Using collocations from comparable corpora to find translation equivalents

Serge Sharoff, Bogdan Babych, Anthony Hartley

Centre for Translation Studies
School of Modern Languages and Cultures
University of Leeds, Leeds, LS2 9JT, UK
{s.sharoff,b.babych,a.hartley}@leeds.ac.uk

Abstract

In this paper we present a tool for finding appropriate translation equivalents for words from the general lexicon using comparable corpora. For a phrase in the source language the tool suggests a range of possible expressions used in similar contexts in target language corpora. In the paper we discuss the method and present results of human evaluation of the performance of the tool.

1. Introduction

One of the most frequent problems that occur in translation practice concerns the choice of the best target word for rendering source expression X in context Y. The translator knows the meaning of each word in a sentence and the standard set of its translations, but cannot find a target expression that is suitable for the current context. The obvious way to find a solution for the word-choice problem is by consulting dictionaries. However, dictionary lookup may fail in two senses: a source expression be missing in available dictionaries, or, worse, a dictionary can mislead the translator by listing a term or source expression with its translation, but the translation is NOT in common use in the target language in the suggested way. The situation is worse for multiword expressions (MWEs). For instance, the Oxford Russian Dictionary (ORD) lacks a translation for the Russian expression четкая программа (‘precise programme’), while the Multitran Russian-English dictionary suggests that it can be translated as *clear programme*. However, it is much less frequent in English, just 2 instances in the BNC, while the Russian expression четкая программа occurs 70 times in a comparable Russian corpus.

On the other hand, there are natural limits on the number of translation equivalents to be listed in a bilingual dictionary, imposed by its size and usability. A printed dictionary cannot afford giving separate translations for derived forms or listing dozens of translation equivalents for a relatively unambiguous word, such as *programme* (for instance, English monolingual dictionaries list no more than 2 or 3 senses for it). As for usability, it is impossible to use a (printed or electronic) dictionary in which the relevant translation is dee pin the long list of potential translation equivalents: a translator or a student is unlikely to find a translation they want. Translations for polysemous words are too numerous to be listed for all possible contexts. For example, ORD already lists 74 translations for *clear* and yet the list does not include many frequent combinations with *clear*, such as *position, distinction, majority*.

With respect to technical terms, dictionaries often lack adequate terminology, especially for rapidly growing domains, such as software or environmental protection. Parallel corpora offer the possibility of searching for examples of translations in context, but they are not as representative as large monolingual corpora and they are not always available in the specific domains needed by a translator. For instance, the Europarl corpus is very large for parallel corpora (its English section contains 18 million words), but it is restricted to the language of parliamentary debates only, e.g. it has no instances of *vent one’s anger*, an expression which is quite frequent in the BNC (26 instances).

There has been surprisingly little research on computational methods for finding translation equivalents of words from the general lexicon. Practically all previous studies concerned detection of terminological equivalence, e.g. (Dagan and Church, 1997; Bennison and Bowker, 2000; Peters et al., 2000). However, words from the general lexicon exhibit polysemy, which is reflected differently in the target language, thus causing the dependency of their translation on corresponding context. Also such variation is not captured by dictionaries. Because of their importance, words from the general lexicon are studied by translation researchers, and comparable corpora are increasingly used in translation practice and training (Varantola, 2003). However, such studies are mostly constrained to lexicographic exercises or analysis of properties of translated texts in comparison to the general language (Hansen and Teich, 2002). Such studies do not provide a computational model for *finding* appropriate translation equivalents for expressions that are not listed or are inadequate in dictionaries.

The paper reports on an ongoing investigation into the detection of translation equivalents in large monolingual corpora for (a) polysemous words that are difficult to translate using decontextualised information in dictionary entries, and (b) technical terminology that is not reflected in dictionaries, but is available in corpora. We use a collection of corpora whose total size is about 300 million words per language, consisting of reference corpora (such as the BNC), newspaper corpora and corpora automatically derived from the Internet (Sharoff, 2006a).

