Evaluation-guided pre-editing of source text: improving MT-tractability of light verb constructions

Bogdan Babych **Anthony Hartley** **Serge Sharoff**
Centre for Translation Studies
University of Leeds, LS2 9JT, UK
{b.babych,s.sharoff,a.hartley}@leeds.ac.uk

Abstract

This paper reports an experiment on evaluating and improving MT quality of light-verb construction (LVCs) – combinations of a 'semantically depleted' verb and its complement. Our method uses construction-level human evaluation for systematic discovery of mistranslated contexts and creating automatic pre-editing rules, which make the constructions more tractable for Rule-Based Machine Translation (RBMT) systems. For rewritten phrases we achieve about 40% reduction in the number of incomprehensible translations into English from both French and Russian. The proposed method can be used for enhancing automatic pre-editing functionality of state-of-the-art MT systems. It will allow MT users to create their own rewriting rules for frequently mistranslated constructions and contexts, going beyond existing systems' capabilities offered by user dictionaries and do-not translate lists.

1 Introduction: Automatic rewriting functionality for MT systems

Current state-of-the-art RBMT systems offer users customisable functionality for the transfer stage, in the form of user-definable do-not-translate lists and user dictionaries. However, the source language analysis capabilities of MT systems remain largely inaccessible to users and such systems still do not offer any support for rewriting at the pre-editing stage. Rewriting offers a way of enhancing the comprehensibility of MT output by more efficiently exploiting existing transfer resources of an MT engine, and greatly extends coverage of construction in a customised way without any changes to the engine.

In our paper we suggest that users will benefit from an integration of source-language rewriting capabilities into MT systems, and from their synergies with existing tools – user dictionaries and do-not-translate lists. Such integration will offer users much greater flexibility, because certain phenomena are much better treated in a monolingual rewriting stage rather than within the dictionary. For example, rewriting can 'repair' intractable word order, or handle discontinuous multiword expressions in a much more principled way than a user dictionary.

We describe an experiment which demonstrates the usefulness of source-rewriting functionality for state-of-the-art MT systems. For our evaluation-guided rewriting experiment we have chosen light-verb constructions (LVCs) – verb phrases consisting of a 'light', i.e., 'semantically depleted' verb and its object. Objects in such constructions are so-called logical predicates – such as names of actions, activities, states, properties, relations – that put forward some situational propositions. The relationship between verbs and complements can be described in terms of lexical functions like $Oper_1$ (Mel'čuk, 1996). LVCs are very common in a variety of languages, e.g., *take action, take part, put pressure, make a decision*} in English, *оказывать давление 'put pressure'* (lit. 'make pressure') in Russian, or *tenir compte de 'take into account'* (lit. 'hold account of') in French (Salkoff, 1999).

These constructions are often mistranslated by state-of-the-art MT systems, since they often require non-literal translation (En: *take part* > Ru: *~'accept part'*). Many of these constructions have synonymous verbs or other phrases, which can facilitate MT: (*take part --> participate*). We used such near-synonyms for rewriting problematic LVCs and evaluated the effect of this rewriting on the comprehensibility of the translations.

For this purpose we selected a group of LVCs for several frequent French and Russian light verbs and assessed the comprehensibility of their English translations. An initial analysis of the comprehensibility of LVCs aimed to identify the

Proceedings of the 13th Annual Conference of the EAMT, pages 36–43,
Barcelona, May 2009

most problematic constructions, whose re-writing could make the biggest impact on MT output quality.

RBMT systems generally have knowledge of the most frequent LVCs. For instance, ProMT does handle constructions with *принимать решение* reasonably, translating it as *make a decision*, not as *accept a decision*. Systran sensibly translates French LVCs like *commettre un crime* as *commit a crime*, and *commettre une erreur* as *make an error*. However, the coverage is not consistent. For instance, for the same directions Systran does not handle *faire en sorte* (*'do in such a way as to'*) in French, while ProMT does not handle *брать штурмом* (*'make an assault'*) in Russian. Hence, a feasible alternative is to rewrite problematic constructions in the source text to produce input that can be more sensibly handled by an individual MT system.

