Corporate Language Resources in Multilingual Content Creation, Maintenance and Leverage

Elia YUSTE
Institute of Computational Linguistics
University of Zurich
Winterthurerstrasse 190
CH-8057 Zürich
Switzerland
yuste@ifi.unizh.ch

Abstract

This paper focuses on how language resources (LR) for translation (hence LR4Trans) feature, and should ideally feature, within a corporate workflow of multilingual content development. The envisaged scenario will be that of a content management system that acknowledges the value of LR4Trans in the organisation as a key component and corporate knowledge resource.

1 Introduction

Corporations willing to go multilingual face two main difficulties, especially at the beginning. The first one is that many organisations are not acquainted with the *processes* behind internationalising their many products, documents, web pages and database interfaces effectively, so they opt to reduce the costs of localisation (and some even do not dare to step in and consequently remain monolingual). The second problem, which may derive from problem number one, relates to the fact that the corporation is then likely to end up hiring the wrong translation team or language service vendor after promising a top quality *product*[1] quickly and inexpensively.

Unfortunately, qualified and truly skilled candidates for posts in translation, localisation, internationalisation, and language-related project management are very difficult to find. Despite the growing competition among language service vendors and providers, the language industry is relatively immature. It is composed of young language service companies that are highly *project-driven*. The lessons learned in one project may be left behind and are often not assembled in a baseline knowledge solution to be retrieved and leveraged later.

The fact that most companies, regardless of whether they outsource[2] their translation jobs or have their own inhouse language service department, pay little attention to the integration, reusability of and interaction with *language resources for translation (LR4Trans)* within a project, let alone from project to project, constitutes a less than desirable panorama for the *creation* of corporate multilingual content.

So far, the *leverage* of LR4Trans has been limited to translation memory systems, where previously translated content is available to the translator through a software tool. This technology is not by any means new or highly sophisticated. While insufficient attention has been given to the integration of translation memories with other language resources and technologies[3] in the workflow, modes of accessing translation memory databases have evolved from purely standalone to distributed data, either synchronised with a central database or as a remotely[4] accessed central database.

[1] Translation (any form thereof, human, machine, technical, scientific, commercial, written, oral, etc.) involves both *process* and *product.* One should not put under scrutiny just the latter and ignore the former.

[2] Since the 1980s and 1990s, outsourcing of translation, as of many highly specialised business processes, has become prevalent. In an attempt to lower translation-mediated communication costs, most activities, even the application of translation memory tools, are managed outside the boundaries of the corporate firewall.

[3] A good start for this, though, are Bruckner & Plitt 2001 and Mügge 2001.

[4] This is the typical situation when outsourcing, whereby the content moves out of the source language

As a result, translation memories have gradually become widely adopted and almost the indispensable *tool* of the trade. Commercial producers of translation memory packages claim that, if properly used and maintained, they are valuable *corporate knowledge resources*.

The question is whether translation memories constitute the only possible corporate language resource containing corporate knowledge, or whether there can be other components, agents and processes that play an important role in multilingual content as well.

In the next section we would like to carefully examine the notion of knowledge in connection with those of LR4Trans and multilingual content.

2 From theory to practice

2.1 Language resources and Knowledge

The breadth and depth of knowledge required today in order to perform a good quality technical or specialised corporate translation relies upon a panoply of language resources (LR) in machine-readable form, which are self-created in the corporation or purchased from external parties (sister organisations, domain-specific specialist groups and societies, applied software and solution companies, etc.). In this panoply of corporate LR4Trans, one may find domain specific terminology, source and target language dictionaries of corporation-dependant word meanings, source and target language structures and rules, a corporate language stylesheet, appendix of phrases and expressions denoting cultural differences within a (multinational) corporation or when attempting global expansion, prescriptive and descriptive notes about the corporation "culture", among others.

