METIS-II: Machine Translation for Low Resource Languages

Vincent Vandeghinste*, Ineke Schuurman*, Michael Carl†, Stella Markantonatou‡, Toni Badia◇

*CCL - Centre for Computational Linguistics, K.U.Leuven
Belgium
vincent, ineke@ccl.kuleuven.be
†IAI - Institut der Gesellschaft zur Förderung der Angewandten Informationsforschung e.V.
an der Universität des Saarlandes - Germany
carl@iai.uni-sb.de
‡ILSP - Institute for Language and Speech Processing
Greece
marks@ilsp.gr
◇UPF - Universitat Pompeu Fabra
Spain
toni.badia@upf.edu

Abstract

In this paper we describe a machine translation prototype in which we use only minimal resources for both the source and the target language. A shallow source language analysis, combined with a translation dictionary and a mapping system of source language phenomena into the target language and a target language corpus for generation are all the resources needed in the described system. Several approaches are presented.

1. Introduction

METIS-II (10-2004 – 09-2007) is a hybrid machine translation system, in which insights from Statistical, Example-based, and Rule-based Machine Translation (SMT, EBMT, and RBMT respectively) are used.

In the European context, the importance of supporting and maintaining a multilingual society is apparent. Machine translation (MT) should be considered a central activity in maintaining such a society. In current NLP technology, however, multilinguality relies heavily on expensive resources, such as large parallel corpora and expensive tools such as parsers and semantic taggers. Consequently, the number of languages that have such advanced technology at their disposal is small. Regarding resources for MT, this diagnosis is correct for many European languages and language pairs, especially those between the smaller languages.

While industrial technologies are mainly rule-based, current research is mainly on data-driven methods (like SMT and EBMT). Both SMT and EBMT systems rely on parallel corpora, and the development of a RBMT system is a tedious and very expensive undertaking. We are looking for a low-cost solution, so we did not consider pure RBMT. Both purely statistical and purely rule-based approaches each have their intrinsic obstacles (other than parallel corpora resp. cost factor mentioned above), cf also Thurmair (2005), suggesting that a hybrid approach is the way to go.

RBMT is only expensive if you try to model fine grained distinctions. Taggers and shallow rule-based parsers are relatively easy to obtain. Similarly many SMT and EBMT approaches are hard tasks since sufficient parallel material is needed to model the whole translation process. On the other hand, monolingual texts are easy to obtain and useful inferences can be drawn that are also helpful for translation.

METIS investigates rule-based and data-driven methods to the extent they can be built and used with relative ease and they complement each other.

Rule-based methods are used where representations and decisions can be determined a-priori with high accuracy, for instance, based on linguistic insight. Corpora serve as a basis to ground decisions where uncertainty remains.

Data-driven methods are used for target language generation, using only a target language corpus and a bilingual dictionary instead of a parallel corpus.

The lack of sophisticated linguistic resources other than parallel corpora for many of the smaller European languages made us develop a scalable system, in which we can refrain from the use of those complex resources. We mainly use basic tools like taggers and chunkers in both source language (SL) input text and target language (TL) corpus. This way, an MT system like METIS might in the end be useful for all European languages. Thus, although for the languages involved (source: Dutch, German, Greek, and Spanish; target: English) more advanced, expensive tools and resources are available, we do not need to use them. The METIS system is scalable and can be upgraded to different degrees of representational richness.

2. Usage of basic resources

The main goal of METIS is to build a translation system without parallel corpora and without an extensive rule-set.

In this section we describe the resources we use for the respective language pairs in the METIS system. Not all language pairs use the same resources, and this shows that the system can be used with a variety of resources, depending on the availability for the languages at hand.

Figure 1 shows the general system flow and which resources are or can be used at which stage in the translation process.