2. Methodology

2.1. Research problem

Our research hypothesis is that it is possibile to use comparable corpora to find linguistic constructions that are used for similar purposes in source language (SL) and target language (TL). Even if the equivalence between constructions in comparable corpora can not be complete, there is sufficient similarity between linguistic resources available in the two languages, such as references to objects and processes, subjective evaluations of a state of affairs, expression of emotions.[1]

Our first attempt at implementing this research programme is devoted to finding the most appropriate translations of collocations and multiword expressions (MWEs). Since many collocations have a more or less fixed meaning, according to the "one-sense-per-collocation" hypothesis (Resnik and Yarowsky, 2000), we can be sure about reliability in their translation. As mentioned above, many frequent collocations are missed in dictionaries, so translators have to rely on their expertise in finding suitable translation. For instance, *daunting experience* is not listed in major dictionaries. In the following examples

(1) *Hospital admission can prove a particularly daunting experience.*

(2) *Even though you knew that what you said didn't matter, it was a daunting experience.*

the expression evaluates an unpleasant experience. We can find a suitable translation by studying similar evaluations in the target language. On the other hand, some collocations cannot be translated independently from the context they are used in. For instance, *no mean feat* is translated in some dictionaries (e.g. ABBYY Lingvo) as стоящее дело (lit. 'worthy deed'), but the suggested translation does not fit into many contexts, such as:

(3) *I did all the cleaning, cooking and kept his books in order, which was no mean feat.*

In this context *no mean feat* does not refer to the worthiness of doing the listed actions, so a translator has to explore various possibilities for rendering the significance of the achievements.

Our goal in this project was to implement a tool that helps translators to find solutions to difficult translation problems. The tool presents the results as lists of suggested translation equivalents (usually 50 to 100 suggestions) ordered alphabetically or by their frequency in target language corpora. Translators can skim through these lists and identify a variant which is mostly appropriate in a given context. From a user perspective the tool works more like a dynamic dictionary or thesaurus, not like a Machine Translation (MT) system. However, unlike dictionaries it can find translation equivalents for words and word combinations that are not explicitly coded as dictionary entries. Very often it successfully suggests translations for idiosyncratic word combinations, e.g., *recreational fear* or соблюдать экологические приличия (lit. 'to observe ecological decency'), which have been created by their authors and are either rare or not present at all in the source language corpus.

The detection methodology, which is implemented in a semi-automatic tool, comprises two stages: generalisation of the context of a problematic expression in the source language, and restriction of the search field in the target language.

2.2. Context generalisation

The problem with using comparable corpora to find translation equivalents is that there is no obvious bridge between two languages. Unlike aligned parallel corpora, comparable corpora provide a model for each individual language, while dictionaries, which can serve as a bridge, are not useful for the task in question, because the problem we want to address involves precisely translation equivalents that are not listed there.

The procedure we use for context generalisation is based on the generation of similarity classes, which consist of words sharing collocations with words in the target expression. This sketches the domain of a lexical item and captures the most important aspects of its use. Optimal feature selection for producing similarity classes has received some attention in recent research. Following (Rapp, 2004) we use the Singular Value Decomposition method to measure the similarity between contexts. For instance, *strong* has the following similarity class: *powerful, weak, strength, potent, heavy, good, overwhelming, intense, robust, tough, weaken, compelling, fierce*. Even if there is no requirement that words in the similarity class are of the same part of speech, it happens quite frequently that most words have the same part of speech because of the similarity of contexts.

2.3. Bridging comparable corpora

At the second stage we generate a translation class by translating words from the similarity class into the target language and producing similarity classes for all translations. The bilingual dictionary resources we use are derived from the source file for the Oxford Russian Dictionary provided by the OUP. Given that a similarity class contains 10-30 words and the dictionary lists 2-3 equivalents for each of them, the procedure typically outputs a list of about 600 words.

In the next step we produce an equivalence class, consisting of translations of words in the similarity class. For instance, the equivalence class of the Russian word опыт (experience) includes:

(4) ability, acquire, aptitude, capability, capacity, competence, courage, evidence, experience, experiment, expertise, feasibility, flair, hypothesis, ingenuity, intelligence, investigation, knowledge, laboratory, learning, method, opportunity, perception, qualification, rat, research, skill, stamina, statistical, strength, study, talent, technique, test, training, vision.

