Harnessing the lawless: using comparable corpora to find translation equivalents

Serge Sharoff
Centre for Translation Studies,
School of Modern Languages and Cultures,
University of Leeds
e-mail: s.sharoff@leeds.ac.uk
tel: +44-113-343 7287
fax: +44-113-343 3287

Abstract

Bilingual dictionaries provide basic translation equivalents for a headword and typically limit the set of equivalents to words of the same part of speech as the headword. However, words taken in their contexts can be translated in many more ways. At the same time, equivalents listed in dictionaries are not adequate in many contexts, because of the contextual and collocational sensitivity of target language expressions. The problem is particularly acute for novice translators who lack the experience for finding contextually-appropriate translations. The paper proposes a methodology for finding translation equivalents in comparable corpora. This helps in training translation students to be aware of the translation potential of polysemous words from the general lexicon.

1 The problem

Di Sciullo & Williams (1987:3) introduced a nice metaphor into linguistic research: "The lexicon is like a prison—it contains only the lawless". Even though their remark was made in the context of typological research to explain the reason why it pays very little attention to lexical items, the same metaphor can be applied to research in translation studies as well, but we will interpret it in a different way: in translation the criminal code is provided by bilingual dictionaries, but words frequently violate it. Dictionaries give translation equivalents for a headword, but words taken in their contexts can be translated in many more ways than indicated in dictionaries. The cause for the problem is that dictionaries cannot address the difference between the way concepts and words are combined in the source language vis-à-vis their potential combinations in the target language.

Some differences are semantic in their nature: the headword and its translation equivalent differ in their meanings, so the suggested 'equivalent' cannot be used in a particular context. There are even claims that it is impossible to establish translation equivalence for culture-specific concepts, for instance, that the notion of the so-called 'basic' human emotions does not exist. Durst (2001) uses a deliberately provocative title in his paper: "Why Germans don't feel 'anger'". The reason for his claim is that even though *anger* in English and *Ärger* in German are etymologically related and can be used as translation equivalents in some contexts, there is a fundamental difference between them, namely, that German expressions using *Ärger* typically occur in the context of a prepositional phrase using *in* (in), *mit* (with) or *über* (about) that expresses the cause of anger, for instance, *Ärger mit dem Chef, Erbe, Hauseigentümer, über die Nachbarschaft haben* (to have anger with the boss, will, landlord, about neighbours). So, according to Durst, *Ärger* refers to the emotional state directed to its cause, whereas *anger* refers to the state of affairs per se without necessarily mentioning its cause.

Also, the semantic field of anger in German is structured differently. In addition to *Ärger*, there are two other nouns *Wut* and *Zorn* with respective adjectives (*wütend* and *zornig*): *Wut* is

used for referring to strong cases of anger, when one looses control over one's behaviour, whereas *Zorn* emphasises the right to have the emotion, so it is frequently used in such collocations as *in gerechtem Zorn* (in righteous anger). Even though English has words that express similar emotions (*rage* and *wrath* respectively), their semantics does not coincide with *Wut* and *Zorn*, which are every-day words frequently used in modern German for expressing respective types of anger. What is more, the two words were used as the two basic designations of this emotional state, while *Ärger* as a more generic designation of the emotional state, was not as common as it became recently (Durst, 2001: 141).

The study by Uwe Durst was a continuation of an earlier study of the Russian word *гнев* made by Anna Wierzbicka (1998). The Oxford Russian dictionary (an authoritative and reliable source) lists *гнев* as the only translation equivalent of *anger*. However, the meanings of the two words differ significantly. Unlike *anger*, *гнев* has the components of 'strong emotion' and 'clash with basic ethical principles', or in Wierzbicka's Natural Semantic Metalanguage 'this person did something bad, if someone does something like this it is bad' (Wierzbicka, 1998: 20). Thus, translating *anger* as *гнев* will drastically change the meaning in cases when those components are not intended in the original.

Another source of differences between translation equivalents suggested in dictionaries and real translations concern collocational restrictions: even if the headword and its translation equivalent share the same meaning in a particular context, the immediate lexicogrammatical environment can reject the equivalent, because it is not customarily used in this context. For instance, the translation of *strong evidence/reason* to German would require *erdrückende/überzeugende Beweise/Argumente*, whereas *stark*, *fest* or other translation equivalents suggested for *strong* in the Oxford-Duden German dictionary do not fit well.

In addition, bilingual dictionaries typically limit the set of translation equivalents to words of the same part of speech (POS) as the headword. However, if the syntactic structures of source and target sentences differ, this may lead to the choice in the target language that is not suggested in a dictionary at all. For instance, *little* in English and *klein* in German are typically listed in dictionaries as translation equivalents. Their meanings are sufficiently close, however, in real translations from English to German little things are frequently translated using suffixes, like *–chen*.

