

## **Controlling Input and Output of MT for Greater User Acceptance**

Arendse Bernth  
IBM T.J. Watson Research Center  
PO Box 218  
Yorktown Heights, NY 10598  
USA

### **Abstract**

We present two tools, EasyEnglishAnalyzer and the Translation Confidence Index, which help a company maximize the benefit of using MT in the information dissemination process by controlling the input (EasyEnglishAnalyser) and filtering the output (the Translation Confidence Index). These tools significantly increase the percentage of useful MT output, thus providing greater user acceptance. We show some examples of how these tools improve the quality of output from IBM's Machine Translation system in translation from English to German.

### **Introduction**

However useful Machine Translation (MT) is for quickly getting the gist of documents in other languages, there is no denying that the state of the art leaves something to be desired when MT is used in the context of professional information dissemination. In this setting, a company typically has a group of professional writers prepare the documents in one language; then a group of professional translators translate the documents into a number of other languages. Obviously, to the degree this process can be successfully automated, the company can save money. Currently, the translation quality of MT output on unrestricted text is not such that translation can be fully automated; however, MT can certainly be of use if applied as part of a suitable process.

We have developed two tools, EasyEnglishAnalyzer (EEA) (Bernth 1997, 1998a, 1998b) and the Translation Confidence Index (TCI) (Bernth 1999) that help a company maximize the benefit of using MT in the information dissemination process. EEA is used by the authors to pre-edit the source documents to conform to a weakly controlled language. This helps IBM's MT systems (McCord and Bernth 1998; Bernth and McCord 1998; Gdaniec 1998) produce better translations. The TCI is used by the translators to filter out potentially bad translations produced by the MT system so that the task of translation is divided up into the tasks of post-editing reasonable MT output and translating the rest from scratch. In this paper we shall look at EEA and the TCI from a user's point of view; for a description of more technical issues we refer the reader to the references mentioned above.

Controlling the input to MT is a well-recognized way of improving the output of MT. Various controlled languages exert varying degrees of control, but often the language is controlled so much that the writers feel too restricted in their writing. A main concern for us has been to make sure that EEA does not impose any unnecessary restrictions on the writers, but rather addresses only the issues of importance to the MT process; these include various

types of ambiguity and complexity. So the approach of EEA is not to specify a fully controlled language that the writers have to adhere to; rather, EEA points out a limited number of problematic constructions in context and gives suggestions for rewording these. We do not absolutely rule out whole classes of phrases or constructions. Rather we just rule them out (or suggest alternatives) when they create an ambiguous situation in the sentence. Thus, most of normal, full English is allowed, with a few exceptions. We will give some examples of how pre-editing with EEA improves the output of IBM's MT system for English-to-German.

In addition to having the writers check the documents with EEA, it can be desirable to have the MT system assign a measure of its confidence in the translation. For this purpose we have developed the TCI. This measure can be utilized in a number of ways. In the context of a professional translator using a translator's workbench, the TCI provides a way of filtering out translations that may be too bad to efficiently post-edit. It is very subjective what constitutes an unuseful translation, and for that reason the TCI engine can read a user profile that indicates the threshold below which the translator does not want to be bothered with MT output. The TCI can also be valuable for the casual user who does not know the source language at all. In this case, an indication of which translations should be taken with a grain of salt can be of value. There are various ways this can be communicated to the user, e.g. by showing the confidence measure as the actual numerical value, or by changing font color, etc. We will show some examples of how the TCI is used for IBM's English-to-German MT system.

### **EasyEnglishAnalyzer**

EasyEnglishAnalyser (EEA) is part of IBM's Information Development Workbench (IDWB) and is interfaced to the ArborText Adept editor and the FrameMaker+SGML editor. EEA points out problems, and, where possible, supplies rewriting suggestions. The editor interfaces show a box with the rewriting suggestions, and the user can choose one and have it substituted automatically in the text by mouse-clicking (see fig. 1); or, if the user does not like any of the suggestions, she or he can of course edit the file directly to make the desired change. The interface also provides statistics on the number of potential problems of each type. Furthermore, the so-called *Clarity Index* provides a measure of the overall quality of the document. A user profile allows the user to turn on or off the individual problem types in order to create the optimal combination of checks for the document and its specific purpose.

