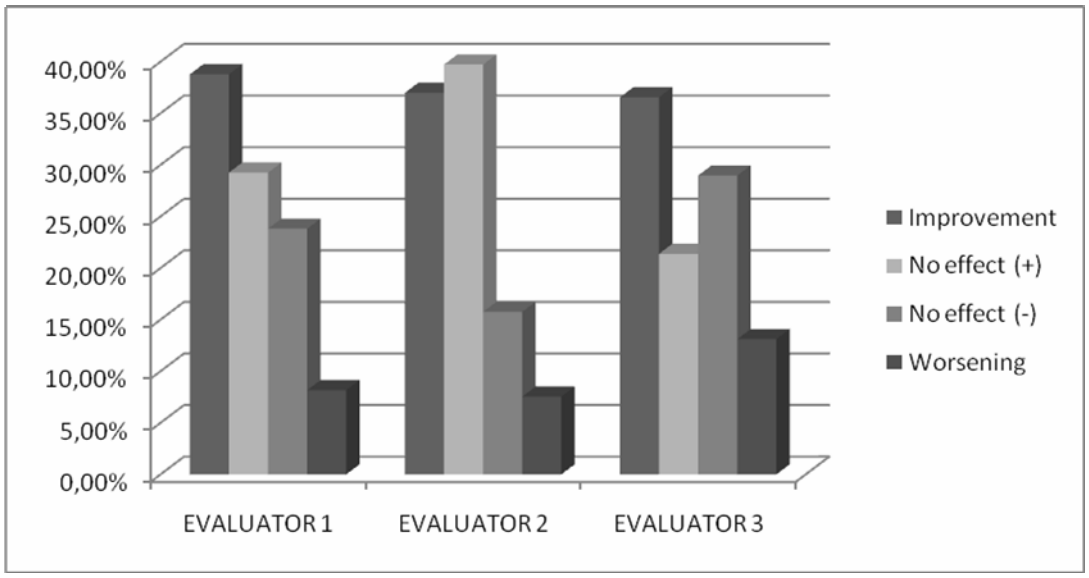

Figure 87: English Test. Relative Frequencies.

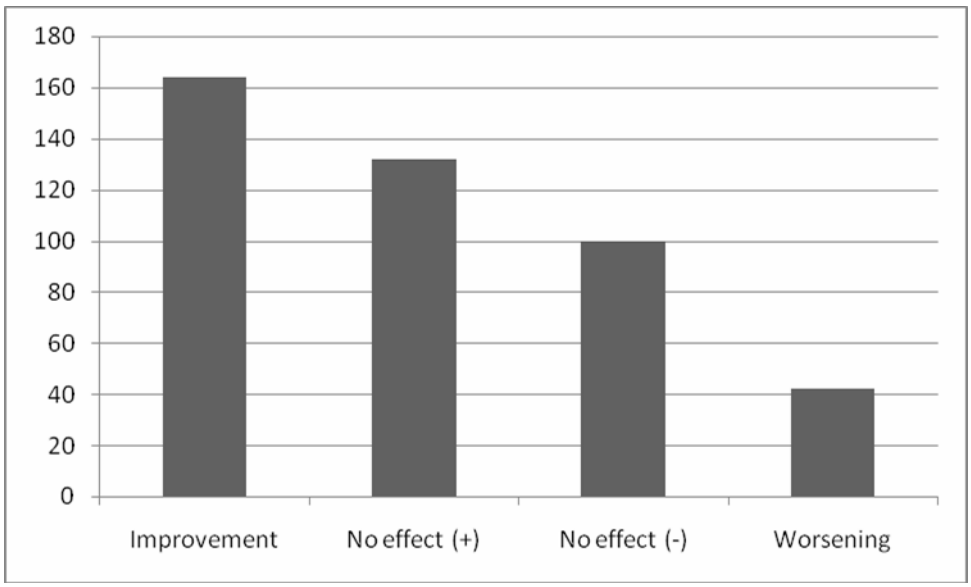

Figure 88: English Test. Total number of sentences. Absolute frequencies.

