# Generalising lexical translation strategies for MT using comparable corpora

**Bogdan Babych, Serge Sharoff, Anthony Hartley**

Centre for Translation Studies, University of Leeds, UK

E-mail: b.babych@leeds.ac.uk, s.sharoff@leeds.ac.uk, a.hartley@leeds.ac.uk

## Abstract

We report on an on-going research project aimed at increasing the range of translation equivalents which can be automatically discovered by MT systems. The methodology is based on semi-supervised learning of indirect translation strategies from large comparable corpora and their application in run-time to generate novel, previously unseen translation equivalents. This approach is different from methods based on parallel resources, which currently can reuse only individual translation equivalents. Instead it models translation strategies which generalise individual equivalents and can successfully generate an open class of new translation solutions. The end goal of the project is integration of the developed technology into open-source MT systems.

## 1. Introduction

Data-driven MT architectures (statistical and example-based) generate translation solutions by reusing translation equivalents which are learnt from aligned parallel corpora. A long-standing problem with this approach (as compared to traditional rule-based MT architectures) is the lack of generality of those equivalents: unlike manually-crafted rules (which normally apply to lexicogrammatic patterns and morphological or semantic classes of words), automatically discovered equivalents are usually expressed as word patterns and do not generalise text words in them beyond the corresponding lemmas.

This leads to two types of disadvantages for practical MT systems. Firstly, from an engineering perspective it becomes difficult to maintain systems with a large number of very specific patterns: e.g., it is virtually impossible to correct translation errors in SMT by manually editing phrase tables. Secondly, from a linguistic perspective the lack of generality in equivalents learnt from parallel data can impose fundamental limits on the range of translation solutions that are generated by the data-driven MT systems. In (Babych et al., 2007a) we discuss the problem of *indirect* translation equivalence, i.e., cases when a word-for-word translation of expressions in a source language (SL) does not render a meaningful expression in a target language (TL). In such cases human translators apply different types of lexical and structural transformations. In extreme cases they have to change the lexical and syntactic structure of a sentence to alter the perspective from which the situation is viewed, as in Example (1).:

*(Ex. 1)* ***Ru**: Механизм принятия решений будет публичным.*

*(1.1)* ***lit**.: 'The mechanism of making decisions will be public'*

*(1.2)* ***human trans**: The answer will come from the people.*

In more technical terms, *indirect translation equivalents* represent *non-compositionality* in the process of translation in the sense that literal translations of individual constituents of a construction (phrase or sentence) do not produce a sensible translation of the whole construction.

The inability to generate indirect translation equivalents for new contexts results in non-fluent and incomprehensible translation or in mistranslation. The following examples were taken from a corpus of interviews originated in French, Russian and Spanish, as available on the Euronews website (www.euronews.net), They were translated into English by the Google statistical MT system (available at google.com/translate). Here we focus on cases where this state-of-the-art SMT system generates incomprehensible translations or comprehensible mistranslations.

*(Ex. 2)* ***Ru**: Из кризисов такого рода как парламентский можно спокойно выходить за счет демократических методов*

*(2.1)* ***lit**.: 'From crises of such type as parliamentary it is possible calmly to go by means of democratic methods*

*(2.2)* ***human trans**: We can escape crises like these through democratic means*

*(2.3)* ***SMT**: This kind of crisis as a parliamentary, can safely go through democratic methods*

In Example (2) the Russian expression *выходить из кризиса (exit from a crisis,* lit.: to go-out from the crisis) can generate a mistranslation if its components are translated directly as in (2.3). The human translation in (2.2) uses an indirect translation equivalent *to escape crises*. The syntactic perspective of the target sentence has also changed to accommodate the structural valencies of this expression, such as the requirement for an animate subject.

*(Ex. 3)* ***Es**: Es verdad que empezamos vacilantes pero era lógico.*

*(3.1)* ***human trans:** Of course we had our doubts to begin with but that's normal.*

*(3.2)* ***SMT**: It is true that we started to waver but was logical.*

In Example (3) the correct translation requires a change in the lexical perspective of the sentence, but the lexical translation equivalent is not covered by the MT system.