An insight into 'post-editing the source text' is offered by Somers (1997). For instance, he suggested re-writing source English country names (ungrammatically) as *the France*, *the Japan* for translation into French to improve the grammaticality of the output (*la France, le Japon*), where an MT system fails to insert the required definite article.

Our experiment evaluates improvements in MT quality, which can be achieved for a particular class of linguistic constructions, if such automatic pre-processing mechanisms are systematically implemented for state-of-the-art RBMT systems.

2 Previous work

Lexical and structural ambiguities in natural language have been traditionally fundamental problems in MT, which seriously limit the quality of its output, especially on unrestricted input text. However, if expressions in the source text are restricted in certain ways, the performance of MT systems can be improved considerably. For general-purpose MT systems, this observation has led to recommendations and tools for MT-oriented authoring and pre-editing of source texts –– so called MTranslability (Bernth and Gdaniec, 2001). For specific technical domains – such as software, aviation or automotive documentation – this observation has led to successful applications of Controlled Language (CL), which minimises post-translation editing (Nyberg et al., 2003) and usually works in conjunction with MT engines customised to match CL specifications.

A disadvantage of this approach is that very few texts are written using CL, so such recommendations are not directly applicable to the majority of texts that have to be translated for assimilation purposes using MT.

The proposed automatic rewriting mechanisms for MT will enable the users to accurately preprocess complex contexts of constructions from general language, which are commonly mistranslated by MT systems.

The remaining sections are organised as follows: In Section 3 we present the design of our experiment for evaluating the improvement achieved by modifying LVCs. In Section 4 we discuss its results and implications for creating re-writing rules for making MT input more tractable. Finally, in Section 5 we discuss other ideas possible in this field.

3 Method

3.1 Identifying LVCs

In our experiment we took the point of view of users of modern commercial MT systems (typically – medium or large translation companies), who want to improve comprehensibility of MT output, focussing on the contexts of particular classes of linguistic constructions, in our case we selected light-verb constructions (LVCs).

Our evaluation-guided procedure for systematically discovering frequent mistranslated contexts of such constructions and designing automatic rules to change them into a more MT-tractable form can be carried out by such users, who would find automatic rewriting mechanism for the pre-processing stage very useful. The procedure consists of the following stages.

As a first stage we generated lists of noun collocations for seven French and eight Russian light verbs, based on a study of LVCs by (Mudraya et al. 2008): French *commettre, donner, faire, mettre, passer, prendre, rendre* and Russian *брать (take), вести (lead), давать (give), делать (make), иметь (have), нести (carry), положить (lay), ставить (put)*. We manually checked the top-ranking collocates sorted by log-likelihood association scores and selected 63 French and 55 Russian LVCs (e.g., *принимать закон/меры/решение* ('pass a law', 'take measures', 'make a decision'). For each LVC we generated concordance lines (in a window of about 20 words) from French and Russian Internet corpora and translated them using three MT engines: French>English Systran 5.0,

Russian>English Systran 5.0 and Russian>English ProMT 8.0.

We randomly selected and analysed up to 25 concordance lines for each construction (the selection was intentionally small to model the real-world scenario of evaluation-guided improvement of MT for potential industrial users of the technology), and we identified those LVCs with the least intelligible translations, e.g.:

(3) Ne **faisons** pas **confiance** aux Anglais.
-> Systran: Let us not make confidence with the English.
Automatically rewritten ST:
Ne **comptons** pas **sur** les Anglais.
-> Systran: Let us not count on the English.

Since for Russian two MT systems were available, we used one of them (Systran) for identifying problematic LVCs, and the other (ProMT) for a 'blind rewriting' experiment, where rules and constructions selected for Systran were also applied to ProMT translation. The performance of ProMT on such constructions was, then, not known in advance.