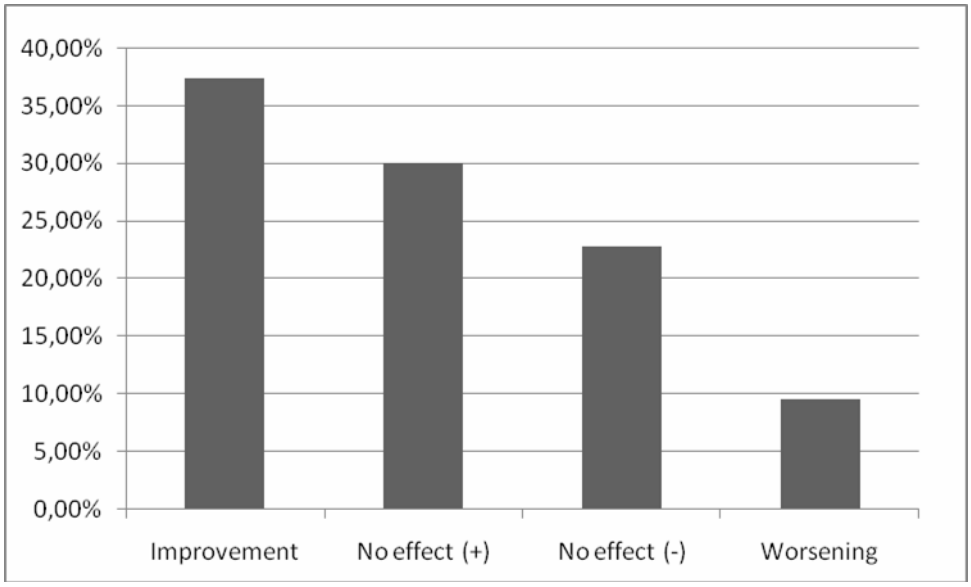

Figure 89: English Test. Total number of sentences. Relative frequencies.

		Kappa Values		
		Percent of overall agreement Po	**Fixed-marginal kappa**	**Free-marginal kappa**
German				
	5 evaluators	0.746032	0.187620	0.661376
	6 evaluators	0.586345	0.156667	0.44846
English				
	3 evaluators	0.606483	0.444616	0.475311

Table 55: Interannotator agreement with Kappa for Phase 2

ANNEX X: PHASE 2 EVALUATION - RESULTS BY CONTROL

ALL CONTROLS	GERMAN	ENGLISH
Worsening (1)	0.00%	0.00%
1.01-1.99	0.00%	8.52%
No effect - (2)	0.90%	13.45%
2.01-2.99	9.42%	22.42%
No effect + (3)	3.14%	16.14%
3.01-3.99	48.43%	13.90%
Improvement (4)	38.12%	25.56%

