

# Removing the Distinction Between a Translation Memory, a Bilingual Dictionary and a Parallel Corpus

Vincent Vandeghinste  
Centre for Computational Linguistics  
Katholieke Universiteit Leuven  
vincent.vandeghinste@ccl.kuleuven.be

## **Abstract**

This paper presents a prototype MT system which does not make the distinction between a dictionary, a sub-sentential aligned parallel corpus, and post-edited information (translators output) like a translation memory. The system is based on the METIS-approach (Vandeghinste et al, 2006), and uses an XML-based dictionary format in which not only simple word-to-word translations can be included, but which also contains complex dictionary entries, including discontinuous entries, like idioms and proverbs. The presented prototype is a system that automatically adapts its dictionary and target language corpus depending on the post-edited output as made by the users of the system, and will therefore have a learning curve in its performance.

# 1 Introduction

In machine translation (MT) research, traditionally a distinction is made between three paradigms: rule-based machine translation (RBMT), statistical machine translation (SMT), and example-based machine translation (EBMT).

RBMT relies on a large set of hand-crafted rules, which makes the development of new language pairs very costly, and improving existing systems becomes a tedious task, as these rule sets can become very large and complex. An RBMT system makes use of a dictionary to bridge the gap between the source language lexical items and the target language, together with a number of transfer rules or a complex interlingual representation, to map the source language sentence structure onto the target language sentence structure.

EBMT and SMT are data-driven approaches, and claim to be much faster in development, as they do not rely on hand-crafted rules. Instead, these approaches rely on *large* parallel corpora, which therefore become the bottleneck in developing new language pairs or new technical domains, as they are often unavailable, or not large enough. A difficult issue in working with parallel corpora is the alignment of the source and the target language. Whereas sentential alignment does not seem to pose too many problems, alignment within a sentence is a much more difficult task.

The ideas behind EBMT are often linked to ideas about translation memories. An EBMT system tries to match its input sentence with the source language side of a parallel corpus, and the aligned target language side of the corpus will generate the target language sentence. The difficulty for such systems lies in how this matching process is done in the case of partial matches, and how to solve overlapping partial matches and recombine the target side of the mapping fragments.

This is not the case for SMT, in which the generated sentence will be based on

statistics derived from the aligned parallel corpus, making abstraction of the cases which are contained in the parallel corpus.

In recent years, hybrid machine translation systems have been starting to emerge. Within the METIS-II-consortium the idea arose to avoid some of the problematic issues of the previous approaches, and develop a prototype for a new translation method (Dologlou et al., 2003; Dirix et al., 2005; Vandeghinste et al., 2006), which relies heavily on the target language generation side, combining techniques from RBMT, SMT, and EBMT. The system was implemented for four language pairs: Dutch to English, Modern Greek to English, Spanish to English, and German to English. A more detailed description is given in section 2.

Another hybrid machine translation system is the Matador system (Habash and Dorr, 2002; Habash, 2003, 2004), which is somewhat similar to the METIS-II approach, in that it does not require parallel data. It translates from Spanish to English, and relies heavily on target language generation. It is aimed at language pairs lacking resource symmetry. It employs symbolic and statistical target language resources, and requires a source language parser and a translation dictionary, but no transfer rules or complex interlingual representation. On the target side, rich symbolic resources like lexical semantics, categorial variations and subcategorization frames are used to overgenerate multiple structural variations from a syntactic dependency representation of the source language sentence, where all terminal nodes are translated by the dictionary. The overgeneration is constrained by several statistical target language models, including surface n-grams and structural n-grams.

*Context-based Machine Translation* (CBMT) as described by Carbonell et al. (2006) is another approach somewhat similar to the METIS approach. It does not require parallel corpora either and relies heavily on the target language side. It has been implemented for Spanish to English translation, and requires an extensive target language corpus, and a full-form dictionary. It does not contain transfer rules

or interlingual representations, but instead relies on long n-grams. The principle is to produce many long n-gram candidate translations by finding those long n-grams that contain as many as possible of the potential word and phrase translations from the dictionary, and as few as possible other content words. These n-grams are matched with the target language corpus, and the highest scoring translation candidate is selected by the decoder. While this is in se a statistical approach, the general idea behind it is not within the classical SMT paradigm, but justifies its classification together with the METIS system and the MATADOR system.

