The Prague Bulletin of Mathematical Linguistics
NUMBER 97 APRIL 2012

EDITORIAL BOARD

Published twice a year by Charles University in Prague

Editorial office and subscription inquiries:
ÚFAL MFF UK, Malostranské náměstí 25, 118 00, Prague 1, Czech Republic
E-mail: pbml@ufal.mff.cuni.cz

ISSN 0032-6585

PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 97 APRIL 2012

CONTENTS

Articles

PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 97 APRIL 2012 5–22

Towards Optimal Choice Selection for Improved Hybrid Machine Translation

Christian Federmann[a], Maite Melero[b], Pavel Pecina[d], Josef van Genabith[c]

[a] DFKI, German Research Center for Artificial Intelligence, Germany
[b] Barcelona Media, Spain
[c] CNGL, Dublin City University, Ireland
[d] Faculty of Mathematics and Physics, Charles University in Prague, Czech Republic

Abstract

In recent years, machine translation (MT) research focused on investigating how hybrid MT as well as MT combination systems can be designed so that the resulting translations give an improvement over the individual translations.

As a first step towards achieving this objective we have developed a parallel corpus with source data and the output of a number of MT systems, annotated with metadata information, capturing aspects of the translation process performed by the different MT systems.

As a second step, we have organised a shared task in which participants were requested to build Hybrid/System Combination systems using the annotated corpus as input. The main focus of the shared task is trying to answer the following question: *Can Hybrid MT algorithms or System Combination techniques benefit from the extra information (linguistically motivated, decoding and runtime) from the different systems involved?*

In this paper, we describe the annotated corpus we have created. We provide an overview on the participating systems from the shared task as well as a discussion of the results.

1. Introduction

Machine translation is an active field of research with many competing paradigms to tackle core translation problems. In recent years, an important focus for research has been investigating how hybrid machine translation engines as well as combination systems including several translation engines can be designed and implemented so that the resulting translations give an improvement over the component parts.

Corresponding author: cfedermann@dfki.de
Cite as: Christian Federmann, Maite Melero, Pavel Pecina, Josef van Genabith. Towards Optimal Choice Selection for Improved Hybrid Machine Translation. The Prague Bulletin of Mathematical Linguistics No. 97, 2012, pp. 5–22. doi: 10.2478/v10108-012-0001-1.

One of the main objectives in our research within the T4ME project[1] is to provide a systematic investigation and exploration of optimal choices in Hybrid Machine Translation supporting Hybrid MT design using sophisticated machine-learning technologies. As a first step towards achieving this objective we have developed a parallel corpus with source data and the output of a number of MT systems, representing carefully selected MT paradigms, annotated with metadata information, capturing aspects of the translation process performed by the different machine translation systems. Including detailed and heterogeneous system specific information as metadata in the translation output (rather than just providing strings) is intended to provide rich features for machine learning methods to optimise combination in hybrid machine translation.

This first version of the corpus is available online under `www.dfki.de/ml4hmt/` and comprises annotated outputs of five machine translation systems, namely Joshua, Lucy, Metis, Apertium, and the Moses-based PB-SMT component of MaTrEx. The language pairs supported by the corpus are: English↔German, English↔Spanish, English↔Czech (all in both directions).

In this paper, we describe the annotated corpus we have created—including the data used to obtain the sample corpus (Section 2), the translation engines applied when building the corpus (Section 3), and the format of the corpus (Section 4). We provide an overview on the challenge (Section 5) and give descriptions of the participating systems from the shared task (Section 6). This includes a comparison to a state-of-the-art system combination system. Using automated metric scores and results from a manual evaluation, we discuss the performance of the various combination systems and their implementations (Section 7). One interesting result from the shared task is the fact that we observed different systems winning according to the automated metrics and according to the manual evaluation. We conclude by summarising our research results and outline future work (Section 8).

2. Data

2.1. Annotation Data

As a source of the data to be included and annotated in the corpus we decided to use the WMT 2008 news test set, which is a set of 2,051 sentences from the news domain translated to the languages of interest (English, Spanish, German, Czech) and also some others (French, Hungarian). This test set was provided by the organizers of the Third Workshop on Machine Translation (WMT) in 2008 as test data for the shared translation task.

[1]EU FP7 funded Network of Excellence, grant agreement no.: 249119.

2.2. Training Data-Driven Systems

Some of the MT systems used in this work are data-driven (Joshua and MaTrEx). They require parallel data for translation phrase pair extraction, monolingual data for language modeling, and parallel development data for tuning of system parameters. Originally we intended to use the Europarl corpus (Koehn, 2005) for training purposes, but since the version of this widely used parallel corpus available at the time when the research reported in this article was carried out, did not include Czech, we used the Acquis (Steinberger et al., 2006) and News Commentary parallel corpora instead.

2.3. JRC-Acquis Multilingual Parallel Corpus

The JRC-Acquis Multilingual Parallel Corpus is an "approximation" of the Acquis Communautaire, the total body of European Union (EU) law applicable in the the EU Member States. It comprises documents that were available in at least in ten of the twenty EU-25 languages (official languages in the EU before 2007) and that additionally existed in at least three of the nine languages that became official languages with the EU Enlargement in 2004 (i.e. Czech, Hungarian, Slovak, etc.).

2.4. WMT News Commentary Parallel Corpus

The WMT News Commentary Parallel Corpus contains news and commentaries from the Project Syndicate and is provided as training data for the series of WMT translation shared tasks (`www.statmt.org`). Version 10 was released in 2010 and is available in English, French, Spanish, German, and Czech.

2.5. Development Data

The development data sets were taken from the WMT 2008 development data package. We chose the `nc-test2007` files, which consist of 2,007 sentences from the news-commentary domain available in English, French, Spanish, German, and Czech. These development sets not overlap with the training set.

3. System and Metadata Descriptions

3.1. Joshua

Description Joshua (Li et al., 2009), referred to as system t1 in the annotated corpus, is an open-source toolkit for statistical machine translation, providing a full implementation of state-of-the-art techniques making use of synchronous context free grammars (SCFGs). The decoding process features algorithms such as chart-parsing,

n-gram language model integration, beam-and cube-pruning and k-best extraction, while training includes suffix-array grammar extraction and minimum error rate training.

Annotation In our metadata annotations, we provide the output of the decoding process given the "test set", as processed by Joshua (SVN revision 1778). The annotation set contains the globally applied feature weights and for each translated sentence: the full output of the produced translation with the highest total score (among the n-best candidates), the language model and translation table scores, the scores from the derivation of the sentence (phrase scores) and merging/pruning statistics of the search process. Each translated sentence, represented by a hierarchical phrase, contains zero or more tokens and points to zero or more child phrases. Finally, the word-alignment of each phrase to the source text, using word indices, is available.

3.2. Lucy

Description The Lucy RBMT system (Alonso and Thurmair, 2003), system t2, uses a sophisticated RBMT transfer approach with a long research history. It employs a complex lexicon database and grammars to transform a source into a target language representation. The translation of a sentence is carried out in three major phases: analysis, transfer, and generation.

Annotation In addition to the translated target text Lucy provides information about the tree structures that have been created in the three translation phases and which have been used to generate the final translation of the source text. Inside these trees, information about POS, phrases, word lemma information, and word/phrase alignment can be found. In our metadata annotations, we provide a "flattened" representation of the trees. For each token, annotation may contain allomorphs, canonical representations, linguistic categories, or surface string.

3.3. Metis

Description The Metis system, system t3, (Vandeghinste et al., 2008) achieves corpus-based translation on the basis of a monolingual target corpus and a bilingual dictionary only. The bilingual dictionary functions as a flat translation model that provides n translations for each source word. The most probable translation given the context is then selected by consulting the statistical models built from the target language corpus.

Annotation Meta-data information for Metis is extracted from the set of final translations ranked by the Metis search engine. For each translation we obtain the score

computed during the search process, together with some linguistic information. The basic linguistic information provided is: lemma, POS tag, and morphological features, including gender, number, tense, etc.

3.4. Apertium

Description Apertium (Ramírez-Sánchez et al., 2006), system t4, originated as one of the machine translation engines in the project OpenTrad, funded by the Spanish government. Apertium is a shallow-transfer machine translation system, which uses finite state transducers for all of its lexical transformations, and hidden Markov models for POS tagging or word category disambiguation. Constraint Grammar taggers are also used for some language pairs (e.g., Breton-French).

Annotation We use Apertium version 3.2. Our metadata annotation includes tags, lemmas and syntactic information. We have used the following commands (in English-to-Spanish): en-es-chunker (for syntax information), en-es-postchunk (for tags and lemmas) and en-es (for the translation).

3.5. MaTrEx

Description The MaTrEx machine translation system (Penkale et al., 2010), system t5, is a combination-based multi-engine architecture developed at Dublin City University exploiting aspects of both the Example-based Machine Translation (EBMT) and Statistical Machine Translation (SMT) paradigms. The architecture includes various individual systems: phrase-based, example-based, hierarchical phrase-based, and tree-based MT. For the corpus data produced here we use the standard MOSES PB-SMT system (Koehn et al., 2007) as integrated into MaTrEx.

Annotation Sentence translations provided by MaTrEx in this work were obtained using the MOSES PB-SMT system decomposing the source side to phrases (n-grams), finding their translation and composing them to a target language sentence which has the highest score according the PB-SMT model. Meta-data annotations for each sentence translated by MaTrEx include scores from each model and is decomposed into phrases each provided with two additional scores: translation probability and future cost estimate. Additionally information about unknown words is also included.

4. Corpus Description

We have developed a new dedicated format derived from XLIFF (XML Localisation Interchange File Format) to represent and store the corpus data. XLIFF is an XML-based format created to standardize localization. It was standardized by OASIS in

2002 and its current specification is v1.2 (docs.oasis-open.org/xliff/xliff-core/xliff-core.html).

An XLIFF document is composed of one or more <file> elements, each corresponding to an original file or source. Each <file> element contains the source of the data to be localized and the corresponding localized (translated) data for one locale only. The localizable texts are stored in <trans-unit> elements each having a <source> element to store the source text and a <target> (not mandatory) element to store the translation.

We introduced new elements into the basic XLIFF format (in the "metanet" namespace) supporting a wide variety of metadata annotation of the translated texts by different MT systems (tools). The tool information is included in the <tool> element appearing in the header of the file. Each tool can have several parameters (model weights) which are described in <metanet:weight>.

Annotation is stored in <alt-trans> element within the <trans-unit> elements. The <source> and <target> elements in the <trans-unit> elements refer to the source sentence and its reference translation, respectively. The <source> and <target> elements in the <alt-trans> elements specify the input and output of a particular MT system (tool). Tool-specific scores assigned to the translated sentence are listed in the <metanet:scores> element and the derivation of the translation is specified in the <metanet:derivation> element. Its content is tool-specific.

The full format specification is available as an XML schema. An example annotation is depicted in Figure 1 at the end of this article.

4.1. Oracle Scores

At first, we compare the performance of the contributing systems (t1–t5) on the sentence level, using two popular metrics, 1-WER and smoothed-BLEU (Lin and Och, 2004). For this initial experiment, we worked on the language pair Spanish→English as this is also used in the ML4HMT shared task.

Table 1 shows the percentage of the cases that a system gave the best translation for a sentence according the two sentence-level metrics in columns 2 and 3.[2] Column 4 shows the overall BLEU score for the individual systems. This indicates that the systems included in the corpus perform complementary to each other.

In table 2 we show what the optimal BLEU performance would be, if a combination system was able to choose the best sentence of each component system according to each of the two metrics. This indicates the possibilities of improvement by a sentence-selection approach given the corpus. We believe that even higher performance would be possible using more sophisticated system combination methods.

[2]Measured over the development set, ties were allowed.

ranked 1st	1-WER[%]	sBLEU[%]	BLEU
system t1	26.44	14.73	12.80
system t2	37.56	38.63	14.94
system t3	5.85	5.85	8.29
system t4	16.00	23.41	13.34
system t5	58.54	29.95	14.47

Table 1. Preliminary investigation for the usability of the corpus for Hybrid MT

Oracle combination	BLEU
sBLEU-based	18.95
1-WER-based	17.62

Table 2. Potential BLEU score reachable using perfect sentence selection.

5. Challenge Description

The "Shared Task on Applying Machine Learning techniques to optimise the division of labour in Hybrid MT" is an effort to trigger systematic investigation on improving state-of-the-art Hybrid MT, using advanced machine-learning (ML) methodologies. Participants of the challenge are requested to build Hybrid/System Combination systems by combining the output of several MT systems of different types and with very heterogeneous types of metadata information, as provided by the organizers. The main focus of the shared task is trying to answer the following question:

> *Can Hybrid Machine Translation algorithms or System Combination techniques benefit from extra information—such as linguistic or linguistically motivated features, decoding parameters, or runtime annotation—from the different systems involved?*

The participants are given a bilingual development set, aligned at a sentence level. For each sentence, the corresponding *bilingual data set* contains:

- the source sentence,
- the target (reference) sentence, and
- the corresponding multiple output translations, annotated with metadata information, from five different systems, based on various machine translation approaches.

5.1. Development and Test Sets

We decided to use the WMT 2008 (Callison-Burch et al., 2008) news test set as a source for the annotated corpus.[3] This is a set of 2,051 sentences from the news domain translated to several languages, including English and Spanish but also others. The data was provided by the organizers of the Third Workshop on Machine Translation (WMT) in 2008. This data set was split into our own Hybrid MT Shared Task development set (containing 1,025 sentence pairs) and test set (containing 1,026 sentence pairs).

6. Combination Systems Participating in the Shared Task

6.1. DCU

The system described in Okita and van Genabith (2011) presents a system combination module in the MT system MaTrEx (Machine Translation using Examples) developed at Dublin City University. A system combination module deployed by them achieved an improvement of 2.16 BLEU (Papineni et al., 2001) points absolute and 9.2% relative compared to the best single system, which did not use any external language resources. The DCU system is based on system combination techniques which use a confusion network on top of a Minimum Bayes Risk (MBR) decoder (Kumar and Byrne, 2002).

One interesting, novel point in their submission is that for the given single best translation outputs, they tried to identify which inputs they will consider for the system combination, possibly discarding the worst performing system(s) from the combination input. As a result of this selection process, their BLEU score, from the combination of the four single best individual systems, achieved 0.48 BLEU points absolute higher than the combination of the five single best systems.

6.2. DFKI-A

A system combination approach with a sentence ranking component is presented in Avramidis (2011). The paper reports on a pilot study that takes advantage of multilateral system-specific metadata provided as part of the shared task. The proposed solution offers a machine learning approach, resulting in a selection mechanism able to learn and rank systems' translation output on the complete sentence level, based on their respective quality.

