THE LOGOS TRANSLATABILITY INDEX

Claudia Gdaniec
Logos Corporation
200 Valley Road, Suite 400
Mt. Arlington, NJ 07856
tel: 201-398-8710
email: cgdaniec@logos-usa.com

Abstract

A serious problem for users of machine translation is estimating the suitability of a particular document for MT and, thus, the translation quality. The available utilities such as style and grammar checkers, well-formedness and readability measures have not proven sufficiently useful in this regard. Researchers at Logos Corporation therefore undertook a study of the feasibility of a translatability index (TI) based on the use of gross sentence properties such as length and other complexity factors; on gross text properties; as well as on the idiosyncrasies of the Logos system. The study was based on English and German test data drawn from a variety of sources and representing a variety of styles and subject areas. The results of this study suggest that a statistically-based special-purpose TI achieves significant and useful correlations with the perceived quality of raw translation, and, further, that improvement in the TI score effected by dictionary updates and pre-edits results in corresponding improvements in output quality. The concrete results of this work are two prototypes that measure and score the suitability of English and German documents for the LOGOS translation system.

1 Issues, Problems, and Assumptions

1.1 Statistical Basis and Scope of Applicability

Logos Corporation has developed a TI to automatically assess the suitability of a given input text for MT. The TI is based on the gross statistical properties of a document rather than on parsing its sentences. This decision was suggested by the fact that there always appeared to be a rough correlation between the quality of raw MT output and certain gross properties of a text, such as sentence length, degree of syntactic complexity, discourse characteristics, etc. It could be shown that such correlation can be found in the finer levels of statistical detail, and that it is indeed sufficiently consistent. Thus, statistical evaluation of a document can provide a measure of the translatability of that document by the LOGOS system. It was not expected, however, that the proposed method for evaluating a document on the basis of sentence properties could provide sentence-specific information with any degree of reliability. The TI applies to the corpus or document as a whole but is not useful in pinpointing problem sentences. For this, full-scale sentence parsing would be required.

1.2 Relative Significance of TI

The TI has **relative** significance for translatability, not absolute significance. It can tell a user that document A is more suitable for translation by the LOGOS MT system than document B. It cannot tell a user whether a given document will produce acceptable or unacceptable translations. This is something users will have to infer for themselves, on the basis of experience with the TI and raw MT output.

The scale for the TI imitates that of our quality index, on a scale from 1 through 7 (For scoring criteria, see the appendix). But, again, whatever the scale, it should be understood that the numbers have only relative significance, reflecting differences among users and the kind of documents they are working with. That is to say, the significance of the numbers would not apply equally to all users in all contexts.

In designing this project, we took into account the fact that measures of translatability and the evaluation of translations are necessarily subjective and therefore, while measures are clearly needed and may indeed be useful, they are impossible to fix in any absolute sense.

1.3 Relative Significance of the Quality Index (QI)

The QI used in this study to express MT output quality suffers from the same subjectivity that characterizes any evaluation of translation. However, the QI was specifically designed to minimize the impact of such subjectivity on the QI ratings. The QI's were assigned by a single evaluator for each target language, someone not connected with this study or with the Logos development team. More critically, the evaluators were asked to focus exclusively on grammatical correctness, understandability, and preservation of information and to ignore style. We assumed that style does not distinguish meaning in the typical document used for MT. On the contrary, the difference between one choice of form or word order and another is either meaningless or reflects the degree of sophistication on the part of the transfer module of the MT system. It was hoped to provide a fairly objective measure of quality against which to develop and test the TI. In fact, this objective was only partially met: experience indicates that the QI assignments do reflect an evaluator's subjective attitude. This shows in the different QI ratings of the German-source translations. The French translations got consistently higher scores than the German-English translations - contrary to the developers' expectations.

The QI has been in use at Logos for a number of years as a way of measuring development progress from release to release. Experience has shown that documents with QI averages of 4 or above are useful as input to post-editing, although we find that some translators are not willing to post-edit text that requires too much correcting, even when doing that might be cost-effective. The threshold of acceptability varies considerably from translator to translator and depends on the culture of the organization. Texts that average 5 or better are widely accepted.