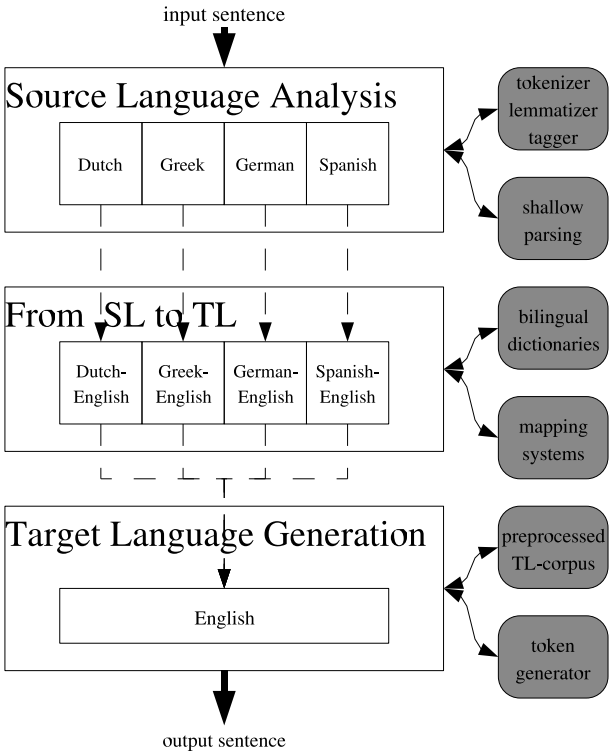

Figure 1: General System Flow and Used Resources

There are a number of structural translation problems, which need to be solved. A non-exhaustive list of some phenomena includes:

Word order issues : English (ENG) has a relatively fixed word order, while German (GER), Greek (GRE) and Dutch (DUT) have a relatively free word order. Spanish (SPA) has a clearly different NP word order.

GRE: all orders below (VSO, VOS and SVO) are acceptable and umarked.
kinigai i alepou tin kota ≈ kiningai tin kota i alepou ≈
i alepou kinigai tin kota
ENG: the fox chases the hen.

SPA: el lápiz azul [the pencil blue]
ENG: the blue pencil

Lexical issues :

- Different complementation:
  GER: Er wartet auf ihn.
  DUT: Hij wacht op hem. (He waits on him)
  ENG: He is waiting for him.

  GRE: mpainv sto domatio (enter-1st in-the room)
  ENG: I enter the room.

  SPA: Juan vio a María (John saw to Mary)
  ENG: John saw Mary

- Intensifying / other collocations:
  GER: harte Kritik ENG: harsh criticism
  GER: starker Raucher ENG: heavy smoker

Syntactical issues :

- Tense / Aspect:
  GER: Er schläft seit zwei Stunden
  (He sleeps since two hours)
  ENG: He has been sleeping for two hours

- Argument switching:
  GER: Das Auto gefällt mir.
  ENG: I like the car.

- Head switching
  GER: Er schwimmt gerne.
  DUT: Hij zwemt graag.
  ENG: He likes to swim.

  GRE: anebike ta skalia trehontas
  (he got up the stairs running)
  ENG: he ran up the stairs.

- Category changes:
  GER: Der vom Baum gefallene Apfel
  ENG: The apple that fell from the tree

- Prodrop:
  GRE: leo tin alithia (speak-1st the truth)
  ENG: I speak the truth.

  SPA: Viajan en coche (travel-3rd by car)
  ENG: They travel by car

- Do-insertion:
  DUT: Wil je nog koffie?
  SPA: Quieres más café?
  (Want you more coffee?)
  ENG: Do you want more coffee?

- Use of article:
  SPA: Los niños son creativos.
  ENG: Children are creative.

