ELECTRONIC CORPORA IN TRANSLATION
BOOTCAT-BOOTSTRAPPING CORPORA AND TERMS FROM THE WEB

Mariana Coanca[1]
Elena Museanu[2]

Abstract

In the new world of technology, the translation profession, like other disciplines, cannot be deprived of modern tools such as electronic corpora. Recently, large monolingual, comparable and parallel corpora have played a crucial role in solving various problems of linguistics, including translation. During recent years, a large number of studies within the discipline of translation studies have focused on corpora and their applications in translation classes. Such studies mainly look into the kind of information trainee translators can elicit from corpora and the effect of using corpus data on the quality of translations produced. Corpora, however, have a lot more to offer to both translation teachers and translation students. Corpus-based translation classrooms, by their very nature, can offer considerable advantages far beyond what traditional translation classes have to offer. This article, in fact, aims to elaborate on advantages of using corpora in translation classrooms for teachers and students of translation. Furthermore, we present types of corpora and a new method of compiling specialized corpora- BootCaT.

1. Introduction

In recent years computers have increasingly found their way into different branches of sciences, including humanities. Language studies are no exceptions in this respect. In this new world of technology, the translation profession, like other disciplines, cannot be deprived of modern tools such as electronic corpora. Constructing as well as exploiting different types of corpora are among the computer applications, available to researchers in various language fields. Recently, large monolingual, comparable and parallel corpora have played a crucial role in solving various problems of linguistics such as language learning and teaching (Aston, 2000; Leech, 1997; Nesselhauf, 2004), translation studies (Mosavi Miangah, 2006), information retrieval (Braschler, & Schauble, 2000), statistical machine translation (Brown et al., 1990) and the like.

Lately, a large number of studies within the discipline of translation studies have focused on corpora and their applications in translation classrooms. Such studies mainly look into the kind of information trainee translators can elicit from corpora and the effect of using corpus data on the quality of translation produced. Corpora, among other things, were shown to provide trainee translators with terminological and conceptual information (Zanettin, 1998), collocational information (Stewart, 2000; Kubler, 2003), phraseological information (Machniewski, 2006), information on cognates, false friends (Zanettin, 2001) and semantic prosody (Bowker, 2000) and contrastive knowledge about the two languages involved (Zanettin, 2001; Schmied, 2002). Furthermore, translations

[1] Mariana Coanca, Romanian-American-University, Bucharest

[2] Elena Museanu, Romanian-American-University, Bucharest

produced with the help of corpora were shown to be of a higher quality in terms of subject field understanding, correct term choice and idiomatic expression compared to translations produced using conventional resources available to translators such as dictionaries (Bowker, 1998).

Corpora, however, have a lot more to offer to both translation teachers and translation students. This article, in fact, aims to elaborate on the benefits of using corpora in translation classes for translation teachers on one hand and translation students in terms of their professional prospects on the other hand. The paper is thus divided into three sections; first section is devoted to the applications of corpora in translation classrooms both for teachers and students, the second section focuses on translation student's professional prospects and what corpora have to offer in this regard, while the third section presents the procedures for compiling specialized corpora by using BootCat.

2. Corpora in translation classrooms. Corpora and trainee translator's professional prospect

A translation class has been traditionally the one in which the teacher was the sole speaker transmitting knowledge to students who were eager to find the answers to their questions in their teacher's words. In such classes, students usually translate a text chosen by teacher and bring it to the class to be discussed. Students read their translations one by one and the teacher passes comments on student's translations and finally the best translation is presented by the teacher to the class. The teacher is, in fact, the absolute authority, the depository of the answers to every question and students are to memorize instances and to reproduce which may overshadow the need to learn the technique. The translation teacher, thus, has a very significant role to play; s/he needs to know the answer to every question/problem students may have , s/he needs to be familiar with translating different texts types and genres or else to pick up a text type and focus on it during his/her professional life and so to acquire knowledge through experience and finally the teacher must be prepared to convince those students who need something more than their teacher's intuition to accept what is acceptable or not. The students, on the other hand, are to listen, memorize and reproduce. As Gonzales Davis (2005: 70) states traditional teacher-based translation classes lack "a motivating component" with teachers being the absolute authority in the class. According to her, in such classes the aim is to "produce an ideal translation sanctioned by the teacher" (ibid).