The result reflects the ambiguity of опыт, which can mean 'experience', as well 'experiment' (hence the presence of *hypothesis, laboratory* and *rat* in the equivalence class); however it does preserve the semantic core of опыт, which is about skills and abilities.

2.4. Filtering multiword collocates

Given that the procedure for finding collocates in the target language produces many irrelevant expressions, we experimented with two techniques for finding the best possible

[1] We do not claim that emotions and evaluation are identical between the two languages. There are good arguments for language-specific differences in the expression of evaluations or emotions (Wierzbicka, 1999); however, a translator has to convey these expressions into the target language.

MWEs: explicit collocation search that restricts collocations to within the translation class; and implicit search that restricts target expressions (which can be single words) by identifying translation similarity classes between the most probable collocations for those expressions. This technique is based on the observation that even if an equivalence class contains some words that are not relevant to the source example (e.g. *hypothesis* or *rat*), those unrelated words create little noise, as they rarely collocate with words in the second equivalence class (e.g. *insurmountable* or *onerous*), which belong to the equivalence class of *daunting*.

2.4.1. Explicit collocation search for multiword equivalents

The procedure of finding translation equivalents for MWEs can be illustrated by the example of the Russian expression спортивный интерес (lit: 'sports interest'). This expression can be used in sports context as well as metaphorically, with a general meaning 'an interest which is not related to profit: something done for fun or pleasure', e.g., Многие хакеры занимаются взломом из спортивного интереса (lit: 'Many hackers break sites out of sports interest'). This expression is moderately frequent in Russian coprora (0.48 items per million words), but it is not found in any mainstream Russian-English dictionaries.

A set of translation equivalents for a multiword expression is found using the following procedure:

Step 1. Each word in query is translated into a TL word literally, using a dictionary. For спортивный интерес the literal dictionary translation is generated: *[sports] [interest]*

Step 2. *TL similarity classes* for these dictionary translations are generated. In our example the classes are:

(5) sports, athletics, badminton, basketball, cricket, football, golf, gymnastics, indoor, leisure, outdoor, racing, rugby, ski, snooker, soccer, sport, sporting, sports, squash, tennis

(6) interest, avid, benefit, concern, curiosity, desire, enthusiasm, excitement, fascinate, importance, influence, keen, pay, share, value

Step 3. *SL similarity classes* for words in the original query are generated and translated into TL using a dictionary. In our example the translated SL similarity classes are:

(7) sports, basketball, biathlon, rolling, skating, ski, olympic, swimming, sailing, competition, sport, sportsman, tennis, figured, football, hockey

(8) interest, kindness, attention, thoughtful, attentive, wish, request, desire, amazement, curiosity, distrustful, hostility, need, attractiveness, fixed, passion, liking, sympathy, respect, satisfaction, pleasure, amusement

Step 4. *TL translation classes* for each word in the query are generated by combining results of stage 2 and stage 3 (a TL similarity class and a translated SL similarity class). In our example these classes are intersections of corresponding sets for *sports* (contains 31 words) and *interest* (contains 35 words).

Step 5. All theoretically possible candidate translations for the query are generated as a cartesian product of the TL translation classes (i.e., as the set of ordered sequences where each word from the first class is combined with each word from the second class, etc.). For спортивный интерес there are $31 \times 35 = 1085$ possible combinations, including

(9) *sports benefit, sports curiosity, sports excitement,... leisure passion, competition concern...,*

most of which never occur in English corpora.

Step 6. Each potential candidate translation generated on stage 5 is checked in the database of TL MWEs. This database is pre-compiled off-line from corpora and includes all N-grams with frequency > 1 which pass a filter of lexical and part-of-speech configuration constraints. On average only 2.0% of the potential candidate translations are found in this database and presented to users of the system. In our example 4.2% (46 out of 1085 expressions) were found in the MWE database. The most frequent expressions are:

(10) *competition concern* (frequency = 60), *sport need (59), football need (43), leisure interest (38), sporting interest (37), sporting passion (23), ...*

Step 7. Human users inspect the list and either select the candidate which fits best into their specific context, or invent a new translation equivalent, using ideas or translation strategies from the TL expressions in the list.