These translation problems are tackled at different places within the METIS-II architecture, corresponding to the sections in this paper: the bilingual dictionaries (2.2.), the expander (2.3.1.) and the search engine (2.3.2.).
Word order issues and *category changes* are solved in the expander.
Intensifying is coded in the lexicon, either by coding the collocation (i.e. starker Raucher - heavy smoker) or by enumerating all possible intensifiers as translation options. In the latter case, the search engine decides which intensifier suits best the modified head word.
Different complementations are best coded in the dictionary i.e. sich erinnern an — remember.
Argument and head switching require both, lexicon coding and expander interaction. The lexicon codes information which triggers a structure modification mechanism in the expander and produces the required changes in the word order. How exactly these features are represented and to what extent they can be computed automatically is to be investigated.
As of now we are uncertain how to tackle *tense and aspect*. We are likely to anticipate an interaction between expander

rules and the search engine here. While the expander would generate several possibilities, the search engine would decide which possibility suits best the empirical data.

2.1. Source Language Analysis

2.1.1. Dutch

For tagging, we use TnT (Brants, 2001), a trigram-based tagger which was trained on the CGN (Spoken Dutch Corpus) (Oostdijk et al., 2002), using the CGN-tagset (Van Eynde, 2004).

For lemmatization we make use of the CGN-lexicon (Piepenbrock, 2004) for known words. This is a full form lexicon containing over 570.000 entries. We envisage the implementation of a rule-based lemmatizer in a later stage, so this would allow lemmatization of unseen word forms.

For shallow parsing we use ShaRPa2.0 (Vandeghinste, 2004), which is a rule-based shallow parser, detecting a.o. NPs and PPs. It also performs subordinate clause detection and subject detection.

A more detailed description of the source language analysis for Dutch is given in (Dirix et al. 2005).

Note that we do not use a full parser like Alpino (Van der Beek et al., 2005), because we want to use only basic resources.

2.1.2. Greek

The annotation of the source language string, being performed on-line, involves its tagging and lemmatising by the respective PAROLE-compatible ILSP tool (Labropoulou et al., 1996) and annotation for its constituent chunks with the ILSP chunker (Boutsis et al, 2000), which yield a sequence of tokens accompanied by grammatical information and organised into labelled chunks.

2.1.3. German

The basis for tagging of German is a morphological analyser (MPRO) and a KURD-based grammar formalism (Carl and Schmidt-Wigger, 1998) for the disambiguation and 'chunking' of German input. Both tools are very mature building the basis for several commercial applications.

The MPRO analysis refers to a lexicon of around 78 000 German morphemes. As all inflection, derivation and compounding is dealt with in a rule-based fashion this comes to hundreds and hundreds of thousands of words that can be recognised. The output of this morphological analysis is a set of feature bundles.

These feature bundles are then processed by a grammar based on a pattern matcher called KURD. The KURD formalism allows to define patterns in form of feature bundles which can be mapped onto the morphologically analysed strings allowing for the manipulation of these strings.

The KURD grammar that defines these manipulations has several components. One of them is a disambiguation module that allows for the resolution of ambiguities stemming from the morphological analysis. Another one is a kind of shallow parser that determines constituents in the sequence of MPRO objects.

The constituents that are determined are NPs, PPs, verbal structures and clauses. The architecture of the German source language modules are motivated by the tasks for which they were originally developed, namely by 'language correction' and 'language control'. The results of the analyses, however, are appropriate for METIS purposes as well.

2.1.4. Spanish

For the pre-processing of the Spanish input, only very basic linguistic resources are needed, namely only a POS tagger and lemmatizer. This means that we are not using any sort of syntactic parser or chunker to process the Spanish input. Our current tagger and lemmatizer is CastCG (Alsina et al., 2002), a shallow morphosyntactic parser for Spanish, based on the Constraint Grammar formalism. It has been built on the Machinese Phrase Tagger from Connexor[1]. The output of the tagger is a string of Spanish lemmas or base forms, with disambiguated POS tags and inflectional information. Morphological disambiguation is performed by selecting the most plausible reading for each word given the context, expressed in linear terms. At a subsequent step, morphological tags are mapped into the Parole/EAGLES tagset used by the dictionary. In this mapping step, information about POS, which will be used during dictionary look-up is separated from inflectional information which will be used only later, in generation.