However, if we return to the prison metaphor, the degree of translational delinquency varies between different types of words. Some words like proper names are minor offenders: there are very few cases when they change their translation in a specific context, though such changes are still possible even for proper names. For instance, the transliteration of foreign names in Russian follows phonetic principles with few exceptions, such as the names of European rulers which are expressed using historical patterns. This means that the Russian transliteration of the name of the current Prince of Wales Charles as *Чарльз* will turn into *Карл*, when he becomes the king.

The cases of contextual sensitivity of translations of names of physical objects are more frequent, because languages structure reality in different ways. The Russian expression *журнальный столик* (lit. a small table for magazines) is not listed in major dictionaries and cannot be translated into English in the word-for-word fashion as *a magazine table*. However, it can be rendered in English as either *a coffee table* or *side table*, depending on its functional purpose, whereas the exact function (and hence translation) of *журнальный столик* can be derived only from the context of the source sentence in Russian.

Words denoting abstract concepts are more hardened offenders, as the example with expressions referring to human emotions shows. However, if we accept the claim that *гнев≠anger≠Ärger* and that they refer to culture-specific concepts, then no translation between English, German and Russian is possible, unless readers of translation know the exact differences between the concepts in respective cultures. The methodology used by Wierzbicka and her colleagues assumes that every word *has* a fixed meaning, which is defined as a dictionary sense and

can be formally defined using Natural Semantic Metalanguage (NSM). In this view *гнев, anger* and *Ärger* refer to pieces of NSM code that are not identical

This view can be compared against the communication-oriented approach, which assumes that the meaning of a word is a function of its use in purposeful communication. The speaker uses resources of language for expressing communicative intentions according to the context of situation, cf. a similar study in (Sharoff, 2004). In this view, the task of a translator is to get the original message across, starting with original meaning intentions realised by the author using resources of the source language and re-realising them using resources of the target language. This does not assume that there is an equivalence between concepts in the two languages, but that there are resources available in the target language that can create a meaning that is functionally adequate within the target language and culture. In such situation *anger* can be indeed fruitfully translated as *гнев*, given a suitable context, in which the emotion is strong and justified:

(1) *she gave a little scream, half of fright and half of anger...* (Alice in Wonderland)
Аня издала легкий крик — не то ужаса, не то гнева ...(Nabokov's translation)

In the following section, we define the exact aim of the study, which consists in finding context-dependent translation equivalents using comparable corpora, and compare it to existing research in the field. Then we propose a four-step methodology for achieving this goal and present a case study of using the methodology to solve a real-life problem occurring in translation of a sentence from English into Russian.

2 The aim of the study and related research

A professional translator has vast experience in finding lexical items that fit well into the context of translation. Some translators maintain "non-systematic" dictionaries (Palazhchenko, 2002), which highlight words that can cause troubles in translation and summarise possible contexts in the source language and their translations. Part of the education of future translators consists in equipping them with a range of resources for finding contextually appropriate translations, or to put it metaphorically, equipping students with skills to harness words that break the dictionary code.

In this paper, we investigate the possibility to detect translation equivalents of polysemous words using comparable corpora. The emphasis of the study is to develop a methodology for training novice translators to produce translations beyond the set of equivalents offered by dictionaries and to increase their awareness of the translation potential of words from the general lexicon. Throughout the paper we use a case study of translation from English into Russian for one word *frustratingly* in the context of a sentence taken from a recent BBC report:

(2) *Mr Blair said that the parties to the Northern Ireland political process were "frustratingly close" to an agreement but had yet to finalise a deal because of uncertainty over the IRA's intentions.*

The word *frustratingly* presents a problem for its translation to Russian for several reasons. First, standard English-Russian dictionaries do not list derived word forms (*frustrated, frustrating, frustratingly*) at all. Second, even though they have headwords for *frustrate* and *frustration* and offer some translation equivalents for them, the translations listed are not adequate in many (if not most) contexts, because of the contextual and collocational sensitivity of the respective Russian expressions and because of the need to transcend POS boundaries in translation. For instance, the Oxford Russian Dictionary lists *разочаровывать, расстраивать, обескуражен* for *frustrate*, and *разочарование, крушение, безысходность, фрустрация* for *frustration*. However, in each of the following examples the most natural translation into Russian uses a word that does not belong to the set listed above:

(3) En: *Saddam's ambition ... is frustrated by the presence of UN inspectors.*
Ru: *Стремлению Саддама ... мешает пребывание инспекторов ООН.*
Gloss: Saddam's ambition ... is hampered by the presence of UN inspectors.

(4) En: *Nixon inherited a society rent by frustration.*
Ru: *Никсон унаследовал общество, раздираемое противоречиями.*
Gloss: Nixon inherited a society torn apart by tensions. (Palazhchenko, 2002: 192)

(5) En: *Yet, frustratingly, her story has no structure or substance.*
Ru: *Но, к сожалению, ее рассказ не выделяется ни формой, ни содержанием.*
Gloss: But, unfortunately, her story is not remarkable for its structure or substance.