For training, due to the lack of human annotations, word-level Levenshtein distance has been used as a (minimal) quality indicator, whereas a rich set of sentence

[3] We deliberately did not use the WMT 2009 (Callison-Burch et al., 2009) news test set as there had been quality issues with this data set during the 2009 shared task.

features was extracted and selected from the dataset. Three classification algorithms (Naive Bayes, SVM and Linear Regression) were trained and tested on pairwise featured sentence comparisons. The approaches yielded high correlation with original, Levenshtein-based rankings ($\tau = 0.52$) and selected the best translation on up to 54% of the cases.

6.3. DFKI-B

The authors of Federmann et al. (2011) report on experiments that are focused on word substitution using syntactic knowledge. From the data provided by the workshop organisers, they choose one system to provide the "translation backbone". The Lucy MT system was suited best for this task, as it offers parse trees of both the source and target side, which allows the authors to identify interesting phrases, such as noun phrases, in the source and replace them in the target language output. The remaining four systems are mined for alternative translations on the word level that are potentially substituted into the aforementioned template translation if the system finds enough evidence that the candidate translation is better. Each of these substitution candidates is evaluated considering a number of factors:

- the part-of-speech of the original translation must match the candidate fragment.
- Additionally they may consider the 1-left and 1-right context.
- Besides the part-of-speech, all translations plus their context are scored with a language model trained on EuroParl.
- Additionally, the different systems may come up with the same translation, in that case the authors select the candidate with the highest count ("majority voting").

The authors reported improvements in terms of BLEU score when comparing to the translations from the Lucy RBMT system.

6.4. LIUM

Barrault and Lambert submitted results from applying the open-source MANY (Barrault, 2010) system on our data set. The MANY system can be decomposed into two main modules.

1. The first is the alignment module which actually is a modified version of TERp (Snover et al., 2009). Its role is to incrementally align the hypotheses against a backbone in order to create a confusion network. Each hypothesis acts as backbone, yielding the corresponding confusion network. Those confusion networks are then connected together to create a lattice.
2. The second module is the decoder. This decoder is based on the token pass algorithm and it accepts as input the lattice previously created. The costs computed

in the decoder can be expressed as a weighted sum of the logarithm of feature functions. The following features are considered in decoding:

- the language model probability, given by a 4-gram language model,
- a word penalty, which depends on the number of words in the hypothesis,
- a null-arc penalty, which depends on the number of null arcs crossed in the lattice to obtain the hypothesis, and
- the system weights: each word receives a weight corresponding to the sum of the weights of all systems which proposed it.

7. Evaluation Results

To evaluate the performance of the participating systems, we computed automated scores, namely BLEU, NIST, METEOR (Banerjee and Lavie, 2005), PER, Word error rate (WER) and Translation Error Rate (TER) and also performed an extensive, manual evaluation with 3 annotators ranking system combination results for a total of 904 sentences.

7.1. Automated Scores

The results from running automated scoring tools on the submitted translations are reported in Table 3. The overall best value for each of the scoring metrics is printed in **bold face**. The lower half of the table presents automated metric scores for the individual systems in the ML4HMT corpus, also computed on the test set. These scores give an indicative baseline for comparison with the system combination results. Again, the overall best value per column is printed in **bold face**. TER values are not available for the baseline systems as we initially did not intend to use this metric.

7.2. Manual Ranking

The manual evaluation is undertaken using the Appraise (Federmann, 2010) system; a screenshot of the evaluation interface is shown in Figure 2. Users are shown a reference sentence and the translation output from all four participating systems and have to decide on a ranking in *best-to-worst order*. Table 4 shows the average ranks per system from the manual evaluation, again the best value per column is printed in **bold face**. Table 5 gives the statistical mode per system which is the value that occurs most frequently in a data set.

7.3. Inter-annotator Agreement

Next to computing the average rank per system and the statistical mode, we follow Carletta (1996) and compute κ scores to estimate the inter-annotator agreement. In our manual evaluation campaign, we had $n = 3$ annotators so computing basic, pairwise

System	BLEU	NIST	METEOR	PER	WER	TER
DCU	**25.32**	**6.74**	56.82	**60.43**	**45.24**	**0.65**
DFKI-A	23.54	6.59	54.30	61.31	46.13	0.67
DFKI-B	23.36	6.31	**57.41**	65.22	50.09	0.70
LIUM	24.96	6.64	55.77	61.23	46.17	0.65
Joshua	19.68	6.39	50.22	47.31	62.37	–
Lucy	**23.37**	6.38	**57.32**	49.23	64.78	–
Metis	12.62	4.56	40.73	63.05	77.62	–
Apertium	22.30	6.21	55.45	50.21	64.91	–
MaTrEx	23.15	**6.71**	54.13	**45.19**	**60.66**	–

Table 3. Automated scores for participants and baseline systems on test set.

System	Annotator #1	Annotator #2	Annotator #3	Overall
DCU	2.44	2.61	2.51	2.52
DFKI-A	2.50	2.47	2.48	2.48
DFKI-B	**2.06**	**2.13**	**1.97**	**2.05**
LIUM	2.89	2.79	2.93	2.87

Table 4. Average rank per system per annotator from manual ranking of 904 (overlap=146) translations.

System	Ranked 1st	Ranked 2nd	Ranked 3rd	Ranked 4th	Mode
DCU	62	79	**97**	62	3rd
DFKI-A	73	65	**82**	80	3rd
DFKI-B	**127**	84	47	42	1st
LIUM	38	72	74	**116**	4th

Table 5. Statistical mode per system from manual ranking of 904 (overlap=146) translations.

Systems	π-Score	Systems	π-Score	Annotators	π-Score
DCU, DFKI-A	0.296	DCU, DFKI-B	0.352	#1,#2	0.331
DCU, LIUM	0.250	DFKI-A, DFKI-B	0.389	#1,#3	0.338
DFKI-A, LIUM	0.352	DFKI-B, LIUM	0.435	#2,#3	0.347

Table 6. Pairwise agreement (using Scott's π) for all pairs of systems/annotators. Note that scores in the last column are computed using all pairwise annotations available; these can be more than the overlapping N=146.

```
<trans-unit id="s71">
  <source xml:lang="es">El paciente fue aislado.</source>
    <target xml:lang="en">The patient was isolated.</target>
    <alt-trans rank="1" tool-id="t3">
      <source xml:lang="es">El paciente fue aislado.</source>
      <target xml:lang="en">The paciente was isolated .</target>
      <metanet:scores>
        <metanet:score type="total" value="-60.4375047559049"/>
      </metanet:scores>
      <metanet:derivation id="s71_t3_r1_d1">
        <metanet:phrase id="s71_t3_r1_d1_p1">
          <metanet:string>The</metanet:string>
          <metanet:annotation type="lemma" value="the"/>
          <metanet:annotation type="pos" value="AT0"/>
          <metanet:annotation type="morph_feat" value=":m:sg:"/>
          <metanet:alignment from="0" to="0"/>
        </metanet:phrase>
```

Figure 1. Example of annotation from the ML4HMT corpus.

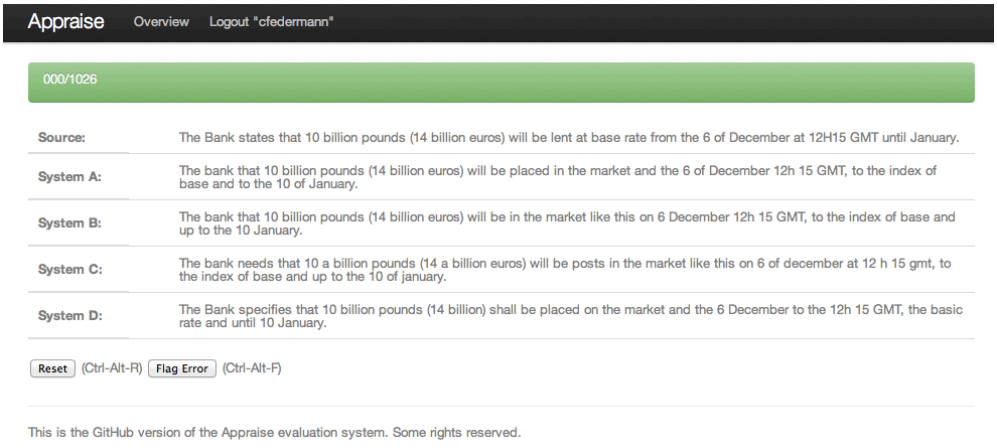

Figure 2. Screenshot of the Appraise interface for human evaluation.

annotator agreement is not sufficient—instead, we apply Fleiss (1971) who extends Scott (1955) for computing inter-annotator agreement for $n > 2$.

Annotation Setup As we have mentioned before, we had $n = 3$ annotators assign ranks to our four participating systems. As ties were not allowed, this means there exist $4! = 24$ possible rankings per sentence (e.g., *ABCD, ABDC, etc.*).[4] In a second evaluation scenario, we only collected the *1-best* ranking system per sentence, resulting in a total of four categories (A: "*system A ranked 1st*", etc.). In this second scenario, we can expect a higher annotator agreement due to the reduced number categories. Overall, we collected 904 sentences with an overlap of $N = 146$ sentences for which all annotators assigned ranks.

Scott's π allows to measure the pairwise annotator agreement for a classification task. It is defined as

$$\pi = \frac{P(A) - P(E)}{1 - P(E)} \tag{1}$$

where $P(A)$ represents the fraction of rankings on which the annotators agree, and $P(E)$ is the probability that they agree by chance. Table 6 lists the pairwise agreement of annotators for all four participating systems. Since we did not allow ties in the ranking process and because our ranks are not absolute categories, but can only be interpreted relatively to each other, we essentially have two categories for each pair of translations which are equally likely. Assuming $P(E) = 0.5$ we obtain an overall agreement π score of

$$\pi = \frac{0.673 - 0.5}{1 - 0.5} = 0.346 \tag{2}$$

which can be interpreted as *fair agreement* following Landis and Koch (1977). WMT shared tasks have shown this level of agreement is common for language pairs, where the performance of all systems is rather close to each other, which in our case is indicated by the small difference measured by automatic metrics on the test set (Table 3). The lack of ties, in this case might have meant an extra reason for disagreement, as annotators were forced to distinguish a quality difference which otherwise might have been annotated as "equal". We have decided to compute Scott's π scores to be comparable (or at least similar) to WMT11 (Bojar et al., 2011).[5]

Fleiss κ Next to the π scores, there also exists the so-called κ score. Its basic equation is strikingly similar to (1)

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)} \tag{3}$$

[4]Given this huge number of possible categories, we were already expecting resulting κ scores to be low.

[5]Note that we do not allow ties for the rankings and do not include the reference in the process, though.

with the main difference being the κ score's support for $n > 2$ annotators. We compute κ for two configurations. Both are based on $n = 3$ annotators and $N = 146$ sentences. They differ in the number of categories that a sentence can be assigned to (k).

1. *complete* scenario: $k = 24$ categories. For this, we obtained a κ score of

$$\kappa_{complete} = \frac{0.1 - 0.054}{1 - 0.054} = 0.049 \tag{4}$$

2. *1-best* scenario: $k = 4$ categories. Here, κ improved to

$$\kappa_{1-best} = \frac{0.368 - 0.302}{1 - 0.302} = 0.093 \tag{5}$$

It seems that the large number of categories of the *complete* scenario has indeed had an effect on the resulting $\kappa_{complete}$ score. This is a rather expected outcome, still we report the κ scores for future reference. The *1-best* scenario supports an improved κ_{1-best} score but does not reach the level of agreement observed for the π score.

It seems that DFKI-B was underestimated by BLEU scores, potentially due to its rule-based characteristics. This is a possible reason for the relatively higher inter-annotator agreement when compared with other systems. Also, DCU and LIUM may have low inter-annotator agreement as their background is similar. It is worth noting that METEOR was the only automated metric correlating with results from the manual evaluation.

Due to the fact that κ is not really defined for *ordinal data* (such as rankings in our case), we will investigate other measures for inter-annotator agreement. It might be a worthwhile idea to compute α scores, as described in Krippendorff (2004). Given the average rank information, statistical mode, π and κ scores, we still think that we have derived enough information from our manual evaluation to support for future discussion.

Cohen's κ As the results for Fleiss κ were disappointing for both settings, we also compute pairwise Cohen κ scores. Interestingly, we can again report *fair agreement* between the annotators, achieving κ scores similar to the π scores in table 6:

$$\kappa_{(\#1,\#2)} = 0.336 \qquad \kappa_{(\#1,\#3)} = 0.312 \qquad \kappa_{(\#1,\#3)} = 0.331 \tag{6}$$

The average Cohen's κ score is:

$$\kappa = 0.327 \tag{7}$$

8. Conclusion

We have developed an Annotated Hybrid Sample MT Corpus which is a set of 2,051 sentences translated by five different MT systems[6] (Joshua, Lucy, Metis, Apertium, and MaTrEx) in six translation directions (Czech→English, German→English, Spanish→English, English→Czech, English→German, and English→Spanish) and annotated with metadata information provided by the MT systems.

Using this resource we have launched the Shared Task on Applying Machine Learning techniques to optimise the division of labour in Hybrid MT (ML4HMT-2011), asking participants to create combined, hybrid translations using machine learning algorithms or other, novel ideas for making best use of the provided ML4HMT corpus data. The language pair for the shared task was Spanish→English.

Four participating combination systems, each following a different solution strategy, have been submitted to the shared task. We computed automated metric scores and conducted an extensive manual evaluation campaign to assess the quality of the hybrid translations. Interestingly, the system winning nearly all the automatic scores (DCU) only reached a third place in the manual evaluation. Vice versa, the winning system according to manual rankings (DFKI-B) ranked last place in the automatic metric scores based evaluation, with only one automated metric choosing it as winning system. This clearly indicates that more systematic investigation of hybrid system combination approaches, both on a system level and with regards to the evaluation of such systems' output, needs to be undertaken.

We have learnt from the participants that the metadata annotations provided by our ML4HMT corpus are possibly too heterogeneous to be used easily in system combination approaches; hence we will work on an updated version for the next edition of this shared task. Also, we will further focus on the integration of advanced machine learning techniques as these are expected to support better exploitation of our corpus' data properties.

Acknowledgments

This work has been funded under the Seventh Framework Programme for Research and Technological Development of the European Commission through the T4ME contract (grant agreement no.: 249119) and was supported by Science Foundation Ireland (Grant No. 07/CE/I1142) as part of the Centre for Next Generation Localisation (`www.cngl.ie`) at Dublin City University and partially supported by the Czech Science Foundation (grant no. P103/12/G084). The authors would like to thank Felix Sasaki and Eleftherios Avramidis for their collaboration on the ML4HMT corpus and the shared task. Also, we are grateful to the anonymous reviewers for their valuable feedback and comments.