This is how a translation class has been run for years. Now let us have a comparison between a traditional translation classroom as portrayed above and a modern corpus-based translation class. In a typical corpus-based translation class, the teacher who is already competent enough to work with corpora and corpus analysis tools may prepare the corpus before the beginning of the course or may decide to bring students into a corpus compilation experience by asking each one of them to take part in the process of corpus building. Students may discuss with their peers and their teacher about the type of corpus which best suits their course needs and later the texts which qualify for inclusion in the corpus. This indirectly attracts the student's attention towards the characteristics of different texts and text types and may eventually increase their knowledge of text types.

Furthermore, this practice can prepare students to build their own corpora in future to help them with other courses or later with the translation jobs at hand.

Students are then asked to translate the text given by the teacher with the help of the corpus along with other resources they usually use. During this stage, students use the corpus analysis tools to elicit the kind of data they need from the corpus. Depending on the information they are looking for, they may use concordances or word lists. Concordances, for example, can be used by students to acquire information about the co-occurrence patterns or phraseology of the text type under study, while word lists can be used to see whether the equivalents found are common in the target language or not. Students then hand in their translations to their peers to be assessed and revised. During this stage, students again use the corpus but this time to revise and edit the translations produced by their peers. The two students, then, agree on changes and only the final version is delivered to the teacher. It is worth mentioning that depending on teachers and students' preferences, this practice can also be done with students put in the groups of three or four. Activities such as this, in fact, not only increase student's assessment and editing skills, but also encourage group learning and cooperation among them. Students learn how to revise and edit translations and how to be open to the changes made to their translations.

In a typical corpus-based translation class, in contrast to the traditional translation class, the teacher is not the absolute authority and the depository of the answers to all questions. It is the corpus which is used to answer student's questions and solve translation problems. The teacher, in fact, acts as an assistant who helps learners use the data offered by the corpus to do their translation assignments and find the answers to their questions. The role the teacher assumes in this class is thus far from being the sole problem solver. The Teacher acts as a guide helping student learn how to query corpora to find answers to their questions and the students learn to take part in their own learning process and to act independently. So whenever they have problem, they do not go directly to the teacher to get the answer. They go to the corpus instead and delve into it and even if they can not find the answers they were looking for, they have already benefited a lot from their query in the corpus. (See Chance Discoveries, Aston: 1999 & Zanettin: 2001).

Now let us see what the implications of this autonomy for translation teachers are. This situation implies that teachers are no longer supposed to know the answers to all questions and to come up with the best solutions to every problem students may have. Instead of providing learners with ready made answers, teachers teach students how to find answers by themselves and definitely the teacher is always ready to help students during the process. So in such corpus-based translation classes, the pressure on the translation teacher is reduced in the sense that the teacher is no longer the supposedly best translator to provide students with the final words. It implies that teachers do not need to be worried about not being able to provide the answer to student's questions on the spur of the moment.

Moreover, there are always some students who find it difficult to accept something which does not seem acceptable to them. In such cases the teacher can easily refer to the corpus to provide such students with hard evidence. Furthermore, it is often the case that each

translation teacher teaches translation of certain text types during his/her professional life. A translation teacher, for example, may teach translation of journalistic texts for years and may never have a class on translation of technical texts. One of the reasons behind this situation is that the teachers may not have full mastery over translation of all text types and genres included in the syllabus. So they usually put their energy into a few of them. However, with corpora of specialized texts at their disposal translation teachers can venture into teaching translation courses they may not have full mastery over and experience a cooperative learning environment with their students. This is especially useful for young translation teachers with limited experience of translation of various text types and genres.

