METIS-II: Low Resources Machine Translation

Michael Carl
Institut für Angewandte Informationsforschung, Martin-Luther Str. 14, 66121 Saarbrücken, Germany

Maite Melero, Toni Badia
GLiCom (Fundaci Barcelona Media - UPF) Avinguda Diagonal, 177, Barcelona, 08002 Spain

Vincent Vandeghinste, Peter Dirix, Ineke Schuurman
KU Leuven - Centrum voor Computerlinguïstiek, Blijde Inkomststraat 13, 3000 Leuven, Belgium

Stella Markantonatou, Sokratis Sofianopoulos, Marina Vassiliou, Olga Yannoutsou
Institute for Language and Speech Processing, Artemidos 6 & Epidavrou, 15125, Greece

October 29, 2008

Abstract.

METIS-II was a EU-FET MT project running from October 2004 to September 2007, which aimed at translating free text input without resorting to parallel corpora. The idea was to use 'basic' linguistic tools and representations and to link them with patterns and statistics from the monolingual target-language corpus. The METIS-II project has four partners, translating from their 'home' languages Greek, Dutch, German, and Spanish into English.

The paper outlines the basic ideas of the project, their implementation, the resources used, and the results obtained. It also gives examples of how METIS-II has continued beyond its lifetime and the original scope of the project. On the basis of the results and experiences obtained, we believe that the approach is promising and offers the potential for development in various directions.

Keywords: Low resource MT, Statistical MT, Pattern-based MT, Shallow linguistic processing for MT

1. Introduction

Starting in October 2004, METIS-II was the continuation of METIS-I (IST-2001-32775) (Dologlou et al., 2003). Like METIS-I, METIS-II aims at translating free text input by taking advantage of a combination of statistical, pattern-matching and rule-based methods. The METIS-II project has four partners, each translating from their 'home' languages Greek, Dutch, German, and Spanish into English.

The following goals and premises were defined for the project:

1. use 'basic' NLP tools and resources,

2. use bilingual hand-made dictionaries,
3. use a monolingual target-language corpus,
4. use translation units within the sentence boundary,
5. allow different tag sets for SL and TL possible,

Crucially, parallel corpora are not required, and their usage was excluded within METIS-II. The rationale behind this was to develop prototypes of MT systems which would be suitable to translate 'small languages', i.e. language pairs for which parallel texts are difficult to come by. A basic set of NLP tools is nonetheless required for these languages, albeit very basic. The availability of the monolingual target language corpus, from which statistical language models are computed, makes METIS-II a data-driven MT system. These facts set METIS-II apart from mainstream SMT/EBMT systems.

With these goals and requirements, a number of implementations are possible. The METIS-II partners decided therefore to test and compare various implementations of the ideas, which will be outlined in this paper.

Hence, METIS-II consists of a number of modules which can be investigated horizontally, from source language to target language, or vertically, dividing the task into source-language analysis, lexical transfer, target language word-order generation and word-token generation. While the development of the four horizontal translation directions are to a large extent free-standing and independent efforts of the respective METIS-II partners, the consortium has also developed an exchange and interface format to communicate intermediate, (i.e. vertical) processing results between the different parallel modules (METIS-II, 2006; METIS-II, 2007)

In this paper we aim at presenting METIS-II from a 'vertical' perspective. We discuss each of the parallel processing steps for all language modules involved, thereby showing their common and diverging characteristics.

The project has used a broad set of tools for source text analysis that were available or else easily obtainable by the partners. The Spanish analysis module experiments on using as few linguistic resources as possible - essentially only a lemmatizer and PoS tagger. The Dutch module adds a shallow parser to detect phrases and clauses while the German module includes also "topological" information. The Greek module seeks a more complete syntactic analysis of input. These monolingual processing tools are described in detail in section 3, together with a reversible lemmatizer for English.

Section 4 describes the bilingual dictionary resources for the four modules. The Spanish module uses only a bilingual dictionary that had been extracted from a printed Spanish-English dictionary. The Dutch-English dictionary was also compiled from external sources and the Greek-English dictionary was compiled from preexisting machine-readable dictionaries and augmented manually by the most frequent entries from the Hellenic National Corpus[1]. The German-English dictionary is the largest of all the reported sizes and has been collected from unnamed sources over a long period of time. It covers words and both continuous and discontinuous phrases. Unlike other dictionaries, the German dictionary is preprocessed before use essentially through morphological analysis and generation of variants.

Section 5 describes the main resources used for generation and section 6 explains the way(s) how translations are generated in METIS-II. METIS-II follows a "generation-heavy" approach (Habash, 2004), where most of the hard translation issues are addressed during the generation phase.

The basic resource for generation are target language models, which are extracted from a huge target language corpus (the BNC) and which assist in selecting — and in some cases also in generating the word order of — the best translations. In this respect, the METIS-II core approach resembles Whitelock's (1991, 1992) 'shake-and-bake' method where the "target texts are constructed from a bag of TL basic expressions, whose elements are derived from the analysis of the source text and a set of equivalences of basic expressions"(Whitelock, 1991, p:1). However, while Whitelock uses logical and semantic constraints for 'baking' a target text from the basic expressions, METIS-II relies on statistical and pattern-based language models extracted from the target corpus to consolidate and verify target sentences.

Section 5 shows how the target-language corpus was preprocessed and how language models were conceptualized and extracted from the corpus. These models are built in idiosyncratic ways, with significant differences across language pairs. The Spanish module uses sequences of lemma/tag to validate insertions, deletions and permutations of words, the Greek and Dutch modules consolidate TL word order based on patterns and templates and the German module uses statistical n-grams.

Section 6 deals with the actual translation, the "decoding" of the source language. The overall translation method in METIS-II is creating a set of possible translation solutions and then using statistical methods to find the most probable translations. The language models

[1] `http://hnc.ilsp.gr/`

play a crucial role the selection process. Section 6.5 provides a detailed comparison on the differences and similarities across these modules.

Section 7 presents an evaluation of the translation systems using two test sets, the test suites used during development and a EUROPARL fragment, using BLEU, NIST and and TER. Results for each language pair, using a well consolidated system such as Systran, are used as topline reference measure to gauge METIS-II results.

Beyond the actual scope of the METIS-II project, several attempts were made to scale up the system to various directions. In section 8.1 we describe how a new language pair can be developed in METIS-II. Section 8.2 reports on tuning the system to a particular domain (the Europarl text) by assigning weights to the dictionary entries.

2. Background of METIS-II Implementations

In this section we briefly describe the basic ideas behind the implementations of the four translation directions. A linguistically minimal approach is favoured by the Spanish module, while the other modules employ a shallow parser to detect phrases and clauses. The Dutch and Greek modules assume some kind of structural isomorphism of phrases and clauses between the source and the target language, while the German module employs flat re-ordering rules.

2.1. Spanish to English

The approach followed by the Spanish-to-English METIS-II system strives to use as little linguistic resources as possible. The motivation in this case is not the lack of resources for processing Spanish but the desire to experiment in the leanest possible conditions, so that our findings can be applied to other, possibly smaller languages with fewer resources available. Consistently with this purpose, the preprocessing of the Spanish input requires only a tool able to lemmatize and assign morphological tags to each word of the sentence. The Spanish sentence is thus tokenized, tagged and lemmatized, but it is not chunked or analyzed in terms of constituency.

2.2. Dutch to English

For the Dutch-to-English translation pair we have chosen an approach that requires a number of tools in order to perform a shallow source language analysis: a tagger, a lemmatizer, and a shallow parser (including a clause detector). We required the target-language corpus to be

preprocessed with the same means, so equivalent tools for the target language are needed off line (Vandeghinste, 2008).

2.3. Greek to English

What is crucial within the Greek-to-English METIS-II approach is the notion of pattern, that is, phrasal segments that serve as the basis for modelling both the source (SL) and the target (TL) languages. The patterns roughly correspond to phrasal constituents of a varying size and type, ranging from clauses to sub-clausal level patterns (chunks and contained tokens). This approach, because it reflects the recursive character of natural language is expected to assist more effectively the translation process. Besides, even within the Statistical Machine Translation paradigm that strictly aimed to avoid using phrasal segments, the potential beneficial role of phrase-based models has now been recognized (Carpuat and Wu, 2007).

2.4. German to English

The German METIS-II architecture uses rule-based techniques to generate a graph of partial translation hypotheses and employs statistical techniques to rank the best translation(s) in their context. Word tokens are generated for the n-best translations.

The core idea is similar to Brown and Frederking (1995) who use a statistical English Language Model to combine partial translations produced by three symbolic MT systems. In contrast to their approach, we build the search graph with flat re-ordering rules.

The re-ordering rules generate an acyclic AND/OR graph which allows for compact representation of many different translations. A beam search algorithm tries to find most likely paths in the AND/OR graph. A similar idea for generation was suggested by Langkilde and Knight (1998) who use 2-gram language models to find the best path in a word lattice. Unlike a usual statistical decoder (Germann et al., 2001; Koehn, 2004), our ranker, hence, does not modify the graph and it does not generate additional paths which are not already contained in the graph.

3. Monolingual Language Tools

Each of the source languages modules in METIS-II has their individual preprocessing and SL analysis tools which are described in this section. In line with the requirements and philosophy of the project, all language modules use a lemmatizer and PoS tagger to process the

source language input. In addition Dutch and Greek use a shallow parser to detect phrases and clauses and German recognizes topological fields. Besides the source language analysis, we have also implemented a reversible lemmatizer for the target language (English) which was used throughout for generation in METIS-II.

3.1. Spanish Tagger and Lemmatizer

The tagger and lemmatizer that was used in the course of the METIS-II project is a shallow rule-based parser for Spanish, developed by Connexor, based on the methodological principles of the Constraint Grammar (CG) (Karlsson, 1995). The output of this tool consists of a file containing the processed sentences. Each sentence is separated by a line containing two ⟨s⟩ tags. The sentence is represented in one token per line with the following information: token number, word form, lemma, syntactic relations, syntactic tags and morphological tag (i.e. PoS and morphological info). Since we purposely do not make use of any syntactic information, the only information that is passed on the following steps is the lemma and the morphological tag. Information about PoS, which will be used during dictionary lookup is separated from inflectional information which will be used only later, in token generation.

