Evaluation of a Machine Translation System for Low Resource Languages: METIS-II

Vincent Vandeghinste[*], Peter Dirix[*], Ineke Schuurman[*], Stella Markantonatou[†], Sokratis Sofianopoulos[†], Marina Vassiliou[†], Olga Yannoutsou[†], Toni Badia[‡], Maite Melero[‡], Gemma Boleda[‡], Michael Carl[◇], Paul Schmidt[◇]

[*] Centre for Computational Linguistics - KULeuven, Belgium
vincent.vandeghinste, peter.dirix, ineke.schuurman@ccl.kuleuven.be

[†] Institute for Language and Speech Processing, Athens, Greece
marks, s_sofian, mvas, olga@ilsp.gr

[‡] GliCom Fundació Barcelona Media - UPF, Spain
toni.badia, maite.melero, gemma.boleda@upf.edu

[◇] IAI - Saarbrücken, Germany
carl,paul@iai.uni-sb.de

Abstract

In this paper we describe the METIS-II system and its evaluation on each of the language pairs: Dutch, German, Greek, and Spanish to English. The METIS-II system envisaged developing a data-driven approach in which no parallel corpus is required and in which no full parser or extensive rule sets are needed. We describe the evaluation on a development test set and on a test set taken from Europarl, and compare our results with SYSTRAN. We also provide some further analysis, namely researching the impact of the number and source of the reference translations and analysing the results according to test text type. The results are expectably lower for the METIS system, but not at an unattainable distance from a mature system like SYSTRAN.

1. Introduction

Within the European context, the importance of supporting and maintaining a multilingual society is apparent, and machine translation should be considered an important activity in such a society. Therefore, a need arises to develop machine translation systems between all European languages. Some of these languages are rather small, and for these languages not many resources or tools are available.

Current approaches to machine translation in industry are still mainly rule-based (RBMT), requiring lots of expensive manual labour in building parsers and transfer rules. It is not economically viable to develop a full RBMT system for smaller languages, although the first steps might be rule-based.

In academia, most current approaches are data-driven (statistical and example-based MT). These data-driven approaches require large parallel corpora, so they offer no solution for building MT systems for smaller languages, since these parallel corpora are simply unavailable, too small, or too restricted.

In the METIS-II project[1], we envisaged developing a data-driven approach in which no parallel corpus is required, and in which no full parser or extensive rule sets are needed, so that the approach can be used for lower resource languages. The main idea was first investigated in Dologlou et al. (2003) while the system has been described more extensively in Vandeghinste et al. (2006).

We have built a prototype system for Dutch, German, Greek, and Spanish as source languages, and English as target language. Although for these languages quite some expensive tools and resources are available, we did not use them.

A somewhat similar approach to translation is the MATADOR system (Habash and Dorr, 2002, 2003; Habash, 2003, 2004). The main difference between MATADOR and METIS-II is the fact that MATADOR aims at language pairs with resource asymmetry: low resources for the source language, and high resources for the target language, whereas the METIS-II approach aims at low resources on both sides. We use much less resources on the target side than MATADOR.

Besides this, MATADOR uses a deep parser for the source language whereas METIS-II uses at most only a shallow parser.

2. The METIS-II system

In this section we describe the METIS system in general terms. The system has been described more extensively in Vandeghinste et al. (2006). For the different language pairs, different experimental conditions were investigated. These are described in detail in Melero and Badia (2007) for Spanish to English, in Carl (2007) for German to English, in Markantonatou et al. (2007) for Greek to English, and in Dirix et al. (2006) and Vandeghinste (2008) for Dutch to English.

2.1. Source Language Analysis

The source language analysis that is performed consists minimally of part-of-speech tagging and lemmatization.