Third, the word *frustratingly* is derived from a common Latin root, *frustra* (in vain), which forms are also used in Russian, e.g. *фрустрирующе* (frustrate+active participle+adverb), but it is a potential 'false friend' for *frustratingly*, because the Latin-derived words are used in modern Russian mostly as terms in psychology and psychiatry, for instance:

(6) *насколько удовлетворяющей или фрустрирующе-депривирующей оказывается новая реальность?*
To what extent is the new reality satisfying or frustratingly-depriving? (used as a term)

The problem with translation of the word *frustration* is also discussed by Wierzbicka:

> *frustration* is a highly culture-specific concept, very characteristic of modern Anglo culture, with its emphasis on goals, plans and expected achievements. In other languages, the concept of "frustration" exists only as a relatively recent loan from English (*Frustration* in German, *frustracja* in Polish, *frustracija* in Russian…) (Wierzbicka, 1999: 72).

First, Wierzbicka is wrong in assuming that *фрустрация* in Russian has the same function as *frustration* in English. The Latin-based loan word is mostly used in science, so its currency is severely restricted: it is appropriate in (6), but cannot be used to render *frustration* in the Russian translations of (3)-(5). She is right in claiming that there is no good translation equivalent for *frustration* in German, Polish and Russian. Yet, a feeling of disappointment because "you are unable to do anything about the problem" (the definition of *frustration* from the Collins-COBUILD English Dictionary) is common to many cultures and by no means recent in Russian:

(7) *Он уверен в самом себе … и совершенно незнаком с сомнениями и разочарованиями, от которых седеют таланты* (А.Чехов, Скучная история)
He believes in himself … and knows nothing of the doubts and frustrations that turn the hair of talent grey. (A.Chekhov, A Dreary Story, translated by Constance Garnett)

Even if *разочарование* is not fully equivalent to *frustration*, it can function as its perfect translation equivalent in the context of example (7). In the same way, *гнев* can be used for translating *anger* in (1), even if they are not fully equivalent.

Parallel corpora consisting of original texts aligned with their translations offer the possibility to search for examples of translations in their context, so parallel corpora provide a useful supplement to decontextualised translation equivalents listed in dictionaries. However, parallel corpora are not representative: millions of pages of original texts are produced daily by native speakers in major languages, such as English and Russian, while translations are produced by a small community of trained translators from a small subset of source texts. The imbalance between original texts and translations is also reflected in parallel corpora, which are simply too small for studying moderately frequent words: *frustrate* occurs 631 times in 100 million words of the British National Corpus (BNC). This gives in average about 6 uses in a maximally conceivable English-Russian parallel corpus of one million words. The frequency of *frustratingly* in the BNC is even below one instance per million words (42 uses in the BNC) and, anyway, no parallel English-Russian corpus of this size is currently available for students.

The use of corpora is popular in lexicography and translation studies as well as in computational linguistics, however, the methodology we are going to propose addresses aspects of corpora and translation training, hitherto not discussed. Some researchers in translation studies, like Baker (1996), concentrate on using parallel corpora for comparing translations to original texts. Others, like Aston (1999) or Teubert (1996), use corpus methods for studying differences between uses of words in original and translated texts, for instance, by comparing the relative frequency of *Schadenfreude* in original German texts and German translations from English. In translation studies it is common to verify translation hypotheses by checking examples in comparable corpora. For instance, Varantola (2003) discusses compilation of ad hoc corpora and their role in translation viewed as a decision-making process for solving context-dependent problems (this contrasts with context-free descriptions usually available in dictionaries). Aston (1999) illustrates the use of corpora in translation training via several case studies. For instance, he shows the study of the concordance lines of *apartment overlooking X* in the BNC, which detects the positive semantic prosody of corresponding objects, such as mountains, rivers or gardens. This can be compared to the connotations of its Italian translation *dava su*.

Our methodology serves essentially the same purpose of using corpora for solving context-dependent problems in translation. However, it is aimed at a more difficult task of *finding* a translation equivalent, if it is not known for the student and not listed in the dictionary. To achieve this goal, the methodology makes extensive use of computational methods in addition to study of concordance lines. Related work in the computational brand of translation studies has concerned the use of comparable corpora for the detection of equivalents between technical terms, for instance, (Bennison & Bowker, 2000, Dagan & Church, 1997). However, our approach addresses the needs of practising translators and advanced learners of a foreign language in finding appropriate translations for words from the general lexicon.