Thus, the morphological tag provided by the tagger is decomposed in the following way:

– The PoS value is converted to the Parole/EAGLES tagset (1994), used by the dictionary.

Table I. Spanish tag conversion from CG tags to PAROLE PoS of the sentence: Me alojo en la casa de invitados (I stay at the boarding house): morphological tag provided by the CG tagger are converted into a PAROLE PoS tag[2] and an inflection tag.

Form	Lemma	Morph tags	PoS tag	Inflection tag
Me	me	PRON Pers SG P1 ACC	PP	sg:1:acc
alojo	alojar	V IND PRES SG P1	VM	i:p:sg:1
en	en	PREP	SP	
la	la	DET FEM SG	TD	f:sg
casa	casa	N FEM SG	NCF	f:sg
de	de	PREP	SP	
invitados	invitado	N MSC PL	NCC i	m:pl

– The inflectional information is encoded into a character string: Each morphological feature is assigned a low-case alpha-numerical character. All characters representing morphological values are concatenated by means of a colon (:).

Table I illustrates the output and conversion of the tagger. The PoS tag will be used to retrieve a lexical translation form the lemma-to-lemma dictionary. The inflection tag will be read off and interpreted at generation time.

3.2. Dutch Tagging, Lemmatization and Chunking

We use the TnT tagger (Brants, 2000), which is trained on internal release 6 of the Spoken Dutch Corpus (CGN) and the CGN tag set (Van Eynde, 2004). Oostdijk et al. (2002) have evaluated TnT trained on the CGN tag set, and have reported an accuracy of 96.2%.

The lemmatizer consists of two main components. The first component is a lexical lookup routine. The second component is a rule-based lemmatizer. All lemmas are looked up in the lexicon. By using the PoS-tag of the word as input argument for the lemmatizer we reduce the ambiguity. When no lemma is found in the lexicon, we switch to LRBL (Szopa, 2007), a rule-based lemmatizer for Dutch. Combining these two components leads to a lemmatization error rate of 3.20%.

For shallow parsing, we use ShaRPa2.1 (Vandeghinste, 2005), a rule-based chunker. It is used for the detection of *noun phrases* (NPs), *prepositional phrases* (PPs) and *verb groups*; and performs head detection as well. It reaches F-scores of 94.13% for NPs and 95.03% for PPs.

Table II. Dutch source-language analysis: "De grote zwarte hond blaft naar de postbode." (The big black dog barks at the postman.)

token	**tag**	**lemma**	**chunk**	**head**
de	LID(bep,stan,rest)	de	NP	
grote	ADJ(prenom,basis,met-e)	groot	||	
zwarte	ADJ(prenom,basis,met-e)	zwart	||	
hond	N(soort,ev,basis,ev,zijd)	hond	||	H
blaft	WW(pv,tgw,met-t)	blaffen	VG	H
naar	VZ(init)	naar	PP	H
de	LID(bep,stan,rest)	de	NP	
postbode	N(soort,ev,basis,zijd,stan)	postbode	||	H
.	LET()	.	—	

The parser is extended with a rule-based clause detector, detecting subordinate clauses and relative phrases. An example of source-language analysis is given in table II.

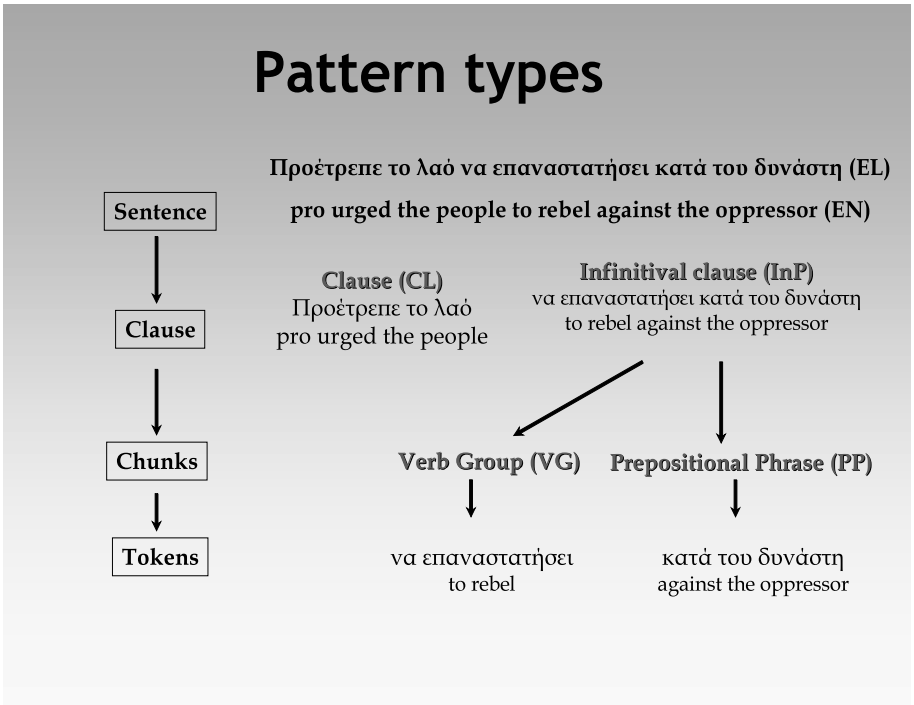

Figure 1. Pattern types used in Greek (SL) and English (TL)

3.3. Greek-English Pattern Matching

The pattern-based language modelling, governs the off-line preprocessing of the English corpus as well as the on-line processing of a given Greek string, the aim being to flexibly match linguistic patterns across the source and the target language (Markantonatou et al., 2006) as shown in figure 1. Minimal, rule-based transformations of the SL patterns are possible in order to improve cross-language matching. These transformations concern addition/deletion of constituents. They do not concern word order because this is directly treated by the pattern-matching algorithm. The processing of the target-language corpus is described in section 5.2

The Greek sentence serving as the input to the core engine, is first tagged and lemmatized (Labropoulou et al., 1996) and then processed with a robust parser (Boutsis et al., 2000) that detects clauses and identifies chunks.

3.4. German Source-Language Analysis

Table III. Analysis for the German sentence "Das Haus wurde von Hans gekauft" (The house was purchased from Hans)

lemma	#	PoS	chunks	clauses
{lu=das,	wnrr=1,	c=w,sc=art,	phr=np;subjF,	cls=hs;vf}
,{lu=haus,	wnrr=2,	c=noun,	phr=np;subj,	cls=hs;vf}
,{lu=werden,	wnra=3,	c=verb,vt=fiv,	phr=vg_fiv,	cls=hs;lk}
,{lu=von,	wnrr=4,	c=w,sc=p,	phr=np;nosubjF,	cls=hs;mf}
,{lu=Hans,	wnrr=5,	c=noun,	phr=np;nosubj,	cls=hs;mf}
,{lu=kaufen,	wnra=6,	c=verb,vt=ptc2,	phr=vg_ptc,	cls=hs;rk}

The German source-language analysis produces a flat sequence of feature bundles which contain chunking and topological information of the sentence (Müller, 2004), similar to the Dutch analysis section 3.2. An example of the German analysis is given in table III.

Among other things, the analysis comprises of a unique word number, the lemma and part-of-speech of the word, as well as morphological and syntactic information. It also contains chunking and topological information. The parser produces a linguistically motivated, flat macro structure of German sentences, as coded by the `cls` feature.

3.5. Reversible English Lemmatization

Within the METIS-II project, we have implemented a reversible lemmatizer for English (Carl et al., 2005) which reads CLAWS5-tagged words and generates a lemma together with two additional features indicating the orthographic properties (O) and the index of the inflection rule (IR). The IR-index serves to memorize the inflection rule which was applied to generate the lemma. Lemmatization rules are used to strip off or modify regular inflection suffixes from word tokens. Table IV plots two lemmatization examples. A lemmatization lexicon is used for the irregular cases.

The lemmatizer uses a single table of 128 lemmatization rules (two of which are shown on the right side in table IV). Each rule specifies the removal or replacement of an ending, conditionally on the TAG of the word and its suffix. Lemmatization and token generation is 100% reversible: a token set {token,TAG} is equivalent to a lemma set {lemma,TAG,O,IR} and both sets can be transformed into each other without loss of information, by reversing the lemmatization rule.

Table IV. Left: input and output of lemmatization and token-generation, Right: corresponding bi-directional inflection rule which can be used for lemmatization and for token generation.

TAG	token	⇔	lemma	TAG_O_IR	IR	suffix mapping		
VVG	sniffing	⇔	sniff	VVG_l_1	1	`ffing`	↔	`ff`
VVG	DRESSING	⇔	dress	VVG_c_3	3	`ssing`	↔	`ss`

However, during token generation, we usually want to produce word forms from incomplete lemma sets {lemma,TAG}, where the inflection rule *IR* is not known. To generate an educated guess which IR would produce the desired word form, we have counted for each lemma suffix the inflection rules which generated the lemma. A word form would then be generated from a lemma by looking at the ending of the lemma and by applying the most likely reversed inflection rule. With slightly more than 20,000 lemma suffixes the reversible lemmatizer achieves a precision of more that 99.5%. In order to achieve this precision we had to add a few additional tags to the original CLAWS5 tagset, and then re-tagged the BNC[3] with the enhanced tagset.

Table IV plots two lemmatization examples.

4. Bilingual Dictionary

Apart from the resources required for the monolingual source language analysis, there are two other types of resources that were used in METIS-II: a bilingual transfer dictionary and the monolingual target-language corpus. This section describes the bilingual resources, how they were obtained, preprocessed and represented, whilst section 5 points out how the target-language corpus was prepared. For Spanish, Dutch and Greek the dictionary was compiled from external resources and adapted to the needs of METIS-II. German used a large inhouse dictionary for which a special compilation procedure was developed.