[1] Supported by the 6th European Framework Programme, FP6-IST-003768

For Greek, German, and Dutch, shallow parsing is performed as well.
Part-of-speech taggers might not be available for any low resource language, but by using a trainable part-of-speech tagger, like TnT (Brants, 2000), and applying it on a tagged corpus, one can obtain a good quality tagger. Of course, this would require a manually corrected part-of-speech tagged corpus. Another solution would be to build a rule-based tagger. Anyhow, the development of such a part-of-speech tagger would be reusable in other NLP applications, and can be considered a basic NLP tool.
We use lemmatization in source language analysis as we use lemmas throughout the translation process. There are two main reasons for this. First, our dictionaries are lemma-based, allowing us to abstract away from specific surface forms of words, so reducing the number of entries in our dictionary compared to a full form dictionary. Second, by matching lemmas with the target language corpus instead of full forms, the data becomes less sparse.
We use part-of-speech tagging as it can help in dictionary lookup, when the dictionary contains part-of-speech information, not confusing between homonyms with a different part-of-speech. The parts-of-speech contain additional information about features like tense, number etc., which are no longer contained in the lemma. As the translations are lemma-based, we need this information to generate the appropriate tokens in the target language.
As sentence lookup in a target language corpus would inevitably result in too sparse data, we chunk up the sentence in smaller bits. For Greek, German, and Dutch, linguistically meaningful chunks are used as translation units, whereas for Spanish, n-gram chunks are used.

2.2. From Source to Target Language

The transition from source to target is made through the following channels:

1. Dictionary lookup is performed: all entries from the source language sentence are translated
2. Tag mapping: the source language part-of-speech tags need to be converted into target language part-of-speech tags in order to allow generation of the correct surface form of the lemmas in the target language with respect to features like number, tense, etc.
3. Structure mapping: through a limited set of possibly weighted transfer rules we map the source language structure onto a more appropriate structure for the target language. This is used especially for mapping verb tenses, and phenomena like do-insertion.

While structure mapping is not strictly necessary in the METIS design, Vandeghinste et al. (2007) have shown that it has a positive effect on BLEU and NIST scores.
These transitions result in a number of translation candidates for each chunk and for each sentence. Different translations of a source language word will result in different translation candidates. As regards different word orders, either the core engine is fed with a disjunction of possible word orders (Spanish, Dutch, German $\rightarrow$ English) or final word order is defined based on similarity scores.
According to the richness of the SL analysis a distinction can be made in the way transfer and generation is processed in METIS-II. While for Spanish only single lemmas and POS tags are mapped into the TL, for Dutch and Greek the SL structure is also transferred. Due to the great number of discontinuous constituents, for German we have also experimented with mapping and transferring of discontinuous lexical units. The SL structure and POS tags are not mapped into the TL in our German experiments.

2.3. Target Language Generation

Reordering of the transferred items into TL structure is conceived as a process of hypothesis generation and filtering. For Dutch, Greek, and Spanish we have experimented with a greedy approach, in which a set of partial hypotheses is immediately evaluated and only the (n-) best hypotheses are kept for further investigation and refinement. For German we have tested a beam search algorithm, which stores all partial hypotheses in an AND/OR graph for final evaluation. For all language pairs, filtering (i.e. evaluation) of the hypotheses is based on language models which were previously generated from the BNC. Generation of hypotheses and their greedy filtering is top-down for Greek and bottom-up for Dutch and Spanish. The generation of reordering hypotheses can be rule-based (Dutch and German) or/and it can be pattern-based (Greek, Spanish), while the reordering patterns themselves may be based on information from the SL (Dutch, German) or on their transferred tags and structures (Dutch, Greek and Spanish).
As all processing steps are lemma-based, these lemmas need to be converted to tokens, which is done on the basis of their part-of-speech. For this we use the token generator (reversible lemmatizer in reverse mode) from Carl et al. (2005).

3. Evaluation

3.1. Methodology

The evaluation proceeded by translating each of two 200-sentence test sets with the SYSTRAN and METIS-II MT systems and evaluating the resulting translations with 3 different standard metrics. We next explain each of these points in more detail.

3.1.1. Test Sets

The final evaluation was performed on two test sets, one consisting of data that has been used throughout the project for development purposes and one consisting of unseen data gathered from a previously existing bilingual corpus.