EEA has a large general dictionary; in addition, it is possible to specify special user dictionaries that may be applied in various ways to control vocabulary use. The most common use is to allow most words, but to check for a limited number of restricted or non-allowed words. However, it is also possible to do quite the opposite, i.e. allow only the words specified in the user dictionary. Using EEA in this way makes it a tightly-controlled vocabulary checker for those who need that. In both cases the full general vocabulary is available for the parser during analysis of the text. This makes it possible to give reasonable messages even when the vocabulary is severely restricted and the user violates these restrictions.

Terminology may be collected by using the terms collector ETerms, which gathers candidates for terms along with their frequency of occurrence in the document. The file created by ETerms may be sorted either alphabetically or by frequency. Whereas EEA can be run

successfully on small parts of the document, it makes most sense to run ETerms on as large a scale as possible. A frequently occurring cluster of nouns is much more likely to qualify as a

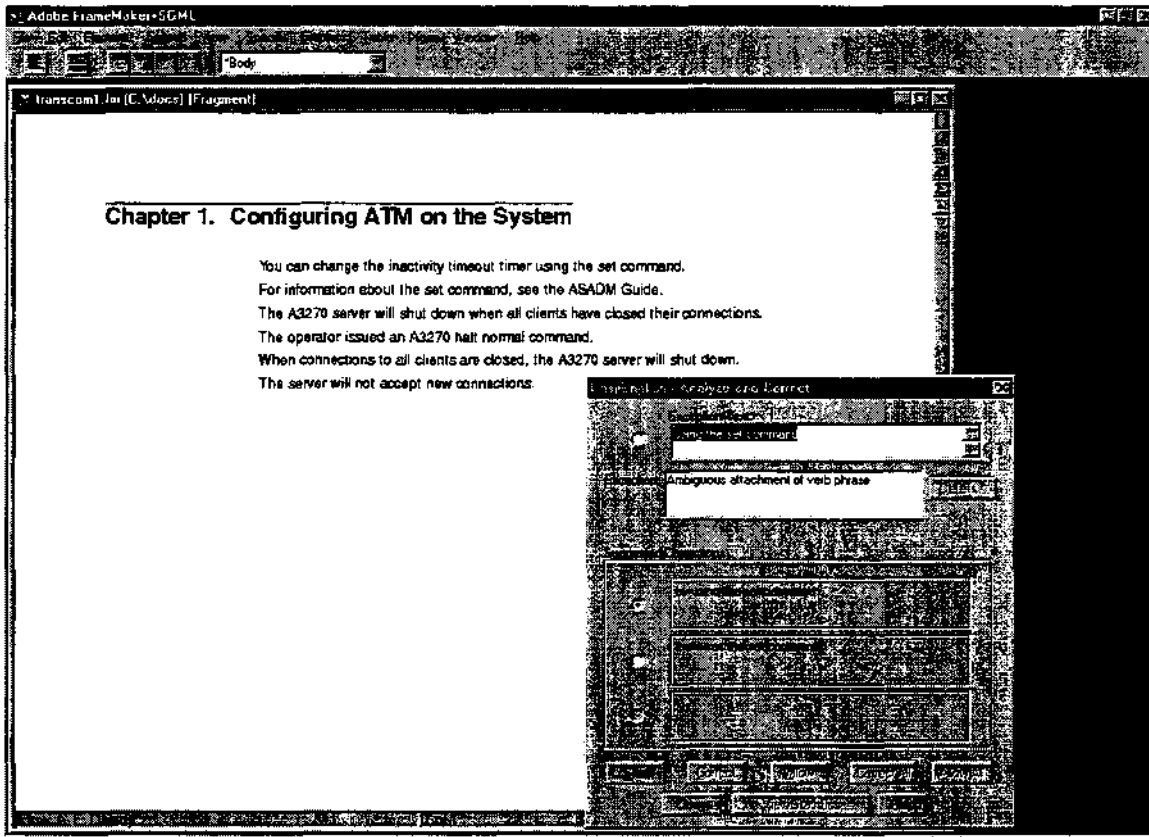


Fig. 1. Pre-editing with EEA

real term than infrequent ones, so it is important to gather the statistics over as large a corpus as possible.