[6] Not all systems available for all language pairs.

Bibliography

Alonso, Juan A. and Gregor Thurmair. The Comprendium Translator System. In *Proceedings of the Ninth Machine Translation Summit*, New Orleans, USA, 2003.

Avramidis, Eleftherios. DFKI System Combination with Sentence Ranking at ML4HMT-2011. In *Proceedings of the International Workshop on Using Linguistic Information for Hybrid Machine Translation (LIHMT 2011) and of the Shared Task on Applying Machine Learning Techniques to Optimise the Division of Labour in Hybrid Machine Translation (ML4HMT)*, Barcelona, Spain, November 2011. META-NET.

Banerjee, Satanjeev and Alon Lavie. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W/W05/W05-0909`.

Barrault, Loïc. MANY : Open-Source Machine Translation System Combination. *Prague Bulletin of Mathematical Linguistics, Special Issue on Open Source Tools for Machine Translation*, 93: 147–155, 2010.

Bojar, Ondrej, Miloš Ercegovčević, Martin Popel, and Omar Zaidan. A Grain of Salt for the WMT Manual Evaluation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 1–11, Edinburgh, Scotland, July 2011. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W11-2101`.

Callison-Burch, Chris, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. Further Meta-Evaluation of Machine Translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 70–106, Columbus, Ohio, June 2008. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W/W08/W08-0309`.

Callison-Burch, Chris, Philipp Koehn, Christof Monz, and Josh Schroeder. Findings of the 2009 Workshop on Statistical Machine Translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 1–28, Athens, Greece, March 2009. Association for Computational Linguistics.

Carletta, Jean. Assessing Agreement on Classification Tasks: the kappa Statistic. *Computational Linguistics*, 22:249–254, June 1996. ISSN 0891-2017. URL `http://dl.acm.org/citation.cfm?id=230386.230390`.

Federmann, Christian. Appraise: An Open-Source Toolkit for Manual Phrase-Based Evaluation of Translations. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valetta, Malta, May 2010. URL `http://www.lrec-conf.org/proceedings/lrec2010/pdf/197_Paper.pdf`.

Federmann, Christian, Yu Chen, Sabine Hunsicker, and Rui Wang. DFKI System Combination using Syntactic Information at ML4HMT-2011. In *Proceedings of the International Workshop on Using Linguistic Information for Hybrid Machine Translation (LIHMT 2011) and of the Shared Task on Applying Machine Learning Techniques to Optimise the Division of Labour in Hybrid Machine Translation (ML4HMT)*, Barcelona, Spain, November 2011. META-NET.

Fleiss, J.L. Measuring Nominal Scale Agreement among Many Raters. *Psychological Bulletin*, 76 (5):378–382, 1971.

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

Koehn, Philipp, Hieu Hoang, Alexandra Birch Mayne, Christopher Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbs. Moses: Open source toolkit for statistical machine translation. In *Proceedings of Annual meeting of the Association for Computation Linguistics (ACL), demonstration session*, pages 177–180, Prague, Czech, June 2007.

Krippendorff, Klaus. Reliability in Content Analysis. Some Common Misconceptions and Recommendations. *Human Communication Research*, 30(3):411–433, 2004.

Kumar, Shankar and William Byrne. Minimum Bayes-Risk Word Alignments of Bilingual Texts. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10*, EMNLP '02, pages 140–147, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics. URL `http://dx.doi.org/10.3115/1118693.1118712`.

Landis, J.R. and G.G. Koch. Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174, 1977. URL `http://dx.doi.org/10.2307/2529310`.

Li, Zhifei, Chris Callison-Burch, Chris Dyer, Sanjeev Khudanpur, Lane Schwartz, Wren Thornton, Jonathan Weese, and Omar Zaidan. Joshua: An Open-Source Toolkit for Parsing-Based Machine Translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 135–139, Athens, Greece, March 2009. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W/W09/W09-0x24`.

Lin, Chin-Yew and Franz Josef Och. Orange: a method for evaluating automatic evaluation metrics for machine translation. In *Proceedings of the 20th international conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics. doi: http://dx.doi.org/10.3115/1220355.1220427. URL `http://dx.doi.org/10.3115/1220355.1220427`.

Och, Franz Josef. Minimum Error Rate Training in Statistical Machine Translation. In *ACL '03: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 160–167, Morristown, NJ, USA, 2003. Association for Computational Linguistics. URL `http://dx.doi.org/10.3115/1075096.1075117`.

Okita, Tsuyoshi and Josef van Genabith. DCU Confusion Network-based System Combination for ML4HMT. In *Proceedings of the International Workshop on Using Linguistic Information for Hybrid Machine Translation (LIHMT 2011) and of the Shared Task on Applying Machine Learning Techniques to Optimise the Division of Labour in Hybrid Machine Translation (ML4HMT)*, Barcelona, Spain, November 2011. META-NET.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: A Method for Automatic Evaluation of Machine Translation. IBM Research Report RC22176(W0109-022), IBM, 2001. URL `http://www1.cs.columbia.edu/nlp/sgd/bleu.pdf`.

Penkale, Sergio, Rejwanul Haque, Sandipan Dandapat, Pratyush Banerjee, Ankit K. Srivastava, Jinhua Du, Pavel Pecina, Sudip Kumar Naskar, Mikel L. Forcada, and Andy Way. MaTrEx: the DCU MT system for WMT 2010. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, WMT '10, pages 143–148, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. ISBN 978-1-932432-71-8. URL `http://dl.acm.org/citation.cfm?id=1868850.1868870`.

Ramírez-Sánchez, Gema, Felipe Sánchez-Martínez, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, and Mikel L. Forcada. OpenTrad Apertium open-source machine translation system: an opportunity for business and research. In *Proceeding of Translating and the Computer 28 Conference*, London, United Kingdom, November 2006. ISBN 0-85142-483-X.

Scott, William A. Reliability of Content Analysis: The Case of Nominal Scale Coding. *The Public Opinion Quarterly*, 19(3):321–325, 1955.

Snover, Matthew G., Nitin Madnani, Bonnie Dorr, and Richard Schwartz. TER-Plus: paraphrase, semantic, and alignment enhancements to Translation Edit Rate. *Machine Translation*, 23:117–127, September 2009. ISSN 0922-6567. URL `http://dx.doi.org/10.1007/s10590-009-9062-9`.

Steinberger, Ralf, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaz Erjavec, and Dan Tufis. The JRC-Acquis: A Multilingual Aligned Parallel Corpus with 20+ Languages. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, pages 2142–2147, 2006.

Stolcke, Andreas. SRILM - An Extensible Language Modeling Toolkit. In *the 7th International Conference on Spoken Language Processing (ICSLP) 2002*, Denver, Colorado, 2002.

Vandeghinste, Vincent, Peter Dirix, Ineke Schuurman, Stella Markantonatou, Sokratis Sofianopoulos, Marina Vassiliou, Olga Yannoutsou, Toni Badia, Maite Melero, Gemma Boleda, Michael Carl, and Paul Schmidt. Evaluation of a Machine Translation System for Low Resource Languages: METIS-II. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May 2008. European Language Resources Association (ELRA). ISBN 2-9517408-4-0. http://www.lrec-conf.org/proceedings/lrec2008/.

Address for correspondence:
Christian Federmann
`cfedermann@dfki.de`

DFKI GmbH
Campus D3 2
Stuhlsatzenhausweg 3
D-66123 Saarbrücken
GERMANY

PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 97 APRIL 2012 23–41

Mapping Semantic Information from FrameNet onto VALLEX

Václava Kettnerová, Markéta Lopatková, Eduard Bejček

Charles University in Prague, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics

Abstract

In this article, we introduce a project aimed at enhancing a valency lexicon of Czech verbs with semantic information. For this purpose, we make use of FrameNet, a semantically oriented lexical resource. At the present stage, semantic frames from FrameNet have been mapped to eight groups of verbs with various semantic and syntactic properties. The feasibility of this task has been verified by the achieved inter-annotator agreement measured on two semantically and syntactically different groups of verbs – verbs of communication and exchange (85.9% and 78.5%, respectively). Based on the upper level semantic frames from the relation of 'Inheritance' built in FrameNet, the verbs of these eight groups have been classified into more coherent semantic classes. Moreover, frame elements from these upper level semantic frames have been assigned to valency complementations of the verbs of the listed groups as semantic roles. As in case of semantic frames, the achieved interannotator agreement concerning assigning frame elements measured on verbs of communication and exchange has been promising (95.6% and 91.2%, respectively).

As a result, 1 270 lexical units pertaining to the verbs of communication, mental action, psych verbs, social interaction, verbs of exchange, motion, transport and location (2 129 Czech verbs in total if perfective and imperfective verbs being counted separately) have been classified into syntactically and semantically coherent classes and their valency complementations have been characterized by semantic roles adopted from the FrameNet lexical database.

1. Introduction

Information on syntactic and semantic properties of verbs, which are traditionally considered as the center of sentence, plays a key role in many rule-based NLP tasks

 Corresponding author: kettnerova@ufal.mff.cuni.cz
Cite as: Václava Kettnerová, Markéta Lopatková, Eduard Bejček. Mapping Semantic Information from FrameNet onto VALLEX. The Prague Bulletin of Mathematical Linguistics No. 97, 2012, pp. 23–41.
doi: 10.2478/v10108-012-0002-0.

as machine translation, information retrieval, text summarizing, question answering, etc. Lexical resources providing such information are designed within different theoretical frameworks and different theoretical assumptions are also reflected in their annotation schemes. As a result, there are great differences between individual lexical resources: each lexical resource captures different types of information. Consequently, interlinking information from several lexical resources represents an effective way of enriching a particular lexical resource.

However, differences in theoretical assumptions reflected in lexical resources bring several difficulties with mapping information: the different level of granularity in word sense disambiguation represents a typical example. Moreover, other requirements for harmonizing linguistic information are imposed on interlinking information from lexical resources of different languages: a fundamental prerequisite for successful mapping lies first of all in an accurate translation.

In this contribution, we introduce a project aimed at enhancing a valency lexicon of Czech verbs, VALLEX (Lopatková et al., 2008), with semantic information from FrameNet (Baker et al., 1998). This project can be seen as a pilot project focusing on mapping information from a lexical resource of a different language, namely from the English lexical resource (FrameNet) onto the Czech lexical resource (VALLEX). VALLEX and FrameNet are based on different theoretical assumptions: VALLEX takes primarily syntactic criteria in describing valency whereas FrameNet adopts more semantically oriented approach to valency. Moreover, in the project, we have to cope with the different levels of granularity in word sense disambiguation made in VALLEX and FrameNet.

The project consists of several steps. First, semantic frames from FrameNet were manually mapped onto valency frames of Czech verbs from the chosen groups of verbs, namely verbs of `communication`, `mental action`, `psych` verbs, verbs of `social interaction`, `exchange`, `motion`, `transport`, `location`. Second, frame elements from the assigned semantic frame were assigned to valency complementations of the given verbs. Then semantic frames from appropriate upper levels of abstraction based on the relation of 'Inheritance' built in FrameNet were used for classifying the verbs of the given groups into more coherent semantic classes. Moreover, frame elements from these upper level semantic frames were assigned to valency complementations of Czech verbs as semantic roles.

Manual annotation, despite being highly time consuming, seems to be indispensable at this stage of research as it brings necessary insight into the problem. Moreover, it allows us to reach the desired quality of the resulting annotation.

Two aspects are addressed in this project: (i) a practical aspect – providing data for NLP tasks, such as generation, information retrieval, or question answering, and (ii) a theoretical aspect – semantic classes allow us to observe the relation between semantic properties of verbs and their syntactic behavior; further, semantic roles enable us to make inference on lexical entailments that verbs impose on their valency complementations.

The present paper is structured as follows: in Section 2 we briefly describe two lexical resources VALLEX and FrameNet; moreover, we provide a motivation for introducing semantic information from FrameNet to VALLEX. Section 3 is focused on our experiment with mapping semantic frames and frame elements onto valency frames and valency complementations, respectively. Evaluation of both annotations is presented. In Section 4, use of the relation of 'Inheritance' built in FrameNet for classifying Czech verbs and assigning semantic roles to their valency complementations is discussed. Finally, the results of this experiment and future work are summarized.

2. Two Lexical Resources: VALLEX and FrameNet

In this section, we briefly characterize two lexical resources used in the project: VALLEX, which takes into account mainly syntactic criteria for the description of valency characteristics of verbs, and the semantically oriented FrameNet.

2.1. VALLEX – Valency Lexicon of Czech Verbs

VALLEX 2.5[1] provides information on the valency structure of verbs in their particular senses: on the number of valency complementations, on their type labeled by functors, and on their morphemic forms (Žabokrtský and Lopatková, 2007). VALLEX 2.5 describes 2 730 verb lexemes containing about 6 460 lexical units (henceforth LUs) typically corresponding to one verbal sense. At present, more than 44% of LUs are divided into heterogeneous 'supergroups', e.g., verbs of `communication`, `contact`, `emission`, `exchange`, `change`, `location`, `mental action`, `motion`, `perception`, `psych` verbs, verbs of `social interaction`, and `transport`, based primarily on similarities in morphosyntactic properties with regard to semantics. Key information on valency is stored in a valency frame.

An example LU entry in VALLEX is structured as follows:[2]

lemmas:	*impf:* ***vymýšlet***; *pf:* ***vymyslet/vymyslit*** 'to think up'
gloss:	*impf: myšlením vytvářet*; *pf: myšlením vytvořit* 'to invent or to imagine something'
frame:	ACT (obligatory) PAT (obligatory) AIM (typical) BEN (typical)
example:	*impf: vymýšlí novou metodu k léčení nádorových onemocnění*; *pf: vymyslel novou metodu k léčení nádorových onemocnění* 'he thinks up a new strategy to neoplasia treatment'
class:	mental action

VALLEX 2.5, which is closely related to the Prague Dependency Treebank 2.0 (Hajič et al., 2006), takes the Functional Generative Description (henceforth FGD) as its

[1] `http://ufal.mff.cuni.cz/vallex/2.5/`

[2] The example is simplified and translated.

theoretical background (Sgall et al., 1986). FGD applies more syntactically oriented approach to valency, see esp. (Panevová, 1994). Valency complementations are sorted out into inner participants (arguments) and free modifications (adjuncts). Both inner participants and free modifications may be obligatory or optional. Five verbal inner participants are determined rather on the basis of syntactic behavior of verbs: 'Actor' (labeled by functor ACT), 'Patient' (PAT), 'Effect' (EFF), 'Addressee' (ADDR) and 'Origin' (ORIG). In contrast to inner participants, free modifications are semantically distinctive, e.g., 'Location', 'Direction-where', 'Temporal-when', 'Cause' or 'Means', see (Mikulová et al., 2006).