Since MT systems can differ in their coverage of problematic constructions, this experiment was designed to assess to what extent the set of rewriting rules is system-dependent. Equally, it sought to establish whether re-writing rules are 'portable' from one system to another, that is, whether certain classes of language constructions are generally intractable for RBMT.

3.2 Rewriting of LVCs

The comprehensibility of certain LVCs clearly benefits from the rewriting of the source. Overall, nine of the LVCs identified for French exhibited this type of problem for at least some of their contexts of occurrence (*faire appel, faire confiance, faire face, faire (en) sorte, donner lieu, donner rendez-vous, mettre fin, prendre conscience, metre (la) main*). These constructions were selected for rewriting and we created rewriting rules for all their problematic contexts. Modifications mostly involved replacing the verb and keeping the noun (as the central meaning component) or replacing the whole construction. For example, the rewriting table for *faire confiance* is presented in Table 2:

Note that in our experiment separate rules were created for each combination of word forms that occurred in concordances, which was supposed only to simulate capabilities of real rewriting mechanisms that can be developed for state-of-the-art commercial MT systems. These rules in practice should be written in a more general way, since the rewriting system can have access to lexical and morphological features of word forms developed for the translation engines, e.g., then the last 3 rules in Table 2 could be merged into a single rule:

(3) [lemma='faire'] pas confiance [lemma=à] [lemma='le']?

The re-writing procedure was applied universally to all examples of LVCs, even if some examples were understandable in their original form.

faisaient au moins confiance à --> comptaient au moins sur
faire davantage confiance à --> compter davantage sur
fais totalement confiance a --> compte totalement sur
fais pas confiance à --> compte pas sur
faisaient pas confiance au --> comptaient pas sur le
faisons pas confiance aux --> comptons pas sur les […]

Table 2. Rewriting table for *faire confiance*

3.3 Evaluation of baseline translation quality for Russian LVCs

For Russian>English Systran 5.0 translations the problems with LVCs were more serious, so we carried out a systematic evaluation of contextual comprehensibility for all 55 LVCs. The comprehensibility of each MT-translated concordance line was annotated on 1-3 scale:

"3" - high confidence I understand correctly
"2" - low confidence I understand correctly
"1" - do not understand at all

The score was given to each concordance line out of 25 randomly selected contexts, and average scores were computed for each source-language LVC that generated these translations. Then these LVCs were ranked by their average scores, and the lowest ranking constructions were identified as those which need to be rewritten. Table 3 shows the numbers of Russian LVCs in the different ranges of comprehensibility scores. In all, 19 LVCs from the three lowest

groups of LVCs were selected for rewriting, since their average scores are centred around 'low confidence' or 'incomprehensible' scores.

Score range	Number of LVCs
[2.6 … 3.0]	23
[2.2 < 2.6)	13
[1.8 < 2.2)	10
[1.4 < 1.8)	5
[1.0 < 1.4)	4

Table 3. Comprehensibility of Russian>English LVCs

This analysis illustrates the extent of the LVC problem for the Russian>English MT system: 19 of 55 frequent LVCs (35%) generate low confidence translations and 9 of them (16%) produce mostly incomprehensible MT output.

Human evaluation scores before rewriting LVCs, and the extent to which LVC rewriting improves these figures are negatively correlated, so it is harder to achieve improvement by rewriting more comprehensible contexts. In particular, Pearson's correlation coefficient *r* between the baseline quality of LVC translations and the extent to which the quality could be improved via rewriting for individual LVCs is -0.71, and for averages for the ranked groups of six LVCs it becomes -0.99.

Therefore, we chose to rewrite the 19 LVCs with lowest evaluation scores, which should clearly benefit from rewriting.

3.4 Evaluators and evaluation packs

The results of re-writing were tested in an evaluation experiment. The comprehensibility of LVCs was judged by 16 native English speakers (Masters students in translation), who did not see the source text. The judges completed a questionnaire like that shown in Figure 1.