### The Translation Confidence Index

The TCI is a measure of IBM's MT system's confidence in its own translation. This confidence is based on a number of factors, some of which depend on the language pair involved in the translation; the *language distance* (Odlin 1989)<sup>1</sup> makes a noticeable difference. For example, a segment that cannot be given a complete parse may translate better into a target language that has a word order similar to the source language than into a target language with a very different word order. For this reason, the TCI makes use of a language-pair specific profile. We have integrated the TCI with IBM's English-German MT system and calibrated the profile to yield a general precision of around 70-75 percent.

An important issue is how to best calibrate the TCI. There is a definite trade-off between precision (how few false hits we get) and recall (how many problems we can identify). If we want to be sure that we do not assign a too low TCI to any segment that gets an acceptable

---

<sup>1</sup>Odlin's (and others') description of the impact of language distance relates to *human* use of language, but is even more true of *machine* translation.

translation, we will also, as an unwanted side effect, get a too high score for some bad translations. It has been our decision to err on the side of losing some good translations rather than getting too many bad translations since it is our impression that a professional translator gets more upset and insulted by seeing bad MT output than by having to write a few extra translations from scratch.

### **Example From an IBM Manual**

In this section we show an example of applying EEA and the TCI to a text taken from an IBM manual. First we show the actual, non-EEA'ed text and the translation produced by IBM's English-German MT system. Then we show the pre-edited text and the resulting translation. For this second version of the translation we also show which sentences would be filtered out by the TCI and thus not shown to the human translator.

Problems flagged by EEA are underlined and given an identifying number in parenthesis to which we will refer below. The same numbers are used to identify the corresponding parts of the text in the German translations, and in the EEA'ed version of the text.

Let us first look at the original, non-EEA'ed text in fig. 2. This example is from an IBM hardware manual. The document uses the Bookmaster mark-up language<sup>2</sup>, and the mark-up is preserved during translation so that the translated document can easily be formatted in the same way. The text used in our example is written as a numbered list, and we show the formatted version in order to display the document in true form; we hope this additional numbering will not cause any confusion to the reader.

---

<sup>2</sup>In addition to Bookmaster, IBM's MT system and EEA also process SGML and HTML.

1. Only the top (green) light comes on. This indicates that the port or cable is good and there is no open shield problem on the port or cable you are testing.

NOTE: There may still be open shield problems somewhere on the port you are testing. To ensure there are no open shield problems, you must test the end of every cable at the input to each device that is attached to the port you are testing. (1)

2. Only the top (green) and the middle (yellow) lights come on. This indicates that there is an open shield between the port tester and the device or workstation controller that the cable is attached to.

NOTE: The following open shield indications are not considered problems:

- o The twinaxial test may detect an open shield problem if testing a fiber optic to twinaxial adapter.
- o The twinaxial test will indicate an open shield problem if testing a telephone-twisted pair to twinaxial adapter.

3. Only the bottom (yellow) light comes on. This indicates that the wires in the cable are reversed somewhere between the port tester and the workstation controller. After solving this problem, you should test the cable again for open shield problems.

4. None of the lights come on. This indicates that either there is an open circuit or short circuit in the wires, a cable impedance problem caused by low signal strength, or the port tester is not working correctly. (2)

This can also indicate that: (3)

- o The selector switch is in the wrong position.
- o The port tester is attached to a cable that is not connected to the system.
- o The attachment to the port or cable is not secure.

If this is a valid indication, or your port tester is not working correctly, call your service representative.

5. All of the lights come on. This indicates that either the selector switch is in the self test position or the port tester is not working correctly. If the port tester is not working correctly, call your service representative.

6. Any other combination of lights not covered (4) (5) previously. This indication is not valid. Perform the test again. If you get another indication that is not valid, call your service representative.

Fig. 2. Example Text before EEA

There are several problems in this text. (1) and (2) are very long sentences. Long sentences can be difficult to analyze correctly. So they need to be rewritten either by simply getting rid of verbosity or by splitting them up into more sentences. In (3) we have a sentence that leads into a bulleted list. Such sentences should preferably be complete sentences because this ensures that the individual list elements also are complete sentences. Most MT systems, IBM's included, process one sentence at a time. In other words, there is no way that the MT system knows that the individual bullets in (3) are parts of a larger whole and need to be

translated as subordinate clauses. For German, this has an impact on word order (verb at the end). (4) illustrates a non-allowed word, *cover*. This word is very hard to translate automatically, because it is hard to distinguish between the physical sense *to put a cover on something* and the sense *to describe something*. In addition, *covered* (5) represents the problem of a nonfinite construction; this phrase will be easier to translate well if it is converted into a finite construction with an overt subject and a finite verb.