Development test set A parallel development test set was established for all language pairs. This test set consisted of 200 sentences, with material evenly distributed among four different categories:

- 56 sentences illustrating grammatical phenomena (defined by each site), for instance for German:
  - lexical translation problems: separable prefixes, fixed verb constructions, degree of adjectives and adverbs, lexical ambiguities, and others.

– syntactic translation problems: nominalisation, determination, word order, different complementation, relative clauses, tense/aspect, head switching, prepositions, category change, and others.

- 48 sentences from newspapers;
- 48 sentences from encyclopaedia articles, or similar sources of non-specialised scientific text;
- 48 sentences from technical manuals, or similar sources of technical text.

Vocabulary and syntactic constructions used in these sentences belongs to general language (as opposed to being exclusively technical, for example), but do not necessarily appear in the target corpus used for the system (BNC) or in each of the bilingual dictionaries.
Each site had three different human translators prepare three English reference translations of the test material for evaluation purposes.

Unseen Data: Europarl test set As the development test set has been used to fine-tune the systems throughout the project, we have also developed an independent test set using data from an already existing bilingual corpus, namely Europarl (Koehn, 2005). This corpus consists of transcriptions of debates in the European Parliament. It is chosen because it is widely used by the MT research community, particularly to build statistical MT systems, and because it contains material in all the language pairs involved in the project.
The corpus was found to contain noisy data, particularly wrong alignments. Therefore, the material was subject to manual validation. We chose 200 sentences from the test set used in Koehn et al. (2003), corresponding to the Q4/2000 portion of the data (2000-10 to 2000-12), that had correct alignments for the 4 languages of the project.
Each consortium partner had a professional translator translate the sentences 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, which facilitates the proper use of the evaluation metrics described next. Note that the number of reference translations is higher for the Europarl test set than for the development test set, which should, in principle, favour the scores of the Europarl test set (see section 3.2.3. for an analysis of this issue).

3.1.2. Comparing with SYSTRAN

We chose SYSTRAN 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.. SYSTRAN 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.

3.1.3. Used Metrics

As automated metrics we use BLEU (Papineni et al., 2002), NIST (Doddington, 2002), and TER (Snover et al., 2006).

3.2. Results

3.2.1. 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

Dutch $\rightarrow$ English Table 1 and table 2 show the scores for the Dutch to English language pair, for the development and Europarl test sets, respectively.

Table 1: NL-EN: Results for development test set

	METIS-II	SYSTRAN
BLEU	0.2369	0.3777
NIST	6.19	7.28
TER	59.52	38.81

Table 2: NL-EN: Results for Europarl test set

	METIS-II	SYSTRAN	Z&D
BLEU	0.1925	0.3828	0.2070
NIST	5.98	7.99	–
TER	60.92	44.66	–

The development test set contains sentences for which we adapted the dictionary, and it was used for debugging purposes, but it still contains several phenomena which are not yet covered by the current implementation.
For the Europarl test set, no adaptations to our system were made, and no dictionary entries were added or changed.
For both test sets, the result on SYSTRAN (Professional version) outperform the results in our approach, but of course, SYSTRAN is much more developed (cf. section 3.1.2.).
The SYSTRAN results show that there is no difference in translation difficulty in the sentences in our development set vs. our training set, as they have more or less the same BLEU and NIST scores.
A more fair comparison can be made with the work presented by Zwarts and Dras (2007; Z&D column in table 2). They have trained a statistical MT system on the Europarl corpus, and have extracted a test set from that corpus. They

report a BLEU score of 0.207. It is not clear whether they excluded their test set from their training set.

When we compare these results with the results we had on our 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.

It should be admitted that there are still some bugs in our prototype, which we will try to solve in the future, so better results can still be expected, without making changes to the techniques applied. Extra evidence for this claim are the results that were presented in Vandeghinste et al. (2007) which were calculated on a test set which was extensively used for debugging, and for which we found BLEU scores between 0.3024 and 0.3486, depending on the experimental condition.

German → English Table 3 shows the scores for the German to English language pair computed on the development test set.