A somewhat different approach using ideas from both SMT, RBMT, and EBMT is called *Data-oriented Translation* (DOT), which was first proposed by Poutsma (1998), and the first large scale implementation of this approach was done by Hearne (2005). DOT still requires parallel data, but this time, it concerns parallel treebanks, in which alignments have been made on several levels in the trees. By using linguistically motivated trees, combined with using translation examples, and statistical techniques like data oriented parsing (Scha, 1990; Bod, 1992), this approach borrows from the three different MT paradigms.

The prototype presented in this paper is based on the METIS-II approach, which will be described in the next section.

## **2 The METIS-II approach**

The aim of the METIS-II approach is to allow development of MT systems for low resource languages. Therefore, we restricted ourselves to using only limited tools, which are available for lots of languages, or which can be easily adapted to the languages in focus.

The METIS-II approach also tries addressing some of the weak points of the classic approaches. Only a limited set of rules is used, and no parallel data, ex-

cept for the dictionary. All entries in the dictionary are lemmas, which is a useful abstraction over word forms, as it reduces dictionary size and data sparsity.

The METIS-II system is a hybrid system based on the ideas of EBMT systems, but without using a parallel corpus.

In a first step, shallow source language analysis is applied. The sentence is tokenized, tagged, lemmatized, and chunked. This results in a shallow parse tree.

Then the lexical entries in the sentence are looked up in a bilingual dictionary. Special care is taken in the lookup of complex entries, which might be discontinuous. Separable verbs are also looked up (which is not only relevant for Dutch, but also for German). In this process, only lemmas and part of speech tags are used.

Through a limited (<50) number of transfer rules, the sentence structure is mapped onto the target language, generating a number of translation candidates.

These candidate translations are weighed by matching them with the target language corpus. For METIS-II, we used the British National Corpus. First we try to find matches for the lowest level chunks. The corpus lookup provides us with information about lexical selection and word order.

Each chunk from the source language tree is considered a *bag*: an unordered list. We retrieve from the corpus all chunks with the same chunk type containing as many lemmas as possible from the bag. According to how well these chunks match the bag, we give them a weight. By considering several translation alternatives and matching them with the corpus, different alternatives get different weightings, allowing us to select the one with the highest weight. Word order is determined by the matched corpus chunks. For words that are in the bag but not in the corpus chunk, we look for a slot in the corpus chunk with the same part of speech, and replace that word with the word from the bag.

At higher levels in the shallow parse tree, we use the heads of lower level chunks, which should have been resolved at this point.

Matching the sentence with a monolingual target language corpus has the advantage of not needing a parallel corpus, which is a scarce resource for most language pairs. Whereas the difficulty of EBMT systems is to find matching source language fragments, this is moved to finding matching corpus segments in the target language, based on the translated words in the dictionary. The difficult issue of sub-sentential alignment is avoided.

For a detailed description of the METIS-II approach, take a look at Dirix et al. (2006) for the description of the Dutch to English approach. Other language pairs developed in METIS-II are German to English (Carl et al., 2007), Spanish to English (Badia et al., 2005), and Greek to English (Markantonatou et al., 2006).

In this paper we describe a new prototype, which is based on the METIS-II paradigm, but whereas METIS-II was restricted to using only limited resources, as it was developed as a new methodology for MT systems for lesser resourced languages, this new system will use all resources that are available for the language pair at hand.

Another difference with the METIS-II system lies in the integration of a post-editing interface with the system, such that each of the human-edited translations is fed back into the system, using that information for future translations.

### **3 The new prototype**

In the prototype which we describe in this paper, we want to scale up the METIS-II approach. We will not limit ourselves anymore to using only linguistic resources. This time, we will use all resources we can get.

We will still use the METIS-II approach, but add a lot of information coming from parallel aligned treebanks, in a similar way as the DOT approach described earlier.

In figure 1, we show the general architecture of the new prototype, which is called PaCo-MT (Parse and Corpus based MT).

This prototype is built for the language pairs Dutch-English and Dutch-French, in both directions. In the rest of this paper, we will only focus on the Dutch-English language pair, but the same principles are applied when translating from Dutch into French or vice versa.

When a sentence is entered into the system, this sentence goes through a source language parser, resulting in a full parse tree analysis of the source language.