*(Ex. 4)* ***Fr** : Son équipe a découvert qu'il serait allé s'entraîner au Mexique. Dans le même temps, David Cassani l'a vu s'entraîner en Italie.*

*(4.1)* ***hum trans**: his team found out that at the time*

*he told them he was in training in Mexico he was seen by David Cassini, training in Italy*

*(4.2)* ***SMT****: His team discovered that he would have to train in Mexico. At the same time, David Cassani has been training in Italy.*

Example (4.2) is a comprehensible mistranslation. The challenge for MT here is to convey that, contrary to someone's belief, the event did not take place.

Given a large parallel corpus, an SMT/EBMT system can estimate probabilities of more frequent indirect translation equivalents in this corpus. However, large parallel corpora are relatively rare. They are not available for many language pairs, e.g., we are not aware of any for the German-Russian or Chinese-Ukrainian language pairs. Even if they are available, they tend to be specialised, e.g., the Europarl corpus for major European languages (Koehn, 2005). Thus the accuracy of models trained on them drops when the models are transferred to another text type in the same domain (Babych, et al, 2007). Moreover, the large (potentially infinite) number of cases like (1) or (2) can render an aligned parallel corpus of the Europarl size too sparse to give an acceptable coverage of problematic phrases possible even in a single domain.

Even if a problem requiring indirect equivalents occurs only a few times in a corpus, in each case it can be solved by humans in an idiosyncratic way. However, in the case of SMT these solutions are outweighed by more frequent direct equivalents which lexically intersect with them. In the case of Example-Based MT it is not easy to separate them from their contexts (since usually there is no generalisation in example patterns beyond lemmas). Therefore in their non-generalised form these solutions can correctly be applied only in very specific contexts, or can even generate undesirable errors if their non-compositionally translated components are aligned separately, e.g.: in (1) *механизм ('mechanism') ≠ answer* and in (2) *выходить ('to go-out') ≠ escape.* In the current EBMT framework such examples are treated as low-adaptable, and it may be beneficial not to use them for training the system (Collins and Somers, 2003).

## 2. Our approach

The goal of this paper is to investigate the possibility of automatic detection of indirect translation equivalents for previously unseen phrases using comparable corpora. Parallel texts are produced by a small number of highly skilled translators, while a much greater volume of texts on a wider variety of topics is produced every day by native speakers. Such texts are readily available on the Web for a large number of languages. Topic- or genre-specific sets of comparable corpora can be easily collected and used (Sharoff, 2006).

Our approach aims at getting information about suitable translations from much bigger comparable corpora using distributional *similarity classes* for key words and existing bilingual lexicons. It builds on results of our project ASSIST (Sharoff et al., 2008), where comparable corpora are used in a decision-support system that helps human translators to discover indirect translation equivalents. The purpose of our new project is to adapt ASSIST technology for MT, assimilating and extending its ability to generate indirect equivalents for previously unseen phrases.

This approach is different from the one used by (Munteanu and Marcu, 2006), who attempt to extend parallel corpus for training SMT by using very similar comparable corpora to discover aligned phrases in texts written about the same events, for instance, in the English and Romanian sections of the BBC website. In our case we work with comparable corpora which contain texts on similar topics, written in similar genres and originating from similar periods, such as 100 million words of the British and Russian national corpora, 200 million words of British and 70 million words of Russian newspaper text, or 200 million word collections of English and Russian webpages. Suggestions to use comparable corpora for MT include (Rapp, 1995; Fung 1998), both focusing on extracting general lexicon, and (Morin et al. 2007), focusing on extracting terminology.

In contrast, our approach does not use any explicit alignments of comparable corpora. In ASSIST indirect equivalents are generated in response to Multiword Expression (MWE) queries typed by users. The system extends each content word in a query using a set of distributionally similar words on the source side (words which have maximally similar collocation vectors in monolingual comparable corpora). This expansion of the query space is based on a procedure designed by (Rapp, 2004). For instance, for the query word *mechanism* the system automatically generates the following similarity class (with cosine similarity scores given in brackets): *device (0.374), framework (0.306), interaction (0.300), induce (0.295), process (0.291)...* Similarly, for the Russian word *публичный* ('public') the similarity class contains *открытый (0.356) ('open'), откровенный (0.267) ('candid', 'frank'),* etc. These word lists are then translated into a target language using available bilingual dictionaries (Oxford Russian Dictionary was used in ASSIST). In addition, for dictionary translations of query words we generate the lists of distributionally similar lexical items also on the target side, and we add it to the translation lists for each word in the query. The words from each of these translation lists are then recombined (as a Cartesian product). Finally, the combinations are filtered to eliminate any that do not actually co-occur in TL corpora.