For example they may note that in certain context the expression *leisure interest* can express the desired idea, especially where contexts are not directly related to sports competitions. Then translators can re-organise the TL sentence around such solution, or think about a similar solution which uses another non-literal translation strategy inspired by the examples presented.

2.4.2. MWE database

The MWE database is the central component in filtering out potential translation candidates for multiword queries. An alternative solution would be to query corpora directly for presence and frequencies of MWEs; however, its implementation would be very slow under any corpus search engine, including the CWB.

The database contains the list of N-grams for corpora in each language filtered by a set of constraints on lexical and part-of-speech features. We used a *permissive* principle of filtering – everything is allowed except that which is explicitly forbidden – which is more flexible and economic than the standard *prudent* method of filtering (Manning and Schütze, 1999; Justeson and Katz, 1995), under which everything is forbidden except that which is explicitly allowed. For example, to exclude N-grams with an undesirable (incomplete) feature combination from the MWE database, such as
weapon_NN of_IN mass_JJ,
student_NN from_IN the_DT poor_JJ,

	British news	Russian news
no of words	217,394,039	77,625,002
no of types	877,566	433,391
REs in filter	25	18
		N-gram types pass RE filter
2-grams	6,361,596	5,457,848
3-grams	14,306,653	11,092,908
4-grams	19,668,956	11,514,626
		N-gram types Pass frq >1
2-grams	2,176,849 (34.2%)	1,786,171 (32.7%)
3-grams	2,869,617 (20.1%)	1,756,200 (15.8%)
4-grams	2,100,598 (10.7%)	924,626 (8.0%)

Table 1: MWEs in News Corpora

run_VV a_DT completely_RB ethical_JJ

the following RE constraints are used: _IN _JJ$ _DT _JJ$. The numbers of N-grams which pass the filter and then the frequency threshold, as well as the percentage of items which pass the frequency threshold, are presented in Table 1.

It is an interesting fact that the filter and the frequency threshold balance each other with respect to MWEs of differnent length, so in general the number of MWEs of length 2, 3 and 4 is not very much different. There are more longer expressions which pass through the RE filter, but fewer of them pass through the frequency threshold, so roughly equal numbers of MWEs of different length are included in the database.

2.4.3. Implicit collocation search for single-word equivalents

For single-word queries only one TL translation class can be generated. The query does not contain other words which can be used as explicit collocation filters.

Note that in our method multiword queries take advantage of internal consistency of multiword queries, in a sense that they represent complete constituents with internally coherent syntactic and semantic structure: users usually don't ask about *weapons of mass*, but rather about *mass destruction* if they experience difficulties with the phrase *weapons of mass destruction*. In fact, multiword queries in our method benefit from the results of such intuitive and highly accurate human parsing of sentences and phrases which contain difficult fragments.

Also, multiword queries have some relative contextual independence: usually they can be used as a unit in a wider variety of contexts. So they do not predict which word or phrase will necessarily preceede or follow them; there is much greater variation outside the muliword query. Therefore, multiword queries provide very reasonable boundaries for accurately expanding and limiting collocation filters to necessary contexts.

Single-word queries do not have these advantages. Although the returned TL translation classes for single words are much shorter and can be inspected more easily compared to multiword queries, it is also desirable to find a method to filter out spurious elements of these classes which would not pass collocation filters in TL corpora. For example, the translation class for the Russian word востребованный ('requested') contains 18 words, which can be easily inspected by human translators. But can some non-relevant items be filtered out from this list automatically?

We suggest that the list of 5 top single-word collocates of the single-word query can be collectively used as a reasonable approximation for such a TL collocation filter. Even though there is no guarantee that the returned collocations concatenated with the query will produce internally consistent and contextually independent MWEs, these properties may be correctly *guessed* by at least by one or a few collocates, so there can be a reasonable *implicit* collocation filter for the single-word query.