It is worth mentioning that the roles teachers and students assume in a corpus-based translation class are in line with the social constructivist approach to translator education adopted by Kiraly which places emphasis on student's autonomy and cooperation (2000). Kiraly, in fact, asks for translation classrooms in which the teacher helps students to learn the task at hand through hands-on experience.

Apart from that, as mentioned by Coffey (2002), corpora can be exploited by translation teachers to create teaching and testing materials. Translation assignment given to students, for example, can be drawn from source language corpora or from the source language component of parallel corpora in order to have comparison between student's productions with the work of experienced professionals and further to draw student's attention to strategies adopted by professional translators (Pearson, 2003).

In order to succeed in the present competitive market, translators need to expand both their linguistic and extra linguistic skills. Among other things, they need interpersonal skills to get the job and maintain a good relation with clients, subject experts and fellow translators; they need computer skills to work with translation software, translation memories and corpora; they need editing skills to provide their clients with an acceptable final version and finally they need an encyclopedic knowledge of various subjects to work with various texts.

Now let us see what a corpus-based translation class has to offer to trainees in terms of the above-mentioned skills. As Fawcett (1987) says the purpose of translator education is to equip trainees with skills transferable to any text, on any subject and a corpus-based teaching by its very nature can provide trainees with such skills. In a corpus-based translation class, once students learn about corpora, corpus analysis tools and their applications for translation, they can compile and use corpora for any kind of text they may encounter in future. Making students familiar with DIY corpora, for example, enables them to build disposable corpora for any type of text they may come across during their education or later when they enter the translation market. In fact, working with corpora during education gives trainees both the courage and the experience they need to continue using corpora on their own. Such students would definitely have a better chance in the market as they do not need to limit themselves to few text types. They know how to compile (disposable) corpora and how to extract information from different types of corpora and without doubt they are more confident to work on a variety of texts and text types. Receiving corpus-based training can also be beneficial for those translation

graduates who enter other markets such as technical writing and editing. As stated by Bowker and Pearson (2002) LSP corpora can well be used as a writing guide to write in a particular style or to produce technical texts.

Furthermore, a corpus-based translation class provides trainees with a favorable opportunity to work together and to experience positive cooperation and group work. In such classes, students can use the corpus to assess and revise translations done by their peers and they can back up their criticisms of their peers' translations with convincing evidences from the corpus. Such kind of practices can help students develop their interpersonal skills and prepare them to deal with future clients and fellow translators. Apart from that using corpus to revise and edit translation is itself a good practice for students to improve their editing skill which is highly needed in today's market.

Another benefit of using corpora to teach translation has to do with student's computer skills. Working with corpora demands computer literacy and having basic computer skills. Exposure to computers and corpora at undergraduate level would help trainee translators to acquire basic computer skills during their education and later they can better develop their computer skills depending on the market's needs. Kubler (2003: 41)

Learning to use corpora and corpus-analysis tools can give future translators the technical skills that were usually not associated with translation, but which seem to be more and more necessary especially in technical translation.

Furthermore, early familiarization with corpora helps students to build corpora of their own translations during their education and later they can do the same for the translation projects they get. They can further expand their repositories of their translations and develop them into translation memories of different genres to be used as a translation resource. This in long term can definitely increase their productivity and the quality of the translations they produce.

3. By the way.....What Is a Corpus?

Generally, a corpus can be defined as a collection of naturally occurring examples of language. A corpus includes no new information about language, but it gives new perspectives to linguistic researches and helps in the development of different processes such as language learning and teaching and translation.

Depending on the purpose and the form, different types of corpora may be distinguished.

3.1. Specialized corpus

Specialized corpus is a corpus which includes a particular type of texts. This specialization has no definite boundaries, but some criteria that specify the type of the text in question should be considered. Such corpora may contain either some texts specialized in terms of a particular timeframe (texts from 1822 to 1876) or a particular subject (art, politics, medicine) or some other factors. Some famous LSP (Language for Special Purposes) corpora are the 5-million word Cambridge and Nottingham Corpus of

Discourse in English (CANCODE) and the Michigan Corpus of Academic Spoken English (MICASE).