2.2. FrameNet

FrameNet[3] is an on-line lexical database documenting semantic and syntactic combinatory possibilities (valences) of each word in each of its senses (Baker et al., 1998). FrameNet is based on frame semantics (Fillmore et al., 2003) and its annotation is supported by corpus evidence: each LU evokes a particular semantic frame (SF) underlying its meaning. Each SF is conceived as a "conceptual structure describing a particular type of situation, object, or event" (Ruppenhofer et al., 2006). Each SF contains the so-called frame elements (FEs), i.e., semantic participants which are understood as components of such situations. FrameNet contains more than 12 thousand LUs in 1 126 semantic frames, exemplified by more than 160 thousand lexicographic annotation sets.

FrameNet builds a wide network of hierarchical relations between SFs and their FEs. For the purpose of enhancing VALLEX with semantic information, we use the transitive relation of 'Inheritance', which is informally described as follows: "Inheritance – everything which is true about the semantics of the parent frame holds for the semantics of its child frame(s). Each FE from the parent frame (except for extrathematic FEs) is related to a relevant FE in the child frame" (Ruppenhofer et al., 2006).

2.3. Motivation for Introducing Semantic Information to VALLEX

In this section, we discuss the motivation for enhancing VALLEX with missing semantic information, namely semantic classes and semantic roles.

Semantic classes. Semantic classes provide information on relations between LUs. At present, VALLEX does not offer sufficient insight into the way a particular LU relates to another LU(s). For illustration, LUs sharing the same morphosyntactic characteristics may have the same valency frame. Thus they remain indistinct with respect to the valency structure, despite being semantically different, see the pairs of sentences (1)-(2) and (3)-(4).

[3] https://framenet.icsi.berkeley.edu/fndrupal/

(1) *Radní*.ACT *vymysleli nový plán*.PAT *rozvoje města.*
Eng. Councilmen.ACT thought a new plan.PAT for development of the city.

(2) *Turisté*.ACT *vyšli kopec*.PAT
Eng. The tourists.ACT climbed the hill.PAT

(3) *Matka*.ACT *vyprávěla dětem*.ADDR *pohádku*.PAT
Eng. The mother.ACT told the children.ADDR the fairy-tale.PAT

(4) *Jana*.ACT *přinesla otci*.ADDR *dárek*.PAT
Eng. Jane.ACT brought the father.ADDR a gift.PAT

Classifying LUs into semantic classes makes it possible to differentiate between semantically different verbs that exhibit a similar syntactic behavior. For instance, assigning SFs to the pairs of verbs in examples (1)-(2) and (3)-(4) allows us to differentiate between the given LUs: the verb *vymyslet* 'to think up' is classified as belonging to the SF 'Coming_up_with', example (1), whereas the SF 'Intentional_traversing' is assigned to *vyjít* 'to climb', example (2). Similarly, different SFs, the SF 'Statement' and the SF 'Bringing', correspond to the LUs from examples (3) and (4), respectively.

Further, semantic classes make it possible to generalize about syntactic behavior of LUs with similar semantic properties. We suppose that verbs that fall into the same class exhibit similar syntactic behavior, see also (Levin, 1993). For illustration, the verb *vystoupat* 'to ascend' and the verb *vyjít* 'to climb' appertaining to the SF 'Intentional_traversing' share the same valency frame. Similarly, other verbs, e.g. *navrhnout* 'to devise', *formulovat* 'to formulate', and *vynalézt* 'to invent' evoking by the SF 'Coming_up_with' are described by the same valency frame.

Semantic roles. Semantic roles represent one of the oldest linguistic constructs associated with a huge variety of sets of roles. These sets range from verb-specific roles, such as the 'Perpetrator' and 'Victim' for the verb 'to rape', or domain-specific roles, such as the 'Cook' and 'Produced_food' for the verbs 'to cook' or 'to bake', to general roles, such as the 'Agent', 'Theme', 'Beneficiary', or ''protoroles'', Proto-Agent and Proto-Patient, see (Dowty, 1991). FGD – using five functors for inner participants and more semantically specific functors for free modifications – lies in between these approaches, see Section 2.1. We suppose that identifying more specific semantic roles for valency complementations allows us to determine which role an individual complementation plays in a situation portrayed by a LU. Moreover, they enable us to draw inferences on lexical entailments imposed by LUs on their complementations.

For illustration, the verb *vymyslet* 'to think' in (1) is classified as belonging to the SF 'Coming_up_with' and thus the valency complementations 'Actor' and 'Patient' are mapped onto the FEs 'Cognizer' and 'Idea', respectively; whereas in case of the verb *vyjít* 'to climb' in (2) appertaining to the SF 'Intentional_traversing', these complementations are interlinked with the FEs 'Self_mover' and 'Path', respectively. Similarly, the valency complementations 'Actor', 'Addressee' and 'Patient' are described by the FEs 'Speaker', 'Addressee' and 'Message' from the SF 'Statement' in case of the verb

vyprávět 'to tell' in (3) , and by the FEs 'Agent', 'Goal' and 'Theme' from the SF 'Bringing' in case of the verb *přinést* 'to bring' in (4), respectively.

3. Mapping Semantic Information from FrameNet onto VALLEX

In this section, we report on mapping semantic information from FrameNet onto VALLEX, namely interlinking Czech LUs in VALLEX with SFs from FrameNet (Section 3.1) and their valency complementations with FEs from these SFs (Section 3.2).

3.1. Mapping Semantic Frames onto Czech Lexical Units

As the first step, we translated each LU belonging to groups of verbs of communication (C), mental action (MA), psych verbs (P), verbs of social interaction (SI), exchange (E), motion (M), transport (T), and location (L) from Czech into English.[4] The total number of annotated Czech LUs was 1 881 (341 verbs of communication, 308 verbs of mental action, 83 psych verbs, 85 verbs of social interaction, 129 verbs of exchange, 347 verbs of motion, 189 verbs of transport, and 399 verbs of location).[5]

Then the annotators had to indicate an appropriate SF (unambiguous assignment of SF) or more than one SF (ambiguous assignment of SF) for these LUs in FrameNet. The annotators could also conclude that no SF corresponds to a given Czech LU. For the overall statistics see Table 1.

Group of verbs	C	MA	P	SI	E	M	T	L
Total Czech LUs for annotation	340/341	308	83	85	129/129	347	189	399
Czech LUs without SF	66/77	125	29	54	21/27	100	34	178
Czech LUs with SF	274/264	183	54	31	108/102	247	155	221
ambiguous assignment	100/57	74	14	6	27/35	157	87	90
unambiguous assignment	174/207	109	40	25	81/67	90	68	131
SFs evoked by English LUs	415/338	292	73	38	150/140	566	279	337
Unambiguous assignments of SF	174/207	109	40	25	81/67	90	68	131
Ambiguous assignments of SF	241/131	183	33	13	69/73	476	211	206

Table 1. Annotated data size and overall statistics on the annotations of SFs.

The most frequent SFs assigned to Czech LUs include the following ones:

- communication: 'Statement', 'Request', 'Telling', 'Communication_manner', Reporting, 'Attempt_suasion';

[4] The on-line dictionary available at http://www.lingea.cz/ was used. The annotators were instructed to use all translations of a given LU provided by the lexicon.

[5] Verbs of communication and exchange were annotated by two annotators in parallel.

- `mental action:` 'Cogitation', Coming_to_believe', 'Becoming_aware', 'Assessing', 'Scrutiny', 'Grasp', 'Awareness', 'Experiencer_subj', 'Categorization', 'Hear';
- `psych:` 'Experiencer_obj', 'Cause_to_experience', 'Experiencer_subj', 'Prevarication', 'Attempt_suasion', 'Subjective_influence', 'Suasion', 'Perception_ _body', 'Objective_influence', 'Influence_on_event_on_cognizer';
- `social interaction:` 'Congregating', 'Forming_relationships', 'Residence', 'Personal_relationship', 'Make_acquaintance', 'Getting', 'Contacting', 'Be_in _agreement_on_assessment', 'Visiting', 'Temporary_stay';
- `exchange:` 'Giving', 'Getting', 'Commerce_pay', 'Theft', 'Receiving', 'Exchange', 'Commerce_buy', 'Bringing', 'Supply', 'Transfer'';
- `motion:` 'Self_motion', 'Motion', 'Arriving', 'Traversing', 'Departing', 'Body _movement', 'Operate_vehicle', 'Motion_directional', 'Path_shape', 'Ride_vehicle';
- `transport:` 'Cause_motion', 'Bringing', 'Removing', 'Cotheme', 'Sending', 'Placing', 'Delivery', 'Smuggling', 'Import_export', 'Taking';
- `location:` 'Placing', 'Attaching', 'Removing', 'Cause_motion', 'Change_posture', 'Theft', 'Residence', 'Being_located', 'Temporary_stay', 'Posture'.

Inter-annotator agreement. The feasibility of the assignment of SFs to Czech LUs was confirmed by the achieved inter-annotator agreement (IAA) measured on the groups of verbs of `communication` and `exchange`; these groups of verbs were chosen with respect to their different syntactic and semantic properties (Kettnerová et al., 2008b,a). Table 2 summarizes the inter-annotator agreement (IAA) and Cohen's κ statistics, see (Carletta, 1996), on the total number of SFs assigned to verbs of `communication` and `exchange`.

Match of SFs	IAA	κ
`C (communication)`	85.9%	0.82
`E (exchange)`	78.5%	0.73

Table 2. Inter-annotator agreement and κ statistics (considering the annotations of individual SFs for a given Czech LU as independent tasks).

Ambiguous assignments of SFs. Ambiguous annotations (i.e., annotations where the annotator has indicated more than one SF to a particular LU) draw attention to the divergence in granularity of word sense disambiguation adopted by VALLEX and FrameNet, which represents a great setback in any project dealing with mapping lexical resources.

First, let us focus on the cases in which two (or more) SFs mapped to a single Czech LU are connected by the hierarchical relation of 'Inheritance' – in general, there are these cases that reveal the finer granularity of senses applied in FrameNet. For instance, the SFs 'Bringing' and 'Smuggling' are assigned to the single Czech LU *převézt* ^{pf}, *převážet* ^{impf} 'to transport' / 'to smuggle', as in *They transported grapes to the wine lodges* and *They smuggle cocain from Peru to Britain*, respectively. The SF 'Smuggling' inherits the characteristics from the SF 'Bringing', its ancestor in the relation of 'Inheritance'; i.e., although the LU 'to smuggle' from the SF 'Smuggling' is semantically more specified – the transport is typically illegal – it inherits semantic properties from the LU 'to transport' evoking the SF 'Bringing'. We will return to the problem of different level of granularity of word sense disambiguation in Section 4.1 where we propose a method of overcoming this difficulty. This method also settles the ambiguous annotation in which sibling SFs in the relation of 'Inheritance' (or SFs with a common ancestor on an appropriate level, see below) are assigned to a single Czech LU.

Second, the ambiguous annotations of SFs that do not arise from the finer granularity (see above) may reveal mistakes in word sense disambiguation made in VALLEX. For instance, the SFs 'Grant_permission' and 'Permitting' are assigned to the Czech LU *dovolit* ^{pf}, *dovolovat* ^{impf} 'to allow', as in *Peter has allowed me to smoke here* and *This program allows data checking*, respectively. Although the verbal occurrences appear to be semantically close, the SFs evoked by them are not in the relation of 'Inheritance'. Thus this Czech LU represents a candidate for being split into two distinct senses. As a consequence, the FrameNet data can be used for checking word sense disambiguation in VALLEX.

3.2. Mapping Frame Elements onto Valency Complementations

If the human annotators indicated an appropriate SF for a Czech LU, they assigned the FE(s) from this SF to the valency complementation(s) (VCs in the following table) of the given Czech LU. Similarly as in case of mapping of SFs, more than one FE could be assigned to a single valency complementation ('Ambiguous annotation of FEs'). When no FE corresponded to a particular complementation, the annotators concluded that the given FE was missing. For the overall statistics see Table 3.

Inter-annotator agreement. As in case of SFs, the inter-annotator agreement (IAA) and κ statistics measured on the FEs assigned to the valency complementations of the verbs of `communication` and `exchange` gave satisfactory results, see Table 4.

Ambiguous assignments of FEs.
Type A. The first type of ambiguous assignments of FEs represents cases when an annotator concluded that more than one FE from a single SF corresponded to a single valency complementation due to a variety of lexical entailments imposed by a verb on such valency complementation. We can illustrate this case by the verb *zkontrolo-*

Group of verbs	C	MA	P	SI	E	M	T	L
Total VCs for annotation	1139/1142	861	259	215	522/522	1412	1024	1176
VCs without FEs (and without SF)	216/257	326	90	136	73/98	366	168	488
VCs without FEs (but with a SF)	30/22	32	8	2	38/37	32	26	26
VCs with FE(s)	893/863	503	161	77	411/387	1014	830	662
Unambiguous assignments of FE	427/534	242	116	59	276/211	275	271	289
Ambiguous assignments (type A)	212/194	98	8	2	50/75	268	218	189
Ambiguous assignments (type B)	351/195	219	46	16	112/150	680	497	279
Ambiguously assigned FEs (type A)	566/456	232	27	5	125/177	941	600	462
Ambiguously assigned FEs (type B)	952/526	537	111	32	277/309	2397	1341	735

Table 3. Annotated data size and overall statistics on the annotations of FEs.

Match of FEs	IAA	κ
C (communication)	95.6%	0.95
E (exchange)	91.2%	0.91

Table 4. Inter-annotator agreement and and κ statistics concerning assignment of FEs.

vat translated as 'to check', which belongs to the (only one) SF 'Inspecting' but has an ambiguous assignment of FEs, namely 'Patient' is labeled both with the FE 'Desired_state' and with the FE 'Ground', see (5)-(6):

(5) *Před odchodem zkontrolujte, (zda jsou zhasnutá světla)*.PAT-Desired_state
Eng. Before leaving check (that the lights are switched off).PAT-Desired_state

(6) *Zkontrolujte zámek*.PAT-Ground, *zda není porušen.*
Eng. Check the lock.PAT-Ground whether it is not damaged.