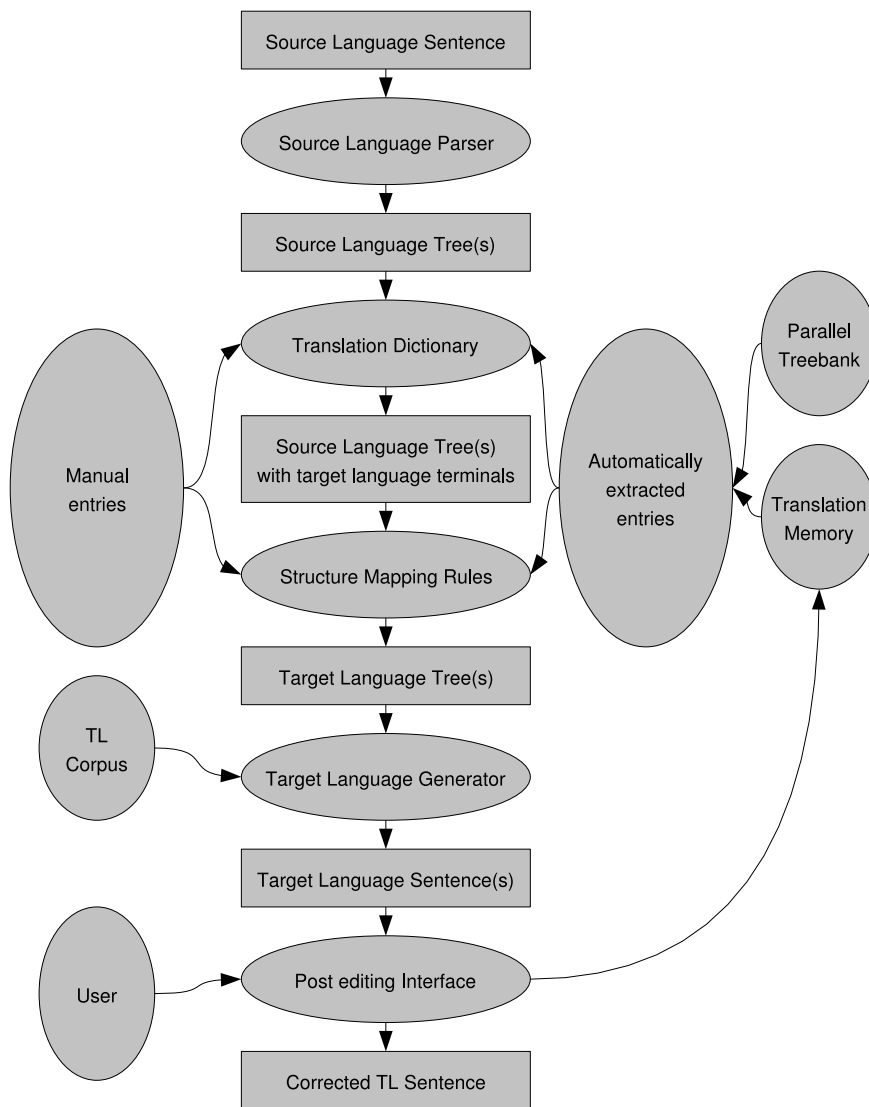
The Alpino parser (Van der Beek et al., 2005; Van Noord, 2006) is used when translating from Dutch. It is chosen because it is a wide-coverage parser with a high accuracy and it is freely available.

The PET parser (Callmeier, 2000) combined with the ERG-grammar (Copestake and Flickinger, 2002) will be used when translating from English. The PET parser is chosen because it is optimized for speed, and they are both freely available. Besides this, they are based on the HPSG paradigm (Pollard and Sag, 1987, 1994), and so is Alpino, which will result in comparable structures for both source language and target language trees.

To go from source to target language we use two paths:

1. We make use of a translation dictionary (containing words, phrases, clauses, and full sentences), which is based on two sources: manual entries and automatically extracted information coming from parallel corpora and translation memories. We use a Dutch-English dictionary from the METIS-II project. We are collecting parallel corpora like Europarl (Koehn, 2005), Acquis Communautaire (Steinberger et al., 2006), and the Dutch Parallel Corpus (which will be available soon). A translation company provides us with real translation memories and translated texts.