We suggest the following procedure for implicit collocation filtering of TL translation classes:

Step 1. Two sets of 5 collocates ranked best on their Log Likelihood score are returned for the immediate right and immediate left context.

Step 2. For each of these, collocated TL translation classes are generated according to the procedure for explicit collocation filtering described in the previous section.

Step 3. TL translation classes are combined into one TL translation class – separately for the right and for left set of collocates.

Step 4. The set of rightmost words in the TL translation class of the left collocate and the set of leftmost words of the TL translation class of the right collocates are intersected. The resulting list contains only words which are present in both lists. In our example, the resulting list for the word востребованный ('requested') is:

(11) advantageous, attractive, claim, competitive,
dynamic, popular, productive, profitable, receptive,
susceptible, technically, topical, unprofitable,
vulnerable, winning

Items, which have been fitered out by the implicit collocation filter are:

(12) complaisant, disadvantageous, talkative

We see that in our example the filter correctly elliminates non-relevant items (in this case the precision is high), but the recall still needs to be improved: some non-relevant items still pass the implicit collocation filter and are present in the returned list, e.g., *claim, receptive, susceptible, technically, unprofitable, vulnerable*. This can be due to the fact that there is less internal consistency and contextual independence for implicit collocates as compared to the explicit collocation search.

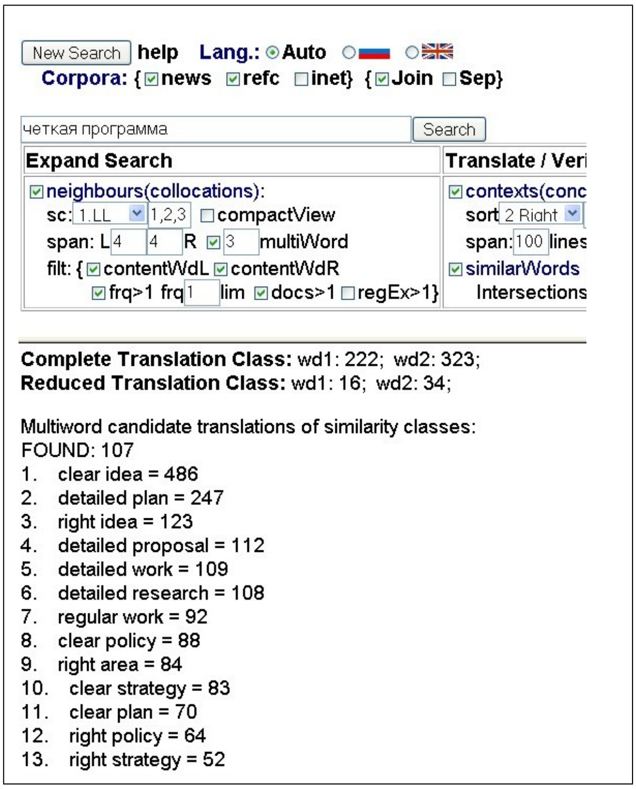
Surprisingly, single-word equivalents pose an even harder problem for our method of finding translation equivalents, since they do not have other components which can act as their explicit collocation filters.

3. Usability of the tool

3.1. Interface

The interface to the tool is powered by IMS Corpus Workbench (Christ, 1994) and itself presents a customisation of a more generic interface to CWB (Sharoff, 2006b). In addition to standard options for making lists of concordance lines and collocations, it provides options for making lists of similarity classes, finding MWEs in the two languages, and choosing the strategy for detection of translation equivalents. For instance, the translator can enter a search term, such as четкая программа (‘precise programme’)[2] and check the resulted list of English expressions, as shown in Figure 1. One of the suggested options that appeal to professional translators is *clear strategy*, which can be used in the following smooth translation:

(13) *This team should be put together by responsible politicians, who have a clear strategy for resolving the current crisis*



New Search help Lang.: ● Auto ○ ○
Corpora: {☑news ☑refc □inet} {☑Join □Sep}

четкая программа Search

Expand Search | Translate / Veri

☑neighbours(collocations):
sc: 1.LL 1,2,3 □compactView
span: L 4 4 R ☑ 3 multiWord
filt: {☑contentWdL ☑contentWdR
☑frq>1 frq 1 lim ☑docs>1 □regEx>1}