This case of the ambiguous assignment of FEs often results from the different approach to in/animateness which FrameNet and VALLEX take: VALLEX does not take into account in/animateness of the first and second inner participants, so 'Actor' and 'Patient' are often assigned ambiguously (in contrast to more semantically based valency complementations), see examples (7)-(8) in which the FEs 'Speaker' and 'Medium' are mapped onto 'Actor' of the verb *diktovat* 'to dictate':

(7) *Vzbouřenci*.ACT-Speaker *diktovali vládě*.ADDR-Addressee *své požadavky*.PAT-Message
Eng. The rebels.ACT-Speaker dictated their requirements.PAT-Message to the government.ADDR-Addresee

(8) *Mnichovská dohoda*.ACT-Medium *diktovala Československu*.ADDR-Addressee *(postoupit Německu pohraničí)*.PAT-Message

Eng. The Munich agreement.ACT-Medium ordered Czechoslovakia.ADDR-Addressee (to hand the border region over to Germany).PAT-Message

In case of verbs of `communication`, the ambiguous assignment of FEs to 'Patient' often follows from the fact that in Czech one abstract entity can express both 'theme' and 'what is said about the theme', see example (9):

(9) *Zprávy*.ACT-Medium *mluvily (o strašném zemětřesení, které zasáhlo v pátek ráno Turecko)*.PAT-Topic, Message
Eng. The news.ACT-Medium talked (about the horrible earthquake that struck Turkey on Friday morning).PAT-Topic, Message

Moreover, in Czech both 'Topic' and 'Message' can be expressed separately within a single structure (Daneš and Hlavsa, 1987), see example (10).

(10) *Cizinci*.ACT-Complainer *si stěžují starostovi*.ADDR-Addressee *na obchodníky*.PAT-Topic, *(že užívají dvojí ceny)*.EFF-Complaint
'foreigners – refl – complain – city mayor – about – sellers – that – use – double – prices'
Eng. The foreigners complain to the city mayor that the sellers use double prices.

Type B. The second type of the ambiguous assignment of FEs arises from the ambiguous assignment of SFs. In case that more than one SF were assigned to one Czech LU, the valency complementations of such Czech LU got FEs from all these SFs. For illustration, the Czech verb *dodat*$^{\mathrm{pf}}$, *dodávat*$^{\mathrm{impf}}$ translated by English verbs 'to supply' and 'to deliver' falls into the SFs 'Supply' and 'Delivery'. Thus the valency complementations of this verb are linked both with the FEs 'Supplier, 'Theme' and 'Recipient' belonging to the SF 'Supply' and with the FEs 'Deliverer', 'Recipient', and 'Theme' coming from the SF 'Delivery', see example (11a)-(11b):

(11) a. *Farmáři*.ACT-Supplier *dodávali obchodníkům*.ADDR-Recipient *čerstvou zeleninu*.PAT-Theme ('Supply')
Eng. a. The farmers.ACT-Supplier supplied the retailers.ADDR-Recipient with fresh vegetable.PAT-Theme ('Supply')
b. *Farmáři*.ACT-Deliverer *dodávali obchodníkům*.ADDR-Recipient*čerstvou zeleninu*.PAT-Theme ('Delivery')
Eng b. The farmers.ACT-Deliverer delivered fresh vegetable.PAT-Theme to the retailers.ADDR-Recipient ('Delivery')

Similarly as for SFs, the affected FEs may be connected by the relation of 'Inheritance' (as a result of SFs being in this relation). These cases arise from finer-grained granularity of word sense disambiguation in FrameNet. We will focus on them in Section 4.

In cases when ambiguously assigned FEs do not come from SFs connected by the relation of 'Inheritance' they may point out to mistakes in word sense disambiguation in VALLEX. These cases are left aside here.

4. Enhancing VALLEX with Semantic Information

In this section, we propose a method of enriching VALLEX with semantic classes (Section 4.1) and with semantic roles (Section 4.2) based on upper level SFs and their FEs from the relation of 'Inheritance'.

4.1. Enhancing VALLEX with Semantic Classes

In classifying Czech LUs into semantic classes and assigning semantic roles to their valency complementations, the semantic relation of 'Inheritance' plays a key role. This relation links such SFs which share basic semantic properties: each child frame inherits semantics from its parent frame(s). As for semantic classes, SFs from the appropriate upper level of this relation are chosen (top level SFs – represented by non-lexical and abstract SFs or SFs indicating a very general event – were disregarded); i.e., each Czech LU was classified according to the selected ancestor of the assigned SF. This method allows us to overcome the problem with coarser level of granularity made in VALLEX.

Let us demonstrate the principles of this classification on the verb *vyhnout se*$^{\mathrm{pf}}$, *vyhýbat se*$^{\mathrm{impf}}$ 'to sidestep'. This verb belongs to the SF 'Dodging' whose upper level ancestor SF in the relation of 'Inheritance' is represented by the SF 'Avoiding'. Thus to the given Czech LU, the SF 'Avoiding' is assigned as a semantic class. The same class is assigned also to the verbs belonging to the other descendant SF of 'Avoiding', namely 'Evading' (e.g., *uhnout*$^{\mathrm{pf}}$, *uhýbat*$^{\mathrm{impf}}$ 'to dodge'). See Figure 1 displaying the relation of 'Inheritance' of the SFs 'Avoiding', 'Dodging' and 'Evading'.

However, in case a Czech LU exhibits different morphosyntactic properties than LUs assigned by the relevant ancestor SF, we use the SF from an appropriate lower level of the relation of 'Inheritance'. E.g., the verb *doprovodit*$^{\mathrm{pf}}$, *doprovázet*$^{\mathrm{impf}}$ 'to accompany' belongs to the SF 'Cotheme' with the ancestor SF 'Self_motion'. Since in Czech this verb has different valency frame (obligatory 'Patient') than verbs onto which the SF 'Self_motion' was mapped (e.g., *běhat* 'to run', *kráčet* 'to march', *létat* 'to fly'), the SF 'Cotheme' from the lower level of the relation of 'Inheritance' was used as semantic class.

We set 81 SFs in total as candidates for semantic classes for verbs from the above mentioned eight groups of verbs, the entire list can be found in Appendix A.

The coverage of selected groups of verbs with these semantic classes is summarized in Table 5 (the numbers indicate a percentage of the annotated verbs from individual 'supergroups' to which semantic classes based on the selected SFs were as-

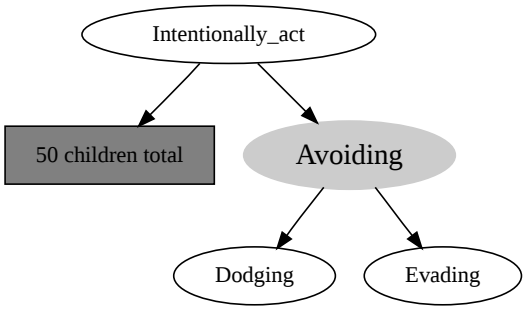

Figure 1. The relation of 'Inheritance' linking the SFs 'Avoiding', 'Dodging', and 'Evading'.

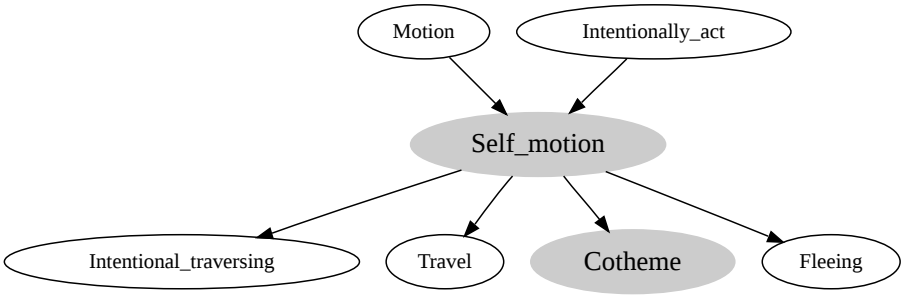

Figure 2. The relation of 'Inheritance' linking the SFs 'Cotheme' and 'Self_motion'.

signed). The differences in coverage are given primarily by the different coverage of the relation of 'Inheritance' in FrameNet.

The proposed method consisting in attributing the ancestor SFs of the assigned SFs as semantic classes allows us to overcome the problem with different granularity of verb senses in FrameNet and VALLEX. This method results in a usable set of syntactically and semantically homogeneous verb classes. Moreover, it represents also a solid basis for semantic classification of valency complementations, which is addressed in the following section.

4.2. Assigning Semantic Roles to Valency Complementations

Based on SFs mapping, we enhanced the valency lexicon with semantic roles. For this purpose, we use FEs from the ancestor SFs of the relation of 'Inheritance' that were

Groups of verbs		Coverage
C	(communication)	57/51%
MA	(mental action)	69%
P	(psych verbs)	13%
SI	(social interaction)	42%
E	(exchange)	58/56%
M	(motion)	88%
T	(transport)	84%
L	(location)	60%
Overall		76%

Table 5. Coverage of semantic classes.

chosen as semantic classes. For illustration, the valency complementations of the verb *vyhnout se*$^{\mathrm{pf}}$, *vyhýbat se*$^{\mathrm{impf}}$ 'to sidestep', included in the semantic class 'Avoiding' (representing the selected ancestor for the assigned SF 'Dodging') were labeled with FEs belonging to the SF 'Avoiding', namely 'Agent', 'Undesirable_situation', and the others, see Figure 3.

We obtained 327 FEs in total as candidates for semantic roles for the mentioned 8 'supergroups' of Czech verbs (only core FEs[6] as the most important ones are counted). The entire list can be found in Appendix B.

Similarly as in the case of semantic classes, there are differences in coverage of semantic roles, which are mainly given by the different coverage of the relation of 'Inheritance' in FrameNet.

5. Conclusion

We introduced the project aimed at enhancing the valency lexicon with missing semantic information – semantic classes and semantic roles. For this purpose, we made use of FrameNet data. We proposed a method of overcoming the problem with finer granularity of word sense disambiguation made in FrameNet. This method is based on the relation of 'Inheritance' built in FrameNet. As a result, 8 'supergroups' of Czech verbs, verbs of `communication`, `mental action`, `psych` verbs, verbs of `social interaction`, `exchange`, `motion`, `transport`, and `location` (specifically, 1 270 lexical units covering 2 129 Czech verbs in total if perfective and imperfective verbs being counted separately) were classified into syntactically and semantically coherent classes and

[6] According to (Ruppenhofer et al., 2006), core FEs are those FEs which are conceptually necessary and whose combination is characteristic of a particular SF.

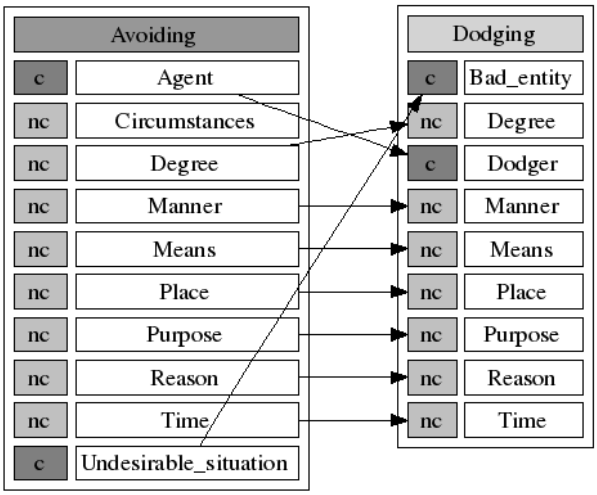

Figure 3. The relation of 'Inheritance' of the FEs belonging to the SF 'Avoiding' and 'Dodging'.

their valency complementations have been characterized by semantic roles adopted from the FrameNet lexical database.

As for future work, we intend to experiment with other groups of verbs and to increase the coverage of semantic information following the progress made in FrameNet.

5.1. Acknowledgement

This work has been using language resources developed and/or stored and/or distributed by the LINDAT-Clarin project of the Ministry of Education of the Czech Republic (project LM2010013). The paper reports on the research supported by the grant of GAČR No. P406/12/0557 and partially by the grant of GAČR P406/10/0875.

Bibliography

Baker, Collin F., Charles J. Fillmore, and John B. Lowe. The Berkeley FrameNet Project. In *Proceedings of the COLING-ACL*, pages 86–90, Montreal, Canada, 1998.

Carletta, Jean. Assessing agreement on classification tasks: The Kappa statistic. *Computational Linguistics*, 22(2):249–254, 1996.

Daneš, František and Zdeněk Hlavsa. *Větné vzorce v češtině.* Academia, Praha, 1987.

Dowty, David. Thematic Proto-Roles and Argument Selection. *Languauge*, 67:547–619, 1991.

Fillmore, Charles J., Christopher Johnson, and Miriam R. L. Petruck. Background to framenet. *International Journal of Lexicography*, 16(3):235–250, 2003.

Hajič, Jan, Eva Hajičová, Jarmila Panevová, Petr Sgall, Petr Pajas, Jan Štěpánek, Jiří Havelka, and Marie Mikulová. *Prague Dependency Treebank 2.0*. Linguistic Data Consortium, Philadelphia, PA, USA, 2006.

Kettnerová, Václava, Markéta Lopatková, and Klára Hrstková. Semantic roles in valency lexicon of czech verbs: Verbs of communication and exchange. In Ranta, Aarne and Bengt Nordström, editors, *Advances in Natural Language Processing (6th International Conference on NLP, GoTAL 2008)*, volume 5221 of *LNAI*, pages 217–221, Berlin / Heidelberg, 2008a. Springer.

Kettnerová, Václava, Markéta Lopatková, and Klára Hrstková. Semantic Classes in Czech Valency Lexicon: Verbs of Communication and Verbs of Exchange. In Sojka, Petr, Aleš Horák, Ivan Kopeček, and Karel Pala, editors, *Proceedings of Text, Speech and Dialog International Conference, TSD 2008*, volume 5246 of *LNAI*, pages 109–116, Berlin / Heidelberg, 2008b. Springer.

Levin, Beth C. *English Verb Classes and Alternations: A Preliminary Investigation*. The University of Chicago Press, Chicago and London, 1993.

Lopatková, Markéta, Zdeněk Žabokrtský, and Václava Kettnerová. *Valenční slovník českých sloves*. Nakladatelství Karolinum, Praha, 2008.

McConville, Mark and Myroslava O. Dzikovska. Using inheritance and coreness sets to improve a verb lexicon harvested from framenet. In *Proceedings of the Second Linguistic Annotation Workshop (LAW II)*, Marrakech, 2008. LREC.

Mikulová, Marie, Alevtina Bémová, Jan Hajič, Eva Hajičová, Jiří Havelka, Veronika Kolářová, Lucie Kučová, Markéta Lopatková, Petr Pajas, Jarmila Panevová, Magda Razímová, Petr Sgall, Jan Štěpánek, Zdeňka Urešová, Kateřina Veselá, and Zdeněk Žabokrtský. Annotation on the tectogrammatical level in the prague dependency treebank. annotation manual. Technical Report TR-2006-30, ÚFAL MFF UK, Prague, 2006.

Panevová, Jarmila. Valency Frames and the Meaning of the Sentence. In Luelsdorff, Philip A., editor, *The Prague School of Structural and Functional Linguistics*, pages 223–243. John Benjamins Publishing Company, Amsterdam, Philadelphia, 1994.