4.1. The Spanish-to-English Dictionary

Lexical translation is performed by a lemma-to-lemma dictionary, which contains information about the PoS of both the source and the target word. The Spanish-English METIS-II dictionary has been automatically extracted from a commercial machine-readable dictionary, the

[3] Section 5 gives more information on this corpus.

Concise Oxford, which has 32,653 entries in the Spanish to English direction with an average of 4 translations per headword.

In the process of extracting the METIS-II dictionary, the PoS of the SL word is converted to PAROLE format. As for the PoS of the translation, it needs to be calculated from scratch because, as it is usually the case with this type of dictionary, the original Spanish-English dictionary does not provide this information. For that reason, the PoS of the translation is automatically assigned on the basis of the PoS of the SL word and is subsequently validated on the target corpus. The initial dictionary coverage has been enlarged, using automatic procedures, with entries coming from the reverse direction (English-Spanish) as well as from terminological glossaries. Orthographic and regional variants, such as British and American spellings have also been added, as well as compounds, which appear in the original dictionary as secondary entries under the main head word. At run-time, every word in the Spanish sentence is looked up in the dictionary and one or more translations are retrieved for each word. The possibility that the word is part of a compound is always considered. In that case, the translation for the whole compound is given, together with the word-by-word translation.

The translation model in the Spanish-English prototype is very simple: since we are not using any structure transfer rule, it consists only of the dictionary. The dictionary in METIS-II effectively functions as a flat translation model and no complex operations can take place in it. However, some translation divergences between SL and TL are actually dealt within the dictionary, such as the following:

1. Category change, e.g. `mundial (ADJ)` translated by `world (N)` (as in "economia mundial" (world economy));
2. A single SL word is translated into a fixed TL multi-word expression, e.g.: `acequia` translates into `irrigation ditch` and `muchos` translates into `a lot of`;
3. A SL compound has a single-word translation, e.g.: `abeto falso` translates into `spruce`.

The output of the Spanish sentence preprocessing and dictionary lookup is a set of translation candidates in form of strings of English lemmas and PoS tags, ordered according to Spanish-like syntax.

4.2. The Dutch-to-English Dictionary

The initial dictionary was compiled from various sources, like the Ergane Internet Dictionaries[4] and the Dutch WordNet (Vossen et al., 1999), and was manually edited, adapted, and improved. It contains around 37,000 entries with an average of about 3 translation alternatives per entry. Table V shows a Dutch-English dictionary entry for a single word-to-word translation with two translation alternatives.

Entries consist of a source-language part and a target-language part. The source-language part may be composed of a series of chunks and tokens, where each token consists of its lemma and its part-of-speech tag. A chunk consists of its chunk name and references to its daughters. The target-language part can consist of a series of translation units, which each can consist of a series of chunks and tokens. The target part also contains information in which is stated which target-language part corresponds to which source-language part(s).

Table V. Excerpt of the Dutch-to-English dictionary

```
<dict-entry id="19">
    <source>
        <token id="1" pos="ADJ" lemma="zwart"/>
    </source>
    <target>
        <trans-unit id="1">
            <token id="1" pos="AJ?" link="1" lemma="black"/>
        </trans-unit>
        <trans-unit id="2">
            <token id="1" pos="AJ?" link="1" lemma="gloomy"/>
        </trans-unit>
    </target>
</dict-entry>
```

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, but the dictionary format can still handle these, as shown in Vandeghinste (2008).

4.3. The Greek-to-English Dictionary

The METIS-II Greek-English dictionary is a lexical database consisting of about 30,000 Greek entries in lemmatized form, their English translation equivalents and their respective parts-of-speech. The database was constructed by combining existing bilingual lexica of ILSP with

[4] Travlang Inc., `http://www.travlang.com/Ergane`

a list of the most frequent lemmas of the Hellenic National Corpus (Modern Greek). BNC conventions of multi-word units were adopted. The main criteria for selecting the translation(s) to be assigned to a lemma were (i) different translations as opposed to different senses of the same translation (ii) 'formal' usages as opposed to 'colloquial' ones. Entries were listed without a priority code.

4.4. The German-to-English Dictionary

The German-English METIS-II dictionary contains more than 629,000 entries collected over the past 20 years. In its editable form, dictionary entries are represented as full forms and both language sides are independent. That is, a single word can translate into a single word, a phrase or a discontinuous phrase as in table VI. The German verb *einsperren* for instance, translates into a discontinuous English verb *lock ⟨so.⟩ away*. Entries are coded as flat trees: while the word(s) of the entries represent the leaves of the tree, the features *DE* and *EN* in table VI are their 'mother nodes', which provide information about the type of the entry.

Table VI. Examples from the German-to-English dictionary

German	*DE*	English	*EN*
einsperren	verb	lock ⟨so.⟩ away	verb
Anweisung ausführen	verb	execute statement	verb
von ⟨etw.⟩ Kenntnis nehmen	verb	take note of	verb

The dictionary undergoes a number of preprocessing steps before the entries can be mapped on a German lemmatized and analysed sentence. The source and the target language sides of the dictionary pass through a multi-layered fully automatic compilation step. For the SL side this involves:

4.4.1. *Morphological analysis and lemmatization of the 'leaves':*

With the lacking context of words in a dictionary, the morphological analyser MPRO(Maas, 1996) provides the following ambiguous readings for the word *ausführen*.

lemma	PoS	agreement	morph. structure
ausführen	noun	sg, acc;dat;nom, neut	aus_$führen
ausführen	verb, fin	plu, 1;3, pres	aus_$führen
ausführen	verb, inf		aus_$führen
ausfahren	verb, fin	plu, 1;3, past, subj	aus_$fahren

The symbol '_$' marks the detachable prefix *aus*, and thus illustrates the structure of the word. These readings are then disambiguated and filtered based in the type of the entry.

4.4.2. *Checking internal consistency of the entries*
By means of a set of patterns we control whether the analyses of the words (i.e. the leaves of the entry, as in the table above) are consistent with its type. A dictionary entry is consistent iff at least one of its readings can be consolidated by a pattern associated to its type; otherwise the entry will be marked obsolete. This process also disambiguates readings and filter those readings that are intended by its type (e.g keeping only the *verb,inf* reading of *ausführen*). The process makes sure that the representations of the entries are consistent with the analysed words of an input text.

4.4.3. *Variant generation*
Variants are generated to extend the coverage of the dictionary for nominal and verbal expressions. A variant is an additional translation relation that covers a different realization of a dictionary entry. The verb *ausführen*, for instance, matches a main-clause verb in a non-compositional tense while the variation *führen ... aus* matches in a subordinate clause. For nominal expressions morpho-syntactic variation for compounding, as e.g.: *Abfertigung des Gepäcks* → *Gepäckabfertigung* but also coordination, and synonyms are generated (Carl and Rascu, 2006).

5. Target Language Modelling

We have experimented with various ways to use the implicit knowledge encoded in the monolingual target language corpus, and generated different language models. All language models are based on the BNC[5]. The BNC is a tagged collection of texts making use of the CLAWS5

[5] The British National Corpus (BNC) consists of more than 100 million words in more than 6 million sentences `http://www.natcorp.ox.ac.uk/`

tagset which comprises roughly 70 different tags. As pointed out in section 3.5, to ensure reversibility of the lemmatized forms we had to add a few tags to the tagset and re-tag the BNC accordingly. The re-tagged BNC was then lemmatized before building the language models. For target language modelling there were, thus, three types of information available: (i) the original word form, (ii) the lemma and (iii) the PoS tag of the words. The next sections describe how this information was used in the METIS-II modules.

5.1. Spanish: Structure changing operations subsumed by the language models

As mentioned in section 3.1, the METIS-II Spanish-to-English approach is very minimalist in the use of linguistic resources, both corpus based and manually created. This means, among other things, that there are no transfer rules to deal with structural divergences between the two languages: the target corpus is the basis both for lexical selection and for structure construction. Translations that imply changes of structure are among the main difficulties of using a bilingual dictionary instead of a true translation model. These structure changes can ultimately be reduced to:

1. local movement of Content Words (CW);
2. deletion and insertion of Function Words (FW)[6];
3. movement of sentence constituents.

Our strategy, which makes crucial use of the distinction between function and content words provided by the PoS tagger, is based on the use of the target-language model to validate any change of structure occurring between SL and TL, instead of writing source-language dependent mapping rules (Badia et al., 2005; Melero et al., 2007).

A series of target language models are built by indexing all the n-grams as retrieved from the tagged and lemmatized BNC for $1 \leq n \leq 5$. An n-gram[7] can belong to one of the following types:

- a sequence of lemma/tag (e.g. always/ADV + wear/VV + a/AT + hat/NN)

[6] The following parts-of-speech are typically considered to be function words: articles, conjunctions, determiners, pronouns, prepositions and, specific to English, the existential (there) and the infinitive marker (to).

[7] The 5-gram model is used only to build the insertion and deletion models.

– a sequence of lemma/tag except for one position of tag alone (e.g. ADV + wear/VV + a/AT + hat/NN)

During the indexing process, tokens are usually indexed as either lemma/tag or tag alone. Exceptions are:

– personal pronouns (PNP) which are always lemma/tag

– cardinals (CRD), ordinals (ORD) and unknown words (UNC) which are always indexed as tag alone.

In order to optimize the indexing and the search process, only n-grams with words appearing among the 30K more frequent are indexed. The models are stored in Berkeley databases, one for each value of n.

To account for structure modifications, we allow permutation of content words (CWs) between two consecutive boundaries[8], as well as insertion and deletion of a predefined set of functional words (FWs).

In order to deal with structure changes, a deletion and an insertion model are created. To build the deletion model, for every n-gram (for n between 3 and 5) containing functional words in any position, excluding the first and the last, the n-gram resulting from deleting the functional word(s) is looked up in the TL model. If the resulting n-gram has a frequency a fixed times greater than the original n-gram frequency, then the original (longer) n-gram is linked to the new (shorter) n-gram in the deletion model, which is also implemented as a Berkeley hash table. The insertion model is built in much the same way.