Evaluators judged concordances for LVCs in 10 different evaluation packs. Each evaluation pack contained 47 pairs of contexts for comparison: exactly one pair of randomly selected contexts for each LVC came from each of the three MT engines: French>English and Russian>English Systran 5.0 and another state-of-the-art Russian>English MT system ProMT 8.0. This system was used for 'blind' rewriting: the baseline performance of ProMT 8.0 on LVCs was unknown to us, and rewriting was done exactly as for Systran, without any preliminary system-specific tuning. The order within each pair of LVC contexts – left/right *vs* before/after rewriting – was also randomised, so evaluators did not know which context was the baseline, and which was experimental.

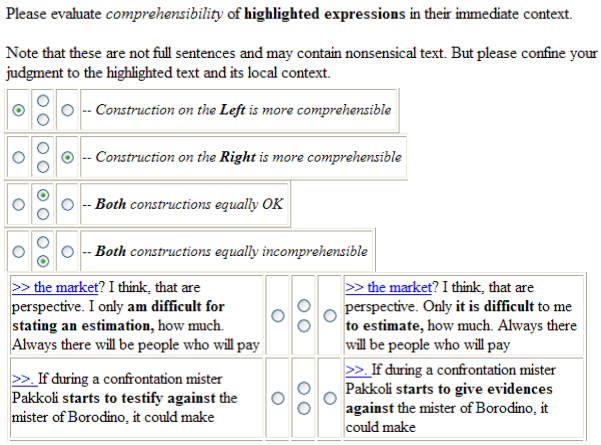

Figure 1. Evaluation questionnaire

There were two independent judgements for each context in the first six evaluation packs, and one judgement for contexts in the remaining packs. Evaluators gave 700 independent comparison judgements in total: 300 for each Russian>English system and 100 for French>English Systran.

Evaluation scores were converted to a numeric scale as shown in Table 4:

	Score Before RW	Score After RW
Before RW more comprehensible	+1	–1
After RW more comprehensible	–1	+1
Both equally comprehensible	+1	+1
Both equally incomprehensible	–1	–1

Table 4. Numeric conversion of evaluation scores

Numeric values were used for computing average scores for evaluators, MT systems and contexts of the same LVC, and for measuring the degree of improvement in these cases.

In our experiment average inter-annotator agreement measured by Cohen's kappa coefficient was around 0.28, which is a typical figure for human MT evaluation (Ye at al., 2007: 242). Still, our experiment was different from traditional MT evaluation, because human judges did not see complete sentences. We specifically asked our evaluators to confine their judgements to highlighted LVCs and their local context.

4 Results

4.1 Overall system evaluation

Chart 3 shows the overall number of comprehensible / incomprehensible translations before and after rewriting for Russian>English Systran 5.0.

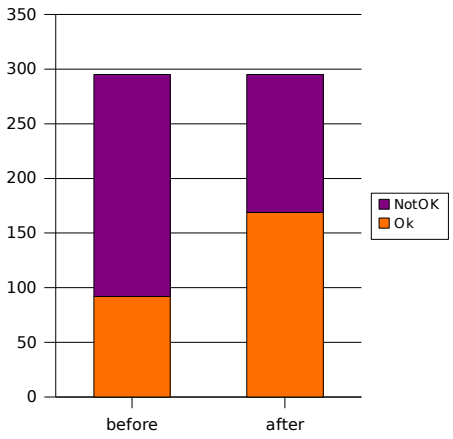

Chart 3. Results of LVC rewriting: Ru>En Systran5

The chart shows that rewriting of problematic contexts for this MT system gives a 38% reduction in incomprehensible translations and an 84% increase in comprehensible translations. French>English Systran showed even better improvement, while for Russian>English ProMT, which used blind rewriting, the improvement was smaller. Tables 5 and 6 summarise these improvement figures. Table 5 represents average evaluation scores on the [-1...+1] comprehensibility scale and proportional change in the number of comprehensible / incomprehensible contexts after rewriting.