Figure 1: Architecture of the prototype



We apply sub-sentential alignment (Gildea, 2003; Tiedemann, 2003) on these parallel data so we can automatically extract dictionary entries at word, phrase, clause and sentence level, which will be added to the already existing dictionaries.

2. We make use of structure mapping rules, which are also based on two sources: manual entries (avoiding the black-box of statistical MT systems) and automatically extracted information coming from parallel corpora and translation memories, as is done in Lavoie et al. (2002), Probst et al., (2002), and Quirk et al. (2005). We generate parallel parse forests, and we automatically extract structure mapping rules, that describe the transformation from the source language structure onto the target language structure. These automatically extracted rules can be augmented with manually defined rules to address remaining translation issues.

When a source language sentence is analysed, this can result in several parse trees. We match these trees and their subtrees with the source language side of the dictionary / parallel corpus / translation memory. By abstracting over some features, and by allowing partial matches, replacements, and insertions / deletions in the matching element, this part of the system functions as an intelligent innovative translation memory, that not only matches sentences or parts of sentences in a parallel corpus, but is also able to combine the outcome of the parallel data with the machine translation engine for parts of the sentence which are not matched in the parallel data. We allow separate parallel corpora and translation memories to be activated, depending on the user profile, such that translations corrected by one user can be kept separate from corrections by another user.

The bilingual dictionary can be considered a parallel corpus available to all users, and mainly contains single words. Full phrases, multi-words and discon-

tinuous dictionary entries are allowed as well. Translations leading to structural changes in the dependency tree can be coded through dictionary entries.

At this point in the processing procedure, we have an intermediate tree representation in which all leaf nodes (or sometimes higher level nodes) are translated into the target language. Parts of this tree structure are already in the target language structure, as they result from the target side of the parallel corpus or the dictionary, while other parts of the tree structure are still in the source language structure. These parts should be converted into the target language structure, through the structure mapping rules.

Alternative target language trees are generated in the previous processing steps, each with a weight representing the confidence. These weights are adapted through information gathered from the monolingual target language corpora: how well does the tree fit the target language?

The target language corpus is preprocessed with the respective parser and grammar for that language, resulting in a target language treebank. For Dutch as target language, we use the Lassy treebank (van Noord et al., 2006) and the Alpino treebank, (<http://www.let.rug.nl/~vannoord/trees/>) which are both publicly available. For English as TL, we use the Redwoods treebank (Oepen et al., 2002). We extend these treebanks with the respective sides of the parallel data, and will possibly extend them with more automatically annotated monolingual treebanks, as the need arises.

Lexical selection amongst several translation alternatives is based on co-occurrence metrics (Dunning, 1993; Church and Hanks, 1990; Evert 2004; Evert and Krenn, 2004), and frequency metrics taking into account the syntactic environment of the word (which are similar to what we already did in Dirix et al. (2006)). This allows us to decide which of the translation alternatives for e.g. an adjective are most likely to go together with a specific noun, etc.

Target language generation needs to be performed based on the obtained target language trees. In the target language corpus database, we store (sub-)tree structure patterns combined with surface string information like word order for these trees, allowing us to generate target language word order as derived from the target language corpus. The better a tree matches a tree in the target language corpus, the higher the weight this translation will get, in the list of generated translations. This is a refined version of the target language generation component in Dirix et al. (2006).

## **4 Human Post-Editing**

The output of our system is sent to a post-editing interface, in which the human post-editor can adapt the translation:

- The post-editor can choose another translation candidate, be it on the sentence level or on any lower levels in the tree
- The post-editor can make changes to the text, by simple typing
- The post-editor can move words or phrases

Because the sentence was automatically generated, and by tracking the post-editor's changes to the sentence, we have a number of ways in improving our system automatically. The newly generated parallel sentence, which is aligned at sub-sentential level can be fed back into our translation dictionary. Like in a traditional translation memory, this sentence will now be automatically generated when the same input sentence is given. The different phrases of this sentence will also be put into the dictionary. As the sentence was automatically generated, we have a detailed sub-sentential alignment which allows this. Apart from that, the corrected

target language sentence will also be added to the target language treebank, so this information becomes available when trying to match future similar sentences.