5.2. Dutch and Greek: Pattern-based Language modelling

In the Dutch-to-English and in the Greek-to-English modules, both the target-language corpus and the source language string are processed likewise. This facilitates the mechanism that matches the source sentence that is fed into the system, with the best sentential and sub-sentential English patterns from the BNC corpus. Processing involves clause splitting, lemmatizing, tagging, chunking and storage in a database etc.

Broad patterns are obtained from the tagged and lemmatized BNC, by splitting each sentence into the clauses contained with a purpose-built tool for clause detection. Finally, each clause is chunked with ShaRPa 2.0 (Vandeghinste, 2005), in order to obtain sub-sentential

[8] Boundary detection is performed on the basis of the PoS information at hand. A boundary is defined by a pair of adjacent PoS tags (e.g. NounArticle), which are considered to unambiguously indicate a transition between two consecutive constituents.

level patterns, i.e. VGs (verbal chunks), NPs (noun chunks), PPs (prepositional chunks), AJPs (adjectival chunks), InPs (infinitival chunks). Subclausal patterns consist of tokens. Single tokens are treated as the smallest patterns.

In order to speed up and facilitate the search for the best match, the TL corpus is stored into a relational database that contains (a) clauses indexed by their main verb and the number of their chunks and (b) NP and PP chunks classified according to their label and head.

5.3. German: Statistical n-gram Language Modelling

In the German-to-English module, we have generated statistical n-gram language models. The language models (LMs) were generated using the CMU language modelling toolkit[9] or SRILM toolkit. The functions provided with these toolkits were adapted and integrated into a beam search algorithm as described in section 6.4. We have experimented with the following parameters:

- number of sentences abitrarily extracted from the BNC:
  - 100K, 1M, 2M and 5M
- different kinds of statistical language models:
  - token-based LM: using the surface word forms
  - lemma-based LM: using the lemmatized word forms
  - tag-based LM: using the CLAWS5 tags
  - lemma-tag co-occurrence statistics
- 3 and 4-gram for token and lemma LMs and 4 to 7-gram CLAWS5-tag LMs

6. Translating with METIS-II

In line with the different philosophies and the variety of resources described in the previous sections, decoding works differently for each of the language pairs. This section illustrates how translations are actually produced for the METIS-II languages. The underlying translation models share a number of common features despite significant differences across the modules. We dedicate a specific section to each of the

[9] This toolkit can be downloaded from `http://www.speech.cs.cmu.edu/SLM_info.html`

language modules on its own, section 6.5 resumes and compares their similarities and differences.

6.1. Spanish to English

During the Spanish to English translation process, translation candidates resulting from dictionary lookup of input SL words are validated against the target language models, taking into account possible structure alterations. In a first version of the implementation, scoring of the translations was performed on the total list of all expanded and permuted translation candidates. This had performance problems due to a combinatorial explosion in the expansion step. The current approach combines both candidate expansion and scoring, by incrementally building the search space, following (Koehn, 2004). This is done by using a stack of a given depth (100 is the default) that only allows candidates scoring over a certain threshold. In the each (partial) candidate is expanded into a set of (partial) candidates, via the structure modifying operations applying to it, plus the different translation options provided by the dictionary. Candidates are then ranked and pruned up to a given stack depth. The scoring of each partial translation is accumulatively computed using the already computed scorings, according to the equation (1).

$$score(pTr, w) = score(pTr) + \sum_{i=1}^{i=n} \lambda_i \; cost_i(end_i(pTr), w) \qquad (1)$$

Candidate scoring follows a logarithmic progression based on length and frequency of the n-grams, complemented with a negative scoring on the pieces that remain untranslated. The rationale of the negative scoring is that even if a long n-gram has been identified as a good candidate for translation, if the remaining pieces are unfrequent PoS tag combinations, there is a penalizing score that counterbalances the positive scoring. At the N-th step (the source sentence contains N tokens) the decoding process stops. We get a ranked stack with the translation candidates. The better-ranked candidate is chosen and undergoes token generation. This is done in two successive steps. In a first step, the extended CLAWS5 tag for each lemma is calculated, by combining the reduced tag (used in the lemma-to-lemma dictionary and lemmatized TL corpus) with the relevant part of the SL morphological information. In a second step, the reversible tokenizer, previously used in the other direction to lemmatize the TL corpus, is applied. The final output is the full-formed English translation of the original sentence.

6.2. Dutch to English

After source-language analysis (cf. section 3.2), all words and all word combinations are looked up in the dictionary, and all translation options are retrieved from it. We consider the translations of a chunk a *bag*. Since the source-language analysis allows embedded chunking, this results in a number of *bags of bags*. Using bag representations all permutations within a chunk (or a sentence) are considered translation candidates.

We assign word order by using the word order information from the best matching corpus fragment. Matching a bag with the corpus results in a number of permutations with different matching scores, because each permutation matches with corpus chunks to a different degree. The closer a permutation matches a chunk in the corpus, the higher the score will be. When a perfect match is found, the score will be one (1).

A limited number of transfer rules are applied to expand the search range in the target-language corpus. For example, in order to deal with *do* insertion, in negative and interrogative sentences, the verb *to do* is introduced in all the bags of translation options.

When we move up one level in the tree representation, we still want to use the same matching mechanism. Higher level chunk re-ordering is based on the lemmas of the heads of all the candidate translations of a lower level chunk are matched.

The matching procedure also performs lexical selection (besides the co-occurrence metric), because not every bag alternative matches with the same accuracy, leading to translation candidate selection when a certain combination of words occurs in the corpus.

A bag element is matching a corpus element when the lemma (or the lemma of the head of the constituent) matches. The accuracy of matching on any level in the tree representation is calculated according to equation (2),

$$a_{m_i} = \frac{m_i}{n_i + p_i^2} \tag{2}$$

where m_i is the number of matching bag elements, n_i is the total number of bag elements, and p_i is the number of elements in the corpus chunk which are not in the bag, and which cannot be replaced by one of the elements in the bag. When $m_i < n_i - 4$, the bag is not retained as a possible solution, because experiments with different values for this parameter show that the number of insertions in the corpus as compared to the bag becomes too large to have a useful match.

Apart from the matching accuracy, we also take into account the relative frequency of the corpus chunk with respect to the total frequency of all corpus chunks in which the same or a higher number of elements match, as in equation (3):

$$g_{m_i} = a_{m_i} \cdot \sqrt{\frac{f_{m_i}}{\sum_{k=m_i}^{q} f_k}} \tag{3}$$

where f_{m_i} is the frequency of the matching corpus chunk, and $\sum_{k=m_i}^{q} f_k$ is the total frequency of all matching corpus elements with an m greater than or equal to the number of matching corpus elements in this chunk, and q is the highest m_i found for the chunk at hand.

Permutations that do not match with any corpus fragments are no longer considered, allowing us to move back from a *bag* representation to a *chunk* or *sentence* representation (in which the order is fixed). A detailed example of this matching procedure is explained in Vandeghinste (2008).

The result of all the previous steps is a set of shallow trees representing the target language sentence candidates, where the terminal nodes contain lemma and tag information, but no tokens.

The tokens can be generated from the lemma and its tag, using the reversible lemmatizer in reverse mode (from lemma to token), as described in section 3.5.

6.3. Greek to English

The Greek to English approach is based on pattern matching techniques that exploit the patterns stored in the English corpus. The mapping of a Greek sentence to structured set of English sentential and sub sentential patterns is executed with (a) the bilingual dictionary (b) the English corpus and (c) a pattern-matching algorithm. The English corpus contains chunked English clauses, and the bilingual dictionary is a minimal resource which contains tuples of the sort:

<Greek lemma, Greek tag, English lemma, English tag>.

The translation process starts after the dictionary lookup that returns the English equivalents for each Greek token. The translation process takes as input a string consisting of Greek lemmas, their (possibly multiple) English translations and syntactic information (chunks and tags). It can be broken down to 2 distinct steps:

1. Retrieval of translation candidates from the TL corpus:
   The most appropriate English clauses are retrieved from the corpus. We consider as relevant all clauses where (i) the verbal chunk head

is the translation of the SL clause, and (ii) the number of chunks in the SL and in the retrieved corpus clause ranges within $[n,\ n+2]$ where n is the number of chunks in the input sentence. At this step, all retrieved clauses that satisfy the search criteria have the same probability of being selected as the basis of the final translation. If no matches are found, the algorithm moves on to the next clause of the SL text. To speed up retrieval, the verbs in the corpus have been indexed.

2. Ranking of translation candidates according to their similarity with the SL clause:
   The retrieved English clauses are ranked according to their similarity with the SL clause. Ranking procedes top-down at two levels. To calculate the similarity, we employ a general pattern-matching algorithm and a series of weights that mainly reflect grammatical information. The pattern-matching algorithm is an implementation of the Hungarian algorithm, also known as Kuhn-Munkres algorithm, initially designed by Kuhn (1955) and later revised by J. Munkres (1955). The weights provide information about the similarity of SL PoS tags and chunk labels with the corresponding TL ones. By addressing this matching problem as a general, weighted assignment problem, METIS-II manages to resolve translation issues without resorting to any linguistic TL generation and transfer rules.

At the first level of comparison, the SL clause (with the English lexical information) is compared with all the retrieved TL clauses at clause level, in order to establish the correct order of chunks within the clause and to disambiguate the possible multiple translations of the head tokens. The clause level comparison takes into account only general chunk information about chunk labels and chunk head tokens and is performed by the pattern-matching algorithm and the set of similarity weights concerning chunk labels and PoS tags.