Corpora are also widely used in lexicography (Landau, 2001), including bilingual lexicography (Atkins, 1994). The proposed methodology may contribute towards improvements in dictionary making, for instance, by suggesting another equivalent for *anger* in the Oxford Russian Dictionary or a new entry for *frustratingly*. However, 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 *frustration* (for instance, English monolingual dictionaries list no more than two-three senses for it). As for usability, it is impossible to use a (printed or electronic) dictionary in which the relevant translation is buried in the long list of potential translation equivalents: a translator or a student will not find a translation they want. There are also limits on the structure of a dictionary entry. For instance, the requirement of using identical POS tags in the source and target words is justified by the need for substitutability, as "the dictionary should offer real lexical units of the target language that, when inserted in the context, produce a smooth translation" (Zgusta, 1987). The requirement is sensible, but it is exactly one of the reasons for decontextualisation of translation equivalents. The purpose of the proposed methodology is to give students skills in finding translations grounded in the context.

3 The methodology

In this study we show how to find contextually appropriate translation equivalents of words from the general lexicon using large comparable corpora. The first two steps in the proposed methodology rely on *source* language corpora to study the polysemy of the word that causes the problem and to detect stable 'islands' that identify the context and can be easier to translate into another language. The third step identifies the bridge between the two languages using translation equivalents for those stable islands. This step may require the use of machine readable bilingual dictionaries or other lists of translation equivalents. In the fourth step we use *target* language

corpora to analyse typical contexts of stable translation equivalents in an attempt to find contexts similar to the original.

The application of the proposed methodology is based on several assumptions. The methodology assumes that trainee translators have skills in the vertical reading and analysis of concordance lines, as the methodology crucially depends on their ability to notice and describe lexical patterns in raw data, i.e. original concordance lines without filtering or preselection by the teacher. Skills for vertical reading of concordance lines sorted around a keyword are different from those required for horizontal reading of a continuous text. This aspect should be explicitly taught in the classroom.

The methodology also assumes the existence of sufficiently large source and target language corpora that are representative for the respective languages in general or representative for a specific domain, such as the BNC as a general-purpose English corpus or the Reuters corpus for newswire texts. The corpora have also to be equipped with an interface to produce concordances and collocation lists. These facilities are treated as standard in modern corpus querying software. In addition to them the methodology uses powerful computational techniques for detection of "similarity classes", i.e. groups of words with lexically similar behaviour, and "equivalence classes", i.e. groups of words in the target language that correspond to similarity classes of the source language.

Corpora for the two languages should be sufficiently large to provide interesting collocations and similarity classes for moderately frequent words, like *frustrate* or *agreement.* For English we use the BNC and a subset of the Reuters corpus (Rose, 2002), consisting of newswire texts annotated with general topic codes only (news from the markets have been excluded). In total this gives the English corpus of about 190 million words (MW). The Russian corpus used in the study is the pilot version of the Russian Reference Corpus consisting of about 50 MW, a balanced collection of written modern Russian. In other experiments our students accessed equally large Chinese and German corpora. The corpora have a uniform interface which can produce concordances, lists of collocations, similarity classes and equivalence classes (Self-identifying reference). The query language is based on the IMS Corpus WorkBench (Christ, 1994).

What is the rationale for using collocations and similarity classes? Concordances and collocations define the syntagmatic relationship between a word and its lexical environment. Computationally the strength of a collocation is described by its frequency relative to the frequency of individual lexical items. In the present study we measured the strength using the log-likelihood score (Manning, Schütze, 1999: 172), according to which the collocation *vent one's anger* is stronger than *in anger* (their log-likelihood scores are respectively 102 and 27), even if the latter combination of words is much more frequent (32 vs. 169 uses in the BNC). This reflects the fact that the proportion of occurrences of *vent one's anger* among other uses of *vent* is much higher than the one for *in*.

The similarity class of a word defines the paradigmatic relationship between it and other words that can appear in similar contexts. The semantic distance between words in a similarity class depends on the number of collocates they share (Manning, Schütze, 1999: 294). This is cognate to the definition of the relationship of synonymy in a thesaurus, but there is a difference, in that the notion of similarity classes pays greater attention to the affinity between the contexts in which the words occur. For instance, the words *compromise* and *negotiation* cannot be listed in a dictionary as synonyms to *agreement.* However, the word *compromise* has such collocates as: *reach, find, seek, accept, negotiate,* while the word *negotiation* collocates with *peace, union, contract, political, general, trade, government.* As both of the two lists contain collocations for the word *agreement*, the two words *compromise* and *negotiation* are listed in its similarity class.

Words in the similarity class can also generalise the meaning of a word in question. For instance, *green* has the following similarity class: *red, blue, yellow, brown, white, pink, black, coloured, grey, purple, dark,* while we cannot make the assumption that *green* is synonymous to *red,*

coloured or *dark*. At the same time, similarity classes produced automatically are not always useful. For instance, the word *new* produces the following similarity class: *first, change, time, modern, old, see,* etc. These words indeed share many collocates with the word *new*, but they cannot be used to indicate contexts that are similar to those in which it is used.