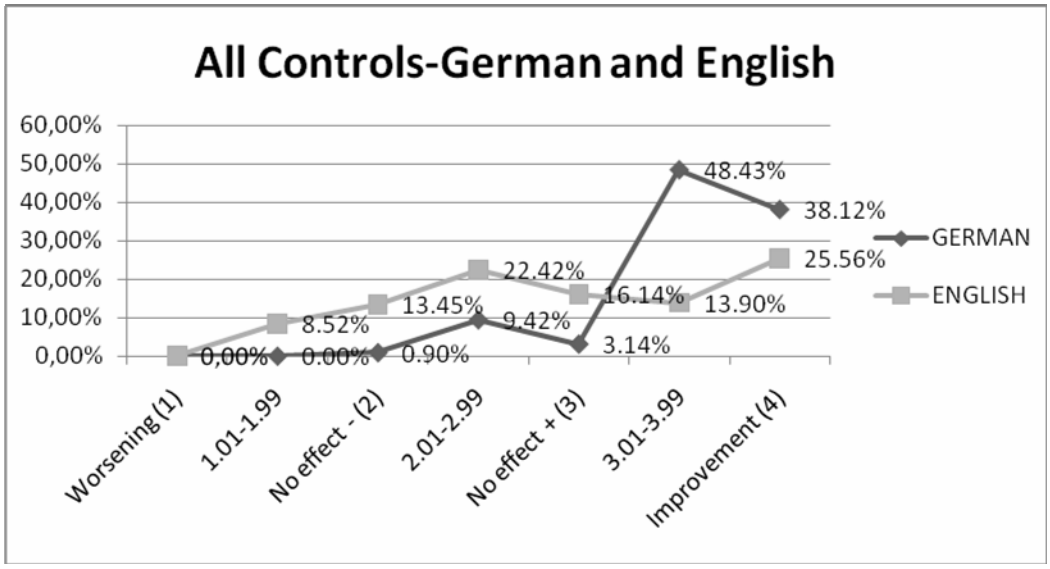

Figure 90: All Controls. Phase 2 Evaluation

GRAMMAR	GERMAN	ENGLISH
Worsening (1)	0.00%	0.00%
1.01-1.99	0.00%	7.50%
No effect - (2)	2.50%	12.50%
2.01-2.99	5.00%	22.50%
No effect + (3)	7.50%	12.50%
3.01-3.99	35.00%	17.50%
Improvement (4)	50.00%	27.50%