2.2. From Source Language To Target Language

The dictionaries that we use are flat bilingual dictionaries, consisting of at least a pair of lemmas and their part-of-speech tag.

We are working with lemmas throughout the whole translation process in order to reduce data sparseness issues. Lemmas have a much higher frequency than word tokens, especially in the case of inflected words.

Looking up lemmas and idioms in the dictionary is not sufficient to have good quality translations. We need to address language-pair specific phenomena, like the issues listed in the beginning of section 2..

Four types of operations seem to be necessary to re-order the translation units so that the sequences of translation options (TOs) in successive translation units (TUs) approach the TL syntax:

1. move, swap or permute TUs on the same level
2. copy and move TUs into embedded TOs
3. insert and delete TUs or TOs
4. copy entire translation phrase if alternate transformations must be carried out

The tagset which is used in the source language and the tagset used in the target language can be (very) different. To overcome these differences we map the source language tags to their equivalent target language tags. Some features of the source language tokens are underspecified for the lemma (e.g. number). To allow transfer of these features, we need to map them onto the target language tagset.

[1] http://www.connexor.com/

2.2.1. Dutch-English

For Dutch to English, we use a bilingual dictionary which was compiled from various sources, like the Ergane Internet Dictionaries (Travlang Inc., http://www.travlang.com/Ergane) and the Dutch WordNet (Vossen et al., 1999). We are still manually editing and improving this dictionary. It contains about 37000 different source language lemmas, with an average of more or less three translations.

We use the CGN-tagset (Van Eynde, 2004), which is a form-based tagset. For English we use the CLAWS5-tagset, which is function-based. This leads to a many-to-many mapping between these two tagsets.

To handle structural relation problems we use a hybrid approach: making use of a limited set of rules for structural issues which can be generalised over a large number of cases, and making use of collocation and co-occurrence statistics in other cases.

2.2.2. Greek-English

The Greek-English bilingual dictionary comprises about 15000 lemmata and 3000 expressions. Both Greek and English data are morphologically annotated using the PAROLE and CLAWS5 tagset respectively. The dictionary is constantly edited and enriched on the basis of lemma frequencies encountered in the ILSP corpus (http://hnc.ilsp.gr/) which consists of about 35 million words. The Greek ILSP/PAROLE tags are mapped onto to the English CLAWS5 tag set on the basis of predefined correspondences (for instance, Ad* maps to AV0, No* maps to N*, Vb* maps to V* etc).

Tag similarity scores are also employed for the Greek-English pair, when comparing the input string to the core translation engine with the sentences retrieved from the BNC. For tags of the same category (e.g. NN1, NN2) a similarity scored is calculated:

$$SimilarityScore = \frac{SameCharacters}{NrOfCharactersOfSmallerTag}$$

A set of predefined similarity scores is employed in order to compare tags belonging to different categories (e.g. NN with AJ).

2.2.3. German-English

Dictionary lookup is best understood as an instance of abductive reasoning[2]: dictionary entries are considered facts; matching a sentence on the dictionary is a process of proving or disapproving the presence of these facts in the sentence. From the perspective of the sentence it is investigated which translation relations fit best the whole of the sentence. If no exact matching entries are found, those translation relations are kept and processed further that provide the best explanation for the observations in the sentence.

A major aspect of dictionary lookup is how to deal with incompleteness. Even the most complete dictionary is likely to contain translation relations only for a subset of words in a language. Particularly for German, due to inflection, derivation and compounding, one cannot even expect all lemmas to be enumerated in the dictionary.

During dictionary compilation, all features that are important to prove the presence of the lexical fact in the input sentence are made explicit and available. In addition, a set of rules are used at runtime to consolidate or disapprove entries by examining the context of the matched items.

For instance, a matching nominal multi-word entry is disapproved if the components of the entry are not all within the same nominal chunk of the sentence.