The equivalence class of a word consists of translations of words in the similarity class. For instance, the equivalence class of the Russian word *зеленый* (green) consists of *blossom, blossoming, blue, bright, brown, colour, crimson, dull, flower, gold, golden, green, grey, lucid, orange, pink, raspberry, red, rosy, white, yellow* (the list was produced automatically using the Oxford Russian Dictionary). The result reflects the ambiguity of some words in the original Russian similarity class, e.g. *цвет* (colour) also means *blossom* and *flower;* or *малиновый* (crimson) also means *raspberry,* though this does not destroy the semantic centre of the equivalence class, which is about the colour.

Now we will take a closer look at the steps in the methodology. In the first step we produce the list of occurrences of a word in question (*frustratingly* in our case) and study the concordance to identify the functions performed by this word in contexts similar to those used in the original expression. In doing this we can also identify contexts in which the function of the problematic word is quite different from the one in the original sentence. Since its translation in such contexts is also likely to be different, these contexts define negative restrictions for further search in target language corpora. Note that the first step assumes that the search term is moderately frequent: the number of concordance lines should be sufficiently large to detect stable patterns of uses of the search term, at least 30. This is typically true for words from the general lexicon queried in large corpora; however, the first step can be skipped, provided we can describe the function of the word in question on the basis of the problematic example alone.

The second step is to generalise the context of the original example by identifying other words which can have more obvious translation equivalents. For instance, in example (2) such words are *close* and *agreement*. They are important for three reasons. First, they occur in the immediate context of *frustratingly* in the example. Second, they directly contribute to the function of *frustratingly*: the example is about expressing a feeling due to the impossibility to reach an agreement of some sort (this is unlike the references to particular political entities, such as *Tony Blair* or *the Northern Ireland political process* in the same example). Third, they form an island of stability for translation, because their Russian translations in this context are relatively easy to guess and they are also listed in bilingual dictionaries: *close* in this sense will be most probably translated as *близкий, agreement* as *соглашение* or *согласование*. This will provide anchors for finding similar contexts in a target language corpus. The two words (*close* and *agreement*) are also more frequent than *frustratingly,* so we can detect their most significant collocations, as well as their similarity classes. A computed list of semantically similar words can also be compared against the list of synonyms in a thesaurus and extended or amended as the result of comparison. The generalised context for example (2) thus reveals itself to be about evaluation of a state of affairs of being close to reaching an agreement, seeking a compromise or finding a solution through negotiations.

It is very unlikely that the second step will produce no results because of a lack of generalisable context, i.e. because all words are specific to the current example. For instance, in example (3) the anchors are *ambition* (with the similarity class including *aspiration, desire, pursuit,…*) and *presence* (*absence, existence, appearance, …*). Among examples (2)-(7) only one example (5) is harder to generalise, because the function of *frustratingly* is not constrained by specific lexical items occurring in the sentence. However, there are two ways to generalise the context here. First, *frustratingly* used in this sense is typically thematic, so we can check words that frequently appear in the context when *frustratingly* is used at the beginning of a sentence. They are *but* and *however*, which can be more or less unambiguously translated. Second, we can generalise the word itself using the readily available list of modal adjuncts of desirability: *(un)fortunately, regrettably, to my delight/distress…*(cf. Table 3(3) in Halliday, 1994).

The third step is to create equivalence classes of stable words using translations of words in their similarity classes. This will provide us with a bridge between monolingual corpora in the two languages. This step can be facilitated by the availability of a large-scale bilingual dictionary in machine-readable form, in order to produce equivalence classes without human intervention. In our experiments we used Oxford University Press bilingual dictionaries for German, Russian and Spanish. In some cases, a completely automatic procedure brings adequate results, such as in the case of *green* discussed above. In other cases, the equivalence class may include more irrelevant terms, for instance, the equivalence class of *гнев* (anger) consists of *admiration, agitation, anger, awfully, choppiness, delight, despair, disdain, disgust, dissatisfaction, disturbance, fear, fright, fury, hatred, horror, hydrophobia, indignation, insult, irritation, malice, risk, shame*. So the list is not consistently about emotions and requires manual filtering. The main aim of this step is to find reliable indicators of situations in the target language that are cognate to situations discussed in the original example.

The final step in the methodology is to study the results of a number of queries in the target language that consist of words in the equivalence class in order to find lines which suggest suitable translation equivalents. Since typically the number of concordance lines in such cases is large (equivalence classes consist of a number of frequent words), it is easier to study the most significant collocations for words in the equivalence class and then to study patterns consisting of those words with their collocations. For instance, if we wish to translate the Russian version of example (3) into English and have problems with the verb *мешать* (hamper), we can study verbs which can appear in the vicinity of two equivalence classes *attempt, desire, dream, intention, wish* (corresponding to the Russian word *стремление*, ambition) and *absence, arrival, departure, entry, presence, return, stay, trip, visit* (corresponding to *присутствие*, presence). The result lists *affected, blocked, discouraged, frustrated, hampered, impaired, mitigated, prevented, ruined.* It includes not only *frustrated* and *hampered,* the verbs used in the original sentence and its gloss, but also a range of other verbs that can be potentially used in similar situations.