These problems cause less-than-desirable output from IBM's MT system. In fig. 3 we show the result of translating the un-EEA'ed text into German. The text has been annotated with a native German speaker's judgement of the quality. For purposes of this paper we are only interested in two categories: *not useful* (indicated by **(nu)** before the segment) and *useful* (indicated by **(u)** before the segment). The latter category also includes quite good translations, so *useful* and *not useful* refer to the potential use a professional human translator may make of the translations. As before, the problems that EEA pointed out for the source text are underlined and identified by a number immediately following the problem.

For the translation given in fig. 3, which is based on the raw (un-EEA'ed) English source text, the percentage of useful translations is 68%.

1. **(u)** Nur das oberste (grün) Licht geht an. **(nu)** Dies zeigt an, daß, die Sie testen, auf dem Anschluß oder dem Kabel der Anschluß oder das Kabel gut sind und es kein Problem einer offenen Sicherung gibt.

NOTE: **(u)** Es kann sein, daß es immer noch Probleme einer offenen Sicherung irgendwo auf dem Anschluß gibt, den Sie testen. **(u)** Um zu gewährleisten, daß es keine Probleme einer offenen Sicherung gibt, müssen Sie das Ende jedes Kabels am Eingang zu jedem Gerät testen, das an dem Anschluß angeschlossen ist den Sie testen. (1)

2. **(nu)** Nur die Oberseite (Grün) und die mittleren (gelb) Lichter gehen an. **(u)** Dies zeigt an, daß es eine offene Sicherung zwischen dem Anschlußprüfgerät und der Geräte- oder Datenstationssteuereinheit gibt, an denen das Kabel angeschlossen ist.

NOTE: **(u)** Die folgenden Anzeigen einer offenen Sicherung werden nicht für Probleme gehalten:

- o **(u)** Es kann sein, daß der twinaxiale Test ein Problem einer offenen Sicherung entdeckt, wenn er einen Adapter zwischen Lichtleitfaser und Twinaxial testet,
- o **(u)** Der twinaxiale Test wird ein Problem einer offenen Sicherung anzeigen, wenn er einen Adapter zwischen paarig verdrehtem Telefonkabel und Twinaxial testet.

3. **(u)** Nur das untere (gelb) Licht geht an. **(u)** Dies zeigt an, daß die Drähte im Kabel irgendwo zwischen dem Anschlußprüfgerät und der Datenstationssteuereinheit umgekehrt werden. **(u)** Nachdem Sie dieses Problem lösen, sollten Sie das Kabel wieder nach Problemen einer offenen Sicherung überprüfen.

4. **(u)** Keines von den Lichtern geht an. **(nu)** Dies zeigt an, daß entweder es ein ausgeschalteter Schaltkreis oder einer Kurzschluß in den Drähten, ein Kabelimpedanzproblem gibt, das von niedriger Signalstärke verursacht wurde, oder das Anschlußprüfgerät nicht richtig funktioniert. (2)

**(u)** Dies kann auch anzeigen, daß: (3)

- o **(nu)** Der Wählschalter ist in der falschen Stellung.
- o **(nu)** Das Anschlußprüfgerät ist an einem Kabel angeschlossen, das nicht mit dem System verbunden ist.
- o **(nu)** Das Zusatzgerät für den Anschluß oder das Kabel ist nicht sicher.

**(u)** Wenn dies ein gültiger Hinweis ist oder Ihr Anschlußprüfgerät nicht richtig funktioniert, rufen Sie Ihren Kundendienst an.

5. **(u)** Alle Lichter gehen an. **(u)** Dies zeigt an, daß entweder der Wählschalter in der Selbstteststellung ist oder das Anschlußprüfgerät nicht richtig funktioniert. **(u)** Wenn das Anschlußprüfgerät nicht richtig funktioniert, rufen Sie Ihren Kundendienst an.

6. **(nu)** Jede andere Kombination nicht vorher bedeckter (4) (5) Lichter (5). **(u)** Dieser Hinweis ist nicht gültig. **(u)** Führen Sie den Test wieder durch. **(u)** Wenn Sie einen anderen Hinweis bekommen, der nicht gültig ist, rufen Sie Ihren Kundendienst an.