Equation (4) is used for measuring the similarity of two chunks, drawing on general chunk information.

$$ChunkScore_n = bcf_n \cdot LabC_n + tcf_n \cdot TagC_n + lcf_n \cdot LemC_n \quad (4)$$

Chunk similarity at clause level ($ChunkScore$) is calculated as the weighted sum of the chunk label comparison score ($LabC$), the chunk head lemma comparison score ($LemC$) and the chunk head tag comparison score ($TagC$). Each discrete chunk label has been preassigned a set of similarity weights. These weights are: the chunk label cost factor

(bcf), the chunk head tag cost factor (tcf) and chunk head lemma cost factor (lcf), where $tcf + bcf + lcf = 1$.

After calculating the similarity score of each SL-TL chunk pair, we calculate the clause similarity as a weighted sum of the similarity score of each SL-TL chunk pair (see equation (5)). The term ocf (overall cost factor) is the cost factor weight of the SL chunk in the chunk pair and $ChunkScore$ is the corresponding score (we have already described how we calculate it). Each chunk label is preassigned a different cost factor that reflects the importance of a chunk label over other chunk labels.

$$ClauseScore = \sum_{n=1}^{m} ocf_n \cdot \frac{ChunkScore_n}{\sum_{j=1}^{m} ocf_j} \tag{5}$$

At the second-level comparison, we apply the same process at chunk level aiming at establishing the correct word order within each chunk and select the correct translations for words. For each clause pair, each SL chunk is compared to the TL chunk on which it has been mapped at the clause level comparison procedure. Similarity of the words contained in the chunks is calculated drawing on lemma and part-of-speech tag information. The comparison is performed with the same pattern-matching algorithm as before, however, this time only the PoS tag similarity weights are taken into account. Equation (6) illustrates the mechanism for calculating word similarity: it is the weighted sum of the lemma and the PoS tag comparison. The tcf weight is the tag cost factor of the PoS tag of the SL word.

$$TokenScore_n = (1 - tcf_n) \cdot LemC_n + tcf_n.TagC_n \tag{6}$$

We calculate the total chunk similarity as the weighted sum of all token scores.

Having established the similarity score of each SL-TL chunk pair, clause similarity is calculated as the weighted sum of the similarity score of each SL-TL chunk pair. The TL clause with the highest final score is selected as the basis for translation, while chunk and token order has already been established. Nevertheless, the final translation is derived from the specific corpus clause, only after the contained chunks have been processed with the purpose of eliminating any mismatches. The necessary actions are performed in a final step: chunks are either modified or substituted for other chunks in order to, eventually, form the final translation. For a detailed description of the aforementioned process see (Tambouratzis et al., 2006).

6.4. German to English

In the German-to-English approach, rule-based devices generate an acyclic AND/OR graph, which allows for compact representation of many different translations. A statistical beam-search tries to find the best translation in that graph. Starting from a SL sentence, the graph is constructed in three rule-based steps. The graph is then traversed and translations are ranked. Finally word tokens are generated for the n-best translations. The architecture consists of the following five steps:

6.4.1. *German SL Analysis*

The *Analyser* lemmatizes and morphologically analyses the SL sentence. It produces a (flat) grammatical analysis of the sentence, detecting phrases and clauses and potential subject candidates as described in section 3.4, table III.

6.4.2. *Dictionary Lookup*

The analysed SL sentence is then matched on the transfer dictionary. The procedure retrieves ambiguous and/or overlapping entries and stores them in the graph. Matching proceeds on morphemes and lemmatized forms and suited to retrieve discontinuous entries, cf. section 4.4.

Due to the complexity of discontinuous matches, we only allow discontinuous matches for verbal and nominal entries. In Carl and Rascu (2006) we have described various strategies to reject matched entries if they do not obey a predefined set of criteria.

For verbal entries, various permutations of the words are possible, according to whether the entry occurs in a subordinate clause or in a main clause. We use the field and chunk annotation in the German analysis to validate and filter or reject the matched entries. These criteria are further developed in Anastasiou and Culo (2007) making use of the German topological fields.

To account for a maximum number of different contexts, the dictionary generates all translation hypotheses which are then filtered and graded by the *Ranker* in the context of the generated sentence.

6.4.3. *Word-Order Generation*

This step inserts, deletes, moves, and permutes items or chunks in the AND/OR graph according to the TL syntax by means of a rule-based device. The rules take into account phrase and clause segmentation of the SL language sentence as well as word grouping resulting from the dictionary lookup. The modifications in the graph are such that each path containes exactly once the translation(s) of all the words of the source language sentence.

As in the so-called "generation-heavy" translation (Habash, 2004), the rules produce numerous partial translation hypotheses. For our German-to-English module we have currently ca. 50 rules, which are described in more detail in Carl (2007). This "symbolic overgeneration" is then constrained by a statistical ranker making use of several statistical feature functions.

6.4.4. *Ranking and Translation Selection*
In this step, the AND/OR graph is traversed to find the most likely translations as a path through the graph. Ranking is a beam search algorithm which estimates each node in the path with a set of feature functions (Och and Ney, 2002) and keeps those target sentence $\hat{e}$ with the highest probability according to equation (7).

$$\hat{e} = \texttt{argmax} \sum_{n} \sum_{m} w_m h_m(\cdot) \tag{7}$$

In equation (7), h_m is a feature function and w_m is a weighing coefficient, while n is the number of non-overlapping translation units matching the SL sentence (including those inserted or deleted in the generation module). Given the rich annotation of our data, there are numerous possibilities for the selection of feature functions, some of which are described in section 5.3. In the METIS-II evaluations reported in sections 7 and 8.2 we experimented with features representing word token, lemmas, PoS tags, lexical weights and co-occurrences of lemmas and PoS tags.

6.4.5. *Token Generation*
This step (cf. section 3.5) generates surface word-forms from the lemmas and PoS tags.

6.5. Comparison of decoders

The previous sections illustrate each of the METIS-II decoder individually. In this section we resume and compare the characteristics of these modules by looking at how hypotheses about TL word order are generated and how the most likely translation is selected.

6.5.1. *Greedy vs. exhaustive translation modelling*
Spanish, Dutch and Greek follow an incremental, non-monotonic approach to 'shake-and-bake', where the target sentence is piece by piece constructed from portions of the 'bag of TL expressions' (Whitelock, 1991) and each portion is in itself locally validated through the target language model. In contrast, the German decoder first produces all

possible translation hypotheses in a compact graph representation and then uses language models and a beam searcher to select the best translation as a path through the graph.

6.5.2. *Algorithmic vs. rule-heuristic word re-ordering*
The Dutch and German modules employ rules to generate hypotheses of possible TL word-order — particularly for long distance movements. Spanish and Greek chose an algorithmic way to permute TL expressions. The latter approach have potentials of making the systems more language independent, while it is hard to correctly produce discontinuous translation in an algorithmic manner, which seems to be particularly important for Dutch and German.

6.5.3. *Isomorphism vs. local changes*
Dutch and Greek assume structure-isomorphism of phrases and clauses in the source and target language, while Spanish and German rely on local re-arrangements of the TL expressions. The former method requires a synchronization of the source- and target language resources, while for the latter, in principle, SL and TL resources may be processed and prepared independently.

6.5.4. *SL vs. TL information for word order hypotheses*
Permutation and re-arrangement of TL expressions for the German module is based exclusively on SL information from which these expressions were derived, while for Spanish TL word order hypotheses are based only on the TL information of the expressions. Due to the isomorphism assumption, the Dutch and Greek modules hypothesize TL word order based to some extend on the correlation of SL and TL information.

6.5.5. *Top-down vs. bottom-up vs. flat re-ordering*
The Greek module generates translations top-down by applying first larger, more abstract clause pattern models and then establishing the correct word order within each chunk. The Dutch module proceeds bottom-up incrementally consolidating word order from lower level phrases to higher level phrases. Spanish and German use flat re-ordering rules.

7. Evaluation of METIS-II

The evaluation of METIS-II was performed on two test sets, one consisting of data that had been used throughout the project for development purposes and one consisting of unseen data gathered from a

previously existing bilingual corpus (Vandeghinste et al., 2008). To measure results we used BLEU (Papineni et al., 2002), NIST (Doddington, 2002) and TER (Snover et al., 2006). The first two metrics measure edit distance using n-grams, while TER (Translation Error Rate) measures the amount of editing that a human would have to perform to get the translation right.

Each language group constructed a development set consisting of 200 sentences, with material evenly distributed among four different categories: 56 sentences illustrating grammatical phenomena (defined by each site), 48 sentences from newspapers; 48 sentences from encyclopedia articles, or similar sources of non-specialized scientific text; 48 sentences from technical manuals, or similar sources of technical text. Each site had three different human translators prepare three English reference translations of the development material for evaluation purposes.

As the development test set had been used to fine-tune the systems throughout the project, an independent test set was also used to evaluate METIS-II. This data came from an existing bilingual corpus, namely Europarl (Koehn, 2005). Europarl is a multilingual corpus of transcriptions of debates in the European Parliament, so that a parallel sentence aligned text exists which contains translations into our 5 languages. In addition, each METIS-II consortium partner had a professional translator translate the sentences that were in the respective source languages (Greek, Dutch, German and Spanish) into English. Together with the original English sentence from the corpus, this procedure yielded 5 reference translations for each of the sentences in the Europarl test set.

We chose Systran as a state-of-the-art system for comparison because it is one of the better known and most widely used MT systems (e.g., by the European Commission and the United States Department of Defense) and it is available for all the language pairs to be evaluated, which provides a homogeneous evaluation framework. This does not mean that Systran is equally developed for all language pairs, but that the underlying technology, and therefore its strengths and weaknesses, is the same. Systran is a syntactic transfer, rule-based MT system that has been under development since 1968, with a huge amount of funding from companies and institutions and large development teams. It uses large repositories of rule sets, large dictionaries, full parsers, elaborated algorithmic principles, etc. METIS-II, on the other hand, has been built in 3 years within 4 university groups, as an exploratory effort to build a hybrid MT system with no parallel corpus. Its architecture and components have been subject to much experimentation during the process. It is therefore reassuring that its results, though clearly

worse than those obtained with Systran, stand up to the comparison. Section 7.1 compares performance of METIS-II with Systran on the development set and the Europarl test set, while 7.2 highlights the differences in performance for the development set and the Europarl test set

7.1. Results METIS-II vs. Systran

In what follows, we will provide two summary tables per language pair, one corresponding to the development test set and one to the Europarl test set

7.1.1. *Spanish to English*

Results of the Spanish-English METIS-II on the two test sets show that the system's output is quite stable across test sets, with a difference of 0.016 points for BLEU, 0.17 for NIST and 4.5 for TER. On the other hand, the differences between METIS-II and Systran are quite large: on average Systran performs between 30 and 40% better than METIS-II. It is worth noting that the Spanish-English language pairs is one of Systran's more mature systems, and one that performs particularly well, as the scores show.