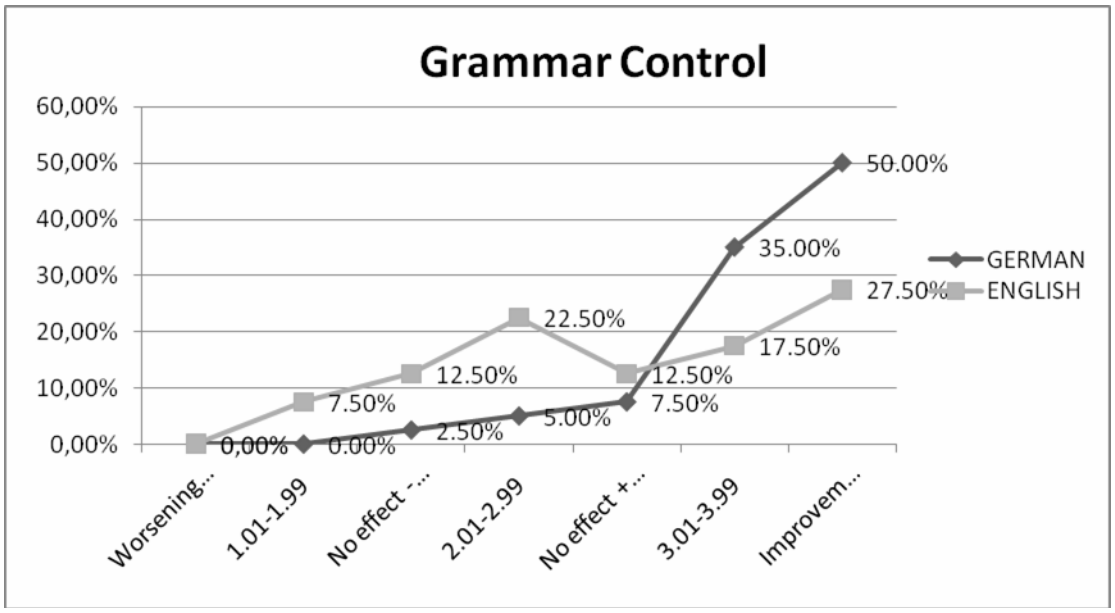

Figure 91: Grammar Control-Phase 2 Evaluation

	Total number of sentences affected	Improvement (4)		3.01-3.99		No effect + (3)		2.01-2.99		No effect - (2)		1.01-1.99		Worsening (1)	
		DE	EN	DE	EN	DE	EN	DE	EN	DE	EN	DE	EN	DE	EN
Orthotypography. Lower case after colon,	1						1			1					
Orthotypography. Bracket missing.	1			1			1								
Repeated word.	1	1			1										
Orthotypography. Hyphenation.	1					1							1		
Orthotypography. Comma between words.	1	1							1						
Orthotypography. Coma beween main and relative sentence.	1	1							1						
Orthography. Confusion between "dass" and "das".	2	1	2			1									
Orthotypography. Delete comma.	2	1		1					1		1				
Grammar. Concordance between subject and predicate.	2	1		1	1						1				
Grammar. Words should be written together.	2	2	1						1						
Orthotypography. Fixed space between number and measure.	3	2	1	1							1		1		
Orthotypography. Comma between main and subordinate clause.	4	1	1	2					1	1	2				
Grammar. Inflection (word ending)	6	4	2	2	3				1						
Grammar. Inflection.	7	2	1	4	1			1	4		1				
Orthotypography. Hyphenation between number, abbreviation and word.	9	4	2	4	2	1	3		1				1		

Table 56: Evaluation of sentences and grammar rules

ORTHOGRAPHY	GERMAN	ENGLISH
Worsening (1)	0.00%	0.00%
1.01-1.99	0.00%	6.12%
No effect - (2)	0.00%	12.24%
2.01-2.99	6.12%	26.53%
No effect + (3)	2.04%	8.16%
3.01-3.99	44.90%	8.16%
Improvement (4)	46.94%	38.78%

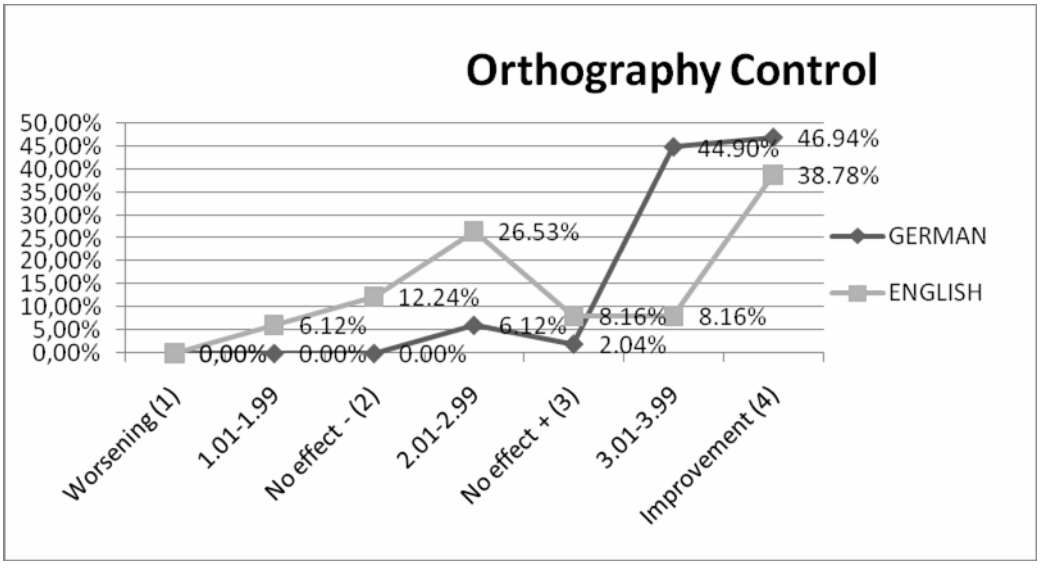

Figure 92: Orthography Control-Phase 2 evaluation

	Total number of sentences affected	Improvement (4)		3.01-3.99		No effect + (3)		2.01-2.99		No effect - (2)		1.01-1.99		Worsening (1)	
		DE	EN	DE	EN	DE	EN	DE	EN	DE	EN	DE	EN	DE	EN
The word is either a proper noun or is a misspelling	29	16	14	13	3		2		6		3		1		
A fixed space should be placed among elements of a multiword acronym	8	1		4		1	2	2	2		2		2		
Capitalise or lowercase the word	4	2	2	2	1				1						
Word has been written incorrectly	3	1		1				1	2		1				
Word has been written incorrectly regarding the new orthographic rules	4	2	2	2					2						
Word is a wrong compound	2	2	1						1						