	Before	**After**	**InC**	**C**
fr>en-Systran	-0.73	0.02	-44%	+283 %
ru>en-Systran	-0.38	0.15	-38%	+ 84 %
ru>en-ProMT*	-0.37	-0.03	-25%	+ 54 %

Before: average score before rewriting
After: average score after rewriting
InC: change in incomprehensible
C: change in comprehensible

Table 5. Average evaluation scores and changes in number of contexts

Note that for both Systran engines (Russian>English and French>English) evaluation-guided rewriting brought average scores above zero, which can be viewed as a comprehensibility threshold. Blind rewriting for Russian>English (ru>en-promt*) also brought an increase in average scores, but not enough to cross the threshold.

Table 6 shows proportions of scores in each category for the 3 MT engines.

	bothX	**before+**	**after+**	**both+**
fr-en-systr	0.42	**0.07**	**0.44**	0.07
ru-en-systr	0.29	**0.14**	**0.40**	0.18
ru-en-promt*	0.29	**0.23**	**0.40**	0.09

bothX: both Not OK
before+: before rewriting more comprehensible
after+: after rewriting more comprehensible
both+: both OK

Table 6. Improvement across MT systems

Again, for blind rewriting there is a much greater proportion of contexts which were judged as being 'better before rewriting'.

4.2 Construction-level evaluation

An evaluation of individual constructions provides a finer-grained analysis of the effect of LVC rewriting on comprehensibility. Charts 4 and 5 represent average scores before and after evaluation-guided rewriting, for each of French and Russian LVCs translated by Systran.

It can be seen from these charts that evaluation-guided rewriting normally increases comprehensibility of LVC contexts. Only 11% to 16% of LVCs show slight degradation in comprehensibility or no change.

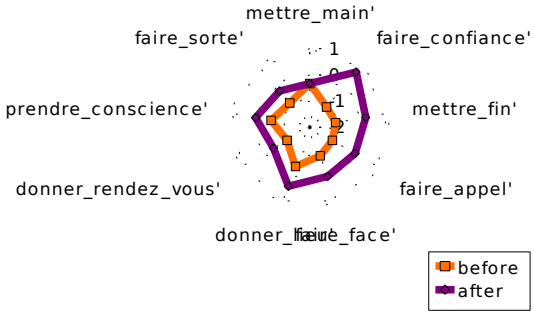

Chart 4. Average scores for Fr LVCs; Fr>En Systran 5.0

However, for blind rewriting 37% of LVCs showed a decline or no change in comprehensibility. Chart 6 illustrates these results. On this chart there is an area where the 'before rewriting' line is outside from the 'after rewriting' line, meaning that there is degradation of comprehens-

ibility for some constructions. Note that the majority of LVCs in this group are highly comprehensible before rewriting, and in this case rewriting decreases comprehensibility.

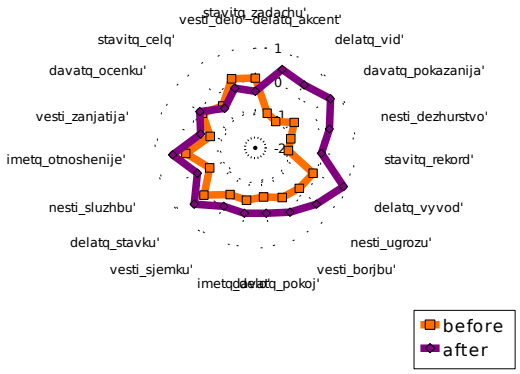

Chart 5. Average scores for Ru LVCs; Ru>En Systran 5.0

Surprisingly, two LVCs showed slight deterioration after evaluation-guided rewriting, but high improvement after blind rewriting: *ставить задачу (set a task): -0.4 vs +0.7* and *ставить цель (set a goal*): *-0.1 vs +1.11*, and only one Russian LVC showed no improvement at all for both engines: *вести дело (carry out business):* average scores changed by *-0.3* and *-0.2*