3.2. General corpus

This is a type of corpus which includes various types of texts, either written or spoken, on a variety of subjects. Sometimes it is called "reference corpus" concerning its function as a reference material for language learning, translation, etc. Some of the best-known general corpora are the 100-million words British National Corpus (BNC) and the 400-million Words Bank of English.

3.3. Comparable corpus

A corpus consisting of texts of the same type and content in different languages (e.g. legal contracts in English and French), or articles about linguistics from English and Persian journals. The ICE corpus (International Corpus of English) is a one-million word comparable corpus of different varieties of English.

3.4. Parallel corpus

Parallel corpora are those consisting of texts with their translations into two or more languages, eg. a medical article translated into Spanish, Finnish, and French. They can be of great help in searching equivalent expressions in each language and investigating the differences between languages by translators and learners.

3.5. Learner corpus

A collection of texts—essays, for example—produced by learners of a language (Hunston, S. 2006). This corpus is prepared to help to find the differences between texts produced by the learners and text produced by native speakers. the International Corpus of Learner English (ICLE) with 20,000 words and Louvain Corpus of Native English Essays (LOCNESS) are the examples of numerous well-known learner corpora.

3.6. Pedagogic corpus

Pedagogic corpus is a corpus consisting of all texts to which a learner has been exposed (Hunston, S. 2006). A pedagogic corpus collected by a teacher or researcher may consist of all course books, readers, etc. used by a learner and the tapes they have listened to. This includes all instances of a word or phrase that learners encounter in different contexts, to improve their knowledge of language.

3.7. Historical and diachronic corpus

This is a corpus which includes texts belonging to various periods of time, to show the development of language over a specified timeframe. The most famous English historical corpus is the Helsinki Corpus with 1.5-million words .

3.8. Monitor corpus

This is a corpus which consists of texts of the same type to trace the changes in the language by adding to it annually, monthly, even daily. So the texts of one year (month or day) can be compared to those of another, similar, period.
Different types of corpora may be annotated differently in accordance with the needs of the researchers. Some types of information, which are encoded in a corpus and are effective in translation tasks are parts of speech (POS), syntactic structure, parsing, word senses, and anaphoric relation

4. Related Work

In recent years, the importance of corpora in the field of translation has become noticeable to trainers and researchers. Therefore, some researchers believe that the analysis of corpora should be integrated into translator education. There have been a number of studies on monolingual corpora (general and specialized) and various kinds of exploitation of such corpora like extraction of collocations.

The website "Gateway to corpus linguistics on the Internet" at http://www.corpus-linguistics.de/ is a proper reference for obtaining information about many of best-known corpora and their features such as their size, content, and accessibility as well as when and by whom they were compiled.

Most of the latest research in translation knowledge acquisition is based on parallel corpora (Brown et al.1993). However, since large aligned bilingual corpora are hard to obtain, some researches have tried to exploit translation knowledge from non-parallel corpora such as comparable corpora or monolingual corpora. One of the best known large-scale monolingual corpora is the British National Corpus (BNC), a 100 million-word collection of samples of written and spoken language from wide range of sources. However, the BNC has, despite its large size, serious limitations as a translation aid if you are translating contemporary specialized text (Wilkinson, M. 2006).

In a pilot experiment, Bowker (1998) found that learners using a specialized corpus of texts in the target language (their L1) showed greater correct term choice than a matched group using bilingual dictionaries alone. In his study, Bowker determined that a specialized monolingual native-language corpus assists translators to improve two of the most important criteria to produce high quality translation: subject-field understanding and specialized native-language competence (Bowker, L. 1998).