☑contexts(conc
sort 2 Right
span: 100 lines
☑similarWords
Intersections

Complete Translation Class: wd1: 222; wd2: 323;
Reduced Translation Class: wd1: 16; wd2: 34;

Multiword candidate translations of similarity classes:
FOUND: 107
1. clear idea = 486
2. detailed plan = 247
3. right idea = 123
4. detailed proposal = 112
5. detailed work = 109
6. detailed research = 108
7. regular work = 92
8. clear policy = 88
9. right area = 84
10. clear strategy = 83
11. clear plan = 70
12. right policy = 64
13. right strategy = 52

Figure 1: Filtered list collocations

3.2. Legitimate translation variation

For each difficult translation problem the system returns multiple translation variants, some of which are potentially useful for translators. We carried out a number of case studies in order to find out whether translators tend to prefer some of these variants and disprefer the others, that is, whether there exists some optimal translation solution for each of such translation problems. We asked several professional translators to score the usefulness of system output for several problems on 5-point scale.

The results were surprising in so far as, for the majority of problems, translators preferred very different translation solutions and did not agree in their scores on the same solutions. In general, the average standard deviation of the responses of different judges is 1.06 , which means that if we assume Gaussian distribution, only about 68% of scores are the same or differ just by 1 point, but about one third of responces differ by 2 points or more on the 5-point scale.

For instance, for the English phrase *recreational fear* in the sentence:

(14) *Patrick West recently claimed that Britain’s mourning for Princess Diana was ‘recreational grief’. Maybe we also suffer from* ***recreational fear****.*

the Russian solution generated by our tool спортивный интерес *’leasure interest’* received the following set of scores: *4, 1, 1, 1, 1, 1, 3*. Interestingly, the translator who gave the score 4 also provided the following comment:

> спортивный интерес (‘sports interest’) is very good for translating ‘recreational grief’, though I would suggest to use a set Russian phrase ради спортивного интереса (‘for the sake of sports interest’).

Other evaluators went for other alternatives and did not see this solution at all, which explains the low scores for it. Note that for this example the score 1 indicates a lack of attention to a potentially good solution in the presence of alternative equally good solutions, but not the quality of the evaluated example – a kind of a *masking effect*.

In general, such divergence cannot be expected under the assumption that the scores reflect some genuine, objective *quality* of translation solutions. This fact can be explained rather by the phenomenon of a legitimate translation variation, e.g. (Papineni et al., 2001), which is even higher for salient lexical items that are often most difficult to translate (Babych and Hartley, 2004). Translators tend to agree on *easy* bits in translation, but have much greater disagreement about *difficult* bits, which also tend to be more central to the general content of the text. However, sentence translation has to be internally consistent, so the choice about such difficult problems has to be made first, and more trivial fragments are built around these solutions.

Our tool gives translators an idea about possible non-literal translation solutions, but in many cases this involves extensive revision of the general structure of the sentence. In this sense the tool is more than just a dictionary: it can point the translators to potentially very good and contextually appropriate suggestions, which usually come at later stages of revision of the draft translation. Translators can organise their texts around such solutions more efficiently compared to the usual way of revising initial non-literal translation (Shveitser, 1988).

4. Conclusions

Future work will involve: extending the suggested approach to wider classes of translation equivalents; implementing alternative automatic search scenarios for these

[2] It appeared in the context of Собрать эту команду должны ответственные люди, имеющие четкую программу выхода из кризиса.

types; developing semi-automatic translation lookup tools which go beyond the limits of standard bilingual dictionaries; testing the system with trainee and professional translators; and applying these methods to MT development. The current implementation works for translations between English and Russian, but can be extended to other languages, for which the necessary bilingual dictionaries and large monolingual corpora are available.