This procedure generalises the context of the initial query beyond what is possible with the direct lexical transfer. The remaining combinations are ranked using distributional similarity scores and information from a dictionary of semantic classes. The result is displayed to users, who manually search through the lists and find combinations suitable for particular contexts. For instance, for the Russian query *четкая программа* (lit. 'precise programme') ASSIST suggests *clear idea, detailed plan, detailed proposal, clear strategy,* all of which may be suitable in a particular context.

ASSIST can even suggest translation equivalents for novel, creative SL expressions that do not occur in SL corpora, for instance, *recreational fear* in Example (5):

(5) *Patrick West recently claimed that Britain's extravagant mourning for Princess Diana and Holly and Jessica was 'recreational grief'. Maybe we also suffer from recreational fear.*

The system suggests phrases that help the translator to produce a translation capturing the gist of the idea in the source, such as *страх ради спортивного интереса*, i.e., '*fear for the sake of leisure (lit. sports) interest*'.

## 3. Methodology for MT

The extension of this approach to MT is based on the idea that these idiosyncratic solutions can become a valuable and reusable translation resource if they are discovered and generalised automatically in a way that they can be accurately applied within a wider range of contexts.

The goal of our project is to develop an automatic method for generalising indirect translation equivalents and for using them within a data-driven MT architectures, such as SMT and EBMT. In this paper we present a proof of concept and a feasibility study for integrating the translation strategy-based approach into data-driven MT architectures.

The main contribution is the concept of a lexical translation strategy – a generalised lexical pattern of indirect translation equivalents. These equivalents are automatically discovered in Russian–English aligned parallel corpora of newspaper articles (approx. 700k words) using a entropy-based approach, proposed for identifying idiomatic expressions (Villada Moirón and Tiedemann, 2006).

ASSIST's capacity for discovering previously unseen solutions comes from the fact that, while parallel alignment-based approaches try to accumulate individual equivalents, our system instead models a more general translation strategy that can generate novel, non-trivial solutions in line with the creative decision-making process used by human translators. In this case it uses a near-synonym, or “a near TL equivalent to a SL word in a context, where ... there is no one-to-one equivalent [and] literal translation is not possible” (Newmark, 1988: 84).

For the query in Example (1) *публичный механизм* (‘*public mechanism’*) a non-literal translation *open process* is found by ASSIST. These words co-occur in the English corpus and are sufficiently similar to the original query, so human translators are able to use the generated phrase to build a non-literal solution for (1): *There will be an open decision-making process.*

However, the translation in (1.2) *The answer will come from the people,* as created by a human translator, is still out of reach for ASSIST. The system cannot currently perform distant lexical transformations that are outside the similarity classes of individual lexical items. This is because we use just a single fixed translation strategy that bridges comparable corpora only via dictionary translations. This strategy is very productive, since it discovers about 20% of translation solutions to problems non-trivial for professional translators. But the remaining cases represent other types of indirect strategies, e.g., those which in Vinay-Darbelnet's (1995) model are described as *transposition* (change of a syntactic perspective), or *modulation* (change of a lexical point of view), see (Munday 2001: 57).

Importantly, in our new approach the range of generated equivalents now becomes larger, since comparable corpora are bridged not only via bilingual dictionaries, but also via generalised indirect solutions discovered by human translators. For instance, the generalisation of translation equivalents in

*(2.3) We can escape crises like these through democratic means*

allows the system to cover not just the phrase *to escape crisis*, but also the following: *to escape conflict/ controversy, to flee difficulty, to capture problem, to survive disaster/ situation/ difficulty/ tragedy/ scandal...*

Similarly, generalisation of the human solution to Example (1) in (1.2) now covers not only the Russian equivalent for публичный механизм ('public mechanism') ‘open process’, but also a wider range of phrases: {*открытый механизм ('open mechanism'), публичный процесс ('public process'), открытая система ('open system'), открытый способ ('open mode') ...*} mapped to a range of English solutions {*open structure , open operation, wide system, wide stage...*}. Each of these bilingual pairs of lexical items can become a focal point for a novel indirect translation equivalent that is not present in the parallel corpus used for training the system.