All these resources contain precious corporate knowledge that should be taken into consideration and be made accessible to all corporate members and partners accordingly. Tagging or flagging the knowledge in those language resources will be extremely useful for optimising - on a constant basis - not only the resources themselves but the whole of the multilingual content production process. Tags or flags, normally called content properties, content attibutes or metadata, are aimed at retrieving a content unit when necessary and preventing loss of content. It is precisely thanks to these attributes, often visualised to the user by means colours or other agreed conventions, that a content management system can *manage* content even if it moves across multiple languages or sites.

Capturing that knowledge will thus be helpful when developing scalable and adaptive applications for *managing* corporate multilingual content.

2.2 From LR4Trans to knowledge repositories and content management systems

A corporate knowledge-geared multilingual content strategy is open to a varying degree of automation, in terms of not only linguistic processing but also in content transaction[5] operations, on the basis of the type of documentation, business conditioning factors, etc. It usually combines tightly integrated translation technologies (and maybe other kind of human language technologies) with human specialist intervention, i.e. unique[6] language work processes, which have to be driven by highly skilled linguists.

This form of knowledge-based translation work aims to bridge the gap between low cost, poor output machine-only translation and costly high-quality human-only translation. Although this could be seen as a type of machine-aided human translation (MAHT), we would like to emphasise the issue of knowledge, corporate knowledge in particular, which precisely ought to be captured into the translation system's knowledge base.

This corporate knowledge base, characterised for being configurable and updatable, will detect and classify the knowledge present in the language resources into: general knowledge, domain-specific knowledge and, knowledge specific to each individual customer or department within the organisation.

The knowledge base will nonetheless be acting as a single repository with the following possible functions[7]: automated identification of terms that

content repository into an external translation process, and then returns in one or more new languages – a further challenge, especially if there is not yet an effective content management system in place.

[5] Transaction costs can outweigh translation costs, especially when the creation and maintenance of multilingual content is required for e-learning or e-customer support.

[6] Ideally tailor-made and customisable, that is, conceived for the corporation or the client they work in or for.

[7] These functions will be linked to one another and called according to the stage of multilingual content creation we are in. A function or component may be called more than once within the multilingual content

are candidates for once-only translation; spotting of translation for terms from previously translated, aligned texts; semi- or automated creation of domain and/or customer-specific terminology, dictionaries and glossaries; creation and regular update of domain and/or customer-specific language rules; implementation of domain and/or customer-specific translation memories; dynamic and integrative machine translation, making use of customised dictionaries (lexicons) and language rules; translation and edition application, ideally increasing ease of use by showing colour-coded aligned bi-texts (bilingual) or multi-texts (multilingual) with a context expansion feature and highlighting terminology; and, most importantly, automated and user-dependant feedback of new knowledge into the knowledge base.

3 The envisaged scenario: workflows, content management systems, and agents

Having the corporate knowledge base linked to various LR4Trans, as presented in section 2 makes us think of a procedural and very agile multilingual content workflow. But let us examine it in greater detail starting to look at the simplest of workflows first (figure 1):

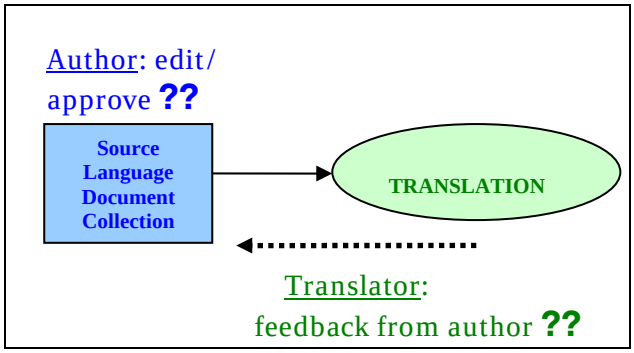

Figure 1.- Typical Workflow diagram

Typically authors do not edit or approve content for translation. The texts to translate, usually from an unstructured document collection, are handed over to the translator, who does translate without having the chance to get feedback from the authorship department. There is neither an obvious use of language resources in machine-readable form nor a corporate knowledge detection and exploitation strategy in operation. This poor production process will have negative consequences in terms of the quality of the product translation (e.g. lacking consistency, frequent content losses, etc.) and costs, particularly in the long run.