(1) `Abbau der Ozonschicht`
`(ozone depletion)`
(2) `Abbau der arktischen Ozonschicht`

Assume the dictionary entry (1). The head of the term Ozonschicht can be modified in the matched sentence, for instance by adjectives as in example (2). While we would like to approve the entry despite the intervening adjectives arctic, we want to reject the entry if the words co-occur "by accident" in the same sentence and are actually unrelated. This would be the case, for instance, if the words occurred in different noun phrases.

2.2.4. Spanish-English

Lexical translation is performed by a lemma-to-lemma dictionary, which has information about the POS of both the source word and the target word. Mapping from source to target is always one-to-one, meaning that entries with more than one translation, or words with more than one POS, become different entries.

Our bilingual dictionary has been extracted from a commercial machine readable dictionary, the Spanish-English Concise Oxford. The Oxford dictionary has a total of 32,653 entries, with between 3 and 4 translations per headword, on average. There are plans to enlarge the initial coverage with entries coming from the reverse direction (English-Spanish), spelling variants, and compounds (which are secondary entries in the original dictionary) as well as from other terminological glossaries.

The POS tags used for Spanish come from the Parole/EAGLES[3] tagset, while for English, we use CLAWS5, which is the same tagset used to tag the BNC.

The output of the dictionary look-up is a set of translation candidates, i.e. strings of English lemmas, plus POS tags, ordered according to Spanish-like syntax.

2.3. Target Language Generation

At this point, we are in the Target Language Generation module (cf. figure 1).

We are currently investigating different approaches to expanding the list of translation candidates, using the preprocessed target language corpus as a search space: we generate additional translation candidates, based on the input of the target language generation.

The main target language resource used is the target language corpus. As this corpus needs to be preprocessed in an equivalent way as the source language input sentence, a number of target language processing tools are needed.

We use the British National Corpus (BNC) (reference) as the target language corpus. The BNC is already a tagged

[2] Abduction is often defined as inference to the best explanation, see discussion on http://www.cs.bris.ac.uk/ flach/ECAI96/ECAI96report.html

[3] http://www.lsi.upc.es/ nlp/freeling/parole-es.html

corpus, using the CLAWS5 tagset. We applied lemmatization using the lemmatizer described in Carl et al. (2005). Chunking was performed using ShaRPa2.0 (Vandeghinste, 2004), with grammars adapted to the CLAWS5 tagset. Subordinate clause detection and subject detection were both performed using rule-based tools.

A more detailed description of the corpus preprocessing is given in Dirix et al. (2005).

2.3.1. Translation Candidate Expansion

When translating *Dutch to English*, we consider the input of the expansion module as a structured bag of bags, representing the structure of the sentence after all the source to target language mapping has been applied. We want to convert this structured bag into a sentence, by resolving each sub-bag by looking it up in the TL-corpus (depth first): we try to find a matching phrase that holds all the elements of the bag. Depending on how well the corpus phrase matches the bag elements, a score is given, resulting in a ranking of permutations, which are considered translation candidates, which get a final score from the search engine.

The *Greek-English* pair employs a pattern-recognition based approach. The pre-processed TL corpus clauses are indexed according to their main verb. Next, the pattern matching algorithm retrieves those clauses that better match with the TL-like input string in terms of number and types of chunks as well as lemmata and tags within each chunk. In case that a sentence very similar to the input one is retrieved from the corpus, it forms the basis of the final translation and is sent to the synthesising algorithm. If the best matching sentences are not very similar to the TL-like input string, the second step is activated. Again, a pattern-matching based algorithm searches the BNC to retrieve chunks similar to the ones in the TL-like input that are extracted from different sentences. The retrieved chunks replace the mismatching chunks of the best translation sentence. No mapping rules are used to reorder lemma or chunks in the TL-like string as similarity is defined in terms of compatibility of grammatical categories while the algorithm takes advantage of word order data in the corpus.