We represent these bilingual lexical patterns as a set of *contextual descriptors* – content words which are lexically central for a given context and can undergo indirect transformations in the process of translation. Fore example., the descriptor representation for Example (2) would be *{выходить ('exit') + кризис ('crisis')}* aligned with the target set of descriptors *{escape + crisis}*. For identifying contextual descriptors on the source and target side we use standard frequency-based methods of extracting discontinuous MWEs from monolingual corpora. Then the descriptors within indirect equivalents are automatically aligned using a combination of initial Giza++ alignment and the entropy-based approach.

In each pattern contextual descriptors on the source and target side are generalised using their distributional similarity classes. The viability of any combination of descriptors in source and target languages is tested against monolingual comparable corpora, as is done in ASSIST.

## 4. Experiments

The ASSIST technology proves that the range of indirect translation equivalents generated by using comparable corpora is larger than the range of equivalents generated from parallel corpora aligned by Giza++. The following table illustrates the numbers of indirect equivalents correctly matched to human solutions by ASSIST in two different corpora. The first one – the corpus of translated Russian and English newspapers (approx. 700k words) – was used for training Giza++. The other corpus – Euronews interviews (approx. 100k words) – was not seen by Giza++ and was used for testing. For evaluation of ASSIST we extracted about 400 problematic pairs of descriptors from the newspaper corpus and about 200 from the Euronews corpus. The technology showed superior performance on both sets of data (Table 1).

| | **Training-Giza** | **Test-Giza** |
|---|---:|---:|
| **Bilingual dict.** | 6.7% | 4.6% |
| **Giza++** | 13.9% | 3.4% |
| **ASSIST** | 21.9% | 19.5% |

Table 1: Coverage of indirect equivalents

We are aiming to improve the coverage figures for ASSIST with our new approach of generalising human solutions. Our evaluation efforts focus on improvements for incomprehensible translations and mistranslations, like those shown in examples above. Our initial case studies look promising, since human indirect solutions are generalisable within the proposed framework and generate a range of adequate and non-trivial translation equivalents e.g.:

*(Ex. 3)* ***MT****: Es verdad que empezamos vacilantes pero era lógico. (lit: started hesitant)*

> *(3.1)* ***human translation:*** *Of course we had our doubts to begin with but that's normal.*

Contextual descriptors: empezar vacilante ~ begin doubt

The automatically generated indirect equivalents on the English side are:

> *we had our fears/ doubts to start with*
> *we began with fear/ scepticism/ worries...*
> *we had our doubts/ suspicions early/ first*
> *we were not convinced then*
> *we worried first/ then*
> *after our early scepticism*

The system even generates descriptors which change lexical perspective (modulation):

> *we were soon/gradually/quickly convinced*

Each of these equivalents avoids the problem of incomprehensibility of SMT output caused by an unacceptable literal translation of *vacilantes*.

*(Ex. 4) Fr: Son équipe a découvert qu'il* **serait allé** *s'entraîner au Mexique.*

> *(4.1) his team found out that at the time he* **told** *them he was in training in Mexico he was seen by David Cassini, training in Italy*

Contextual descriptors: *serait allé ~ tell*

The automatically generated descriptors generalise and successfully convey counterfactual meaning:

> *his team was sure / understood / thought he was training...*
> *he explained / said to his team he was training*

It is interesting that the distributional model for the word *tell* predicts not only synonyms for its denotative meaning, but also cases when it is used to express facts that are contrary to someone's belief.