Table 3: DE-EN: Results for development test set

	exp1	exp2	SYSTRAN
BLEU	0.1861	0.2231	0.3133
NIST	5.4801	5.3193	6.3644
token model	6M-n3	5M-n3	–
tag model	6M-n3	5M-n7	–

We report the results of the system in two different experiments. In the first experiment (exp1 in table 3), we took the expander rules from the basic system and varied feature weights between 0.01 and 10, using lemma language models with 3- and 4-grams and tag language models with 4-, 5-, 6-, and 7-grams. With the best combination of language models and weights we obtained a BLEU value of 0.1861. On a 1GB/2.8GHz, single core Linux machine it takes less than 4 minutes to translate the 200 sentences.

In the second experiment (exp2 in table 3), we further developed and refined some expander rules for handling adverbs and negation particles, such as 'never', 'usually', extraposition of prenominal adjectives (e.g., "der vom Baum gefallene Apfel" would become "The apple fallen from the tree"), and "um ... zu" constructions. We used 50 sentences from an earlier experiment for fine-tuning the system and tested on the development set of 200 sentences. The BLEU score increased to 0.2231; however, as can be seen in the table, NIST values decreased slightly.

The public version of SYSTRAN (Babelfish), however, outperforms our efforts. Their results on the same test set can be seen in the last column in table 3.

Table 4 shows the scores for the German to English language pair computed on the Europarl test set.

In the Europarl test set, SYSTRAN clearly outperform METIS-II, i.e. SYSTRAN's BLEU score of 0.3958 is about 30% better than METIS-II BLEU of 0.2816.

Table 4: DE-EN: Results for Europarl test set

	METIS-II	SYSTRAN
BLEU	0.2816	0.3958
NIST	6.6854	8.0473
TER	55.97	42.93

Greek → English Table 5 illustrates the scores obtained for the Greek to English language pair when evaluating the output for the development test set.

Table 5: EL-EN: Results for development test set

	METIS-II	SYSTRAN
BLEU	0.3661	0.3946
NIST	7.2645	7.7041
TER	48.256	37.258

According to the BLEU and NIST metrics, it is evident that both systems generate translations of a broadly comparable quality. When using the TER metric, SYSTRAN gives better translations than METIS-II, receiving a total score of 37.258. In 48 cases METIS-II achieves better scores, while in 41 cases the same scores are obtained for both systems. In total, METIS-II outperforms SYSTRAN in 24% of the sentences of the test set, while SYSTRAN generates better translations in 56% of the sentences.

Table 6 illustrates the scores obtained for the Greek to English language pair when evaluating the output for the Europarl test set.

Table 6: EL-EN: Results for Europarl test set

	METIS-II	SYSTRAN
BLEU	0.1861	0.3132
NIST	6.1658	7.6867
TER	64.959	50.747

According to the BLEU and NIST metrics, SYSTRAN outperforms METIS-II. This behaviour is more pronounced in the case of BLEU, while for NIST the results are more similar.

According to the TER metric, SYSTRAN again scores better than METIS-II, receiving a total score of 50.747. In 42 cases METIS-II achieves better scores, while in 26 cases the same scores are obtained for both systems. In total, METIS-II outperforms SYSTRAN in 21% of the sentences of the test set, while SYSTRAN generates better translations in 66% of the sentences.

According to table 5 and 6, it can be seen that in general SYSTRAN has a higher performance than METIS-II. The relatively poor performance of METIS-II in both test sets is probably attributable to several reasons. First of all, its output is compared to an established commercial MT system,

which has obviously been developed more extensively (cf. section 3.1.2.).

Second, METIS-II is still a prototype version, and thus has some difficulty in handling unrestricted text of high complexity, while its lexical coverage is still relatively limited. Its core engine has so far been designed, and respectively developed, with the view to handle specific syntactic phenomena. Therefore, it is only expected to be outperformed by mature MT systems such as SYSTRAN.