We can also adapt the weights in our dictionary for the current user / text to improve consistent translation. The automatically extracted transfer rules (or their weights) can also be updated as a consequence of human post-editing, in a similar way as Font Llitjós et al. (2007).

## 5 Removing the Distinction Between a Translation Memory, a Dictionary, and a Parallel Corpus

In the METIS-II system, we started the development of a translation dictionary format in which we could not only represent single word entries, but also complex entries leading to structural changes in the target language. For this purpose, we use XML.

In example 1 you can see how a single word entry, with only one translation alternative looks in our dictionary.

```
(1) <dict-entry id="19">
    <source>
        <token id="1" pos="ADJ" lemma="blauw"/>
    </source>
    <target>
        <trans-unit id="1">
            <token id="1" pos="AJ?" link="1" lemma="blue"/>
        </trans-unit>
    </target>
</dict-entry>
```

Each dictionary entry consists of two main parts. In the `<source>` part, the source language side of the entry is described, through one or more `<token>`-tags, which can be grouped to higher level linguistic units in `<chunk>`-tags.

In the `<target>` part, the target language side of the entry is described, consisting of one or more `<trans-unit>`s, each representing a translation alternative for the source language token(s).

In the `<token>` entries, you can see a `pos` feature. On the source side this contains the restrictions to which the entry must comply in order to generate the translation candidates. On the target side, this contains part of speech information which applies to the token. In the source language tag set (Van Eynde, 2005), features are represented between brackets. For instance, a singular (`ev`) non-diminutive (`basis`) common (`soort`) noun (`N`) in standard case (`stan`) gets the tag `N(soort, ev, basis, zijd, stan)`. When no brackets are used, this indicates that there are no restrictions on the features of the source side.

On the target side, question marks are used for underspecification, as the CLAWS5 tagset uses the third character to represent the features. For instance, `NN1` represents singular nouns, whereas `NN2` represents plural nouns.

Underspecifications are due to the fact that we use lemmas, which are underspecified for features like number, case, etc.

The `link` feature is used on the target side only, to indicate which part of the target side is a translation of which part in the source side.

For the translation of more complex entries, for instance, the translation of Dutch *'s morgens* into English *in the morning* things are a bit more complicated. As shown in (2), the Dutch phrase is an NP chunk, which is represented by the `<chunk>`-tag. This tag contains the feature `ref` to indicate which `<token>`s belong to the chunk. This Dutch NP is translated into an English PP, consisting of the preposition *in* and the NP, consisting of the tokens *the* and *morning*. To

indicate that the target language PP is a translation of the source language NP, the `link` feature in the target language PP refers to the source language NP. Note that the source language `pos` features contain the restriction that they need to be in *genitive* case, which avoids translating Dutch *de morgen* (in standard case: *the morning*) into English *in the morning* which would be incorrect.

```
(2) <dict-entry id="4">
    <source>
        <token id="1" pos="LID(gen)" lemma="de"/>
        <token id="2" pos="N(gen)" lemma="morgen"/>
        <chunk id="c1" ref="1-2" label="NP" head="2"/>
    </source>
    <target>
        <trans-unit id="1">
            <token id="1" pos="PRP" lemma="in"/>
            <token id="2" pos="AT0" lemma="the"/>
            <token id="3" pos="NN1" lemma="morning"/>
            <chunk id="c1" ref="1-c2" label="PP" link="c1"
head="1"/>
            <chunk id="c2" ref="2-3" label="NP" head="3"/>
        </trans-unit>
    </target>
</dict-entry>
```

Thinking about how to represent this structural change in the METIS-II dictionary, resulted in a format in which we represent both sides of an entry as a tree. The link values allow us to represent lower level alignment in these trees.

If our dictionary can be used to map a source language tree onto a target lan-

guage tree, which is what we do, then it would also be possible to use the dictionary to map full source language sentences in much the same way onto full target language sentences, which is necessary when translating idiomatic phrases or proverbs.

And, if we can translate full sentences using the dictionary, why would we not add human post-edited sentences to this dictionary. For these sentences, we have lower level alignments, as they were originally generated by our system. When the post-editor does not make any typing changes, but merely changes the selected translation candidates and moves around words or phrases, this does not pose any problems.

We not only add these full sentences to the dictionary, but also all their aligned parts. When some parts are already in the dictionary, we update the weight for the selection as made by the post-editor. This ensures that the system will learn immediately from the post-editor's behaviour.