1.4 Effect of Different Target Languages on TI/QI Correlation

The correlation between QI and TI does not remain constant for the same source language document over different target language translations. The TI relates to the source document,

while the QI relates to the raw translation. Experience shows that the translation of a given English document into German, French, Spanish, and Italian will yield different QI's. Therefore, the correlation between the TI and the QI will differ depending on the language pair. Some target languages are more forgiving, others less so, when analysis of the English source is poor. The Romance languages have a higher "free ride" factor than does German. This suggests that the TI has to be adjusted for each target to reflect this. An example of a "free ride" is the following: If the boundaries of a relative clause are determined incorrectly in the parse, that does not necessarily show up in the French translation because it follows the English word order. When the system translates into German, however, this wrong parse makes the verb show up in the wrong place and the sentence almost unreadable.

1.5 Importance of the Dictionary in the TI/QI Correlations

The TI presupposes that the LOGOS dictionary will have been adequately updated for the translation task. This assumption is important since a statistically-based TI cannot hope to be sensitive to lexical weaknesses, which can adversely affect the QI. There are, for example, sentences from technical documents that show very low correlations between TI and QI. The explanation of this considerable discrepancy seems to be that the terminology in the dictionary is not adequate. E.g. *Pour neutralisation medium* (for *Neutralisationsmittel einfüllen)* got a QI of 4 but a TI of 7. *In all 6 rinsing procedures one rinses with normal line water* (for *In allen 6 Spülgängen wird mit normalem Leitungswasser gespült)* has a QI of 4, but a TI of 6.4.

1.6 Translatability Index Acquires Value with Usage

The long-range feasibility of a TI hinges on the establishment of an index that takes on significance to the users over time, as they work with it and find it useful. A TI is useful if it can provide the user with a measure that correlates reasonably well with the quality of the MT output over a period of time. In particular, the index will establish its usefulness if, as the users manipulate the source document to improve the index, they experience a corresponding improvement in the quality of the MT output.

2 Description of Work

2.1 Capturing and Interpreting the Statistical Data

For each source language, a set of "negative" sentence properties was identified and a program was written to capture and quantify these properties for any given document. We found these sentence properties by comparing raw output with the input and recording those features and properties that appeared to be associated with bad output. A second program was written to interpret and manipulate the statistical data from the first program. This program assigns weights, examines interrelationships, etc. and produces a TI for each sentence as well as for the document as a whole. For English documents, the text TI is the average of the sentence TI's. For German documents, it is the results of a calculation based on the average sentence TI's and a special document TI.

2.2 Scoring Procedure and Corpora

The procedure for scoring the TI for each sentence is similar to scoring for the QI. The program starts off with a score of 7 and then penalizes the sentence for various negative properties that have previously been recorded and quantified by the first statistical program. In all, there are over 30 sentence properties used in the computation of the TI; some of these are source-language specific, while others are common to both sources. For German source, there is an additional score for the document as a whole that penalizes for certain averages which characterize the text as difficult for the LOGOS system.

We worked with and tested 541 English sentences (translated into German and French) and ca. 1300 German sentences (translated into English and French). The sentences came from various documents (running text) from different subject matter areas and text types. We modified the TI programs until the resulting TI's achieved a high correlation with the QI's that had been assigned to the same sentences. I also looked at 205 German sentences to test the TI program on untrained documents.

2.3 Statistical Data

2.3.1 Average Sentence Length

The first step in the statistical analysis was an investigation into a possible correlation between the average length of the sentence(i.e. number of words) in the document and the quality of the MT output. We found some confirmation of a correlation between average sentence length of a document and its QI score. The correlation was not confirmed for individual sentences. English-source sentences with a QI of 5 can have any length -- from 13 to 33 words in our data. (QI of 5 = rough estimate that its translation is acceptable.) It appears that in order to be assigned a QI of a least 5, the length of a sentence may not exceed a certain number of words above the average sentence length in the document from which it comes.

2.3.2 Type of Text, Syntactic Complexity, and Ambiguities

Francis and Kucera (1982) found that English sentences in informative text are longer but do not contain a significantly higher number of predicates per sentence than in imaginative text, i.e. they are not necessarily more syntactically complex. This was reflected in our corpus: Certain types of text allow for long sentences, others do not.

Our set of "negative" sentence properties contains lists of features of syntactic complexity, features we know the LOGOS system does not handle well, and features that characterize texts that are not purely informative (see discussion of discourse characteristics in section 5).