Table 57: Evaluation of sentences and orthography rules

TERMINOLOGY	DEUTSCH	ENGLISCH
Worsening (1)	0.00%	0.00%
1.01-1.99	0.00%	8.86%
No effect - (2)	1.27%	10.13%
2.01-2.99	11.39%	26.58%
No effect + (3)	1.27%	12.66%
3.01-3.99	59.49%	17.72%
Improvement (4)	26.58%	24.05%

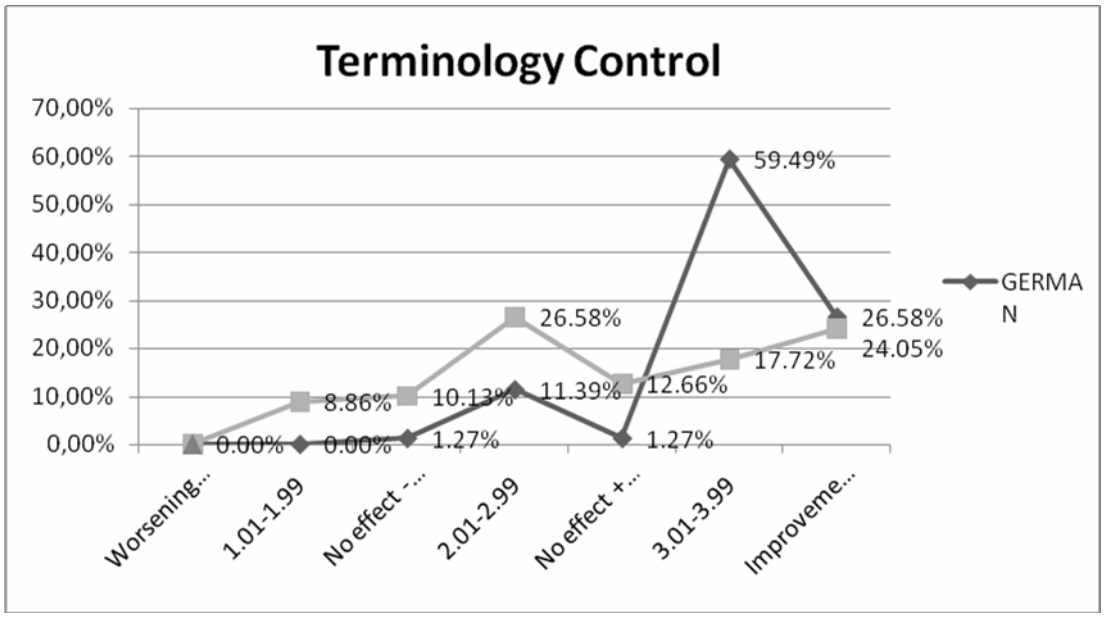

Figure 93: Terminology Control-Phase 2 evaluation

	Total number of sentences affected	Improvement (4)		3.01-3.99		No effect + (3)		2.01-2.99		No effect - (2)		1.01-1.99		Worsening (1)	
		DE	EN	DE	EN	DE	EN	DE	EN	DE	EN	DE	EN	DE	EN
Deprecated terms	36	10	11	22	7		6	3	8	1	2		2		
The term is not stored in the database. Please use this term instead.	20	3	2	14	5	1	2	2	6		3		2		
The term can be deprecated depending on the context	22	9	5	11	3		3	2	6		2		3		
Abbreviation	1				1		1								