In order to streamline the procurement and management of corporate multilingual content we propose the following workflow (see figure 2 in the appendix). Its main assets would be an overall corporate knowledge base linked to various LR4Trans, as appropriate, and maintained by all agents[8] intervening in the workflow, plus a content management system, or CMS, that would reflect the business roles controlling the workflow, data production and update flow, user roles and access privileges, costing rules, etc.

In contrast to figure 1, the following features can be found in the workflow presented in figure 2:

- Cyclic nature of content, from monolingual to multilingual, and back to enhance and expand the first;
- Corporate content is traceable and its state and structure can be followed-up at all times;
- Authors are aware of what happens at the other end, and so are capable of "writing for translation", that is, editing or approving content that will be later received by an audience or market of another language and culture. In other words, the package of the content starts being taken care of from the beginning;
- Translators are connected with the authoring department: the concepts of content negotiation and feedback are essential here. Translators, being intercultural mediators, have a strong say in issues of international content relevancy.

CMS are meant to work seamlessly in the background, automatically identifying changes in the content (e.g. keeping track of the content production or processing stage, keeping a log of agent participation, etc.) by means of a built-in feedback loop mechanism. Besides, a multilingual CMS comes to live action when, as some kind of document gate keeper and donor, passes on the content from one agent to another,

production lifecycle.

[8] By this we mean not only the multiskilled corporate linguist (who could be a translator, terminologist, editor, domain validator, cross-cultural consultant...), but also all those agents that construct and share the knowledge of a corporation, namely decision makers (i.e. management force), marketeers, legal specialists, and so on.

notifying him or her of any vital new piece of information: “a new translation has been received”, “glossary validated by expert XY and saved today at 18:27 hours”, “not possible to close up project before client acceptancy test”.

CMS are usually dependant on the corporate knowledge base. Together, they define the workflow and have interaction capabilities with the various users by means of secure interfaces, usually very similar to a web portal for internal and very often external use, too (mainly for workers or at different sites and clients).

Concerning language work, it is extremely important that both online and offline editing and review of content are allowed. In other words, the corporate knowledge base has to be centralised (online use) and yet distributed at times (offline use). It will be the system, which will manage the synchronisation of content and knowledge base alterations and updates across all the different user types.

The CMS thus relies heavily upon automated mechanisms (e.g. automatic updating of the translation memory once the project translations have gone through the review process) but needs skilled human intervention to improve its efficiency over time.

4 Conclusion

After introducing some commonplace problematic issues surrounding the creation, managing and leveraging of multilingual content, we have analysed the interrelation of corporate knowledge and language resources for translation in a corporate setting. It has been argued that the corporate knowledge residing in the corporation documentation and language resources has to be captured and introduced in a corporate knowledge base, which has to be made accessible to and constantly cared for by all agents intervening in the multilingual content workflow, not only by linguists. We have gone on to underline the importance of having a content management system in place, in order to account for and dynamise the tasks and processes within the workflow. Other relevant issues such as linkage between resources, knowledge base and CMS, and balance between automation and human intervention have been discussed.

5 Acknowledgements

My special thanks go to the two blind reviewers of this paper’s first draft. I would also like to thank my colleagues at the Institute for Computational Linguistics of the University of Zurich for their interesting questions during a recent presentation.

References

D. Barabé. 2003. Soaring demand, shrinking supply in translation: how we plan to make ends meet. *MT Summit IX*, New Orleans, USA, 23-27 September 2003. Presentation slides available at: http://www.amtaweb.org/summit/MTSummit/FinalPapers/MTSummit_Sept2003.ppt [Powerpoint file, last consulted: 20 May 2004]

U. Boehme & S. Svetova. 2001. An integrated solution: applying PROMT machine translation technology, terminology mining, and the TRADOS TWB translation memory to SAP content translation. *MT Summit VIII*, Santiago de Compostela, Spain, 18-22 September 2001; pp. 49-52.