For *German to English* translation, we use a set of handcrafted so-called mapping rules which aim at adjusting major translation divergences between German and English.Mapping rules have access to all pieces of information: information of source language chunking and the English sides of the lexicon entries together with their deltas computed from the original source word(s).

In the *Spanish to English* translation, the idea is to use the target language model to validate changes of structure, instead of writing source language dependent mapping rules. These changes can be reduced to (a) local movement of content words; (b) deletion and insertion of function words (i.e. articles and prepositions); (c) and movements of sentence constituents. By allowing reordering of elements, plus deletions and insertions, the combination of possibilities in the search algorithm explodes. In order to limit the search space in a linguistically principled way, we use a sort of pseudo-chunking strategy by identifying constituents' boundaries on the strings of English lemmas. This boundary detection is performed on the basis of the POS information at hand. Boundaries are used twofold: (i) On the one hand, two consecutive boundaries mark the limits within which content words are allowed to move, and function words can be inserted or removed; (ii) On the other hand, boundaries are used to build a second-level language model (aka syntactic model) needed to handle non-local order changes, such as movements of constituents. This is an n-gram model over sequences of POS tags. The tags in this model are complex tags of the type AT0-AJ0-NN, limited by boundaries. In this way, this model gives us a representation of the syntactic patterns of the target language which is then used to rank all possible permutations of the input tag sequences.

2.3.2. Search Engine

A target language model is built by indexing all the n-grams for $2 \leq n \leq 4$, over lemmas and tags. Thus, an n-gram is a sequence of lemmas, with at most one of the lemmas substituted by its POS tag (possibly none).

The translation candidates obtained from the dictionary are validated against this model, starting with the longest n-grams. In order to reduce the search space, a pre-selection of candidates is optionally performed based on the probability of co-occurrence of content words.

At this point, the set of the remaining candidates, as well as the new candidates resulting from the application of the expansion principles explained above, are ranked according to the evidence found in the target corpus. Scoring follows a logarithmic progression based on lenght and frequency of the n-grams.

Several enhancements are foreseen. Some of them belong to optimization of parameter tuning, such as percent of candidates pre-selected in the Lexical Selection phase or scoring of different types of n-grams in the Candidate Ranking phase. Currently these paramenters are manually fixed but we plan to use Machine Learning techniques to tune them more optimally.

2.3.3. Token Generation

The last step in the translation process is to convert the lemmas with their associated part-of-speech tags into tokens. This is done with the token generator (Carl et al., 2005).

3. Evaluation

We are setting up an experiment in which we compare the different approaches, by separating the source-language dependent parts from the rest of the system. This allows the comparative evaluation of the different approaches to target-language generation, as presented in this paper.

Apart from this comparative evaluation, we want to test the accuracy of each of the approaches in the full translation process.

We are setting up a test-set for evaluation of 1000 sentences, 250 from each source language, and translated (manually) into the three other source languages, and into English (multiple reference translations), to allow the usage of automated evaluation metrics like BLEU (Papineni et al., 2002), NIST (NIST, 2001) and Levenshtein (1966).

4. Conclusions and Future Work

The general design of the METIS-II approach allows for translation without an extensive rule-set or a parallel corpus, although it is hard to draw firm conclusions before a thorough evaluation has been done.

It is clear from the current setup that a number of phenomena are solved by matching the information coming from the dictionary and the SL analysis with the TL corpus. Word order issues e.g. can be solved without rules.

We are investigating a number of different approaches for different problems, and an evaluation experiment will make clear which of these approaches are most suitable for which phenomena, and which phenomena are not covered by any of the approaches.

When building a hybrid system, using rules and statistics, it is important to keep the number of rules limited, to ensure that the system can be transferred to other language pairs, without spending large amounts of time on rule-writing.

Acknowledgements

The METIS-II project is financed under the EU Future and Emerging Technologies Specific Targeted Research Programme (FET - STREP).