Another approach which we investigate within the UK-funded project ASSIST together with the University of Lancaster uses a model of the semantic context of situation, e.g. 'unpleasant situation' in the case of *daunting experience*. This allows less restrictive identification of possible translation equivalents, as well as a reduction in suggestions irrelevant for the context of the current example. This can be achieved by using 'semantic signatures' obtained from USAS, a broad-coverage semantic parser, (Rayson et al., 2004). The semantic tagset used by USAS is a language-independent multi-tier structure with 21 major discourse fields, subdivided into 232 sub-categories, such as `E5-` = Fear (for *daunting*); `X9.1+` = Ability (for *experience*), which can be used to detect the semantic context. Identification of semantically similar situations can be achieved by the use of segment-matching algorithms as employed in Example-Based MT and translation memories (Planas and Furuse, 2000; Carl and Way, 2003).

Acknowledgements

The project is supported by the EPSRC grant EP/C005902.

5. References

Bogdan Babych and Anthony Hartley. 2004. Extending the BLEU MT evaluation method with frequency weightings. In *Proceedings of the 42^{d} Annual Meeting of the Association for Computational Linguistics*, Barcelona.

Peter Bennison and Lynne Bowker. 2000. Designing a tool for exploiting bilingual comparable corpora. In *Proceedings of LREC 2000*, Athens.

Michael Carl and Andy Way, editors. 2003. *Recent advances in example-based machine translation*. Kluwer, Dordrecht.

Oliver Christ. 1994. A modular and flexible architecture for an integrated corpus query system. In *COMPLEX'94*, Budapest.

Ido Dagan and Kenneth Church. 1997. Termight: Coordinating humans and machines in bilingual terminology acquisition. *Machine Translation*, 12(1/2):89–107.

Silvia Hansen and Elke Teich. 2002. The creation and exploitation of a translation reference corpus. In *Proceedings of the Third Language Resources and Evaluation Conference, LREC 2002*, Spain.

John S. Justeson and Slava M. Katz. 1995. Techninal terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, 1(1):9–27.

Christopher Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA.

K. Papineni, S. Roukos, T. Ward, and W. Zhu. 2001. Bleu: a method for automatic evaluation of machine translation. Technical Report RC22176 (W0109-022), IBM Research Division, Thomas J. Watson Research Center.

C. Peters, E. Picchi, and L. Biagini. 2000. Parallel and comparable bilingual corpora in language teaching and learning. In S. P. Botley, T. McEnery, and A Wilson, editors, *Multilingual Corpora in Teaching and Research*, pages 73–85. Rodopi.

Emmanuel Planas and Osamu Furuse. 2000. Multi-level similar segment matching algorithm for translation memories and example-based machine translation. In *COLING, 18th International Conference on Computational Linguistics*, pages 621–627.

Reinhard Rapp. 2004. A freely available automatically generated thesaurus of related words. In *Proceedings of the Forth Language Resources and Evaluation Conference, LREC 2004*, pages 395–398, Lisbon.

Paul Rayson, Dawn Archer, Scott Piao, and Tony McEnery. 2004. The UCREL semantic analysis system. In *Proceedings of the workshop on Beyond Named Entity Recognition Semantic labelling for NLP tasks in association with LREC 2004*, pages 7–12, Lisbon.

Philip Resnik and David Yarowsky. 2000. Distinguishing systems and distinguishing senses: New evaluation methods for word sense disambiguation. *Natural Language Engineering*, 5(2):113–133.

Serge Sharoff. 2006a. Creating general-purpose corpora using automated search engine queries. In Marco Baroni and Silvia Bernardini, editors, *WaCky! Working papers on the Web as Corpus*. Gedit, Bologna.

Serge Sharoff. 2006b. A uniform interface to large-scale linguistic resources. In *Proceedings of the Fifth Language Resources and Evaluation Conference, LREC 2006*, Genoa.

A.D. Shveitser. 1988. *Theory of Translation: Status, Problems, Aspects*. Nauka, Moskow.

Krista Varantola. 2003. Translators and disposable corpora. In Federico Zanettin, Silvia Bernardini, and Dominic Stewart, editors, *Corpora in Translator Education*, pages 55–70. St Jerome, Manchester.

Anna Wierzbicka. 1999. *Emotions across Languages and Cultures*. Oxford University Press, Oxford.