Bowker & Pearson (2002) provide a good experiment on exploiting such monolingual corpora in translating texts on mechanical engineering. They attempt to investigate the term "nut" and its various collocations in the 100-million-word BNC corpus. They found 670 occurrences of this term. However they found most of the concordance lines not helpful, since most of contexts show examples of "nut" being used in other meanings, such as food or eccentric person. Although some of the occurrences describe the type of nuts used in engineering, it takes time to identify them; there is excessive "noise" due to

the fact that "nut" is a homonym—it has various meanings—and so separating the wheat from the chaff is a time-consuming process.

Bowker & Pearson go on to report that a search for the term "nut" in a 10,000-word corpus containing catalogues, product descriptions and assembly instructions from companies in the manufacturing industry generated 49 occurrences. Although this was far fewer than the BNC search, the findings were far more relevant, since the noise was considerably reduced, and it was easy to spot the many different types of "nut" used in manufacturing (e.g. collar nut, compression nut, flare nut, knurled nut, winged nut), as well as the verbs that collocate with nut (e.g. thread, screw, tighten, loosen)

Thus, the role of specialized corpora in translating different types of texts becomes more prominent. Such specialized corpora which are restricted to the language of a particular specialized field and focus on Language for Special Purposes are sometimes referred to as LSP corpus (Wilkinson, M. 2006). Nowadays, specialized corpora play a crucial role in translation. However, due to the unavailability of ready-made LSP corpora, translators can construct their own specialized corpora.

As it is mentioned in the definition of corpus, corpora by themselves are nothing more than collection of examples of language. But beside other tools they become invaluable and find their position in translation task.

5. Compiling and Exploiting Specialized Monolingual Corpora using BootCaT or Bootstrapping Corpora and Terms from the Web

Despite certain obvious drawbacks (e.g., lack of control, sampling, documentation etc.), there is no doubt that the World Wide Web is a mine of language data of unprecedented richness and ease of access (Kilgarriff and Grefenstette, 2003). It is also the only viable source of “disposable” corpora (Varantola 2003) built ad hoc for a specific purpose (e.g., a translation task, the compilation of a terminological database, domain-specific machine learning).
These corpora are essential resources for language professionals who routinely work with specialized languages, where new terms are introduced at a fast pace and standard reference corpora must be complemented by easy-to construct, focused, up-to-date text collections. It is possible to construct a web-based corpus through manual queries and downloads, this process is extremely time-consuming. The time investment is particularly unjustified if the final result is meant to be a single-use corpus.

5.1. BootCaT toolkit

It is a suite of perl programs implementing an iterative procedure to bootstrap specialized corpora and terms from the web, requiring only a small list of “seeds” (terms that are expected to be typical of the domain of interest) as input. The basic idea is very simple: Build a corpus by automatically searching Google2 for a small set of seed terms; extract new (single-word) terms from this corpus; use the latter to build a new corpus via a new set of automated Google queries; extract new terms/seeds from this corpus and so forth. The final corpus and unigram term list are then used to build a list of multi-word terms.

These are sequences of words that must satisfy a set of constraints on their structure, frequency and distribution. As a result, BootCaT is extremely modular: One can easily run a subset of the programs, look at intermediate output files, add new tools to the suite, or change one program without having to worry about the others.

5.2. BootCaT procedure

It can be divided into two main phases: We first use an iterative algorithm to bootstrap corpora
and unigram terms from the web. We then proceed to extract multi-word terms on the basis of the final corpus and unigram term list we extracted in the previous phase. Of course, one can stop after collecting the corpus and unigram list; and, *vice versa*, one can use our multi-word term extraction method on corpora that were not downloaded from the web. We enumerate the steps of the BootCaT procedure as presented by Maroni and Bernardini:
Select Initial Seeds
Run Google Queries
Retrieve Corpus
Extract Seeds (Unigram Terms)
Extract Multi-Word Terms

The bootstrapping process starts with a small list of seeds that are expected to be representative of the domain under investigation. For well-defined specialized domains, a small list of seeds (in the 5-to-15 range) is typically sufficient, and the authors obtained interesting results by starting with as few as two seeds.