C. Boitet. 2001. Four technical and organizational keys to handle more languages and improve quality (on demand) in MT. *MT Summit VIII*, Santiago de Compostela, Spain, 18-22 September 2001. Towards a Road Map for MT.

C. Bruckner & M. Plitt. 2001. Evaluating the operational benefit of using machine translation output as translation memory input. *MT Summit VIII*, Santiago de Compostela, Spain, 18-22 September 2001. Workshop on MT Evaluation.

J. A. Brundage. 2001. Machine translation – evolution not revolution. *MT Summit VIII*, Santiago de Compostela, Spain, 18-22 September 2001; pp.59-62.

A. Clarke. 2000. MT within Productive Translation Workflow. *Fifth EAMT Workshop "Harvesting existing resources"*, May 11 - 12, 2000, Ljubljana, Slovenia; pp.79-81.

M. Franco Sabarís, J.L. Rojas Alonso, C. Dafonte & B. Arcay. 2001. Multilingual authoring through an artificial language. *MT Summit VIII*, Santiago de Compostela, Spain, 18-22 September 2001; pp.99-102.

D. Gervais, 2003. MultiTrans™ system presentation: translation support and language management solutions. *MT Summit IX*, New Orleans, USA, 23-27 September 2003; pp.459-462.

J. Hutchings (ed). 1998. *Translation technology: integration in the workflow environment. EAMT Workshop*, WHO, Geneva, 2-3 April 1998.

C. Hyland. 2003. Testing “Prompt”: the development of a rapid post-editing service at CLS Corporate Language Services AG, Switzerland. *MT Summit IX*, New Orleans, USA, 23-27 September 2003; pp.189-193.

T. Lewis. 2001. Combining tools to improve automatic translation. *MT Summit VIII*, Santiago de Compostela, Spain, 18-22 September 2001; pp.207-209.

E. Maier & A Clarke. 2003. Scalability in MT systems. *MT Summit IX*, New Orleans, USA, 23-27 September 2003; pp.248-253.

E. Maier, A. Clarke & H.-U. Stadler. 2001. Evaluation of machine translation systems at CLS Corporate Language Services AG. *MT Summit VIII*, Santiago de Compostela, Spain, 18-22 September 2001; pp.223-228.

E. Macklovitch & A. Sánchez Valderrábanos. 2001. Rethinking interaction: the solution for high-quality MT? *MT Summit VIII*, Santiago de Compostela, Spain, 18-22 September 2001. Towards a Road Map for MT.

T. Mitamura, K. Baker, E. Nyberg & D. Svoboda. 2003a. Diagnostics for interactive controlled language checking. *Controlled language translation, EAMT-CLAW-03*, Dublin City University, 15-17 May 2003; pp. 87-94.

T. Mitamura, K. Baker, D. Svoboda, & E. Nyberg. 2003b. Source language diagnostics for MT. *MT Summit IX*, New Orleans, USA, 23-27 September 2003; pp.254-260.

T. Murata, M. Kitamura, T. Fukui & T. Sukehiro. 2003. Implementation of collaborative translation environment ‘Yakushite Net’. *MT Summit IX*, New Orleans, USA, 23-27 September 2003; pp.479-482. http://www.yakushite.net/ [last consulted: 20 May 2004]

U. Mügge. 2001. The Best of Two Worlds: Integrating Machine Translation intoTranslation Memory Systems - A universal approach based on the TMX standard. *Language International*, December 2001, John Benjamins, 26-29.

E. Nyberg, T. Mitamura, D. Svoboda, J. Ko, K. Baker, & J. Micher. 2003. An integrated system for source language checking, analysis and term management. *MT Summit IX*, New Orleans, USA, 23-27 September 2003; pp.487-490.