Table 58: Evaluation of sentences and terminology rules

STYLE	DEUTSCH	ENGLISCH
Worsening (1)	0.00%	0.00%
1.01-1.99	1.82%	10.91%
No effect - (2)	0.00%	20.00%
2.01-2.99	10.91%	12.73%
No effect + (3)	3.64%	30.91%
3.01-3.99	45.45%	10.91%
Improvement (4)	38.18%	14.55%

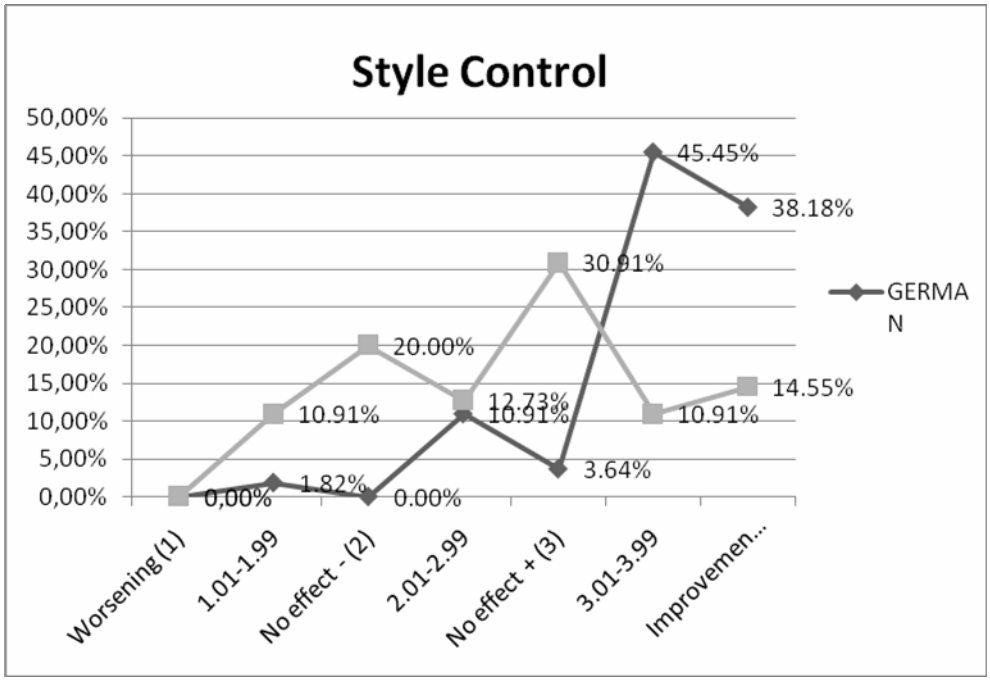

Figure 94: Style Control-Phase 2 evaluation

	Total number of sentences affected	Improvement (4)		3.01-3.99		No effect + (3)		2.01-2.99		No effect - (2)		1.01-1.99		Worsening (1)	
		DE	EN	DE	EN	DE	EN	DE	EN	DE	EN	DE	EN	DE	EN
Avoid the use of "I"	1	1									1				
Substitute the pronoun to avoid ambiguities	5	5	1				2		1		1				
Use a more meaningful verb	3			1	1		1	2			1				
Substitute "ausserdem" for "zudem"	2		1	2			1								
Use "wenn" to express conditional sentences	2	1		1			2								
Reduce the insertion	2	1	1	1	1										
Too many nouns. Paraphrase.	12	4	3	6	1		1	2	1		5		1		
Reduce or split the sentence in two.	9	2	2	5			2	2	3		1		1		
Split the sentence in two if possible	6	4		2	1		4				1				
Formulate the content in brackets in a separate sentence if possible	2			2	1				1						
Limit the number of insertions in brackets to one	1			1			1								
Use the active voice	1			1					1						
Use a verb	1		1	1											
In an instruction, write the verb in the imperative ("commanding") form.	1							1					1		
Represent the enumeration as a list	1							1					1		
Use "nicht" to express negation	1	1	1												
Use demonstrative forms after prepositions or the contracted form (e.g. am, vom)	3	2		1			2				1				
Avoid the use of "im ControlDisplay" and use instead "am ControlDisplay"	4			2		2					1		3		