More information about the METIS-II project can be found at http://www.ilsp.gr/metis2/. As soon as evaluation is done, results will appear here.

References

Alsina, A., Badia, T., Boleda, G., Bott, S., Gil, À., Quixal, M., Valentín, O. (2002). CATCG: a general purpose parsing tool applied. In *Proceedings of Third International Conference on Language Resources and Evaluation*. Las Palmas, vol. III, p. 1130-1134.

Badia, T., Boleda, G., Melero, M., Oliver, A. (2005). An n-gram approach to exploiting a monolingual corpus for Machine Translation. In *Proceedings of Workshop Example-Based Machine Translation*. MT Summit X. Phuket, Thailand.

Brants, T. (2001). *TnT - A Statistical Part-of-Speech Tagger*. Published on line at http://www.coli.uni-sb.de/thorsten/tnt.

Boutsis, S., Prokopidis, P., Giouli, V., Piperidis, S. (2000). A Robust Parser for Unrestricted Greek Text. In *Proceedings of the Second International Conference on Language Resources and Evaluation*, 31 May-2 June, Athens, Greece, Vol. 1, pp. 467-482.

Carl, M. and Schmidt-Wigger, A.. (1998). Shallow Post morphological processing with KURD. In *Proceedings of NeMLaP3/CoNLL98*, Sydney, 1998.

Carl, M., Schmidt, P., and Schütz, J. (2005). Reversible Template-based Shake & Bake Generation. In *Proceedings of Workshop Example-Based Machine Translation*. MT Summit X. Phuket, Thailand.

Dirix, P., Schuurman, I., Vandeghinste, V. (2005). METIS-II: Example-based machine translation using monolingual corpora - System description. In *Proceedings of Workshop Example-Based Machine Translation*. MT Summit X. Phuket, Thailand.

Labropoulou, P., Mantzari, E., Gavrilidou, M. (1996). *Lexicon-Morphosyntactic Specifications: Language Specific Instantiation (Greek)*, PP-PAROLE, MLAP 63-386 report.

Levenshtein V.I. (1966). Binary codes capable of correctin deletions, insertions, and reversals. In *Soviet Physics Doklady*, 10 (8), 707-710.

Markantonatou, S., Sofianopoulos, S., Spilioti, V., Tambouratzis, Y., Vassiliou, M., Yannoutsou, O., Ioannou, N. (2005). Monolingual Corpus-based MT using Chunks. In *Proceedings of Workshop Example-Based Machine Translation*. MT Summit X. Phuket, Thailand.

National Institute of Standards and Technology. (2001). *Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics*. http://www.nist.gov/speech/tests/mt/doc/ngram-study.pdf

Oostdijk, N., Goedertier, W., Van Eynde, F., Boves, L., Martens, J.P., Moortgat, M., Baayen, H. (2002). Experiences from the Spoken Dutch Corpus Project. In *Proceedings of the Third International Conference on Language Resources and Evaluation*, Vol. 1, pp. 340-347.

Papineni, K., Roukos, S., Ward, T., Zhu, W.J. (2002). BLEU: a method for automatics evaluation of machine translation. In *Proceedings of the 40th ACL meeting*, pp. 311-318.

Piepenbrock, R. (2004). *CGN Lexicon v9.3*. Spoken Dutch Corpus.

Thurmair, G. (2005). Improving Machine Translation Quality. In *Proceedings of the Tenth Machine Translation Summit*. Phuket, Thailand.

Vandeghinste, V. (2004). *ShaRPa2.0. Shallow Rule Based Parsing*. Online at http://www.ccl.kuleuven.be/ vincent/ccl/SHARPA/

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. (2004). *Tagging and Lemmatisation for the Spoken Dutch Corpus*. Internal Report.

Vossen, P., Bloksma, L., Boersma, P. (1999). *The Dutch WordNet*. University of Amsterdam.