Table VII shows the scores for the Spanish-to-English language pair, for the development and Europarl test sets.

Table VII. ES-EN results for METIS-II and Systran on the Development and the Europarl test set

	Development set		Europarl test set	
	METIS-II	Systran	METIS-II	Systran
BLEU	0.294	0.463	0.278	0.464
NIST	6.78	8.51	6.61	8.62
TER	49.76	36.16	54.24	37.02

7.1.2. *Dutch to English*

Table VIII shows the scores for the Dutch-to-English language pair, for the development and Europarl test sets. For both test sets, the results on Systran outperform the results in our approach. A more fair comparison can be made with the work presented by Zwarts and Dras (2007). They have trained a statistical MT system on the Europarl corpus, and have extracted a test set from that corpus. When we compare these results (column Z&D) with the results we had on our

Table VIII. NL-EN results for METIS-II and Systran on the Development and the Europarl test sets

	Development set		Europarl test set		
	METIS-II	Systran	METIS-II	Systran	Z&D
BLEU	0.237	0.378	0.193	0.383	0.207
NIST	6.19	7.28	5.98	7.99	—
TER	59.52	38.81	60.92	44.66	—

development test set, we notice that we perform better than Zwarts and Dras. This is not an unfair comparison, as for the development test set we mainly just added the words occurring in this test set to our dictionary, which can be compared with training a translation model based on word alignments. Even the results from the Europarl test set do not score much lower than the results presented by Zwarts and Dras.

7.1.3. *Greek to English*

Table IX. EL-EN results for METIS-II and Systran on the Development and the Europarl test set

	Development set		Europarl test set	
	METIS-II	Systran	METIS-II	Systran
BLEU	0.366	0.395	0.186	0.313
NIST	7.26	7.70	6.17	7.69
TER	48.256	37.258	64.959	50.747

Table IX illustrate the scores obtained for the Greek-to-English language pair. Results on the development test set show that both systems generate translations of a broadly comparable quality. On the other hand, results on the Europarl set are clearly more favourable to Systran, especially measured with BLEU.

7.1.4. *German to English*

In table X we plot the results of the German-to-English METIS-II system in two different experimental settings.

In the first experiment (METIS-II$_1$), we used a basic set of generation rules (cf. section 6.4). In the second experiment (METIS-II$_2$), we further developed and refined some generation rules for handling adverbs and negation particles, such as 'never', 'usually', extraposition

Table X. DE-EN results for METIS-II and Systran on the Development and the Europarl test set

	Development set			Europarl test set	
	METIS-II$_1$	METIS-II$_2$	Systran	METIS-II$_2$	Systran
BLEU	0.186	0.223	0.313	0.282	0.396
NIST	5.48	5.32	6.36	6.68	8.05
TER	—	—	—	55.97	42.93

of prenominal adjectives (e.g., "der vom Baum gefallene Apfel" would become "The apple fallen from the tree"), and "um ... zu" constructions. In the ranker we used lemma language models with 3 and 4-grams and tag language models with 4, 5, 6, and 7-grams. We varied weights between 0.01 and 10 for each of the feature functions and the kept the combination which provided the best results. This setting was also used to evaluate the Europarl test set. The public version of Systran (Babelfish), however, performs even better than our best setting.

7.2. Cross-language results for METIS-II on development & Europarl test sets

For convenience purposes, in this summary we concentrate on the BLEU metric, one of the most used in current MT research.

Table XI. Cross-language results on the development and Europarl test set (BLEU)

	Europarl	development	difference
NL-EN	0.1925	0.2369	0.0444
DE-EN	0.2816	0.2231	-0.0585
EL-EN	0.1861	0.3661	0.1800
ES-EN	0.2784	0.2941	0.0157

Table XI shows that ES-EN is the system that has the most stable performance across test sets, while EL-EN shows the greatest variation. The most surprising result is DE-EN's, which performs better on the Europarl corpus than on the development set. A partial explanation may be that DE-EN has used Europarl type of text to tune lexical weights. Also, the DE-EN development set was chosen to contain hard translation problems so that also Systran performs more poorly on it than on the Europarl test set.

8. Beyond METIS-II

In this section we report on two attempts to push METIS-II beyond the its immediate goals as defined in the project specifications. We show that the METIS-II framework can be upgraded with additional resources which may increase the translation quality and that new language pairs can be quickly developed.

8.1. Rapid development of a new language pair

The METIS-II strategy, compared to other statistical MT systems, implies a shift of emphasis from the translation model to the target language generation. In the particular case of the Spanish-English METIS-II system, the shift is even more pronounced, since we do not use any kind of mapping rules between source and target language structures. This strategy favours modularity and language independence and, thus, should be easily translatable to new language pairs, requiring only very basic linguistic resources. We present here an experimental Catalan-English system, which may serve as benchmark to test the portability of the METIS-II ideas (Badia et al., 2008)[10].

Rapid deployment of a new language pair has been approached by corpus-based systems in the past. Among the most recent attempts in the SMT community, we find Abdelali et al. (2006), Engelbrecht and Schultz (2005), and Lavie et al. (2005). Pinkham and Smets (2002) describe the same thing for a hybrid EBMT system. Both data driven approaches (SMT and EBMT) require large parallel corpora. Parallel corpora simply do not exist for many language combinations, and are scarce even for 'bigger' languages. They are an expensive resource that low-density languages such as Catalan cannot afford.

To overcome this problem, Gispert and Mariño (2006) present a Catalan-English SMT system, which does not use a Catalan-English parallel corpus. What they actually do is use Spanish as a bridge language. They are able to do without a parallel Catalan-English corpus, only by using two other parallel corpora: Catalan-Spanish and Spanish-English. Along similar lines, Pytlik and Yarowsky (2006) use French-English and Italian-English bitexts to train their Spanish-English system.

In our case, the METIS-II approach allows us to build a translation system between Catalan and English without resorting to any bilingual corpus at all. In the experiment that we have envisaged, since we keep

[10] We are currently working on yet another MT system, this time with Catalan as target language, and Spanish as source, which is already yielding some very interesting results.

English as the target, we can optimally reuse our existing Spanish-English system by simply plugging a PoS tagger for Catalan and a bilingual Catalan-English dictionary to the English generation part of the system.

8.1.1. *Catalan-English METIS-II system*
The choice of this particular language pair has been motivated by several factors:

- There are very few Catalan-English systems available. There is a commercial rule-based system (Translendium) and a couple of incipient research systems: the aforementioned Gispert & Mariño's and OpenMT, which is based on open source technologies (Alegria et al., 2005).

- Given our "Generation intensive approach", keeping English as target, already gives us a head start.

- Basic processing tools for Catalan are easily available to us.

The tagger and lemmatizer of our choice is CatCG (Alsina et al., 2002), a shallow morphosyntactic parser for Catalan, based on the Constraint Grammar formalism. It has been built on the Machinese Phrase Tagger from Connexor[11]. The output of the tagger is a string of Catalan lemmas or base forms, with disambiguated PoS tags and inflectional information. After lemmatizing and tagging has taken place, morphological tags are mapped into the Parole/EAGLES[12] tagset used by the dictionary, as explained in section 3.1 for the Spanish-English system. 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 token generation. Since the PoS tagger used for Spanish is also based on the Machinese Phrase Tagger and uses the same tagset, the module that performs the mapping can be reused from the Spanish preprocessing module.

To be able to reuse as much as possible the extraction scripts used for the Spanish-English dictionary, we keep the same format for the Catalan-English METIS-II dictionary, as well as the same tagsets. As an initial source dictionary we have chosen to use DACCO[13], an open-source, good-quality, but not very big Catalan-English dictionary that, at the moment, has 13,384 entries and 16,909 translations.

[11] http://www.connexor.com/
[12] http://garraf.epsevg.upc.es/freeling/doc/userman/parole-es.html
[13] http://www.catalandictionary.org/eng/

Table XII. results of evaluation in terms of BLEU, NIST and TER

Evaluation BLEU	Grammar	News	Science	Tech	All
METIS-II Cat-Eng	0.2059	0.2533	0.2070	0.2365	0.2342
METIS-II Sp-Eng	0.2241	0.3273	0.2876	0.2633	0.2941
Translendium Cat-Eng	0.3334	0.4406	0.4226	0.4264	0.4250
Evaluation NIST					
METIS-II Cat-Eng	4.4252	5.6894	4.7482	4.8189	5.7543
METIS-II Sp-Eng	4.9688	6.3122	5.9071	5.7074	6.7779
Translendium Cat-Eng	4.3013	7.2239	7.1815	7.0841	8.0044
Evaluation TER					
METIS-II Cat-Eng	49.238	56.372	63.358	60.447	51.894
METIS-II Sp-Eng	41.890	47.834	50.754	52.960	49.759
Translendium Cat-Eng	60.085	39.310	41.541	40.689	41.790

8.1.2. *Experiment and evaluation*

In order to test our rapidly assembled Catalan-English system, we use a test set of 200 sentences with a balanced distribution of four different text types (Grammar, Newspaper, Technical and Scientific).

The resulting translations have been evaluated using three automatic metrics: BLEU, NIST and the more recently proposed TER (Snover et al., 2006). The first two metrics measure edit distance between the machine-translated sentence and three human created references, while the TER measures the amount of editing that a human would have to perform to convert the MT output into the reference translation.