Table 59: Evaluation of sentences and style rules

ANNEX XI: OVERVIEW OF MT CASE STUDIES

Company	TMS	Terminology MS	MT System	Language Pairs	Productivity gain (Time)	Savings	Scenario	Local Installation
Baan Development B.V.	TRANSIT (Version 2.7)	TermStar	Logos	De → En	up to 50%	No data	Online Help Texts	NO
CNH	SDLX	TermBase PhraseFinder	SDL KbT (Knowledge based Translation System)	En → Fr, It, De, Es, Nl, Da, Po	60%	50%	Technical Support Database	NO
VW	TRADOS	MultiTerm	COMPRENDIUM	De ↔ En, Es, Fr	No data	No data	Intranet Portal (E-mails, Reports)/ Assembly Instructions	NO
SAP	TRADOS	MultiTerm	LOGOS PROMPT METAL LOGOVISTA	En → Fr, Es En → Ru, Po De → En En → Ja	30%	Up to 40%	Documentation material, training courses "SAP notes" (Technical Support)	MIXED

Table 60: Overview of MT Case Studies

ANNEX XII: ROI CALCULATION

	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
Proposal										
Translation costs with MT	12,081.97 €	20,602.90 €	23,770.65 €	47,623.12 €	35,459.32 €	74,643.34 €	47,147.99 €	101,663.57 €	58,836.67 €	128,683.79 €
Maximal translation costs with MT		16,342.43 €	22,186.77 €	35,696.88 €	41,541.22 €	55,051.33 €	60,895.67 €	74,405.78 €	80,250.12 €	93,760.23 €
Implementation costs						81,385.50 €	25,159.80 €	25,159.80 €	25,159.80 €	25,159.80 €
Cash outflows: implementation costs plus translation costs						-136,436.83 €	-86,055.47 €	-99,565.58 €	-105,409.92 €	-118,920.03 €
Savings (Benefits): cash inflows						14,655.81 €	14,454.37 €	20,453.89 €	18,401.43 €	26,251.97 €
Net cash flow						-66,729.69 €	-10,705.43 €	-4,705.91 €	-6,758.37 €	1,092.17 €
Business as usual										
Translation costs without MT	14,695.16 €	23,662.55 €	30,330.90 €	56,480.85 €	45,966.63 €	89,299.16 €	61,602.37 €	122,117.46 €	77,238.10 €	154,935.76 €
Maximal translation costs without MT: cash outflows					-51,223.74 €	-67,632.89 €	-75,450.76 €	-91,859.91 €	-99,677.78 €	-116,086.93 €
Benefits: cash inflows					0.00 €	0.00 €	0.00 €	0.00 €	0.00 €	0.00 €
Net cash flow					-51,223.74 €	-67,632.89 €	-75,450.76 €	-91,859.91 €	-99,677.78 €	-116,086.93 €
Incremental Cash Flows										
Total incremental inflows						14,655.81 €	14,454.37 €	20,453.89 €	18,401.43 €	26,251.97 €
Total incremental outflows						-68,803.94 €	-10,604.71	-7,705.67 €	-5,732.14 €	-2,833.10 €