A closer examination of METIS-II results (Greek-to-English) has shown that, apart from fine-tuning issues such as the weight adaptation to different registers, there exist quite a few areas that may potentially lead to a substantial improvement.

To this end, we have further studied the output of METIS-II, in order to check to which extent its inferior performance is due to its principles (i.e. core engine) or to peripheral modules/resources such as the token generation and the lexicon. More specifically, for both the test sets all the translations produced have been manually corrected regarding the tokens. Then, a second evaluation experiment has been conducted on the basis of the new translation outputs. The respective scores obtained (table 7 and table 8) indicate that, even though SYSTRAN still performs better as a whole, a substantial improvement is noticeable in the output quality, especially in the case of the development test set.

According to table 7, METIS-II achieves higher scores for both the BLEU and NIST metrics, while it has a still lower, though improved, performance based on TER. On the contrary, for the EUROPARL test set, SYSTRAN outperforms METIS-II for all three metrics. This can be probably attributed to the fact that the specific corpus is unconstrained and contains more diverse phenomena than both those studied during the project lifetime and those included within the development test set.

Table 7: EL-EN: Results for development test set (with corrections)

	METIS-II	SYSTRAN
BLEU	0.4147	0.3946
NIST	7.7962	7.7041
TER	44.857	37.258

Table 8: EL-EN: Results for Europarl test set (with corrections)

	METIS-II	SYSTRAN
BLEU	0.1949	0.3132
NIST	6.3846	7.6867
TER	64.319	50.747

Spanish → English Table 9 and table 10 show the scores for the Spanish to English language pair, for the development and Europarl test sets, respectively.

Table 9: ES-EN: Results for development test set

	METIS-II	SYSTRAN
BLEU	0.2941	0.4634
NIST	6.7779	8.5056
TER	49.759	36.163

Table 10: ES-EN: Results for Europarl test set

	METIS-II	SYSTRAN
BLEU	0.2784	0.4638
NIST	6.6057	8.6241
TER	54.241	37.015

In all conditions, as could be expected, SYSTRAN shows a better performance on the automatic metrics.

A regards the results of the Spanish-English METIS-II on the two testset, it shows a slightly better performance for the development test set than for the Europarl test set. However, the difference is not large (0.016 points for BLEU, 0.17 for NIST, 4.5 for TER), which shows that the system's output is quite stable and not too dependent on fine-tuning for a specific test suite.

On absolute terms, neither of the systems performs satisfactorily as a stand-off tool. Taking BLEU as a reference, SYSTRAN achieves less than half the optimal performance, while METIS-II achieves only slightly more than one quarter of the optimal performance. Both systems, thus, should be regarded as a translation aid, rather than as a translation solution on their own.

The differences between METIS-II and SYSTRAN are quite large: 0.17 points for Bleu, 1.73 for NIST and 13.60 for TER, in the development set. These differences are slighly larger for the Europarl set: 0.19 (BLEU), 2.02 (NIST), and 17.23 (TER). In average, thus, SYSTRAN performs between 30 and 40% better than METIS-II. It is a large, significant difference. However, as mentioned above, we have to take into account the development times of METIS-II and SYSTRAN (cf. section 3.1.2.). SYSTRAN's performance for the Spanish-English pair is particularily good with respect to other language pairs (see section 3.2.2.).

3.2.2. Overall scores: a cross-language summary

The results for each language pair have been independently presented in section 3.2.1.. In this section we put together the different results achieved by SYSTRAN and METIS-II for the different language pairs on both test sets, on four separate tables. We concentrate on the BLEU metric because it is the most standardly used in current MT research (the scores for the other metrics can be checked in the tables in section 3.2.1.).

Table 11 compares the cross-language results for both systems using the Europarl corpus.