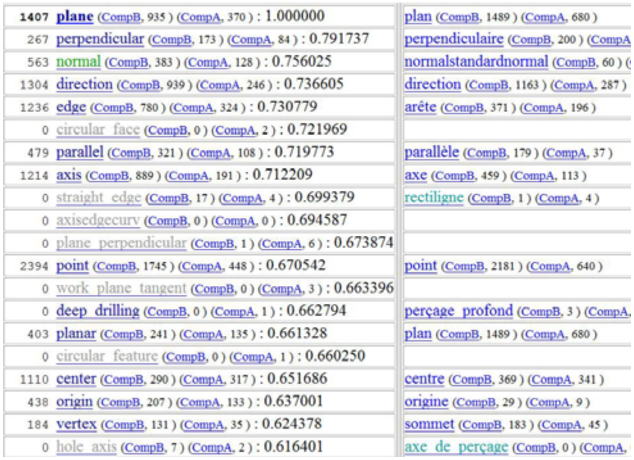
Our ongoing work on evaluating this approach includes the following stages. The system is trained on the Russian-English newspaper corpus (700k words), with Giza++ alignments and dictionary equivalents extended by distributional similarity classes. These classes are computed on comparable, non-parallel Russian and English corpora (each around 200M words). The Euronews Interviews corpus is used for testing. We automatically extract multiword expressions from the Euronews corpus using a part-of-speech filter and frequency threshold, as described in (Babych et al, 2007: 138). The corpus is also automatically aligned at the word level with Giza++. Then MWEs on the Russian and English side are matched using these alignments, and those which contain non-dictionary equivalents are selected as potential cases of indirect translation strategies, which are then manually checked to ascertain whether they really are indirect translation solutions. For evaluation we compute recall and average rank of translation equivalents found by the tested system which match these solutions.

## 5. Applications for terminology research

The proposed distributional similarity framework is designed primarily to deal with the general lexicon, not with terminology. The reason is that terms usually have fixed equivalents, and rarely change lexical or syntactic perspective. In other words, terminological equivalents are usually direct.

However, the framework can be useful for applications which automate terminological research for human translators. In particular, we designed a terminology exploration workbench, where for any given French or English term the tool generates lists of distributionally similar terms (based on co-occurrence in monolingual comparable corpora) with their translations, based on Giza++ alignments. We computed distributional similarity classes and translation equivalents for terminological multiword expressions using two aligned English-French corpora in a specialised domain (2M words in total) provided by two large industrial companies. Some multiword expressions in the corpora were bridged by term tables and some by the automatic Giza++ dictionary, which included MWEs.

The functionality of the tool is based on an observation that translators and interpreters prepare for their work within some domain by reading texts originally written in the target language for that job, looking not just for individual terms, but for a system of related terms used in that field. Our tool partially automates this process by presenting translators with a searchable network of related terms and their translation equivalents, which can be looked up directly or followed via hyperlinks. Figure 1 presents the user interface of the tool, with the query for the English term *plane*. The left column shows the similarity class for the query term, while the right column shows the French translation equivalents.



| | | |
|---|---|---|
| 1407 | plane (CompB, 935 ) (CompA, 370 ) : 1.000000 | plan (CompB, 1489 ) (CompA, 680 ) |
| 267 | perpendicular (CompB, 173 ) (CompA, 84 ) : 0.791737 | perpendiculaire (CompB, 200 ) (CompA |
| 563 | normal (CompB, 383 ) (CompA, 128 ) : 0.756025 | normalstandardnormal (CompB, 60 ) ( |
| 1304 | direction (CompB, 939 ) (CompA, 246 ) : 0.736605 | direction (CompB, 1163 ) (CompA, 287 ) |
| 1236 | edge (CompB, 780 ) (CompA, 324 ) : 0.730779 | arête (CompB, 371 ) (CompA, 196 ) |
| 0 | circular face (CompB, 0 ) (CompA, 2 ) : 0.721969 | |
| 479 | parallel (CompB, 321 ) (CompA, 108 ) : 0.719773 | parallèle (CompB, 179 ) (CompA, 37 ) |
| 1214 | axis (CompB, 889 ) (CompA, 191 ) : 0.712209 | axe (CompB, 459 ) (CompA, 113 ) |
| 0 | straight edge (CompB, 17 ) (CompA, 4 ) : 0.699379 | rectiligne (CompB, 1 ) (CompA, 4 ) |
| 0 | axisedgecurv (CompB, 0 ) (CompA, 0 ) : 0.694587 | |
| 0 | plane perpendicular (CompB, 1 ) (CompA, 6 ) : 0.673874 | |
| 2394 | point (CompB, 1745 ) (CompA, 448 ) : 0.670542 | point (CompB, 2181 ) (CompA, 640 ) |
| 0 | work plane tangent (CompB, 0 ) (CompA, 3 ) : 0.663396 | |
| 0 | deep drilling (CompB, 0 ) (CompA, 1 ) : 0.662794 | perçage profond (CompB, 3 ) (CompA, |
| 403 | planar (CompB, 241 ) (CompA, 135 ) : 0.661328 | plan (CompB, 1489 ) (CompA, 680 ) |
| 0 | circular feature (CompB, 0 ) (CompA, 1 ) : 0.660250 | |
| 1110 | center (CompB, 290 ) (CompA, 317 ) : 0.651686 | centre (CompB, 369 ) (CompA, 341 ) |
| 438 | origin (CompB, 207 ) (CompA, 133 ) : 0.637001 | origine (CompB, 29 ) (CompA, 9 ) |
| 184 | vertex (CompB, 131 ) (CompA, 35 ) : 0.624378 | sommet (CompB, 183 ) (CompA, 45 ) |
| 0 | hole axis (CompB, 7 ) (CompA, 2 ) : 0.616401 | axe de perçage (CompB, 0 ) (CompA, |