							€			
	2002	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**	**2009**	**2010**	**2011**
Net incremental cash flow						-54,148.13 €	3,849.67 €	12,748.22 €	12,669.30 €	23,418.87 €
Cumulative Incremental Cash Flow						-54,148.13 €	-50,298.46 €	-37,550.24 €	-24,880.94 €	-1,462.07 €
Payback Period										
Net incremental cash flow						-54,148.13 €	3,849.67 €	12,748.22 €	12,669.30 €	23,418.87 €
Cumulative Incremental Cash Flow						-54,148.13 €	-50,298.46 €	-37,550.24 €	-24,880.94 €	-1,462.07 €
Payback Period	5.1	Years								
ROI										
						-78.70%	-63.34%	-43.10%	-26.80%	-1.53%

Table 61: Overview of Calculations for ROI

	2012	2013	2014	2015	2016	2017	2018	2019
Proposal								
Translation costs with MT	70.525,34 €	155.704,01 €	82.214,02 €	182.724,24 €	93.902,69 €	209.744,46 €	105.591,37 €	236.764,68 €
Maximal translation costs with MT	99.604,57 €	113.114,68 €	118.959,02 €	132.469,13 €	138.313,46 €	151.823,58 €	157.667,91 €	171.178,02 €
Implementation costs	25.159,80 €	25.159,90 €	25.159,90 €	25.159,90 €	25.159,90 €	25.159,90 €	25.159,90 €	25.159,80 €
Cash outflows: implementation costs plus translation costs	-124.764,37 €	-138.274,48 €	-144.118,82 €	-157.628,93 €	-163.473,26 €	-176.983,38 €	-182.827,71 €	-196.337,82 €
Savings (Benefits): cash inflows	22.348,49 €	32.050,05 €	26.295,56 €	37.848,12 €	30.242,62 €	43.646,20 €	34.189,68 €	49.444,28 €
Net cash flow	-2.811,31 €	6.890,25 €	1.135,76 €	12.688,32 €	5.082,82 €	18.486,40 €	9.029,88 €	24.284,48 €
Business as usual								
Translation costs without MT	92.873,84 €	187.754,06 €	108.509,57 €	220.572,36 €	124.145,31 €	253.390,66 €	139.781,04 €	286.208,96 €
Maximal translation costs without MT: cash outflows	-123.904,80 €	-140.313,95 €	-148.131,82 €	-164.540,97 €	-172.358,83 €	-188.767,98 €	-196.585,85 €	-212.995,00 €
Benefits: cash inflows	0,00 €	0,00 €	0,00 €	0,00 €	0,00 €	0,00 €	0,00 €	0,00 €
Net cash flow	-123.904,80 €	-140.313,95 €	-148.131,82 €	-164.540,97 €	-172.358,83 €	-188.767,98 €	-196.585,85 €	-212.995,00 €
Incremental Cash Flows								
Total incremental inflows	22.348,49 €	32.050,05 €	26.295,56 €	37.848,12 €	30.242,62 €	43.646,20 €	34.189,68 €	49.444,28 €
Total incremental outflows	-859,57 €	2.039,47 €	4.013,00 €	6.912,04 €	8.885,57 €	11.784,61 €	13.758,14 €	16.657,18 €
Net incremental cash flow	21.488,93 €	34.089,52 €	30.308,56 €	44.760,16 €	39.128,19 €	55.430,81 €	47.947,82 €	66.101,46 €
Cumulative Incremental Cash Flow	20.026,85 €	54.116,37 €	84.424,93 €	129.185,09 €	168.313,28 €	223.744,09 €	271.691,91 €	337.793,36 €
Payback Period								
Net incremental cash flow	21.488,93 €	34.089,52 €	30.308,56 €	44.760,16 €	39.128,19 €	55.430,81 €	47.947,82 €	66.101,46 €
Cumulative Incremental Cash Flow	20.026,85 €	54.116,37 €	84.424,93 €	129.185,09 €	168.313,28 €	223.744,09 €	271.691,91 €	337.793,36 €
Payback Period								
ROI								
	20,74%	57,27%	93,30%	154,57%	225,35%	355,69%	552,82%	1039,71%

Table 62: Overview of calculations for ROI