The seed terms are randomly combined and each combination is used as a Google query string. The top *n* pages returned for each query are retrieved and formatted as text. New unigram seeds are extracted from the corpus of retrieved pages by comparing the frequency of occurrence of each word in this set with its frequency of occurrence in a reference corpus. We compare frequencies using the log odds ratio measure (Everitt, 1992).

Random combinations of the newly extracted seed terms are then used for another round of Google queries and a new corpus is created by retrieving and formatting the top *n* pages found in this round. The iterative term extraction/corpus downloading procedure is repeated as many times as desired (e.g., until the corpus reaches a certain size). In their experiment, Marco Baroni and Silvia Bernardini never found the need to repeat the process more than two or three times. The user must control several important parameters, such as the number of queries issued for each iteration, the number of seeds used in a single query, the number of pages to be retrieved, etc.

5.3. Extraction of multi-word terms

The first step of this phase is to extract a list of single and two-word *connectors* from the corpus, by looking for words and bigrams that frequently occur between two single-word

terms (e.g., *of, of the*). We then extract a list of stop words, i.e., words with a very high document frequency that were not identified as connectors.

At this point, we can look for multi-word terms, which we define for our current purposes as sequences of words that satisfy the following constraints:
- they contain at least one unigram term;
- they do not contain stop words;
- they may contain connectors, but these cannot occur at the edges nor be adjacent to each other;
- they have frequency above a certain threshold (dependent on length);
- they cannot be part of longer multi-word terms with frequency above k _ fq, where k is a constant between 0 and 1 (but typically much closer to the upper end of the range) and fq is the frequency of the current term;
- conversely, they cannot contain shorter multi-word terms with frequency above (1=k) _ fq.

The multi-word terms are searched recursively. Starting with bigrams, we look left and right for an n+1gram term containing the current n gram and satisfying the constraints above, except the one banning edge connectors (otherwise, we would not find longer terms with inner connectors). For each seed bigram, the longest well-formed term containing it and without edge connectors is returned (this, of course, can equally be the bigram itself).

Again, the user must set various parameters, such as the minimum frequency for bigram terms and the value of the constant k (the minimum frequency threshold for longer terms will follow from these two parameters). It would be interesting (and relatively straightforward) to add a filter that keeps only bigrams with a high mutual information (or other association measures) as possible starting points for the recursive multi-word term search procedure. If the relevant resources are available, it would also be possible to filter out multi-word terms that do not match certain part-of-speech patterns.

6. Conclusions

According to Larson, to do effective translation one must discover the meaning of the source language and use receptor language forms which express this meaning in a natural way (Larson, M. 1984). So, in addition to other conventional translation tools a translator should use corpora to become more certain that his/her choice is a proper and natural one. According to above explanations, corpora can be of great help in finding suitable collocates and verifying or rejecting the suggested translations by dictionaries. As Varantola states, the general comment made by her students about the corpus evidence: "This evidence helps translators to be less bound to the source material and feel much more confident when deviating from the way things are expressed in the source material if they feel that the changes are justified." (Varantola, 2003, p. 67). Large monolingual as well as bilingual electronic corpora are just recently becoming available to translators, and this is a good opportunity for them to be provided with more precise, natural, and up-to-

date information about words and collocations' senses than before. Open parallel corpora can play their greatest role in resolving different translation problems.

Speaking about Romania, the universities lack the facilities needed to run a corpus-based translation class. Extra funding is needed to equip classes with computers and projectors and to provide access to corpus analysis tools for all students. This especially imposes pressure on those institutions which have a traditional structure with only one or two computer laboratories intended for special courses on listening or audiovisual translation. Finally, most researches done on the applications of corpora in translation classrooms are limited to few countries and deal with few languages. This situation implies that similar researches are needed in different languages addressing the specific needs of educational settings of different countries.

References

Aston, G. (2000). I corpora come risorse per la traduzione e l'apprendimento. In Silvia Bernardini and Federico Zanettin (eds.) *I corpora nella didattica della traduzione.* Bologna: CLUEB, 21-29.