Ruppenhofer, Josef, Michael Ellsworth, Miriam R. L. Petruck, Christopher R. Johnson, and Jan Scheffczyk. *FrameNet II: Extended Theory and Practice*. University of California, Berkeley, 2006. http://framenet.icsi.berkeley.edu/book/book.html.

Sgall, Petr, Eva Hajičová, and Jarmila Panevová. *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. Reidel, Dordrecht, 1986.

Žabokrtský, Zdeněk and Markéta Lopatková. Valency information in VALLEX 2.0: Logical structure of the lexicon. *The Prague Bulletin of Mathematical Linguistics*, (87):41–60, 2007.

Appendix A: List of Semantic Frames Assigned to Lexical Units as Semantic Classes

- **communication:** 'Communication', 'Statement', 'Communication_response', 'Judgment_communication', 'Chatting, 'Prohibiting', 'Request', 'Reporting', and 'Commitment';
- **mental action:** 'Cogitation', 'Assessing', 'Memorization', 'Coming_to_believe', 'Becoming_aware', 'Awareness', 'Categorization', 'Scrutiny', 'Desiring', 'Differentiation', 'Opinion', 'Forgiveness', 'Certainty', 'Purpose', 'Memory', 'Judgment', 'Resolve_problem', 'Attention', and 'Deciding';
- **psych:** 'Cause_to_experience', and 'Eventive_cognizer_affecting';
- **social interaction:** 'Forming_relationships', 'Make_acquaintance', 'Be_in_agreeement_on_assessment', 'Visiting', 'Rewards_and_punishments', 'Hostile_encounter', and 'Finish_competition';
- **exchange:** 'Giving', 'Getting', 'Replacing', 'Exchange', 'Robbery', 'Hiring', 'Transfer', 'Frugality', 'Taking', and 'Supply';
- **motion:** 'Departing', 'Self_motion', 'Motion', 'Traversing', 'Motion_directional', 'Change
_posture', 'Cause_to_move_in_place', 'Avoiding', 'Surpassing',
'Cause_impact', 'Arriving', and 'Touring';
- **transport:** 'Cause_motion', 'Bringing', 'Cotheme', 'Filling', 'Firing', and 'Releasing';
- **location:** 'Placing', 'Attaching', 'Removing', 'Residence', 'Being_located', 'Inhibit_movement', 'Gathering_up', 'Aiming', 'Hiding_objects', 'Appointing', 'Cause_to_amalgamate', 'Being_attached', 'Arranging', 'Preserving', 'Emptying', and 'Amalgamation'.

Appendix B: List of Frame Elements Assigned to Valency Complementations as Semantic Roles

- **`communication`:**
  1. 'Communication': 'Communicator', 'Medium', 'Message', and 'Topic';
  2. 'Statement': 'Medium', 'Message', 'Speaker', and 'Topic';
  3. 'Communication_response': 'Addressee', 'Message', 'Speaker', 'Topic', and 'Trigger';
  4. 'Judgment_communication': 'Communicator', 'Evaluee', 'Expressor', 'Medium', 'Reason', and 'Topic;
  5. 'Chatting': 'Interlocutor_1', and 'Interlocutor_2';
  6. 'Prohibiting': 'Principle', and 'State_of_affairs';
  7. 'Request': 'Addressee', 'Medium', 'Message', 'Speaker', and 'Topic';
  8. 'Reporting': 'Authorities', 'Behavior', 'Informer', and 'Wrongdoer';
  9. 'Commitment': 'Addressee', 'Medium', 'Message', 'Speaker', and 'Topic'.
- **`mental action`:**
  1. 'Cogitation':' Cognizer', and 'Topic';
  2. 'Assessing': 'Assessor', 'Feature', 'Medium', 'Method', and 'Phenomenon';
  3. 'Memorization': 'Cognizer', and 'Pattern';
  4. 'Coming_to_believe': 'Cognizer', 'Content', 'Evidence', 'Medium', 'Means', and 'Topic';
  5. 'Becoming_aware': 'Cognizer', 'Instrument', 'Means', 'Phenomenon', and 'Topic';
  6. 'Awareness': 'Cognizer', 'Content', 'Topic', and Expressor;
  7. 'Categorization': 'Cognizer', 'Criteria', 'Item', and 'Category';
  8. 'Scrutiny': 'Cognizer', 'Ground', 'Instrument', and 'Medium';
  9. 'Desiring': 'Event', 'Experiencer', 'Focal_participant', and 'Location_of_event';
  10. 'Differentiation': 'Cognizer', 'Phenomena', 'Phenomenon_1', 'Phenomenon_2', and 'Quality';
  11. 'Opinion': 'Cognizer', and 'Opinion';
  12. 'Forgiveness': 'Judge', 'Evaluee', and 'Offense';
  13. 'Certainty': 'Cognizer', 'Content', 'Expressor', and 'Topic';
  14. 'Purpose': 'Agent', 'Attribute', 'Goal', 'Means', and 'Value';
  15. 'Memory': 'Cognizer', 'Content', and 'Topic';
  16. 'Judgment': 'Cognizer', 'Evaluee', 'Reason', and 'Expressor';
  17. 'Resolve_problem': 'Agent', 'Cause', and 'Problem';
  18. 'Attention': 'Expressor', 'Figure', and 'Perceiver';
  19. 'Deciding': 'Cognizer', and 'Decision'.
- **`psych`:**
  1. 'Cause_to_experience': 'Agent', and 'Experiencer';
  2. 'Eventive_cognizer_affecting': 'Cognizer', 'Content', and 'Event'.
- **`social interaction`:**
  1. 'Forming_relationships': 'Partner_1', 'Partner_2', and 'Partners';
  2. 'Make_acquaintance': 'Individuals', 'Individual_1', and 'Individual_2';
  3. 'Be_in_agreeement_on_assessment': 'Cognizer_1', 'Cognizer_2', 'Cognizers', 'Opinion', 'Question', and 'Topic';
  4. 'Visiting': 'Agent', and 'Entity';

5. 'Rewards_and_punishments': 'Agent', 'Evaluee', and 'Reason';
6. 'Hostile_encounter': 'Side_1', 'Side_2', 'Sides', 'Purpose', and 'Issue';
7. 'Finish_competition': 'Competition', 'Competitor', 'Opponent', and 'Competitors'.

- **exchange:**
  1. 'Giving': 'Donor', 'Recipient', and 'Theme';
  2. 'Getting': 'Recipient', and 'Theme';
  3. 'Replacing': 'Agent', 'New', and 'Old';
  4. 'Exchange': 'Exchanger_1', 'Exchanger_2', 'Theme_1', and 'Theme_2';
  5. 'Robbery': 'Perpetrator', 'Source', and 'Victim';
  6. 'Hiring': 'Employee', 'Employer', 'Field', 'Position', and 'Task';
  7. 'Transfer': 'Donor', 'Recipient', 'Theme', and 'Transferors';
  8. 'Frugality': 'Behavior', 'Resource', and 'Resource_controller';
  9. 'Taking': 'Agent', 'Source', and 'Theme';
  10. 'Supply': 'Purpose_of_recipient', 'Recipient', 'Supplier', and 'Theme'.
- **motion:**
  1. 'Departing': 'Source', and 'Theme';
  2. 'Self_motion': 'Area', 'Direction', 'Goal', 'Path', 'Self_mover', and 'Source';
  3. 'Motion': 'Area', 'Direction', 'Distance', 'Goal', 'Path', 'Source', and 'Theme';
  4. 'Traversing': 'Area', 'Direction', 'Distance', 'Goal', 'Path', 'Path_shape', 'Source', and 'Theme';
  5. 'Motion_directional': 'Area', 'Direction', 'Goal', 'Path', 'Source', and Theme';
  6. 'Change_posture': 'Protagonist';
  7. 'Cause_to_move_in_place': 'Agent', 'Body_part_of_agent', 'Cause', and 'Theme';
  8. 'Avoiding': 'Agent', and 'Undiserable_situation';
  9. 'Surpassing': 'Attribute', 'Profiled_attribute', 'Profiled_item', 'Standard_attribute', and 'Standard_item';
  10. 'Cause_impact': 'Agent', 'Cause', 'Impactee', 'Impactor', and 'Impactors';
  11. 'Arriving': 'Goal', and 'Theme';
  12. 'Touring': 'Attraction', and 'Tourist'.
- **transport:**
  1. 'Cause_motion': 'Agent', 'Area', 'Cause', 'Goal', 'Initial_state', 'Path', 'Result', 'Source', and 'Theme';
  2. 'Bringing': 'Agent', 'Area', 'Carrier', 'Goal', 'Path', 'Source', and 'Theme';
  3. 'Cotheme': 'Area', 'Cotheme', 'Direction', 'Goal', 'Path', 'Road', 'Source', and 'Theme';
  4. 'Filling': 'Agent', 'Cause', 'Goal', and 'Theme';
  5. 'Firing': 'Employee', 'Employer', 'Position', and 'Task';
  6. 'Releasing': 'Agent', 'Location_of_confinement', and 'Theme'.
- **location:**
  1. 'Placing': 'Agent', 'Cause', 'Goal', and 'Theme';
  2. 'Attaching': 'Agent', 'Connector', 'Item', 'Items', and 'Goal';
  3. 'Removing': 'Agent', 'Cause', 'Source', and 'Theme';
  4. 'Residence': 'Resident', 'Co_resident', and 'Location';
  5. 'Being_located': 'Theme', and 'Location';
  6. 'Inhibit_movement': 'Agent', 'Cause', 'Theme', and 'Holding_location';
  7. 'Gathering_up': 'Agent', 'Aggregate', and 'Individuals';

8. ‘Aiming’: ‘Agent’, ‘Instrument’, ‘Targeted’, and ‘Target_location’;
9. ‘Hiding_objects’: ‘Agent’, ‘Hidden_object’, and ‘Hiding_place’;
10. ‘Appointing’: ‘Selector’, ‘Role’, ‘Official’, ‘Function’, and ‘Body’;
11. ‘Cause_to_amalgamate’: ‘Agent’, ‘Part_1’, ‘Part_2’, ‘Parts’, and ‘Whole’;
12. ‘Being_attached’: ‘Item’, ‘Items’, ‘Goal’, and ‘Connector’;
13. ‘Arranging’: ‘Agent’, ‘Configuration’, and ‘Theme’;
14. ‘Preserving’: ‘Agent’, ‘Medium’, and ‘Undergoer’;
15. ‘Emptying’: ‘Agent’, ‘Cause’, ‘Source’, and ‘Theme’;
16. ‘Amalgamation’: ‘Parts’, ‘Part_1’, ‘Part_2’, and ‘Whole’.

Address for correspondence:
Václava Kettnerová
kettnerova@ufal.mff.cuni.cz
Institute of Formal and Applied Linguistics
Charles University in Prague
Malostranské náměstí 25
118 00 Praha 1, Czech Republic

PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 97 APRIL 2012 43–53

BIA: a Discriminative Phrase Alignment Toolkit

Patrik Lambert[a], Rafael E. Banchs[b]

[a] LIUM, LUNAM Université, University of Le Mans
[b] Institute for Infocomm Research

Abstract

In most statistical machine translation systems, bilingual segments are extracted via word alignment. However, word alignment is performed independently from the requirements of the machine translation task. Furthermore, although phrase-based translation models have replaced word-based translation models nearly ten years ago, word-based models are still widely used for word alignment. In this paper we present the BIA (BIlingual Aligner) toolkit, a suite consisting of a discriminative phrase-based word alignment decoder based on linear alignment models, along with training and tuning tools. In the training phase, relative link probabilities are calculated based on an initial alignment. The tuning of the model weights may be performed directly according to machine translation metrics. We give implementation details and report results of experiments conducted on the Spanish–English Europarl task (with three corpus sizes), on the Chinese–English FBIS task, and on the Chinese–English BTEC task. The BLEU score obtained with BIA alignment is always as good or better than the one obtained with the initial alignment used to train BIA models. In addition, in four out of the five tasks, the BIA toolkit yields the best BLEU score of a collection of ten alignment systems. Finally, usage guidelines are presented.

1. Introduction

Most statistical machine translation (SMT) systems (*e.g.* phrase-based, hierarchical, n-gram-based) build their translation models from word alignment trained in a previous stage. Many papers have shown that intrinsic alignment quality is poorly correlated with MT quality (for example, Vilar et al. (2006)). Accordingly, some research has attempted to tune the alignment directly according to specific MT evaluation metrics (Lambert et al., 2007). Furthermore, although phrase-based transla-

 Corresponding author: patrik.lambert@lium.univ-lemans.fr
Cite as: Patrik Lambert, Rafael E. Banchs. BIA: a Discriminative Phrase Alignment Toolkit. The Prague Bulletin of Mathematical Linguistics No. 97, 2012, pp. 43–53. doi: 10.2478/v10108-012-0003-z.

tion models have replaced word-based translation models nearly ten years ago, word-based models are still widely used for word alignment.

In this paper we present the BIA (BIlingual Aligner) toolkit, a suite consisting of a discriminative phrase-based word alignment decoder based on linear alignment models (Moore, 2005; Liu et al., 2005, 2010), along with training and tuning tools. Thus this toolkit allows one to overcome the limitations of most current word alignment systems: the basic alignment unit is not a single word but a phrase (a group of consecutive words),[1] and it provides tools to tune the alignment model parameters directly according to MT metrics. Although currently these tuning tools are implemented to work with the Moses phrase-based decoder (Koehn et al., 2007), it is straightforward to extend them to work with other MT systems (hierarchical, n-gram-based, etc.).

The paper is organised as follows. In Section 2, we present the alignment algorithm and the tuning procedure. In Section 3, we detail how we implemented the different parts of the alignment system presented in Section 2. Then in Section 4, we report results of experiments conducted on the Spanish–English Europarl task (with three corpus sizes), on the Chinese–English FBIS task, and on the Chinese–English BTEC task. In Section 5, we give instruction for training, tuning and decoding with our toolkit. Finally, some conclusions are provided.

2. Phrase-based Discriminative Alignment System

2.1. Alignment Algorithm

This aligner implements a linear combination of feature functions calculated at the sentence pair level. It searches the alignment hypothesis $\hat{\mathbf{a}}$ which maximises this linear combination, as expressed in (1):

$$\hat{\mathbf{a}} = \arg\max_{\mathbf{a}} \sum_{m} \lambda_m h_m(\mathbf{s}, \mathbf{t}, \mathbf{a}), \tag{1}$$

where $\mathbf{s}$, $\mathbf{t}$ and $\mathbf{a}$ refer respectively to the source sentence, the target sentence and the alignment hypothesis, h stands for the feature functions used and the λs are their corresponding weights. It follows a two-pass strategy, as proposed by Moore (2005). The initial alignment may be computed using BIA with a first set of features (Lambert and Banchs, 2008), or with any other alignment system. In the experiments presented in Section 4, we actually took as initial alignment the combination of the IBM Model 4 source–target and target–source alignments with the "grow-diag-final-and" heuristic (Koehn et al., 2003). This initial alignment was used to calculate the following improved features:

[1] The output of BIA is nevertheless an alignment at the word level, that is, in many SMT systems, the step previous to phrase-pair extraction.