Because of the fact that our dictionary grows fast, we need to have a fast way of matching our sentence with the source language side of the dictionary. We can use classic EBMT methods for this.

And when we can use aligned post-editing information in the improvement of our system, there is no reason why we cannot import already existing translation memories or parallel corpora, hence removing the distinction between a dictionary, a translation memory, and a parallel corpus. For this, we would need to parse both sides of the translation memory or parallel corpus and align them at a sub-sentential level.

## 6 Conclusions

In the METIS-II system, the only parallel data which is used is a bilingual dictionary. Since we wanted to be able to model complex dictionary entries which lead to structural changes in the tree representation of the sentence under translation, we set up an XML dictionary in which we can map source language trees onto target language trees. While this was initially only intended for use with idioms and proverbs, there was no principle reason why we could not use this set up in much the same way as a traditional translation memory.

In the Paco-MT system, which is still in its development phase, we are no longer tied to the METIS-II restriction of using low resources. Therefore we can use existing, available data, and the most logical spot to incorporate this data in our system is in the dictionary.

We also wanted an adaptive system which learns from interaction with the human post-editor. Apart from a growing target language corpus (a by-product of the corrected sentences which are added to this corpus), the amount of parallel data grows as well. Since this parallel data is based on our MT output, it is aligned with the source language. By immediately feeding back this information into the dictionary, the system will learn immediately.

In the way that we represent all parallel data in what was originally our dictionary, we have removed the distinction between a translation memory, an aligned parallel corpus, and a traditional bilingual dictionary.

## References

- Badia, T., Boleda, G., Melero, M., Oliver, A. (2005). An n-gram Approach to Exploiting a Monolingual Corpus for Machine Translation. In *Proceedings of the 2nd Workshop on Example-based Machine Translation*, Phuket, Thai-

land. pp. 1-7.

Bod, R. (1992). A Computational Model of Language Performance: Data Oriented Parsing. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING'92)*. pp. 855-859. Nantes, France.

Callmeier, U. (2000). PET. A platform for experimentation with efficient HPSG processing techniques. *Journal of Natural Language Engineering*, vol. 6(1). pp.99-107.

Carbonell, J., Klein, S., Miller, D., Steinbaum, M., Grassiany, T., and Frei, J. (2006). Context-based machine translation. *AMTA 2006: Proceedings of the 7th Conference of the Association for Machine Translation in the America, "Visions for the Future of Machine Translation"*. pp.19-28. Cambridge, Massachusetts.

Carl, M., (2007). METIS-II. The German to English MT System. In *Proceedings of the 11th Machine Translation Summit*. pp. 65-72, Copenhagen, Denmark.

Church, K., Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, vol. 16(1).

Copestake, A., Flickinger, D. (2000). An open source grammar development and broad-coverage English grammar using HPSG. *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC)* pp. 591-598. Athens, Greece.

Dirix, P., Schuurman, I., and Vandeghinste, V. (2005). METIS-II: Example-based machine translation using monolingual corpora - System description. In *Proceedings of the 2nd Workshop on Example-based Machine Translation*, Phuket, Thailand.

- Dirix, P., Vandeghinste, V., Schuurman, I. (2006). A new hybrid approach enabling MT for languages with little resources. In *Proceedings of the 16th meeting of Computational Linguistics In the Netherlands. CLIN-2005*, Amsterdam.
- Dologlou, Y., Markantonatou, S., Tambouratzis, G., Yannoutsou, O., Fourla, A., and Ioannou, N. (2003). Using Monolingual Corpora for Statistical Machine Translation: The METIS System. In *Proceedings of EAMT - CLAW 2003*, Dublin, pp. 61-68.
- Dunning, T. (1993). Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*.
- Evert, S. (2004). *The Statistics of Word Cooccurrences: Word pairs and collocations*. PhD dissertation. University of Stuttgart.
- Evert, S., Krenn, B. (2004). Methods for the qualitative evaluation of lexical association measures. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*. pp. 188-195. Toulouse, France.
- Font Llitjós, A., Ridmann, W.A. (2007). The Inner Works of an Automatic Rule Refiner for Machine Translation. In F. Van Eynde, V. Vandeghinste and I. Schuurman (eds.) *New Approaches to Machine Translation: METIS-II Workshop*. pp. 47-56. Centre for Computational Linguistics. K.U.Leuven.
- Gildea, D. (2003). Loosely Tree-Based Alignment for Machine Translation. In *Proceedings of the 41st Annual Meeting of the Association of Computational Linguistics (ACL)*. pp. 80-87. Sapporo. Japan.