Fig. 3. Translation without Pre-editing

Above we described the various problems that EEA found in the English source text. If we

edit the English text according to EEA's suggestions we get the text shown in fig. 4. As

<p>1. Only the top (green) light comes on. This indicates that the port or cable is good and there is no open shield problem on the port or cable you are testing.</p> <p><u>NOTE: There may still be open shield problems somewhere on the port you are testing. You must ensure that there are no open shield problems. To do this, test the end of every cable at the input to each device that is attached to the port you are testing. (1)</u></p> <p>2. Only the top (green) and the middle (yellow) lights come on. This indicates that there is an open shield between the port tester and the device or workstation controller that the cable is attached to.</p> <p>NOTE: The following open shield indications are not considered problems:</p> <ul style="list-style-type: none"><li>o The twinaxial test may detect an open shield problem if testing a fiber optic to twinaxial adapter.</li><li>o The twinaxial test will indicate an open shield problem if testing a telephone-twisted pair to twinaxial adapter.</li></ul> <p>3. Only the bottom (yellow) light comes on. This indicates that the wires in the cable are reversed somewhere between the port tester and the workstation controller. After solving this problem, you should test the cable again for open shield problems.</p> <p>4. None of the lights come on. <u>This indicates one of the following problems: (2)</u></p> <ul style="list-style-type: none"><li>o An open circuit or short circuit in the wires.</li><li>o A cable impedance problem caused by low signal strength.</li><li>o The port tester is not working correctly.</li></ul> <p><u>This can also indicate one of the following problems: (3)</u></p> <ul style="list-style-type: none"><li>o The selector switch is in the wrong position.</li><li>o The port tester is attached to a cable that is not connected to the system.</li><li>o The attachment to the port or cable is not secure.</li></ul> <p>If this is a valid indication, or your port tester is not working correctly, call your service representative.</p> <p>5. All of the lights come on. This indicates that either the selector switch is in the self test position or the port tester is not working correctly. If the port tester is not working correctly, call your service representative.</p> <p>6. Any other combination of lights <u>that was (4) not described (5) previously.</u> This indication is not valid. Perform the test again. If you get another indication that is not valid, call your service representative.</p>
---

Fig. 4. Example Text after EEA

before, the areas that EEA were concerned with are underlined and identified by the same numbers as earlier.

The resulting translation into German is given in fig. 5. We notice that pre-editing increased



1. **(u)** Nur das oberste (grün) Licht geht an. **(nu)** Dies zeigt an, daß, die Sie testen, auf dem Anschluß oder dem Kabel der Anschluß oder das Kabel gut sind und es kein Problem einer offenen Sicherung gibt.

NOTE: **(u)** Es kann sein, daß es immer noch Probleme einer offenen Sicherung irgendwo auf dem Anschluß gibt, den Sie testen. **(u)** Sie müssen gewährleisten, daß es keine Probleme einer offenen Sicherung gibt. **(u)** Um dies zu tun, testen Sie das Ende jedes Kabels am Eingang zu jedem Gerät, das an dem Anschluß angeschlossen ist den Sie testen. (1)

2. **(nu)** Nur die Oberseite (Grün) und die mittleren (gelb) Lichter gehen an. **(u)** Dies zeigt an, daß es eine offene Sicherung zwischen dem Anschlußprüfgerät und der Geräte- oder Datenstationssteuereinheit gibt, an denen das Kabel angeschlossen ist.

NOTE: **(u)** Die folgenden Anzeigen einer offenen Sicherung werden nicht für Probleme gehalten:

- o **(u)** Es kann sein, daß der twinaxiale Test ein Problem einer offenen Sicherung entdeckt, wenn er einen Adapter zwischen Lichtleitfaser und Twinaxial testet.
  - o **(u)** Der twinaxiale Test wird ein Problem einer offenen Sicherung anzeigen, wenn er einen Adapter zwischen paarig verdrehtem Telefonkabel und Twinaxial testet.
3. **(u)** Nur das untere (gelb) Licht geht an. **(u)** Dies zeigt an, daß die Drähte im Kabel irgendwo zwischen dem Anschlußprüfgerät und der Datenstationssteuereinheit umgekehrt werden. **(u)** Nachdem Sie dieses Problem lösen, sollten Sie das Kabel wieder nach Problemen einer offenen Sicherung überprüfen.
  4. **(u)** Keines von den Lichtern geht an. **(u)** Dies zeigt eins der folgenden Probleme an: (2)
    - o **(u)** Ein ausgeschalteter Schaltkreis oder einer Kurzschluß in den Drähten. (2)
    - o **(u)** Ein Kabelimpedanzproblem, das von niedriger Signalstärke verursacht wurde. (2)
    - o **(u)** Das Anschlußprüfgerät funktioniert nicht richtig. (2)