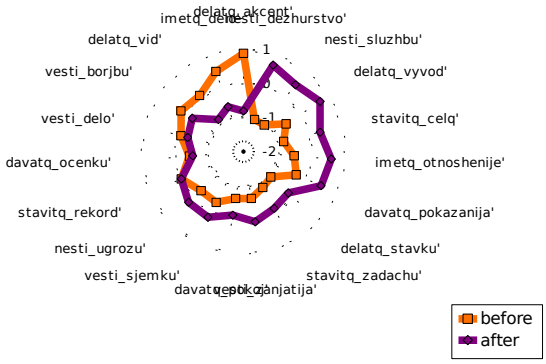

Chart 6. Blind rewriting: Average scores for Ru LVCs; Ru>En ProMT 8.0

5 Discussion

5.1 Automatic rewriting capabilities of MT systems

Our experiment demonstrated that the comprehensibility of translations of certain types of constructions, such as LVCs, can be improved via automatic rewriting of the source text into synonymous and more MT-tractable constructions.

Rewriting gives users gentle control over the analysis stage in the MT architecture, so they do not need to rely excessively on user dictionaries, which intervene at the transfer stage.

A general architectural consensus for RB MT is to try to do as much work as possible at the analysis stage and to keep transfer simple. Rewriting functionality will allow users to classify some problems, such as incomprehensible LVCs, as analysis problems, and deal with them in a more principled way using the existing capabilities of their MT engine, before such constructions are adequately covered by a new release of the engine. Automatic rewriting can also be more cost-efficient, e.g., for organisations which use MT for translating into several languages.

A rewriting mechanism can be more efficient in relying on the source-language analysis modules of the MT engine for extracting morphological and lexical features, multiword expressions, etc., and in allowing users to write rules in a compact way.

Traditionally user dictionaries do not handle discontinuous multiword expressions, even within the top commercial MT systems. Automatic pre-processing functionality can fill this gap, allowing the users to correct more MT problems than they currently handle via updating user dictionaries, e.g.:

(1) Законодательная власть принимает сотни законов ежегодно.
(Literal translation: Legislative power **passes** hundreds of **laws** yearly)
→ Systran: Legislative authority assumes hundreds of laws yearly.

The Russian construction *принимать закон (to pass a law)* has been mistranslated by Systran, so (1) is incomprehensible unless the reader knows Russian. Currently the entries in Systran's user dictionaries fail to match constructions with a gap, e.g., ***vote*** *hundreds* ***of laws***.

Adding a pair "*принимать ~ to pass"* into the MT dictionary will be detrimental for other contexts, because translation of this Russian verb, even in the same domain, can be very different: *to accept, to admit, to join, to take, to pass…*

Automatic rewriting rules on the design stage can take into account the context of rewritten constructions, and consequently – efficiently cover discontinuous multiword expressions. In the previous example when the verb of the source Russian sentence is replaced with a less ambiguous equivalent in a discontinuous context, Systran translates this part of the sentence correctly:

(2) Законодательная власть **вотирует** сотни законов ежегодно.
(Literal translation: Legislative power **votes** hundreds of **laws** yearly)
→Legislative authority votes hundreds of laws yearly.

The Russian input in (2) is much less idiomatic than in (1), but native English speakers find the MT output much easier to understand.

5.2 Evaluation-based requirements for MT-tractable language

Traditionally Controlled Language for MT and MT-tractable language are viewed as universal concepts, which should ideally work for any MT system. Specifications for such language have hitherto been derived from general considerations about MT, language complexity and ambiguity, or from results obtained on test suites (Bernth and Gdaniec, 2001: 177-195), but not from a corpus-level evaluation of particular MT engines. However, it is reasonable to expect that many such requirements would not stand the test of this type of evaluation. On the other hand, new requirements not envisaged by intuitive considerations may be discovered via corpus-based MT evaluation.