In the process of studying concordance lines in the target language, e.g. comparing *mitigated by the presence* vs. *impaired by the presence*, we benefit from the analysis of source language concordance lines performed in the first step, because this allows us to remove from consideration examples which lack functions appropriate to the original sentence, e.g. the disapproval of Saddam's ambitions expressed in the original sentence is unlikely to co-occur with *mitigated*.

4 The case study

The case study is set up by the task of translating *frustratingly* from example (2) into Russian. It starts with the list of the 47 instances of the pattern *frustratingly* followed by an adjective (pos="JJ") in the English corpus (composed of the BNC and Reuters corpus). The following are some lines from the concordance:

(8) *a real do-or-die enthusiasm that's* ***frustratingly absent*** *from so many*. **(J)**

(9) *The notes that send us round the exhibition are* ***frustratingly brief***. **(A)**

(10) *Trust is* ***frustratingly difficult*** *to measure*. **(A)**

(11) *Comets are* ***frustratingly rare*****.** **(A)**

(12) *the re-moulding of mentalities was going to be a* ***frustratingly slow*** *process.* **(J)**

(13) *Yet the* ***frustratingly small*** *amount we are told about the marriage emerges…***(A)**

(14) *Television interviewers find him* ***frustratingly vague***. **(J)**

(15) *It seems the gulf between potential and reality remains* ***frustratingly wide***. **(A)**

It is clear that the basic function of *frustratingly* in such constructions is to evaluate something negatively with a strong emotional charge: 'it is frustrating that X is so badly Y-ish'. Note that in the case of our source expression the state itself is clearly positive (it is good to reach an agree-

ment). The negative evaluation concerns the fact that the parties missed the possibility to reach it: 'it is frustrating that we were at X, but failed to reach Y'. A closer look at the set of concordance lines reveals that there are two basic patterns of evaluation: we can use *frustratingly* to evaluate physical objects and state of affairs or to judge activities of other people and organisations. In the linguistic study of evaluation (Martin, 2000), the difference has also been identified under the names of Appreciation (evaluation of objects) and Judgement (evaluation of behaviour). The third term discussed by (Martin, 2000) is Affect (direct expression of one's emotions), which is the basic way of encoding feelings, whereas Appreciation and Judgement are indirect expressions of feelings in the context of evaluation. Examples of Affect do not occur in our concordance lines composed of *frustratingly* followed by an adjective, but they are typical for other patterns of uses of the word, for instance:

(16) *Frustratingly, we still don't know enough about what brought Behn to feel this exceptional moment suited her.*

Concordance lines from examples (8)-(15) have been marked with codes (J for Judgement, A for appreciation). The use of *frustratingly* in example (2) is also a judgement of the results of negotiation. There is no requirement that a student should know the exact terminology or be aware of appraisal theory in general. The terminology might be useful to present the results of the study, however, even without a theoretical background, it is easy to notice the fact that the function of *frustratingly* in examples (8)-(15) is to evaluate something and this something can be either a state of affairs or people's behaviour, in particular, the progress in negotiations in example (2).

We have already suggested that two words (*agreement* and *close*) are good indications for the context and also provide the islands of stability of translation. Because of the general polysemy of the word *close*, neither its automatically constructed similarity class (17) nor a list of synonyms (18) provide generalisations useful for the current context:

(17) *closed, proximity, intimate, open, shut, closely, distance, confidant, informal, tight, friendship, kissing, friend, contact, closure,*

(18) *near, adjacent, adjoining, at hand, cheek by jowl, handy, impending, nearby, neighbouring, nigh; intimate, attached, confidential, dear, devoted, familiar, inseparable, loving* (from the Collins Gem Thesaurus)

However, the automatic similarity class of *agreement* gives a much more useful word list:

(19) *accord, agree, agreed, arrangement, bilateral, clause, compromise, contract, covenant, deal, declaration, lease, negotiate, negotiation, obligation, pact, sign, treaty*

Many words in the list can be used in the context of evaluation, when one expresses regret for being not able to reach the state of affairs designated by the word. A thesaurus (we consulted the Collins Gem Thesaurus) is less useful here, as for *agreement* it lists three groups of senses:
1. assent, agreeing, compliance, concord, concurrence, consent, harmony, union, unison
2. correspondence, compatibility, conformity, congruity, consistency, similarity
3. contract, arrangement, bargain, covenant, deal, pact, settlement, treaty, understanding

The third group is potentially relevant, but words not included in the automatically produced list (*bargain, settlement* and *understanding*) are less likely to occur in the context of *frustratingly*. On the other hand, *negotiation* is not listed as a synonym to *agreement*, but it evidently belongs to the generalised target context, which can be described exactly as the evaluation of the progress in negotiations.