Figure 1. Terminology exploration interface

A further challenge for our approach is to use it within the MT framework, alongside the described distributional similarity model for the general lexicon.

## 6. Conclusions and future work

The presented methodology uses indirect translation strategies to generate previously unseen translation equivalents. The strategies are learnt from parallel corpora and are generalised using data from much larger comparable corpora. The integration of this methodology into data-driven MT architectures looks promising; it has the potential to improve the quality and comprehensibility of MT output and, in certain cases, to avoid mistranslations. Our model makes testable predictions about whether a particular type of translation strategy finds indirect solutions produced by human translators. Future work will include automatic identification of phrases which need non-literal translation and using the contextual descriptors as a lexical core to automatically build proper translation equivalents around them, e.g., using a freely-available SMT decoder such as Moses. Finally we will evaluate the improvement in coverage of human solutions and the degree of reusability of automatically discovered lexical translation strategies.

## Acknowledgements

The work is partially supported by the Leverhulme Trust. We thank Francis Tyers for his work on extending ASSIST for terminology exploration.

## References

Babych, Bogdan, Sharoff, Serge, Hartley, Anthony and Mudraya, Olga. (2007). Assisting Translators in Indirect Lexical Transfer. In *ACL 2007: the 45th Annual Meeting of the Association of Computational Linguistics*.

Collins, Brona and Somers, Harold. (2003). EBMT seen as case-based reasoning. In *Recent advances in Example-Based Machine Translation*. Carl, Michael and Way, Andy. pp. 115-153.

Fung, Pascale. (1998) A Statistical View on Bilingual Lexicon Extraction: From Parallel Corpora to Non-parallel Corpora. In *Proceedings of the 3rd Conference of the Association for Machine Translation in the Americas (AMTA'98)*.

Koehn, Philipp (2005) Europarl: A Parallel Corpus for Statistical Machine Translation, In *Proceedings of MT Summit X*.

Moiron, Begona Villada and Tiedemann, Joerg. (2006). Identifying idiomatic expressions using automatic word-alignment. In *Proceedings of the workshop on Multi-word-expressions in a multilingual context at EACL06*.

Morin, Emmanuel and Daille, Béatrice and Takeuchi, Koichi and Kageura, Kyo. (2007). Bilingual Terminology Mining - Using Brain, not brawn comparable corpora. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. 2007.

Munday, Jeremy. (2001). *Introducing translation studies*. Routledge.

Munteanu, Dragos Stefan and Marcu, Daniel. (2006). Extracting Parallel Sub-Sentential Fragments from Non-Parallel Corpora. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*.

Newmark, Peter. (1988). *A textbook of translation*. Prentice Hall, New York, London

Rapp, Reinhard (1995) Identifying word translations in non-parallel texts, *Proc. the 33rd ACL*, June 26-30, 1995, Cambridge, Massachusetts

Rapp Reinhard. (2004). A freely available automatically generated thesaurus of related words. In *LREC 2004*.

Sharoff, S., (2006) A Uniform Interface to Large-Scale Linguistic Resources. In *Proc. of LREC2006*, Genoa, May, 2006.

Sharoff, S., Babych, B., Hartley, A. (2008) 'Irrefragable answers' using comparable corpora to retrieve translation equivalents. In *Language Resources and Evaluation Journal*

Vinay, J-P and Darbelnet, J. (1995) *Comparative stylisitcs of French and English*. Trans. J-C Sager and M-J Hamel. Amsterdam: John Benjamins.