- a phrase association score model with relative link probabilities (Melamed, 2000). These links are between phrases (although in practice most phrases are of length one, *i.e.* single words).
- source and target *word* fertility models giving the probability for a given *word* to have one, two, three or more than three links.

These improved features, together with the following features, were used to align the corpus in a second pass:

- A link bonus model, proportional to the number of links in $\mathbf{a}$.
- Two distortion models, counting respectively the number and amplitude (the difference between target word positions) of crossing links.
- A 'gap penalty' model, proportional to the number of embedded positions between two target words linked to the same source words, or between two source words linked to the same target words.

To find the best hypothesis, we implemented a beam-search algorithm based on dynamic programming (see Section 3).

2.2. Weight Optimisation According to BLEU Score

The alignment weights λ of Equation 1 are optimised so as to maximise the BLEU score calculated on a parallel development corpus (with no alignment annotations), as proposed by Lambert *et. al* (2007).

The optimisation algorithm (presented in Section 2.3) adjusts the weights so as to maximise the BLEU score. At each iteration, the training corpus is aligned as described in Section 2.1. This alignment is used to build an SMT system, including bilingual phrase extraction, translation model(s) estimation and MERT (Och, 2003). Then the development corpus is translated with this SMT system and the BLEU score is computed. Two different development sets can be used for the alignment weight optimisation (Dev) and the MERT process performed at each iteration (MERT Dev).

Note that it would be straightforward to introduce MT metrics other than the BLEU score, and that it would be easy to implement a supervised weight optimisation procedure (for example according to F-score).

2.3. Optimisation Algorithm

The available optimiser is the SPSA algorithm (Spall, 1992). The SPSA (Simultaneous Perturbation Stochastic Approximation) is a stochastic implementation of the conjugate gradient method which requires only two evaluations of the objective function, regardless of the dimension of the optimisation problem. The SPSA procedure is in the general recursive stochastic approximation form, as shown in 2:

$$\hat{\lambda}_{k+1} = \hat{\lambda}_k - \alpha_k \hat{\mathbf{g}}_k(\hat{\lambda}_k) \tag{2}$$

where $\hat{g}_k(\hat{\lambda}_k)$ is the estimate of the gradient $g(\lambda) \equiv \partial E/\partial\lambda$ at the iterate $\hat{\lambda}_k$ based on the previous evaluations of the objective function. α_k denotes a positive number that usually decreases as k increases. In the default settings, the gradient is computed with a one-sided approximation which, given $E(\hat{\lambda}_k)$, requires the evaluation of $E(\hat{\lambda}_k + \text{perturbation})$. The original SPSA algorithm has been adapted to achieve convergence after typically 60 to 100 evaluations of the objective function. Note that in general, this algorithm converges to a local minimum.

3. Implementation

The BIA toolkit is implemented in C++ (with the Standard Template Library) and Perl and contains:

- training tools (mostly in C++).
- an alignment decoder (in C++).
- tools to tune the alignment model parameters directly according to MT metrics (in Perl).
- Perl scripts which pilot the training, tuning and decoding tasks.
- a sample (bash) shell script to run the whole pipeline (the same as the one used to produce the results of Section 4, but with sample data).

Although the BIA toolkit uses the Moses toolkit (Koehn et al., 2007) by default for two tasks, it is straightforward to use other tools instead. First, the initial alignment used to train BIA models (see Section 2.1) is by default the "grow-diag-final-and" alignment computed by Moses. However, any other initial alignment may be used instead. Second, in the BIA tuning tools, a function commands the training, tuning and evaluation of an MT system from the output of the alignment decoder and returns an MT score to be optimised. Currently, only a function performing these steps with the Moses phrase-based SMT system is implemented, in which the symmetrised Giza++ alignment is substituted by the BIA alignment. However, the only task required to extend the toolkit to another MT system is to write another function performing the same steps for that MT system.

The toolkit has been only tested in linux, but should be portable to any system compatible with `cmake`. No multi-threading is implemented. However, a parameter for the number of threads available allows the user to divide tasks by forking or submitting jobs to a cluster (via the `qsub` command).

3.1. Training

The main training task is the estimation of the phrase association model and the source and target fertility models (the "improved features" described in Section 2.1). This task consists of counting the number of links and co-occurrences found in the initial alignment for each co-occurring phrase pair (to calculate relative probabilities),

as well as the number of links for each source and target word. It is performed using `hash maps` with a custom hash function.

3.2. Decoding

Before aligning each sentence pair, models are loaded in memory (into `hash maps`). Then, for each sentence pair, a set of links to be considered in search is selected. This set is formed by the n best links for each source and for each target phrase (typically $n = 3$). For each link selected, relevant information (source and target positions, costs, etc.) is stored in a specific data structure. The set of considered links is then arranged in stacks corresponding to each source (or target) word.

Decoding consists in extending alignment hypotheses (that is, sets of links), also called states, by including each link of these link stacks. Note that we use an hypothesis stack for each number of source+target words covered. Decoding is based on a beam-search algorithm as follows:

```
insert initial state (empty alignment) in hypothesis stack
for each stack of links considered in search
* for each state in each hypothesis stack
    for each link in link stack
      - expand current state by adding this link
      - place new state in corresponding hypothesis stack
* perform histogram and threshold pruning of hypothesis stacks
```

Note that having one link stack for each source (or target) word ensures a fair comparison between hypotheses in which this word is covered. Furthermore, multiple hypothesis stacks ensure a fair comparison between hypotheses having the same number of covered words.

3.3. Tuning

At each iteration of the alignment weight tuning procedure (see Section 2.2), an SMT system is build, with which the development set is translated. The feature weights of this SMT system may be kept constant during alignment tuning, or they may be tuned with MERT at each iteration. In the latter case, we restrict the number of MERT runs to 12 iterations (not limited by default) and 10 restarts[2] (20 by default), to limit its maximum processing time. We also increase the minimum required change in weight variable from 0.00001 to 0.0001. Internal experiments on the tasks presented in Section 4 showed consistently that it is better to perform MERT at each iteration than to use a constant set of SMT feature weights during alignment tuning.

[2] According to internal experiments on two tasks, the average and standard deviation after 10 MERT runs is not affected by using 10 restarts instead of 20. In contrast, limiting the number of iterations to 12 may of course affect the results.

3.4. Issues

We had to face some issues during the implementation of the algorithm. First, the alignment result depends on the order of introduction of the links in the alignment hypotheses. Several solutions were envisaged: (i) a future cost; however, it should include the cost of crossing links, which we found no effective way to estimate. (ii) introduce the most confident or less ambiguous links first (iii) start from a non-empty initial alignment (for example, decode along the source side, then along the target side, and finally re-decode taking the intersection as initial alignment). In this configuration, we can expand a state by deleting or substituting a link. (iv) use multiple hypothesis stacks, which help decoding being more stable.

Second, the tuning process is not very stable (the optimisation algorithm can fall into a poor local maximum).

4. Experimental Evaluation

4.1. Data Sets

The experiments were conducted for the following tasks:

- the TC-STAR OpenLab[3] Spanish–English EPPS parallel corpus, which contains proceedings of the European Parliament. The BIA alignment model weights were tuned on two subsets extracted by randomly selecting 100,000 and 20,000 sentence pairs (these subsets will be referred to as 'ran100k' and 'ran20k' respectively). We built SMT systems from the optimum alignment obtained on each of these subsets. We also aligned the whole corpus (referred to as 'full') with the optimum weights obtained by tuning on the ran100k corpus, and built an SMT system from this alignment.
- the Chinese–English FBIS corpus, a collection (LDC2003E14) of texts in the news domain and released by the Linguistic Data Consortium (LDC[4]). We selected 100k sentence pairs as training data. Since the data was released in 2003, we used the test sets of NIST 2001 (nist01), NIST 2002 (nist02) and NIST 2003 (nist03) as development and test data.
- the Chinese–English data provided within the IWSLT 2007 evaluation campaign, extracted from the Basic Travel Expression Corpus (BTEC) (Takezawa et al., 2002). This speech corpus contains sentences similar to those that are usually found in phrase books for tourists going abroad. Training data consisted of the default training set, to which we added the sets devset1, devset2 and devset3.

The characteristics of the training, development and test sets used in each task are indicated in Tables 1, 2 and 3. More specifically, the statistics shown are the number

[3]http://www.tcstar.org/openlab2006

[4]http://www.ldc.upenn.edu

Set	Language	Sentences	Words	Vocabulary	Lmean	Ref.
Train	Spanish	1.27 M	36.2 M	152 k	28.4	1
(full)	English	1.27 M	34.6 M	106 k	27.2	1
Train	Spanish	100 k	2.8 M	55 k	28.4	1
(ran100k)	English	100 k	2.7 M	38 k	27.2	1
Train	Spanish	20 k	0.57 M	27 k	28.6	1
(ran20k)	English	20 k	0.55 M	20 k	27.3	1
MERT Dev.	Spanish	892	28.6 k	4.8 k	32.0	2
1st ref.	English	892	28.9 k	3.9 k	32.4	
Dev.	Spanish	1008	25.8 k	3.9 k	25.6	2
1st ref.	English	1008	26.3 k	3.1 k	26.1	
Test	Spanish	840	22.7 k	4.1 k	27.1	2
1st ref.	English	840	22.8 k	3.3 k	27.1	

Table 1. Statistics for the training, development and test data sets for EPPS data (M and k stand for millions and thousands, respectively, Lmean refers to the average sentence length in number of words, and Ref. to the number of available translation references).

of sentences, the number of words, the vocabulary size (or number of distinct words), the average sentence length in number of words and the number of available translation references. As mentioned previously, Dev refers to the development set used to calculate the BLEU score at each alignment optimisation iteration (with the SPSA algorithm), MERT Dev refers to the development corpus used within the internal SMT MERT procedure at each SPSA iteration, and Test refers to the test set used to realise an extrinsic evaluation of the optimal alignment system.

4.2. Results

In this section we compare the BLEU score obtained by SMT systems built from alignments computed with the BIA toolkit and with a number of state-of-the-art alignment systems. The second-pass BIA models were trained on a high-quality alignment computed by combining the IBM Model 4 source–target (s2t) and target–source (t2s) alignments with the "grow-diag-final-and" heuristic (Koehn et al., 2003). We also computed the source–target and target–source alignments of IBM Model 4 (Brown et al., 1993) as implemented by Giza++, and 4 different combinations of these alignments (intersection (I), union (U), grow-diag-final (GDF) and grow-diag-final-and (GDFA) heuristics (Koehn et al., 2003)). In addition, we used an HMM-based joint training model with posterior decoding (Liang et al., 2006) and an HMM-based model which explicitly takes into account the target language constituent structure (DeNero

Set	Language	Sentences	Words	Vocabulary	Lmean	Ref.
Train	Chinese	100 k	3.1 M	30.4 k	30.7	1
	English	100 k	3.7 M	56.2 k	37.3	1
MERT Dev.	Chinese	935	27.9 k	4.6 k	29.9	3
1st ref.	English	935	28.9 k	4.9 k	30.9	
Dev.	Chinese	993	26.7 k	4.7 k	26.9	8
1st ref.	English	993	29.1 k	4.9 k	29.3	
Test	Chinese	878	25.4 k	4.3 k	28.9	5
1st ref.	English	878	28.2 k	4.8 k	32.1	

Table 2. Basic statistics for the training, development and test data sets for FBIS data.

Set	Language	Sentences	Words	Vocabulary	Lmean	Ref.
Train	Chinese	41.5 k	362 k	11.4 k	8.7	1
	English	41.5 k	389 k	9.7 k	9.4	1
MERT Dev.	Chinese	500	6.1 k	1.3 k	12.1	7
1st ref.	English	500	7.3 k	1.2 k	14.7	
Dev.	Chinese	489	5.7 k	1.1 k	11.7	7
1st ref.	English	489	6.4 k	1.0 k	13.4	
Test	Chinese	489	3.2 k	0.9 k	6.5	6
1st ref.	English	489	3.7 k	0.8 k	7.6	

Table 3. Basic statistics for the training, development and test data sets for BTEC data.

and Klein, 2007), both implemented in the Berkeley word alignment package,[5] and referred to as "bk" and "syn-bk", respectively. Finally, we computed alignments with the Posterior Constrained Alignment Toolkit (PostCAT[6] (Graça et al., 2010)).

Table 4 shows, for each of the five considered tasks, the BLEU score of the SMT systems built from four types of alignments: (i) BIA alignment, (ii) the best alignment(s) (named in parenthesis) among the nine other alignment systems, (iii) the Moses default alignment (GDF), and (iv) the initial alignment used to train BIA models (GDFA). These BLEU scores are an average obtained over four MERT runs with different random seeds. We can make a number of interesting observations from these results. First, in all cases, the score achieved via BIA alignment was at least as good as the score achieved via the initial alignment used to train BIA models. Second, with respect to the Moses default alignment scheme, using BIA yielded a lost of 0.1 BLEU point in one task, and gains of 0.5, 0.4, 1.3 and 1.2 BLEU points in the other tasks. Fi-

[5] http://nlp.cs.berkeley.edu/pages/WordAligner.html

[6] http://www.seas.upenn.edu/~strctlrn/CAT/CAT.html

	EPPS				
Alignment	Full	Ran100k	Ran20k	FBIS	BTEC
BIA	56.2	**51.7**	**46.6**	**23.0**	**35.2**
Best other	**56.7** (U)	51.4 (t2s)	46.2 (U,GDF,GDFA)	**23.0** (GDFA)	34.8 (bk,syn-bk)
Moses default (GDF)	56.3	51.2	46.2	21.7	34.0
Initial (GDFA)	56.2	51.1	46.2	23.0	33.9

Table 4. Extrinsic evaluation (in terms of BLEU score) of the BIA alignment system, compared to the other alignment systems considered.

nally, BIA always yielded the best alignment (in terms of BLEU score) of the set of ten alignment systems when its model parameters had been tuned on the whole corpus. This was the case for the EPPS ran100k and ran20k tasks, and for the FBIS and BTEC tasks. For the EPPS "full" task, the parameters had been tuned on the ran100k task, and the whole corpus had then been aligned with the optimal parameters found.

This last result is problematic for the alignment of large corpora given a limitation of the current version: a full SMT system must be built at each iteration of the alignment parameter optimisation, which would be very costly on a large corpus. Thus the tuning cannot be performed on the whole training corpus, unless it is reasonably small.