The word *close* on the other hand provides a useful list of important adverbial collocates, starting with:

(20) *very, perilously, dangerously, fairly, quite, uncomfortably, pretty, sufficiently, finally, temporarily, reasonably, remarkably, desperately, extremely, agonisingly, tantalisingly, conveniently, particularly, officially, really* [1]

The list contains some other words referring to emotions (they are underlined), which confirms that this word is frequently used in the context of emotive evaluation, even though the exact use of '*frustratingly close*' has not been recorded in the corpus.

According to the third step, it is necessary to produce a list of translation equivalents for words from similarity classes in Russian. The situation with *close* is simple: its similarity class lacks suitable synonyms and it has only one translation into Russian appropriate here, namely *близки.* The rich similarity class of *agreement* gives a large translation equivalence class using translations listed in the Oxford Russian Dictionary:

(21) *аранжировка, аренда, декларация, договариваться, договор, договоренность, двусторонний, клаузула, контракт, меры, обязанность, обязательство, обсуждение, оговорка, отдавать, пакт, подавать, подписывать, предложение, предоставлять, примета, приходить, признак, пробираться, пункт, расписываться, расположение, соглашаться, соглашение, согласие, согласование, согласовывать, сокращать, статья, сжимать, утверждать, вывеска, заявление, заключать, знак*

Because of the inherent polysemy of words in the original lexical similarity list (19), the equivalence class significantly extends the domain of target expressions, including some irrelevant cases, for instance, *sign* from (19) refers not only to the event of signing an agreement, but also to symbols, so the list in (21) also contains *знак* and *примета* (symbol). It is possible to filter the list manually leaving only expressions relevant to the notion of negotiation or its result. Fortunately, the construction of a Russian query using *близки* (the equivalence word for *close*) followed by words from the equivalence class of *agreement* leaves mostly relevant expressions in the output.

The fourth step is to study the contexts in Russian using queries based on words in the equivalence class. It is easy to note that the word *close* in English collocates with modifiers referring to a degree, such as *very, fairly, quite, sufficiently,* etc. This suggests the first attempt to study collocates of the construction *близки к* (close to), which gives some modifiers which can already be used in translations of example (2): *уже совсем* (quite), *почти* (almost), *достаточно* (sufficiently), for instance, we can translate it as:

(22) *Стороны были уже совсем близки к достижению соглашения*
parties were quite close to reaching agreement

(23) *Стороны уже почти достигли соглашения*
parties almost reached agreement

However, they lack the emotional evaluation suggested by *frustratingly.* Given that *близки к* (close to) is very frequent (more than 3000 instances), we can restrict the search space by including the equivalence class of the *agreement.* However, to our disappointment the search finds very few occurrences of this pattern. Fortunately, the results can be corrected, if we pay attention to the difference between the syntactic structure of *close to an agreement* in English and the Russian expression *близки к* with its right collocates. The latter include many words referring to the completion of a process:

(24) [*близки к* (close to)] *достижению* (reaching), *завершению* (finishing), *осуществлению* (performing), *окончанию* (ending), *удаче* (success), *поражению* (failure), *победе* (victory)

If we now check the left contexts of a more complex query:

[1] Words are ordered according to their loglikelihood score.

COMPLETION-CLASS+AGREEMENT-CLASS

which includes Russian words referring to the completion of the process (24) followed by words from the equivalence class of *agreement* (21), then we find such expressions as *досадно* (regrettably), *обидно* (pitifully), *к несчастью* (unfortunately), *так* (so-emphatically). The first two expressions are especially suitable, because they provide standard ways for expressing an evaluation of the behaviour of others in emotional terms. This also finds the suggested translation of the target expression:

(25) *стороны были до обидного близки к достижению соглашения*

using the word *обидный* (pitiful) in a grammatically correct way.

5 Conclusions

The paper describes a methodology for searching comparable corpora to find translation equivalents for words from the general lexicon. The reason why we concentrate on words from the general lexicon instead of terminology is related to the fact that translation of terms is (should be) stable, while general words can vary significantly in their translation. It is important to collect the terminological database for terms that are missed in dictionaries or specific to a problem domain. However, once the translation of a term in a domain has been identified, stored in a dictionary and learned by the student, the process of translation can go on without consulting a dictionary or a corpus. To follow the metaphor from the introduction: terms are obedient citizens. In contrast, words from the general lexicon exhibit polysemy, which is reflected differently in the target language, thus causing the dependency of their translation on a context. It also happens quite frequently that such variation is not captured by dictionaries. Novice translators tend to rely on dictionaries and use direct translation equivalents whenever they are available. In the end they produce translations that look awkward and do not deliver the meaning intended in the original text. Making trainees aware of the translation potential of words from the general lexicon is one of the tasks of their education.