We compare these numbers with:

- The results obtained by the Spanish-English METIS-II system, on a similar test set (Vandeghinste et al., 2008);
- The results obtained on the same test set by the only existing commercial rule-based system for Catalan-English (Translendium).

The tables below show the results for the three metrics. The first row corresponds to the rapidly assembled Catalan-Spanish system; the second row, to the Spanish-English system developed during the METIS-II project and the third row corresponds to the rule-based Translendium.

The results that we have obtained are, as was our expectation, not far from the results obtained by the Spanish-English pair, although

they are not as good as the well-established, rule-based Translendium, which has years of development behind.

In our case, the METIS-II architecture has allowed us to assembly a new language pair, which compares well with the original system, in a very short time. Development time, which has been less than one person month has been employed mostly in obtaining and converting the dictionary and adapting the output of the tagger.

8.2. Adaptation to Europarl Domain

For testing the adaptability of the German METIS-II system to Europarl terminology and text, we extracted a set of 200 test translations and a set of 10,000 training translations from the Europarl corpus[14]. All sentences had at most 32 words. We did not consider aligned sentences pairs where one language side was empty. The 10,000 translations training set was used to estimate and train weights for our available dictionary entries. A further feature function would then take these weight into account to compute the most likely translations.

8.2.1. *Weighing the Dictionary*

Weighing of the transfer dictionary is only crucial for ambiguous entries in order to discriminate between different translation options for a German SL expression g. The weights of all other entries which have one single translation were set to 0.001. For entries with more than one translation option, weights $w(g \rightarrow e)$ were computed as follows. For each of the 10.000 German SL sentences we retrieved all matching entries $g \leftrightarrow e$ from the dictionary. We checked then for the English sides e whether they are covered in the corresponding TL sentence. A hit was assigned for dictionary entries where the German side g matches the SL sentence and the English side e matches the TL sentence of an alignment. We count as noise dictionary entries which match on the SL side but with no realization of the translation e in the TL side of the alignment. We then sum up hits $h(g \leftrightarrow e)$ and noise $n(g \leftrightarrow e)$ for the ambiguous entries over all 10.000 reference sentences.

$$w(g \rightarrow e) = \frac{h(g \leftrightarrow e)}{n(g \leftrightarrow e) + \sum_e h(g \leftrightarrow e)} \tag{8}$$

The weight $w(g \rightarrow e)$ in equation (8) was finally computed as the ratio of hits produced by the entry $g \leftrightarrow e$ divided by the noise of that entry plus the cumulated hits $H(g) = \sum_e h(g \leftrightarrow e)$ of all translation options for g. The weight w is thus a number $0 < w \leq 1$. It is 1 if an

[14] This test set is different from the set we used in the evaluation experiments as described in section 7

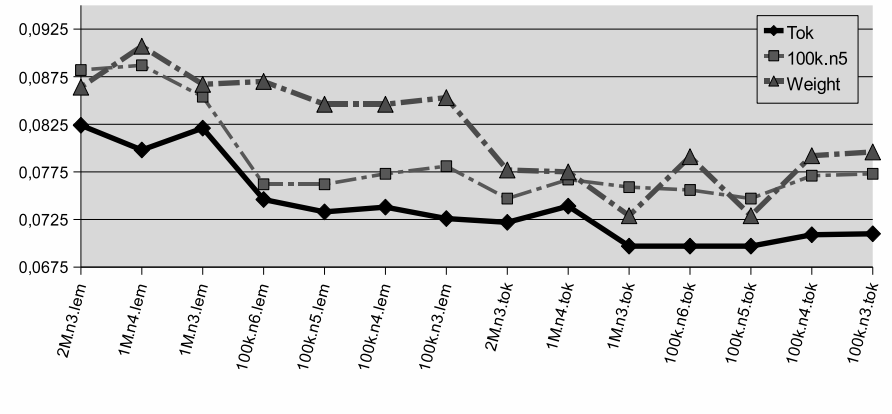

Figure 2. BLEU scores using various lemma (lem) and word-token (tok) based language models. The graph `100kn.5` adds a PoS tag language model to the lemma and token models; the graph `Weight` uses lexical weights.

entry has only hits and no other translation option of g was seen in the data. It is close to 0 if a dictionary entry produces mainly noise.

8.2.2. *Evaluation*

We started the evaluation experiments with using only one feature function, and then incrementally added further feature functions to see whether and how the system output improves. with the token-based and an lemma-based LM[15]. Second we added tag language models. The results can be seen in

Results can be seen in figure 2 which represents BLEU scores as obtained with the Europarl reference translations. The lower of the graphs (Tok) represents BLEU scores when using only one feature function, the lemma-based (lem) on the left side, or word-form based (tok) language models on the right side in the graph. These LMs were trained on sets of 100K, 1M and 2M sentences abitrarily extracted from the BNC, with 3, 4, 5 and 6-grams and they were generated with the SRI toolkit. Best performance is reached with a 3-gram lemma-based language model on 2M sentences, on the right side in figure 2.

The second graph (100k.5n) takes into account also a tag language model. We experimented with CLAWS5 tag LMs using 100K, 1M and 5M sentences and with $n = \{3, 4, 5, 6\}$. The graph plots the results when adding a 100k, 5-gram tag LM to the token models, which provided among the best results. As a tendency we observed that, using larger n-grams for tag-based LMs provides in many cases better results than increasing the size of the training corpus.

[15] See section 6.4 for an overview of these LMs.

The top graph (Weight) in figure 2 shows the impact on BLEU scores when adding lexical weights to the token and tag LMs as a third feature function. As can be seen, better results are produced, particularly in conjunction with the lemma LMs. This behavior could not be observed in another set of test sentences that contained unrelated terminology to the one the weights were trained on. Adding the lexical weights as an additional feature function to these test translations (cf. the development set from section 7) had no impact on the BLEU scores (Carl, 2008).

8.2.3. *Conclusions from statistical language modelling*
We resume our findings:

1. Lemma-based models produce better results than word-form based models. We find that (although not consistently) increasing the size of the training material for lemma models provides better results than increasing the length of the n-gram models.

2. Adding a tag language model improves the output in any case. Contrary to the findings for the word-forms and lemma models, larger values of n (in our case $n = 5$) may be an easier way to increase perform than to increase the size of the training set.

3. Lexical weights are suitable if the training material is similar to the texts to be translated (i.e. they are from the same domain).

9. Conclusions

The paper reports on the underlying ideas, implementation and results of the EU-FET MT project METIS-II running from October 2004 to September 2007. METIS-II aimed at translating free text input using basic linguistic resources and a monolingual target language corpus.

With only a limited amount of work (about 12 man years) we have developed four language pairs, Dutch, German, Greek and Spanish into English. While results of METIS-II are not as good as a well-established MT system such as Systran, which we have chosen as topline reference, they can be considered of an acceptable quality. The paper shows that METIS-II provides a solid framework that can be easily adapted to new language pairs, that can be tuned to particular domains, and that can be upgraded with additional resources as they become available.

The paper describes the language processing tools and bilingual dictionaries of METIS-II which rely on shallow linguistic representations.

Within METIS-II we have developed and explored various innovative language models and the paper points out how the models are exploited during translation. While we also give a comparative evaluation of the modules, we feel it is too early to draw ultimate conclusions on the best parameter settings.

We view METIS-II in the bigger context of self-learning systems that learn to translate from textual resources. Instead of learning relations between surface word forms, we maintain that the learned parameters must include linguistic properties of words and sentences for the system to tackle the hard problems of machine translation. Appropriate adaptive and dynamic representation of these parameters together with suitable reasoning mechanisms will ultimately help overcome the shortcomings of today's SMT systems. METIS-II has explored some of the possible avenues, and pointed to further directions that can be followed.

References

Abdelali, A., J. Cowie, S. Helmreich, W. Jin, M. P. Milagros, B. Ogden, H. M. Rad, and R. Zacharski: 2006, ' Guarani: a case study in resource development for quick ramp-up MT'. In: *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas, "Visions for the Future of Machine Translation"*. Cambridge, Massa-chusetts, USA, pp. 1–9.

Alegria, I., A. D. de Ilarraza, G. Labaka, M. Lersundi, A. Mayor, K. Sarasola, M. L. Forcada, S. Ortiz-Rojas, and L. Padró: 2005, ' An open architecture for transfer-based machine translation between Spanish and Basque'. In: *Proceedings of the X Machine Translation Summit workshop OSMaTran: Open-Source Machine Translation X*. Phuket, Thailand, pp. 7–14.

Alsina, A., T. Badia, G. Boleda, S. Bott, A. Gil, M. Quixal, and O. Valentí: 2002, 'CATCG: a general purpose parsing tool applied'. In: *Proceedings of Third International Conference on Language Resources and Evaluation*. Las Palmas, Spain, pp. 1130–1134.

Anastasiou, D. and O. Culo: 2007, 'Using Topological Information for detecting idiomatic verb phrases in German'. In: *Proceedings of the Conference on Practical Applications in Language and Computers (PALC)*. Lodz, Poland, pp. 49–58.

Badia, T., G. Boleda, M. Melero, and A. Oliver: 2005, 'An n-gram approach to exploiting a monolingual corpus for machine translation'. In: *MT Summit X Workshop on Example-Based Machine Translation*. Pukhet, Thailand, pp. 1–7.

Badia, T., M. Melero, and O. Valentín: 2008, 'Rapid Deployment of a New METIS Language Pair: Catalan-English'. In: *Proceedings of the Sixth International Language Resources and Evaluation (LREC08)*. Marrakech, Morocco, p. 96.

Boutsis, S., P. Prokopidis, V. Giouli, and S. Piperidis: 2000, 'A Robust Parser for Unrestricted Greek Text'. In: *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC)*. Athens, Greece, pp. 467–482.

Brants, T.: 2000, 'TnT – a statistical part-of-speech tagger'. In: *Proceedings of the 6th Applied Natural Language Processing Conference (ANLP)*. Seattle, Washington, USA, pp. 224–231.