Our experiment supports the argument for the development of evaluation-guided methods for deriving specifications for tractable language, since blind rewriting caused deterioration for 37% of rewritten constructions. This result implies that there is no universal concept of MT-tractability, but that 'tractability' depends on the performance of particular MT engines. The proportion of constructions difficult for all MT systems is much smaller than expected and it is hard to justify any system-independent requirements for MT-tractable language. A challenge is to derive such specifications and rewriting rules automatically, based on evaluation of particular MT systems.

5.3 Construction-oriented MT evaluation

In this context, automated evaluation methods should not only give a general indication of MT quality, but also identify poorly-translated constructions, so BLEU-type scores alone will not be sufficient. In fact, the improvements analysed here may have a negligible effect in terms of BLEU scores, but still they can have an impact on the comprehensibility of many frequent translation contexts, so the proposed methodology can help to some extent to automate the process of error analysis for frequently mistranslated linguistic constructions.

In this paper we have demonstrated the use of concordance-based evaluation, which can be also modified for use in an automated evaluation framework, as suggested in (Anon, 2008), and used for assessing translation quality of particular constructions. As a next stage, synonymous constructions which are more tractable for a given MT system can be found automatically using distributional similarity techniques for multiword expressions, such as those proposed in (Anon, 2007). These constructions then become candidates for automatic rewriting.

6 Conclusions

Rewriting of LVCs can greatly improve the comprehensibility of their translations. In our experiment we achieved a reduction in incomprehensible translations of around 40%. The experiment also suggests that there is no universal concept of MT-tractability, so rewriting of contexts for problematic constructions should be guided by evaluation of the performance of particular MT systems for those constructions.

Future work will involve developing an automated approach to identifying ambiguous lexical units and problematic constructions in MT and finding their MT-tractable counterparts with similar distribution.

Automatic rewriting can be developed as a pre-processing functionality for users of state-of-the-art MT systems, and also as stand-alone rewriting applications, e.g., for pivot MT architecture via closely-related languages (Babych et al, 2007), where MT-tractable language can be viewed as closely related to the source.

References

Babych, Bogdan, Anthony Hartley, and Serge Sharoff. 2007. Translating from under-resourced languages: comparing direct transfer against pivot translation. In *Proceedings of the MT Summit XI*, pages 412–418, Copenhagen.

Bernth, Arendse and Claudia Gdaniec. 2001. *MTranslatability. Machine Translation*, 16:175–218.

Mel'čuk, Igor A. 1996. Lexical Functions: a tool for the description of lexical relations in a lexicon. In Leo Wanner, editor, *Lexical Functions in Lexicography and Natural Language Processing*, pages 37–102. John Benjamins.

Mudraya, Olga, Scott S. L. Piao, Paul Rayson, Serge Sharoff, Bogdan Babych, and Laura Lofberg. 2008. Automatic extraction of translation equivalents of phrasal and light verbs in English and Russian. In S. Granger and F. Meunier, editors, *Phraseology: an interdisciplinary perspective*, pages 293–309. John Benjamins.

Nyberg, Eric, Teruko Mitamura, and Willem-Olaf Huijsen. 2003. Controlled language for authoring and translation. In Harold Somers, editor, *Computers and Translation. A translator's guide*, pages 245–281. John Benjamins.

Salkoff, Morris. 1999. *A French-English Grammar: a contrastive grammar on translational principles*. John Benjamins.

Somers, Harold. 1997. A practical approach to using machine translation software. *The Translator*, 3(2):193–212.

Ye, Yang, Ming Zhou, and Chin-Yew Lin. 2007. Sentence Level Machine Translation Evaluation as a Ranking Problem: one step aside from BLEU. In: *Proceedings of the Second Workshop on Statistical Machine Translation,* Prague.