The methodology has been tested within the module of Corpus Linguistics for Translators. The students in the module performed their own case studies with *frustratingly* from (2) and with other examples, such as *absolutely*, *completely* and *entirely* (Partington, 1998) and checked the possibilities of their translations into their other languages. The exercises of this sort are useful for training future translators, even though they are not efficient as a regular procedure of practicing translators, because the study with all associated experiments takes too much time (two or three hours, potentially more, if we study and classify more contexts of uses in the two languages). However, the results are rewarding for trainees, because the final description covers not only the translation of a specific word (e.g. *frustratingly*) in a specific context, but a wider range of contexts in which such words as *close* and *agreement* are used, as well as the variation for expressing the results of negotiations in the target language.

A corpus does not produce translations automagically. Every step in the methodology requires creative thinking for formulating hypotheses and checking the results. The human task is to notice a pattern or suitable expressions in the set of concordance lines in both source and target language corpora. The computer's role is to support pattern discovery by producing and ordering concordance lines, giving collocates, similarity classes and classes of translation equivalence.

Acknowledgements

The current study benefited from two electronic resources, the Reuters Corpus (Rose, et al 2002), and the electronic version of the Oxford bilingual dictionaries. We are grateful to Reuters and Oxford University Press for providing the resources. Special thanks to Bogdan Babych and Tony Hartley for very useful comments on earlier drafts of the paper.

References

Aston, G. (1999). Corpus use and learning to translate. *Textus* 12: 289-314.

Atkins, B.T.S, (1994). A corpus-based dictionary. In: *Oxford-Hachette French Dictionary (Introductory section).* Oxford: Oxford University Press. xix - xxxii

Baker, M.(1996). Corpus-based Translation Studies: The Challenges that Lie Ahead. In Harold Somers (ed.) *Terminology, LSP and Translation: Studies in Language Engineering,* Amsterdam: John Benjamins, 175-186.

Bennison, P. & Bowker, P., (2000). Designing a tool for exploiting bilingual comparable corpora. In *Proceedings of LREC 2000,* Athens, Greece, May 30 - June 2, 2000.

Di Sciullo, A.-M., Williams, E. (1987). *On the Definition of Word*. Cambridge: MIT Press

Christ, O. (1994). A modular and flexible architecture for an integrated corpus query system. In *Proc*. *COMPLEX'94*, Budapest, 1994. Also available from the URL:

http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/Papers/christ:complex94.ps.gz

Dagan, I., Church, K. (1997). Termight: Coordinating Humans and Machines in Bilingual Terminology Acquisition. *Machine Translation*, vol 12:1/2, pp. 89-107.

Durst, U. (2001). Why Germans don't feel 'anger'. In J. Harkins, A. Wierzbicka, (eds.) *Emotions in Crosslinguistic Perspective*. Berlin: Mouton de Gruyter. 115-148.

Halliday, M.A.K. (1994). *Introduction to Functional Grammar.* 2nd edition. London: Edward Arnold.

Landau, S.I. (2001). *Dictionaries: the art and craft of lexicography*. Cambridge: Cambridge University Press.

Manning, C., Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Martin, J. (2000). Beyond Exchange: APPRAISAL Systems in English. In *Evaluation in Text*, Hunston, S. & Thompson, G. (eds), Oxford: Oxford University Press.

Palazhchenko, P. (2002). *Мой несистематический словарь* (My non-systematic dictionary). Moskva: Valent.

Partington, A. (1998). *Patterns and meanings: using corpora for English language research and teaching*. Amsterdam: Benjamins.

Rose, T., Stevenson, M., & Whitehead, M. (2002). The Reuters Corpus Volume 1–from Yesterday's News to Tomorrow's Language Resources. In *Proceedings of the Third International Conference on Language Resources and Evaluation*, Las Palmas de Gran Canaria.

Sharoff, S. (2004). How to handle lexical semantics in SFL: a corpus study of purposes for using size adjectives. In: G.Thompson and S.Hunston, (eds.) *System and Corpus: Exploring Connections*. London: Equinox.

Sharoff, S. (2004). Methods and tools for development of the Russian Reference Corpus. In D. Archer, A. Wilson, P. Rayson (eds.) *Corpus Linguistics Around the World*. Amsterdam: Rodopi, 167—180. The corpus interface: http://corpus.leeds.ac.uk/

Teubert, W., (1996). Comparable or parallel corpora? *International Journal of lexicography* 9(3), 238-264

Varantola, K. (2003). Translators and Disposable Corpora. In F. Zanettin, S. Bernardini, and D. Stewart (Eds.). *Corpora in Translator Education*. Manchester: St Jerome. 55-70

Wierzbicka, A. (1998). Sadness and anger in Russian: the non-universality of the so-called 'basic human emotions'. In: A. Athanasiadou and E. Tabakowska (eds.) *Speaking of Emotions: Conceptualisation and Expression*. Berlin: Mouton de Gruyter. 3-28.

Wierzbicka, A. (1999). *Emotions across Languages and Cultures*. Cambridge: Cambridge University Press.

Zgusta, L. (1987) Translational equivalence in a bilingual dictionary: Bāhukośyam. *Dictionaries* **9**. 1-47