S. Bernardini and M. Baroni. 2004. Web mining in the translation classroom. *Submitted*.

Bowker, L., 1998, Using specialized monolingual native-language corpora as a translation resource: a pilot study, *Meta*, 43/4, pp. 631-651.

Bowker, L. and Pearson, J. (2002). *Working with Specialized Language—A practical guide to using corpora.* London: Routledge, Pp. xiv + 242

Brown P.F., Pietra, S.A.D., Pietra, V. J. D., and Mercer R. L. 1993. The mathematics of machine translation: parameter estimation. *Computational Linguistics*, 19(2): 263-313.

Braschler, M. and Schauble, P. 2000. Using corpus-based approaches in a system for multilingual information retrieval. *Information Retrieval*, 3, PP. 273-284.

Brown, P., Cocke, S., Della Pietra, V., Della Pietra, S., Jelinek, F., Lafferty, J., Mercer, R. & Roosin, P. 1990. A Statistical Approach to Machine Translation. *Computational Linguistics* 16:2, 79-85.

T. Calishain and R. Dornfest. 2003. *Google Hacks*. O'Reilly.

C. Enguehard and L. Pantera. 1995. Automatic natural acquisition of bilingual terminology. *Journal of Quantitative Linguistics*, 2:27–32.

B. Everitt. 1992. *The Analysis of Contingency Tables.* Chapman and Hall, 2nd edition.

S. Evert. 2004. *The Statistics of Word Cooccurrences: Bigrams and Collocations.* Ph.D. thesis (in progress), University of Stuttgart.

Fawcett, P. (1987). Putting translation theory to use. In H. Keith & I. Mason (Eds.), Translation in the Modern Language Degree. London: CILT. pp. 31-18.

W. Fleisher, D. Staley, P. Krawetz, N. Pillay, J. Arnett, and J. Maher. 2002. A comparative study of trauma related phenomena in subjects with pseudo seizures and subjects with epilepsy. *American Journal of Psychiatry*, 159:660–663.

R. Ghani, R. Jones, and D. Mladenic. 2001. Mining the web to create minority language corpora. *CIKM 2001*, 279–286.

Gonzalez Davis, M. (2005). Minding the process, improving the product: Alternatives to traditional translator training. In Tennent, M. (Ed.) Training for the New Millennium.Amsterdam and Philadelphia: John Benjamins, pp. 67-83.

Larson, Mildred L. (1998). *Meaning-based translation: A guide to cross- language equivalence*. Lanham, MD: University Press of America and Summer Institute of Linguistics.
Leech, G. (1997). Teaching and language corpora: A convergence. In: A. Wichmann, S. Fligelstone, T. McEnery & G. Knowles (Eds.), *Teaching and language corpora* (1-23). New York: Addison Wesley Longman
A. Kilgarriff and G. Grefenstette. 2003. Introduction to the special issue on the web as corpus. *Computational Linguistics*, 29:333–347.
H. Kucera and N. Francis. 1967. *Computational Analysis of Present-Day American English*. Brown University Press.
Mosavi Miangah, T. (2006). Applications of corpora in translation. *Translation Studies*, 12, pp: 43-56.
Nesselhauf, N. (2004). Learner corpora and their potential for language teaching. In: J. McH. Sinclair (Ed.), *How to use corpora in language teaching* (125-152). Amsterdam: Benjamins.
P. Pantel and D. Lin. 2001. A statistical corpus-based term extractor. *Proceedings of AI 2001*.
P. Rayson and R. Garside. 2000. Comparing corpora using frequency profiling. *Proceedings of Workshop on Comparing Corpora of ACL 2000*, 1-6.
Varantola, K. 2003. Translators and Disposable Corpora. In Federico Zanettin, Silvia Bernardini and Dominic Stewart (eds.) *Corpora in Translator Education* Manchester: St Jerome, pp 55-70.
Wilkinson, M, (2006). Compiling Corpora for Use as Translation Resources, *Translation Journal*, Vol. 10, No. 1.