Examples:
Words that were not found in the system's dictionary; short parentheses; coordination; homographs; interrogatives; unmatched parentheses; dependent and relative clauses; complement sentences; noun and verb form ambiguities (for German); nested participle

constructions modifying nouns (for German); "suspicious" or difficult pronouns; certain ambiguous words *(allein, gerade, ja, von, erst,* etc. for German; *-ing, as, with,* etc. for English). Specifically for German texts, the list of properties includes the following: certain conjunctions and sentence adverbs *(also, desto, wie,* etc. and *da*-compounds); sentences beginning with *daß;* words that reflect syntactic ambiguities, such as participles; words that can be both pronoun and determiner; inverted sentences functioning as conditionals; certain pronouns and possessive modifiers *(sie, ihm, er, ihr, sein,* etc.); strings of noun-compounded nouns (e.g. *Die Kontrollampe "Enthärter aufbereiten" leuchtet auf).*

Because there are certain lexical and syntactic indicators for complex texts, and because I found that the TI tended to be consistently too high for German-source translations, I used the average sentence TI as only one value. I added a second TI that pertains to the text as a whole using slightly different statistics and penalties. At the end, I computed a TI based on a text value and the average sentence TI to achieve a higher correlation between QI and TI.

3 Findings

3.1 English Source

Table 1: Summary statistics for English source, comparing average QI and TI (English-German translations)

Document & Target	# sentences	average # words in sentence	QI	TI	Correlation
LEGAL	21	26.7	4.47	4.80	93.1%
ELECTR1	32	17.7	5.87	5.94	98.8%
DRILLING	66	27.4	5.03	4.44	88.3%
BANKING	105	29.0	4.38	4.69	93.4%
INFORMA	69	23.3	4.95	5.12	96.6%
ELECTR2	73	19.8	5.65	5.69	99.3%
AVIONIC	48	20.0	5.39	5.27	97.7%
MEDICAL	48	20.7	5.04	5.42	92.8%
CHEMICAL	79	22.0	5.05	5.00	99.0%

Total number of sentences = 541
Average correlation = 95.4%

3.2 German Source

Table 2: **Summary statistics for German source, comparing average QI and Text-TI**

Document & Target	# sentences	average # words in sentence	QI	TI	Correlation
TECH(E)	34	13.4	6.32	5.93	93.8%
MECH2(F)	57	19.9	4.94	5.32	92.8%
ECON(F)	95	23.9	4.76	5.00	95.2%
ECON(E)	95	23.9	4.56	5.00	91.2%
MECH1 (F)	207	22.8	4.57	5.11	89.4%
MEDIC (F)	50	23.3	5.12	5.34	95.8%
MEDIC (E)	50	23.3	5.09	5.34	95.3%
EEC(F)	160	33.5	4.57	4.62	98.9%
EPP(F)	95	16.6	5.47	5.67	96.4%
MRCK(F)	173	19.5	4.60	5.35	85.9%
BANK(F)	47	15.4	5.08	5.35	94.9%
EDITO(F)	100	25.7	4.81	4.67	97%
EDITO(E)	100	25.7	3.21	4.67	68.7% (*)
LUX1 (E)	18	18	4.88	4.63	94.8%
LUX2(E)	19	17.2	5.26	4.93	93.7%
CAN1 (E)	22	18	4.09	4.61	£5,7%
CAN2(E)	26	15.2	4.84	5.12	94.5%
PEET1 (E)	23	32.9	3.15	1.74	55% (*)
PEET2(E)	39	19.2	4.26	4.66	91.4%

Total number of sentences = 1165
Average correlation = 93.5 % (exceptions EDITO and PEET1)
Average correlation French translations = 94%
Average correlation English translations = 92.9% (exceptions EDITO and PEET1)

On the whole, there is a relatively high correlation between TI scores and QI scores. That is, the predictions that the TI makes is reflected in the actual QI assigned to the documents. Where the correlation is less impressive, it seems valid all the same under the premise that the TI has to be seen as **relative,** not absolute. The document that gets the highest TI score (TECH) also has the highest QI score; the document with the lowest TI score (PEET1) also has the lowest QI score. I am not worried about the low correlation (55%) in this one case, since it is corroborated in the relative perspective: the low TI for PEET1 correctly predicts a bad translation. The table reflects two other factors worth pointing out: (a) the QI scores for the French and the English translations of the document EDITO (a newspaper editorial) differ tremendously. The difference may indicate one or both of the following: the subjectivity of the French evaluator who assigned the QI's, and/or differences in the target generation or dictionary modules of the LOGOS system; (b) the document MRCK shows a low correlation between QI and TI (4.60 Vs 5.35). My explanation of this discrepancy is poor terminology. The document is about Pharmaceuticals and chemistry and the transfers in the dictionary are not appropriate. I would predict that with an expert terminology update in the dictionary, the QI would increase.