Brown, R. and R. Frederking: 1995, 'Applying statistical English language modelling to symbolic machine translation'. In: *Proceedings of the 6th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI)*. Leuven, Belgium, pp. 221–239.

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

Carl, M.: 2008, 'Using Log-linear Models for Tuning Machine Translation Output'. In: *Proceedings of the Sixth International Language Resources and Evaluation (LREC08)*. Marrakech, Morocco, p. 49.

Carl, M. and E. Rascu: 2006, 'A Dictionary Lookup Strategy for Translating Discontinuous Phrases'. In: *Proceedings of the European Association for Machine Translation*. Oslo, Norway, pp. 49–58.

Carl, M., P. Schmidt, and J. Schütz: 2005, 'Reversible Template-based Shake & Bake Generation'. In: *Proceedings of the Example-Based Machine Translation Workshop held in conjunction with Machine Translation Summit X*. Phuket, Thailand, pp. 17–26.

Carpuat, M. and D. Wu: 2007, 'How Phrase Sense Disambiguation outperforms Word Sense Disambiguation for Statistical Machine Translation'. In: *Proceedings of the 11th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-07)*. Skövde, Sweden, pp. 43–52.

Doddington, G.: 2002, 'Automatic Evaluation of Machine Translation Quality using N-gram Co-occurrence Statistics'. In: *Proceedings of the second Human Language Technologies Conference (HLT-02)*. San Diego, pp. 128–132.

Dologlou, I., S. Markantonatou, G. Tambouratzis, O. Yannoutsou, A. Fourla, and N. Ioannou: 2003, 'Using Monolingual Corpora for Statistical Machine Translation'. In: *Proceedings of EAMT/CLAW 2003*. Dublin, Ireland, pp. 61–68.

EAGLES: 1994, 'Guidelines, EAG-LWG-T4-2'. Technical report, ILC-CNR, Pisa, Italy.

Engelbrecht, H. and T. Schultz: 2005, 'Rapid development of an Afrikaans English speech-to-speech translator'. In: *International Workshop on Spoken Language Translation: Evaluation Campaign on Spoken Language Translation*. Pittsburgh, PA, US., pp. 24–25.

Germann, U., M. Jahr, K. Knight, D. Marcu, and K. Yamada: 2001, 'Fast Decoding and Optimal Decoding for Machine Translation'. In: *Proceedings of the 39th ACL and 10th Conference of the European Chapter*. Toulouse, France, pp. 228–235.

Gispert, A. and J. B. Mariño: 2006, 'Catalan-English statistical machine translation without parallel corpus: bridging through Spanish'. In: *Fifth International Conference on Language Resources and Evaluation (LREC), 5th SALTMIL Workshop on Minority Languages: "Strategies for developing machine translation for minority languages"*. Genoa, Italy, pp. 65–68.

Habash, N.: 2004, 'The use of a structural n-gram language model in generation-heavy hybrid machine translation'. In: *Proceeding 3rd International Conference on Natural Language Generation (INLG '04), volume 3123 of LNAI, Springer, Germany*. Brockenhurst, UK, pp. 61–69.

Karlsson, F. e. a.: 1995, *Constraint Grammar: A Language-Independent Formalism for Parsing Unrestricted Text*. Berlin/New York: Mouton de Gruyter.

Koehn, P.: 2004, 'Pharaoh: a Beam Search Decoder for Phrase-Based Statistical Machine Translation Models'. In: *Proceedings of AMTA, the Association for Machine Translation in the Americas*. Washington, DC, USA., pp. 115–124.

Koehn, P.: 2005, 'Europarl: A Parallel Corpus for Statistical Machine Translation'. In: *Proceedings of MT Summit X*. Pukhet, Thailand, pp. 79–86.

Kuhn, H. W.: 1955, 'The Hungarian Method for the assignment problem'. *Naval Research Logistics Quarterly* **2**, 88–97.

Labropoulou, P., E. Mantzari, and M. Gavrilidou: 1996, 'Lexicon – Morphosyntactic Specifications: Language-Specific Instantiation (Greek)'. In: *PP-PAROLE, MLAP report*. Athens, Greece, pp. 63–386.

Langkilde, I. and K. Knight: 1998, 'The Practical Value of n-grams in generation'. In: *In Proceedings of the 9th International Natural Language Workshop (INLG '98)*. Niagara-on-the-Lake, Ontario, pp. 248–255.

Lavie, A., E. Peterson, K. Probst, S. Wintner, and Y. Eytani: 2004, 'Rapid prototyping of a transfer-based Hebrew-to-English machine translation system'. In: *Proceedings of the Tenth Conference on Theoretical and Methodological Issues in Machine Translation*. Baltimore, USA, pp. 1–10.

Maas, H.-D.: 1996, 'MPRO - Ein System zur Analyse und Synthese deutscher Wörter'. In: R. Hausser (ed.): *Linguistische Verifikation, Sprache und Information*. Tübingen: Max Niemeyer Verlag.

Majithia, H., P. Rennart, and E. Tzoukermann: 2005, 'Rapid ramp-up for statistical machine translation: minimal training for maximal coverage '. In: *Proceedings of the Machine Translation Summit X*. Phuket, Thailand, pp. 438–444.

Markantonatou, S., S. Sofianopoulos, V. Spilioti, G. Tambouratzis, M. Vassiliou, and O. Yannoutsou: 2006, 'Using Patterns for Machine Translation (MT)'. In: *Proceedings of the European Association for Machine Translation 2006*. Oslo, Norway, pp. 239–246.

Melero, M., A. Oliver, T. Badia, and T. Suñol: 2007, 'Dealing with Bilingual Divergences in MT using Target language N-gram Models'. In: *Proceedings of the METIS-II Workshop: New Approaches to Machine Translation, CLIN 17 - Computational Linguistics in the Netherlands*. Leuven, Belgium, pp. 19–26.

METIS-II: 2006, 'Validation/Evaluation framework'. Public Report, D5.1, European Commission, FP6-IST-003768, Brussels. `http://www.ilsp.gr/metis2/files/Metis2_D5.1.pdf` [25.Aug.2008].

METIS-II: 2007, 'Validation & Fine-Tuning Results for the first Prototype'. Public Report, D5.2, European Commission, FP6-IST-003768, Brussels. `http://www.ilsp.gr/metis2/files/Metis2_D5.2.pdf` [25.Aug.2008].

Müller, F. H.: 2004, 'Stylebook for the Tübingen Partially Parsed Corpus of Written German (TÜPP-D/Z)'. `http://www.sfb441.uni-tuebingen.de/a1/pub.html`[25.Aug.2008].

Munkres, J.: 1955, 'Algorithms for the Assignment and Transportation Problems'. *Journal of the Society of Industrial and Applied Mathematics* **5(1)**, 32–38.

Och, F. J. and H. Ney: 2002, 'Discriminative Training and Maximum Entropy Models for Statistical Machine Translation'. In: *Proceedings of the 40th annual ACL Conference*. Philadelphia, PA, pp. 295–302.

Papineni, K., S. Roukos, T. Ward, and W. Zhu: 2002, 'BLEU: a method for automatic evaluation of machine translation'. In: *Proceedings of the 40th ACL*. pp. 311–318.

Pinkham, J. and M. Smets: 2002, 'Modular MT with a learned bilingual dictionary: rapid deployment of a new language pair'. In: *Coling*. Taipei, Taiwan, pp. 800–806.

Pytlik, B. and D. Yarowsky: 2006, 'Machine translation for languages lacking bitext via multilingual gloss transduction'. In: *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas, "Visions for the Future of Machine Translation"*. Cambridge, Massachusetts, USA, pp. 156–165.

Snover, M., B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul: 2006, 'A Study of Translation Edit Rate with Targeted Human Annotation'. In: *Proceedings of Association for Machine Translation in the Americas (AMTA 2006)*. pp. 223–231.

Szopa, R.: 2007, 'LRBL. A Rule-Based Lemmatizer (with rules for Dutch)'. Technical report, Centre for Computational Linguistics, Leuven, Belgium.

Tambouratzis, G., S. Sofianopoulos, V. Spilioti, M. Vassiliou, O. Yannoutsou, and M. S.: 2006, 'Pattern matching-based system for Machine Translation (MT)'. In: *Proceedings of Advances in Artificial Intelligence: 4th Hellenic Conference on AI, SETN 2006*. Heraklion, Crete, pp. 345–355.

Van Eynde, F. .: 2004, 'Part of Speech Tagging en Lemmatisering van het Corpus Gesproken Nederlands'. Annotation protocol, Centrum voor Computerlinguïstiek, Leuven, Belgium.

Vandeghinste, V.: 2005, 'Manual for ShaRPa 2.1'. User manual, Centre for Computational Linguistics, Leuven, Belgium.

Vandeghinste, V.: 2008, 'A Hybrid Modular Machine Translation System'. Phd thesis, Netherlands Graduate School of Linguistics, Leuven, Belgium.

Vandeghinste, V., P. Dirix, I. Schuurman, S. Markantonatou, S. Sofianopoulos, M. Vassiliou, O. Yannoutsou, T. Badia, M. Melero, G. Boleda, M. Carl, and P. Schmidt: 2008, 'Evaluation of a Machine Translation System for Low Resource Languages: METIS-II'. In: *Proceedings of the Sixth International Language Resources and Evaluation (LREC)*. Marrakech, Morocco, p. 96.

Vossen, P., L. Bloksma, and B. P.: 1999, 'The Dutch Wordnet'. Technical report, University of Amsterdam, Amsterdam, NL.

Whitelock, P.: 1991, 'Shake-and-Bake Translation'. Unpublished Draft.

Whitelock, P.: 1992, 'Shake-and-Bake Translation'. In: *Proceedings of the COLING92*. Nantes, France, pp. 784–791.

Zwarts, S. and M. Dras: 2007, 'Syntax-based Word Reordering in Phrase-Based Statistical Machine Translation; Why Does it Work?'. In: *Proceedings of MT Summit XI*. Copenhagen. Denmark, pp. 559–566.