U. Reuther. 2003. Two in one – Can it work? Readability and translatability by means of controlled language. *Controlled language translation, EAMT-CLAW-03*, Dublin City University, 15-17 May 2003; pp.124-132.

A. Rinsche. 2000. Computer-assisted business process management for translation and localisation companies. *Fifth EAMT Workshop "Harvesting existing resources*", May 11 - 12, 2000, Ljubljana, Slovenia; pp.83-85.

A. Sánchez Valderrábanos, J. Esteban & L. Iraola. 2003. TransType2 – a new paradigm for translation automation. *MT Summit IX*, New Orleans, USA, 23-27 September 2003; pp.498-501.

F. Schäfer. 2003. MT post-editing: how to shed light on the "unknown task". Experiences at SAP. *Controlled language translation, EAMT-CLAW-03*, Dublin City University, 15-17 May 2003; pp.133-140.

J. Senellart, C. Boitet & L. Romary. 2003. SYSTRAN new generation: the XML translation workflow. *MT Summit IX*, New Orleans, USA, 23-27 September 2003; pp.338-345.

R. Smith. 2001. Using information technology to optimise translation processes at PricewaterhouseCoopers Madrid. *MT Summit VIII*, Santiago de Compostela, Spain, 18-22 September 2001; pp.341-344.

N. Underwood & B. Jongejan. 2001. Translatability checker: a tool to help decide whether to use MT. *MT Summit VIII*, Santiago de Compostela, Spain, 18-22 September 2001; pp.363-368.

J. van der Meer. 2003. At last translation automation becomes a reality: an anthology of the translation market. *Controlled language translation, EAMT-CLAW-03,* Dublin City University, 15-17 May 2003, pp. 180-184.

E. Yuste. 2002a. Language Resources and the Language Professional. In E. Yuste (Ed.) *Proceedings of the First International Workshop in Language Resources for Translation Work and Research*. LREC 2002, 28th May 2002. Las Palmas de Gran Canaria (Spain). Paris: ELRA (European Association for Language Resources). More information can be found at the workshop and post-workshop web sites at http://www.ifi.unizh.ch/cl/yuste/LREC/LR4Trans.html and http://www.ifi.unizh.ch/cl/yuste/postworkshop/postworkshop.html

E. Yuste. 2002b. MT and the Swiss language service providers: an analysis and training

perspective. *Sixth EAMT Workshop "Teaching machine translation"*, November 14-15, 2002, UMIST, Manchester, England; pp.23-32.

E. Yuste & S. Cerrella Bauer. 2004. In print**.** Circumstances, challenges and consequences of implementing a quality-geared and technology-aided process of translating: a case study. 90-minute workshop delivered at the *IV Jornadas Internacionales sobre la Formación y la Profesión del Traductor e Intérprete*. UEM,Madrid, 25th-27th February 2004. Paper[9] due to appear later on this year in resulting CD-ROM Proceedings. More info about this international conference is available at http://www.uem.es/traduccion/actividades/jornadas/.

T. Zervaki. 2002. *Globalize, Localize, Translate: Tips and Resources for Success*. Bloomington: 1st Book Library.

[9] The results in this paper are also discussed in two other articles by the same authors, due to appear by invitation in two other professional publications, namely in the *Bulletin of the DÜV (Agentur der Dolmetscher- und Übersetzervereinigung, Switzerland*, http://www.duev.ch), under the title "Implementing a quality-geared and technology-aided process of translating: a case study", and *Hieronymous*, the professional quarterly journal of the ASTTI (Association suisse des traducteurs, terminologues et interprètes, http://www.astti.ch), under the title "Circumstances, challenges and consequences of a quality-geared and technology-aided process of translating: a case study". [last consulted in May 2004]