The third column shows the difference between METIS-II and SYSTRAN for each of the language pairs, measured on the Europarl test set. SYSTRAN visibly outper-

Table 11: Cross-language results on the Europarl test set (BLEU)

	METIS-II	SYSTRAN	difference
NL-EN	0.1925	0.3828	0.1903
DE-EN	0.2816	0.3958	0.1142
EL-EN	0.1861	0.3132	0.1271
ES-EN	0.2784	0.4638	0.1854

form METIS-II for all languages by around a 30% in the BLEU scores, as has become already clear in section 3.2.1.. The differences across languages are quite homogeneous, showing a variation of only 0.0761 between the result of the METIS-II language pair that performs closest to SYSTRAN (DE-EN) and the most distant one (NL-EN).
Table 12 compares the cross-language results for both systems using the development corpus.

Table 12: Cross-language results on the development test set (BLEU)

	METIS-II	SYSTRAN	difference
NL-EN	0.2369	0.3777	0.1408
DE-EN	0.2231	0.3133	0.0902
EL-EN	0.3661	0.3946	0.0285
ES-EN	0.2941	0.4634	0.1693

The third column in this table shows that differences between METIS-II and SYSTRAN are smaller when measured on the development set in all cases. The general impression is that - not altogether unexpectedly- METIS-II performs better (i.e. closer to SYSTRAN) on the development set than on the Europarl test set. This is true for all language pairs but one (DE-EN), as will become apparent in table 13.
Table 13 compares the cross-language results of the METIS-II system on both test sets.

Table 13: Cross-language results for METIS-II on both test sets (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

This table 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 results 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 its weights. Also, the DE-EN development set may be harder to machine translate, since SYSTRAN also performs more poorly on it than on the Europarl test set (see Table 14).

In order to help clarify the whole picture, we present SYSTRAN's performance on the two test sets in 14.

Table 14: Cross-language results for SYSTRAN on both test sets (BLEU)

	Europarl	development	difference
NL-EN	0.3828	0.3777	0.0051
DE-EN	0.3958	0.3133	0.0825
EL-EN	0.3132	0.3946	-0.0814
ES-EN	0.4638	0.4634	0.0004

These results mirror the current state of the MT system, as neither lexicon nor rule set has been equally developed for all language pairs available for the SYSTRAN MT system. It is worth mentioning that, while the performance of SYSTRAN is very stable for ES-EN and NL-EN, it shows a greater variation for the other two language pairs. For German, SYSTRAN, as METIS-II, performs better on Europarl. For Greek, the opposite is true: SYSTRAN (and METIS-II) scores much better on the development test set. No explanation readily comes to mind, apart from the differing characteristics of the languages themselves and the degree of development of SYSTRAN.

3.2.3. Further Analysis

This section contains an analysis of two aspects of the METIS-II system having to do with the final evaluation, beyond the overall scores presented in sections 3.2.1. and 3.2.2. We offer an analysis of the impact of the number and type of reference translations in the evaluation scores of the German-English translation task, and an analysis of the differences in scores across text types in the development test set for the Spanish-English translation task.

Impact of the number and source of reference translations For the German-English translation task, the following five reference translations were available for the Europarl test set:

- *ep*: English translation as provided with the Europarl corpus.
- four manual translation into English for the four source languages were provided by the consortium:
  - *de*: German
  - *es*: Spanish
  - *nl*: Dutch
  - *el*: Greek

In order to observe the impact of the number and origin of the reference translations on the outcome of the BLEU and NIST score, the evaluation was carried out on different sets of these reference translations:

- each of the five reference translations *ep, de, es, nl, el*
- *5-de*: all minus *de* i.e. *ep,es,nl,el*
- *5-es*: all minus *es* i.e. *ep,de,nl,el*

- *5-nl*: all minus *nl* i.e. *ep,de,es,el*
- *5-el*: all minus *el* i.e. *ep,de,es,nl*
- *all*: using all five reference translations *ep,de,es,nl,el*. This is the setting used for the overall evaluation in section 3.2.1..

Table 15 shows the scores obtained under each of the specified conditions.