4 Operational Use and Benefits of a TI Facility

Let's assume that the user sends the documents CAN, LUX, and PEET through the TI programs. If the TI scores are lower than 5, the program will tell her that the translatability of this document is 4.63 (or 4.61 or 1.74). ('This document is not suitable for MT" or "This document is conditionally suitable for MT"). The TI utility would also suggest why a document is not suitable, indicating what could be done to improve the translation quality. For instance:

> "The sentences on the whole are too long.
> Sentence # X is far too long.
> The document contains many words and compounds that are not in the dictionary. Run your document through the LOGOS New-Word-Search utility and update your dictionary.
> The document contains many difficult words such as *ja, gerade,* etc.
> The document contains many pronouns and *da*-compounds."
> etc.

The user can make syntactic changes in the document to decrease both complexity and ambiguities and update new words and compounds. Then the document can be run through the TI utility again, where it will receive a higher score. If it is now in the range that the user accepts for quality in raw output, it can be translated with the expectation of an acceptable output quality.

Thus, the TI utility can provide the user with a measure that not only correlates with the quality of the MT output, but that also helps the user manipulate the source document in such a way as to improve the MT output quality.

5 Pitfalls of Discourse

MT systems deal largely with what M.A.K.Halliday (1967-68) terms the ideational component of language -- in other words, the representation of experience and abstract logical relationships. MT is less capable of dealing with what Halliday would call the interpersonal component, which encodes the social, expressive, and conative functions of language, or the textual component, which creates for a particular text the fabric that holds it together. In addition to the first one, the TI tries to take the latter two characteristics into account as much as possible.

The document PEET1 is particularly difficult for an MT system because of the abundance of interpersonal and textual features and marked word order. When I saw the low TI score, I updated unfound words and compounds in the dictionary, and I also changed the document in various places: I reduced the sentence length and took out those ambiguous interjections and particles that function as interpersonal markers. Then I ran PEET2 through the translation system and asked somebody to score QI's for both texts: PEET1 got a 3.15, PEET2 a 4.26. The corresponding TI scores reflect the difference in translation quality: PEET1 with a 1.74 and PEET2 with a 4.66. Another interesting aspect is that the more appropriate the input for MT, the higher the correlation between TI and QI. I did similar rewrites and dictionary updates with the documents LUX and CAN and achieved similar results: I got higher QI's, higher TI's, and, in general, a higher correlation between QI and TI.

While LUX and PEET both deal with biographical material as their subject matter, LUX uses fewer interpersonal features; it is much more straight-forward syntactically. I wanted to show that it is not so much the subject matter but rather the type of discourse that is relevant for MT output quality. After dictionary updates and minimal rewriting, LUX2 gets the third-highest QI score, while PEET2 is still in the lower grades. More importantly, PEET would need drastic rewriting to make it suitable for MT and this would distort the original style and intention of the author. Other documents that do not rely so extensively on interpersonal and textual markers can be revised with minimal effort and negligible changes in style to yield better MT output.

Acknowledgments

Major credit for our work is due to Bernard Scott, who initiated the project and implemented the TI for English source.

References

Francis, W Nelson and Henry Kucera, 1982: *Frequency Analysis of English Usage.* Boston: Houghton Mifflin

Halliday, M.A.K. 1967-68: "Notes on Transitivity and Theme in English". Parts 1 and 2, *Journal of Linguistics Vol.* 3,37-81; 199-244; Part 3, *Journal of Linguistics* Vol. 4,179-215

Appendix

Quality Index (QI) Definitions and Criteria (for Understandability)

7 No corrections required

6 One or two minor changes are required. No reference to the source is required because the sentence is understandable despite errors.

5 Several minor changes are required. The solution is relatively obvious, but reference to the source may be desired for confirmation.

4 Minor changes are required. There are multiple solutions, or the solution is not obvious from reading the translation. Reference to the source is required to correct the sentence.

3 Major changes are required. Reference to the source is definitely required to make the changes.

2 Large parts of the sentence must be retranslated.

1 The entire sentence must be retranslated, with virtually nothing salvageable from the raw translation.