- Habash, N. (2003). Matador: A Large-Scale Spanish-English GHMT System. In *Proceedings of the MT Summit IX*, New Orleans.
- Habash, N. (2004). The Use of a Structural N-gram Language Model in Generation-Heavy Hybrid Machine Translation. In *Proceedings of the Third International Conference on Natural Language Generation (INLG04)*. Birghton.
- Habash, N., and Dorr, B. (2002). Handling translation divergences: Combining statistical and symbolic techniques in generation-heavy machine translation. In *Proceedings of the 5th Conference of the Association for Machine Translation in the Americas: Machine Translation: From Research to Real Users*, London, UK. Springer-Verlag.
- Hearne, M. (2005). *Data-Oriented Models of Parsing and Translation*. PhD Thesis. Dublin City University.
- Lavoie, B., White, M., Korelsky, T. (2002). *Learning Domain-specific Transfer Rules: An Experiment with Korean to English Translation*.
- Koehn, P. (2005). Europarl: a parallel corpus for statistical machine translation. *Proceedings of MT Summit X*. Phuket, Thailand.
- Markantonatou, S., Sofianopoulos, S., Splioti, V., Vassiliou, M., Yannoutsou, O. (2007). An MT System Embedding Pattern Knowledge. In: Van Eynde, Schuurman, and Vandeghinste (eds.) *New Approaches To Machine Translation: Proceedings of the METIS-II Workshop*. Leuven. pp. 11-18.
- Pollard, C. and Sag, I. (1987). *Information-based Syntax and Semantics*. Stanford. CSLI.
- Pollard, C. and Sag, I. (1994). *Head-driven Phrase Structure Grammar*. Stanford/Chicago: CSLI and University of Chicago Press.

- Poutsma, A. (1998). Data-Oriented Translation. In *Ninth Conference of Computational Linguistics*. Leuven. Belgium.
- Probst, K., Levin, L., Peterson, E., Lavie, A., Carbonell, J. (2002). MT for Minority Languages Using Elicitation-Based Learning of Syntactic Transfer Rules. In *Machine Translation*, vol. 17(4). pp. 245-270.
- Quirk, C., Menezes, A., Cherry, C. (2005). Dependency Treelet Translation: Syntactically Informed Phrasal SMT. In *Proceedings the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*. pp.271-279. Ann Arbor. USA.
- Scha, R. (1990). Language Theory and Language Technology: Competence and Performance. *Computertoepassingen in de Neerlandistiek*. pp. 7-22.
- Steinberger R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufis, D., Varga D. (2006). The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*. pp.2142-2147. Genova, Italy.
- Tiedemann, J. (2003). *Recycling Translations - Extraction of Lexical Data from Parallel Corpora and their Application in Natural Language Processing*, Doctoral Thesis, Studia Linguistica Upsaliensia 1.
- Vandeghinste, V., Schuurman, I., Markantonatou, S., Carl, M., Badia, T. (2006). Machine Translation for Low Resource Languages. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*. pp.1284-1289. Genova, Italy.
- Van der Beek, L., Bouma, G., Daciuk, J., Gaustad, T., Malouf, R., Nederhof, M.-J., Van Noord, G., Prins, R., Villada, B. (2005). *Algorithms for Linguistic*

*Processing*. NWO Pionier. Final Report.

Van Eynde, F. (2005). *Part-of-speech tagging en lemmatisering in D-coi*. Centre for Computational Linguistics. K.U.Leuven.

Van Noord, G. (2006). At Last Parsing Is Now Operational. In: Mertens, P., Fairon, C., Dister, A., Watrin, P. (eds.). *TALN06. Verbum Ex Machina. Actes de la 13e conférence sur le traitement automatique des langues naturelles*. pp. 20-42. Leuven.

Van Noord, G., Schuurman, I., Vandeghinste, V. (2006). Syntactic Annotation of Large Corpora in STEVIN. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*. pp.1811-1814. Genova, Italy