Table 15: DE-EN: Impact of the number of reference translations on the scores for the Europarl test set

		METIS-II		SYSTRAN	
Row	Ref	NIST	BLEU	NIST	BLEU
1	*el*	2.7556	0.0761	3.5443	0.1182
2	*nl*	3.4391	0.0923	4.3278	0.1483
3	*ep*	3.5441	0.1155	4.7178	0.1975
4	*es*	3.6323	0.1199	4.7423	0.1922
5	*de*	5.1899	0.2376	5.7767	0.2912
6	*5-de*	5.1745	0.1803	6.9000	0.3064
7	*5-es*	6.4458	0.2697	7.6893	0.3739
8	*5-nl*	6.4525	0.2750	7.7317	0.3817
9	*5-el*	6.5057	0.2774	7.8182	0.3871
10	*all*	6.6854	0.2816	8.0473	0.3958

The table shows that the worst results are obtained with only one set of reference translations (rows 1-5), while best results are obtained with all 5 reference sets (row 10, corresponding to the overall evaluation in table 4), as could be expected given the way the metrics work.
When looking at single reference sets, it is interesting to note that the *de* set (i.e., the set of manual reference translations produced from the German source sentences) yields for the German METIS-II translations by far the best BLEU results of 0.2376 and 0.2912 for METIS-II and SYSTRAN, respectively, while the worst results are obtained with the el set, with BLEU scores of 0.0761 (METIS-II) and 0.1182 (SYSTRAN).
Provided that the sentences in the different reference sets are all paraphrases of each other (actually they are back-translations from different languages), this indicates that the translations ep,es,nl,el are rather 'free' with respect to the German source sentences, while the de translations are somewhat more similar to the actual output of both MT systems, SYSTRAN and METIS-II. This is to be expected precisely due to the back-translation character of the reference translations.

Results according to text type As for the Spanish-English translation task, the following tables show the results for the development test set for the Spanish-English language pair, broken down by text type as specified in section 3.1.1. above, for the two systems METIS-II and SYSTRAN.
The two systems show a very different behaviour pattern with respect to the four types of text. Thus, while METIS-II performs clearly worse on the grammar set, this is the text where SYSTRAN obtains the best results.
Overall, SYSTRAN is more stable with only slight variations across text types. METIS-II, on the other hand, is sensible to text type and obtains its better results on newspaper text. Press is one of the most neutral, standard genres. In contrast, both technical and scientific texts present a very specific vocabulary and constructions. Therefore, press samples can be expected to be more similar to the texts that form the written portion of the BNC (standard English texts) than other text types, which explains the relatively good performance of METIS in this subset.
As to why results are so different for the grammar subset, we note that it contains a representative sample of grammatical phenomena that diverge from Spanish to English, which are difficult for a statistical system such METIS, but which have long been identified and addressed by SYSTRAN human developers.

Table 16: ES-EN: Results for METIS-II, broken down by text type

	Grammar	News	Scientific	Technical
BLEU	0.22	0.33	0.29	0.26
NIST	4.97	6.31	5.91	5.71
TER	41.9	47.8	52.3	55.2

Table 17: ES-EN: Results for SYSTRAN, broken down by text type

	Grammar	News	Scientific	Technical
BLEU	0.48	0.46	0.47	0.45
NIST	6.09	7.47	7.50	7.39
TER	30.0	36.0	35.0	39.5

4. Conclusions

We have carried out an evaluation of the METIS system on two different sets: one known by the system, one completely new. We have compared the outcome of the evaluation with SYSTRAN, for the four language pairs involved. The results are expectably lower for the still young METIS systems but not at an unatainable distance from the mature system.
The results for all the languages in the project are quite homogeneous, which is encouraging given the differences between them. These differences concern both linguistic aspects (as the languages belong to the Germanic, Romance, and Hellenic families within the Indoeuropean language family) and the computational resources available for each language.
It is to be expected that further development of the METIS system, including parameter tuning, enrichment of the translation model, new target corpora, etc., will be able to, not only shorten distances with SYSTRAN, but also overcome it, and become a real alternative in the MT world.

References

Brants, T. (2000). TnT – a statistical part-of-speech tagger. In *Proceedings of the 6th Applied Natural Language Processing Conference (ANLP)*, pp. 224-231, Seattle, WA, 2000.