**(u)** Dies kann auch eins der folgenden Probleme anzeigen: (3)

- o **(u)** Der Wählschalter ist in der falschen Stellung.
- o **(u)** Das Anschlußprüfgerät ist an einem Kabel angeschlossen, das nicht mit dem System verbunden ist.
- o **(u)** Das Zusatzgerät für den Anschluß oder das Kabel ist nicht sicher.

**(u)** Wenn dies ein gültiger Hinweis ist oder Ihr Anschlußprüfgerät nicht richtig funktioniert, rufen Sie Ihren Kundendienst an.

5. **(u)** Alle Lichter gehen an. **(u)** Dies zeigt an, daß entweder der Wählschalter in der Selbstteststellung ist oder das Anschlußprüfgerät nicht richtig funktioniert. **(u)** Wenn das Anschlußprüfgerät nicht richtig funktioniert, rufen Sie Ihren Kundendienst an.
6. **(u)** Jede andere Kombination von Lichtern, die (4) nicht vorher beschrieben (5) wurde (4). **(u)** Dieser Hinweis ist nicht gültig. **(u)** Führen Sie den Test wieder durch. **(u)** Wenn Sie einen anderen Hinweis bekommen, der nicht gültig ist, rufen Sie Ihren Kundendienst an.

Fig. 5. Translation with Pre-editing

the percentage of useful translations to 93%. Additionally, if we apply the TCI<sup>3</sup>, the number of useful translations goes up to 96%. Do note however, that we also lose a couple of useful translations. It is a personal preference on the part of the translator whether this is better than being subjected to some bad translations.

## Conclusion

We have described the use of a tool for pre-editing the input to MT and a tool for filtering the output of MT. Constraining the input, even as little as EasyEnglishAnalyzer does, has been shown to improve the quality of MT significantly, as illustrated by the examples in this paper. Pre-editing the source documents in order to reduce post-editing of the translations also makes a lot of sense because pre-editing is only done once, no matter how many languages we might want to translate the documents into. However, as much as pre-editing helps, there are limits to what it can do for us; for this reason filtering of the MT output is also a good idea. Including both of these two processes in the overall information dissemination process ensures that the MT output is of such a quality that a professional translator can benefit from using MT without being put off by having to deal with bad quality.

## References

- Bernth, A. 1997. "EasyEnglish: A Tool for Improving Document Quality." *Proceedings of the Fifth Conference on Applied Natural Language Processing*. Association for Computational Linguistics, pp. 159—165.
- Bernth, A. 1998a. "EasyEnglish: Preprocessing for MT." *Proceedings of the Second International Workshop On Controlled Language Applications*. Carnegie-Mellon University, Pittsburgh, pp. 30—41.
- Bernth, A. 1998b. "EasyEnglish: Addressing Structural Ambiguity." *Proceedings of AMTA-98*. Association for Machine Translation in the Americas, pp. 164—173.
- Bernth, A. 1999. "A Confidence Index for Machine Translation." *Proceedings of the 8th International Conference on Theoretical and Methodological Issues in Machine Translation*. Chester College, England, pp. 120—127.
- Bernth, A. and M. C. McCord. 1998. "LMT at Tivoli Gardens". *Proceedings of the 11th Nordic Conference on Computational Linguistics*. Copenhagen, pp. 4—12.
- Gdaniec, C. 1998. "Lexical Choice and Syntactic Generation in a Transfer System: Transformations in the New LMT English-German System". *Proceedings of AMTA-98*. Association for Machine Translation in the Americas, pp. 408—420.
- McCord, M.C and A. Bernth. 1998. "The LMT Transformational System." *Proceedings of AMTA-98*. Association for Machine Translation in the Americas, pp. 344—355.
- Odlin, T. 1989. *Language Transfer*. Cambridge Applied Linguistics Series. Cambridge University Press.

---

<sup>3</sup>The translations that would have been filtered out by the TCI are italicized.