APPENDIX

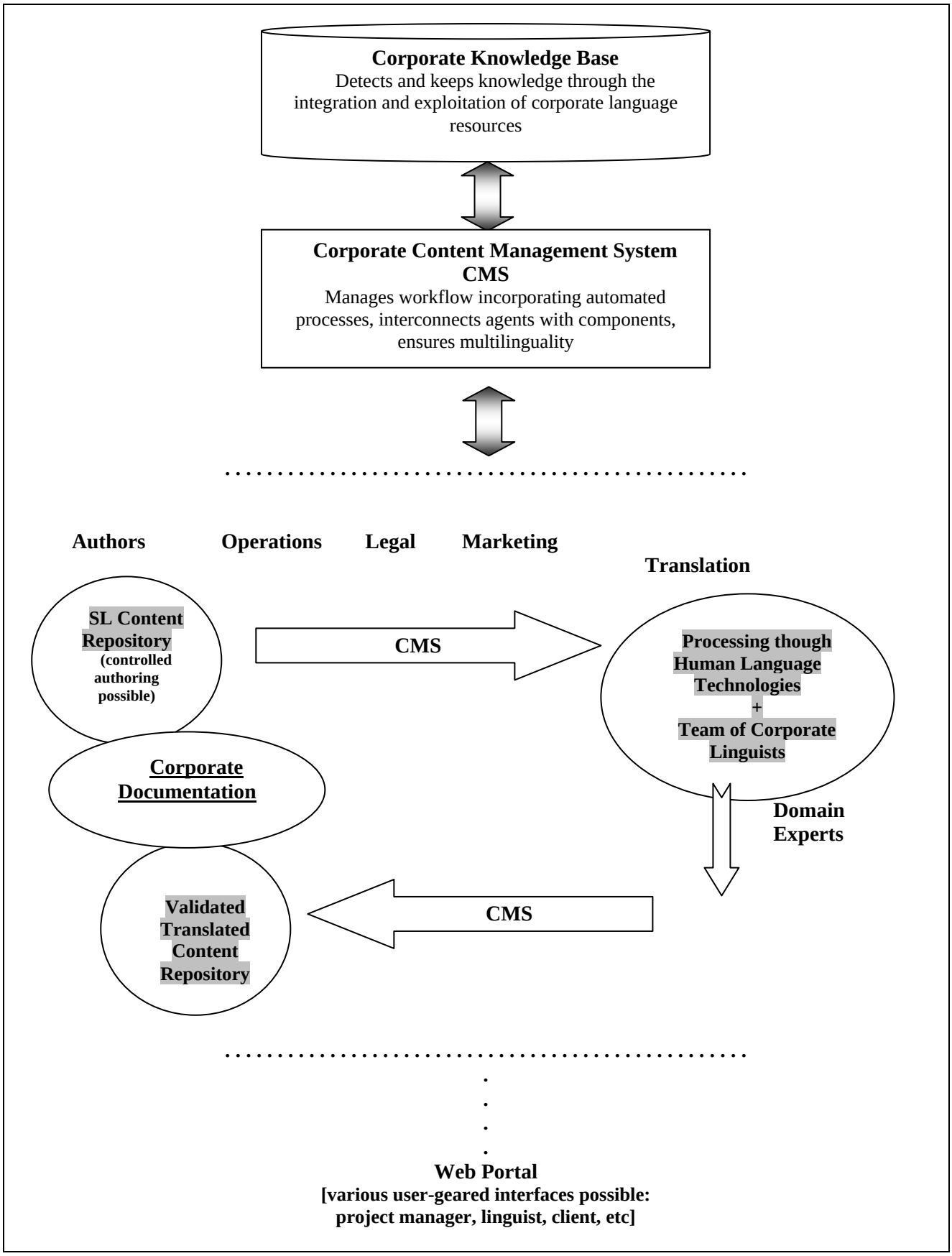

Figure 2.- Envisaged Corporate Multilingual Content Development with Knowledge Base and Content Management System