Carl, M., Schmidt, P., and Schütz, J. (2005). Reversible Template-based Shake & Bake Generation. In: *Proceedings of the 2nd Workshop on Example-Based Machine Translation held in conjunction with the 10th Machine Translation Summit*, pp. 17-26, Phuket, 2005.

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

Dirix, P., Vandeghinste,V., and Schuurman, I. (2006). A new hybrid approach enabling MT for languages with little resources. In: *Proceedings of the 16th Meeting on Computational Linguistics in the Netherlands (CLIN)*, Amsterdam, 2006.

Dologlou, Y., Markantonatou, S., Tambouratzis, G., Yiannoutsou, O., Fourla A., and Ioannou, N. (2003). Using Monolingual Corpora for Statistical Machine Translation: The METIS System. in: *Proceedings of the EAMT-CLAW'03 Workshop*, pp. 61-68, Dublin, 2003.

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.

Habash, N. (2003). Matador: A Large-Scale Spanish-English GHMT System. In *Proceedings of the 9th Machine Translation Summit of the International Association for Machine Translation*, pp. 149-156, New Orleans, LA, 2003.

Habash, N. (2004). The Use of a Structural N-gram Language Model in Generation-Heavy Hybrid Machine Translation. In *Proceedings of the Third International Conference on Natural Language Generation (INLG04)*, Brighton, 2004.

Habash, N. and Dorr, B. (2002). Handling translation divergences: Combining statistical and symbolic techniques in generation-heavy machine translation. In S. Richardson (ed.), *Proceedings of the Fifth Conference of the Association for Machine Translation in the Americas (AMTA): Machine Translation: From Research to Real Users*, pp. 84-93, Tiburon, CA, 2002. .

Habash, N. and Dorr, B. (2003). A Categorial Variation Database for English. In *Proceedings of the Human Language Technology Conference and North American Chapter of the Association for Computational Linguistics Annual Meeting (NAACL)*. Association for Computational Linguistics. Edmonton, 2003.

Koehn, P. (2005). Europarl: a parallel corpus for statistical machine translation. In *Proceedings of the 10th Machine Translation Summit of the International Association for Machine Translation*, pp. 79-87, Phuket, 2005.

Koehn, P., Och, F., and Marcu, D. (2003) Statistical Phrase-Based Translation. In *Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference (NAACL/HLT)*, Edmonton, 2003.

Markantonatou S., Sofianopoulos,S., Spilioti, V., Vassiliou, M., and Yannoutsou,O. (2007). An MT System Embedding Pattern Knowledge. In: *Proceedings of the METIS-II Workshop: New Approaches to Machine Translation*, pp. 11-18, Leuven, 2007.

Melero, M., Oliver, A., Badia, T., and Suñol, T. (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*, pp. 19-26, Leuven, 2007.

Papineni, K., Roukos, S., Ward, T., Zhu, W.J. (2002). BLEU: a method for automatic evaluation of Machine Translation. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 311-318, Philadelphia, PA, 2002.

Snover, M., Dorr, B., Schwartz, R., Micciula, L., Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In: *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA)*, pp. 223-231, Cambridge, MA, 2006.

Vandeghinste, V. (2008). A Hybrid Modular Machine Translation System. PhD thesis. K.U.Leuven.

Vandeghinste, V., Dirix, P., and Schuurman, I. (2007). The effect of a few rules on a data-driven MT system. In: *Proceedings of the METIS-II Workshop: New Approaches to Machine Translation*, pp. 27-34, Leuven, 2007.

Vandeghinste, V., Schuurman, I., Carl, M., Markantonatou, S., and Badia, T. (2006). METIS-II: Machine Translation for Low-Resource Languages. In: *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pp. 1284-1289, Genoa, 2006.

Zwarts, S., and Dras, M. (2007). Syntax-based Word Reordering in Phrase-Based Statistical Machine Translation; Why Does it Work? In *Proceedings of the 11th Machine Translation Summit of the International Association for Machine Translation*, pp. 559-566, Copenhagen, 2006.