5. Usage Instructions

Training, tuning and decoding instructions are available on the BIA aligner project website,[7] from where the source code can also be freely downloaded. The sample shell script also gives usage examples. Several wrapper scripts were implemented to make training, tuning and decoding easier. To train the alignment models, use:
`training/train-models.pl`
To tune the alignment feature weights, use:
`tuning/tune-model-weights.pl`
and finally to run the alignment decoder, use the following binary:
`bia`

6. Conclusions and Further Work

We presented the BIA toolkit, a suite consisting of a discriminative phrase-based word alignment decoder based on linear alignment models, along with training and

[7] http://code.google.com/p/bia-aligner/

tuning tools. The tuning of the model weights may be performed directly according to MT metrics.

We reported results of experiments conducted on the Spanish–English Europarl task (with three corpus sizes), on the Chinese–English FBIS task, and on the Chinese–English BTEC task. The BLEU score obtained with BIA alignment was always as good or better than the one obtained with the initial alignment used to train BIA models. In addition, BIA always yielded the best alignment (in terms of BLEU score) of a set of ten alignment systems when its model parameters had been tuned on the whole corpus. In one task, the corpus was too large to perform the tuning on all the data and thus tuning was performed on a subset of it (less than 10% of its size).

In the future we want to develop a new tuning procedure whose required computing time would be independent from the size of the training corpus.

Acknowledgements

This research is supported by the European Union under the EuroMatrix Plus project (http://www.euromatrixplus.net, IST-2007.2.2-FP7-231720) and by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (www.cngl.ie) at Dublin City University.

Bibliography

Brown, Peter F., Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311, 1993.

DeNero, John and Dan Klein. Tailoring word alignments to syntactic machine translation. In *Proc. of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 17–24, Prague, Czech Republic, June 2007.

Graça, João V., Kuzman Ganchev, and Ben Taskar. Learning tractable word alignment models with complex constraints. *Computational Linguistics*, 36(3):481–504, 2010.

Koehn, Philipp, Franz J. Och, and Daniel Marcu. Statistical phrase-based translation. In *Proc. of the Human Language Technology Conference of the NAACL*, pages 48–54, Edmonton, Canada, 2003.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open source toolkit for statistical machine translation. In *Proc. of the 45th Annual Meeting of the Association for Computational Linguistics (Demo and Poster Sessions)*, pages 177–180, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P/P07/P07-2045`.

Lambert, Patrik and Rafael E. Banchs. Word association models and search strategies for discriminative word alignment. In *Proc. of the Conference of the European Association for Machine Translation*, pages 97–103, Hamburg, Germany, 2008.

Lambert, Patrik, Rafael E. Banchs, and Josep M. Crego. Discriminative alignment training without annotated data for machine translation. In *Proc. of the Human Language Technology Conference of the NAACL (Short Papers)*, pages 85–88, Rochester, NY, USA, 2007.

Liang, Percy, Ben Taskar, and Dan Klein. Alignment by agreement. In *Proc. of the Human Language Technology Conference of the NAACL*, pages 104–111, New York City, USA, June 2006.

Liu, Yang, Qun Liu, and Shouxun Lin. Log-linear models for word alignment. In *Proc. of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 459–466, Ann Arbor, Michigan, June 2005.

Liu, Yang, Qun Liu, and Shouxun Lin. Discriminative word alignment by linear modeling. *Computational Linguistics*, 36(3):303–339, 2010.

Melamed, I. Dan. Models of translational equivalence among words. *Computational Linguistics*, 26(2):221–249, 2000.

Moore, Robert C. A discriminative framework for bilingual word alignment. In *Proc. of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 81–88, Vancouver, Canada, October 2005.

Och, Franz J. Minimum error rate training in statistical machine translation. In *Proc. of the 41th Annual Meeting of the Association for Computational Linguistics*, pages 160–167, 2003.

Spall, James C. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Trans. Automat. Control*, 37:332–341, 1992.

Takezawa, Toshiyuki, Eiichiro Sumita, Fumiaki Sugaya, Hirofumi Yamamoto, and Seiichi Yamamoto. Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world. In *Proc. of Third International Conference on Language Resources and Evaluation 2002*, pages 147–152, Las Palmas, Canary Islands, Spain, 2002.

Vilar, David, Maja Popovic, and Hermann Ney. AER: Do we need to "improve" our alignments? In *Proc. of the Int. Workshop on Spoken Language Translation, IWSLT'06*, pages 205–212, Kyoto, Japan, 2006.

Address for correspondence:
Patrik Lambert
patrik.lambert@lium.univ-lemans.fr
LIUM, University of Le Mans
Avenue Laënnec, 72085 Le Mans Cedex 9, France

PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 97 APRIL 2012 55–82

Le syntagme nominal défini en arabe standard contemporain

Mahmoud Fawzi Mammeri, Nacereddine Bouhacein

Laboratoire ÉTAL (Études Théoriques et Appliquées en Linguistique arabe et générale)
Département de la Langue Arabe, Faculté des Lettres et Sciences Sociales
Université Saad Dahlab, BP 270 RP, Blida, Algérie

Abstract

Cet article propose une analyse HPSG du syntagme nominal en arabe standard moderne. L'accent y est mis spécialement sur le phénomène de la définitude. Pour cela, nous examinons un certain nombre de phénomènes qui lui sont liés dont le marquage de la définitude au sein du syntagme nominal, et en particulier dans le syntagme génitival, l'accord en définitude entre les différents éléments du syntagme nominal, le statut affixal de l'article défini, la combinaison de l'article défini avec les éléments de l'état construit et l'héritage de la définitude au niveau de la projection maximale. L'analyse que nous développons repose sur deux hypothèses fondamentales. La première stipule que l'article défini *al-* de l'arabe est un affixe flexionnel, qui, par conséquent, trouve sa place dans la morphologie (ou le lexique) plutôt que dans la syntaxe. Pour affirmer cette hypothèse, nous apportons des arguments en appliquant les critères communément appelés les critères de Zwicky. De plus, notre analyse, qui a été implémentée dans un cadre lexicaliste, se base sur une deuxième hypothèse qui reconnaît le nom comme tête du syntagme nominal ; par conséquent, nous considérons les syntagmes nominaux comme étant des NPs et non des DPs.

Introduction

Le présent article est consacré à la grammaire du syntagme nominal en arabe standard moderne (ASM). L'ASM est la langue de la littérature arabe moderne depuis plus d'un demi-siècle. Elle est aussi la langue utilisée par les médias. C'est la langue la plus largement comprise par les locuteurs arabes et c'est aussi la langue enseignée dans les écoles. L'ASM est aujourd'hui la langue officielle de presque une vingtaine de pays (Kouloughli, 1994c, pp :14–15). Cette langue est parlée à l'ouest du continent asiatique et au nord et à l'est du continent africain (Kouloughli, 1994c, p :7). L'analyse que

Corresponding author : Mahmud.mammeri@gmail.com
Cite as : Mahmoud Fawzi Mammeri, Nacereddine Bouhacein. Le syntagme nominal défini en arabe standard contemporain. The Prague Bulletin of Mathematical Linguistics No. 97, 2012, pp. 55–82.

nous développerons dans ce texte est implémentée dans le cadre lexicaliste des grammaires syntagmatiques endocentriques (Head-driven Phrase Structure Grammar, ou HPSG). HPSG, décrite dans Pollard (1994) et Sag (1999), théorie issue de plusieurs courants théoriques modernes, relève des grammaires d'unification. C'est une grammaire bidirectionnelle[1] se proposant de fournir un cadre de modélisation de principes grammaticaux universels. Ce qui différencie HPSG des autres modèles, est sa volonté de donner des descriptions uniformes des différentes strates du langage. Cette uniformité de la modélisation se manifeste en ce que le modèle de toute unité est construit sur le même patron quelle que soit sa taille. En d'autres termes, il s'agit d'utiliser les structures de traits comme cadre unique pour représenter des informations linguistiques de natures aussi hétérogènes que phonologiques, syntaxiques, sémantiques, etc. Ainsi, un mot (i.e. une unité du lexique) est représenté de la même manière qu'un syntagme ou une phrase, voire un discours ; tous ces différents objets, tout en étant des signes, ne sont que des structures de traits typées (Typed Feature Structures, ou TFSs). En plus, et c'est ce qui rend la grammaire HPSG plus attractive, les règles de grammaires, les principes généraux et les grammaires elles mêmes ne sont, eux aussi, que des structures de traits typées. Le principal centre d'intérêt de cet article est le phénomène de la *définitude*. Nous présenterons une analyse du syntagme nominal arabe qui rend compte de la définitude au sein de ce dernier. Le marquage de la définitude est assez particulier en arabe et dans d'autres langues sémitiques, surtout en hébreu moderne (HM) et en amharique. Cette spécificité nous a poussés à mener une exploration du phénomène dans ces deux langues de la même famille avant d'avancer une analyse pour l'arabe. Pourquoi une telle exploration ? Les langues sémitiques, surtout l'ASM et l'HM (et avec un degré moindre, l'amharique), partagent certaines caractéristiques linguistiques qui leur sont particulières. Nous passerons ici en revue les importantes ressemblances et les différences mineures entre l'ASM, l'HM et l'amharique en ce qui concerne la définitude au sein du syntagme nominal. Notre analyse repose sur deux hypothèses principales. La première, consiste à stipuler que l'article défini *al-* en arabe est un affixe flexionnel, par conséquent il trouve sa place tant en morphologie (ou lexique) qu'en syntaxe. Pour affirmer cette hypothèse, nous apportons des arguments en appliquant les critères communément appelés *Critères de Zwicky* (Zwicky, 1983, 1985a ; Miller, 1992). La seconde hypothèse, que nous supposerons affirmée, et que nous ne discuterons pas ici par faute d'espace, stipule que la tête du syntagme nominal est un nom. Cet article est organisé en trois sections. La section 1 décrit la définitude au sein du syntagme nominal arabe à partir de données empiriques assez représentatives. Ces données nous permettront de montrer la complexité du phénomène et nous aideront à développer notre analyse ultérieurement. La section 2 discute de la nature affixale de l'article défini en arabe, et répond ainsi à une question fondamentale à ce travail qu'on peut formuler ainsi : l'article défini *al-* en arabe est-il un affixe ? Dans la section 3, nous développerons une analyse du syn-

[1] Les grammaires écrites dans ce modèle sont utilisées en analyse comme en génération.

tagme nominal arabe, basée sur les contraintes, qui permettra d'expliquer les données présentées dans la section 1.

1. Une étude descriptive de la définitude au sein du SN en arabe

Certaines langues telles que le français et l'anglais marquent les noms définis et les noms indéfinis au niveau de la syntaxe. La langue arabe, ainsi que d'autres langues sémitiques telles que l'hébreu et l'amharique, le font au niveau de la morphosyntaxe (i.e. à l'interface morphologie-syntaxe) et d'une manière assez différente. L'arabe présente un seul article défini, *al-*. *Al-* ne subit aucune flexion et est préfixé à certains types de mots, jamais à des syntagmes. Il peut se combiner avec la plupart des nominaux : noms communs, certains noms propres, adjectifs et numéraux (i.e. nombres ordinaux et nombres cardinaux). De plus, les syntagmes nominaux définis sont dits *poly-définis* ; la plupart des éléments du syntagme 'défini' doivent être explicitement définis et il y a une condition stricte quant à l'accord entre ces différents éléments en définitude pour que le syntagme soit grammatical.(Wintner, 2000, p :323) L'arabe ne possède pas d'article indéfini, telle que l'ont d'autres langues sémitiques ; voir Wintner (2000) pour l'hébreu. L'indéfinitude est portée par le suffixe *-n* (dit *tanwīn*) attaché au nominal sans l'article défini ; les deux affixes *al-* et *-n* sont toujours en exclusion mutuelle même s'ils n'occupent pas la même position dans le schème du nominal (cf. l'agrammaticalité en (1.c)).

1.1. Le marquage de la définitude

Dans cette sous-section nous répondrons à deux questions essentielles. Comment se fait le marquage de la définitude au niveau du syntagme nominal de l'arabe ? Et, puisque, comme nous le verrons ci-dessous (dans la présente sous-section), les syntagmes nominaux peuvent être définis sans l'utilisation explicite de l'article défini, comment se matérialise alors cette définitude au sein du syntagme nominal arabe ? Le syntagme nominal (in)défini élémentaire en arabe est présenté en (1). Dans les SNs indéfinis de l'arabe (1.a), et contrairement à l'anglais et au français, la définitude n'a pas besoin d'être spécifiée par un déterminant ; ce qui veut dire que des noms nus peuvent constituer à eux seuls des SNs indéfinis, et peuvent, donc, servir comme arguments syntaxiques et recevoir ainsi une interprétation indéfinie. En (1.b), la définitude est marquée par l'affixe *al-*, préfixé au nom *kitāb* 'book'. Le syntagme en entier *al-kitāb-u* est alors défini.

(1) a. kitāb-u-n
book-Nom-Indef
'a book' (كتابٌ)
'un livre'

b. al-kitāb-u
DEF-book-NOM
'the book' (الكتابُ)
'le livre'

c. * al-kitāb-u-n
DEF-book-NOM-INDEF

Dans l'exemple (1.b) ci-dessus la définitude a été marquée par la présence de l'article défini, ce qui n'est pas toujours le cas : les syntagmes nominaux peuvent être définis sans l'utilisation explicite de l'article défini. C'est le cas des noms propres, des possessifs et des états construits [2]. En (2.a), malgré que le nom propre *karīm* ne porte pas le préfixe *al-*, sa définitude est confirmée par celle de l'adjectif *al-kabīru* qui le modifie, qui, lui, est défini par l'article *al-* ; dans la sous-section suivante, nous montrerons que l'adjectif suit toujours le nom qu'il modifie en définitude. En (2.b), la phrase a été marquée comme agrammaticale du fait que la construction *karīmun kabīrun* n'est pas acceptable comme syntagme nominal. Il s'agit ici d'une phrase thématique et non d'un syntagme nominal ; le nom *karīmun* est le thème de la phrase (dit *mubtada'* dans la tradition grammaticale arabe), alors que l'adjectif *kabīrun* est le prédicat ou l'attribut (dit *ẖabar*). Le deuxième type de SN défini sans l'utilisation de *al-* consiste en un nom auquel est suffixé un pronom possessif tel qu'en (3). Un suffixe du possessif, en l'occurrence le morphème */-ī/*, est attaché au nom radical, et ajoute l'information [1.PER.SG] ; il introduit, alors, un possesseur, et *kitāb* 'book' devient *kitāb-ī* 'my book'.

(2) a. karīm-u al-kabīr-u
karim-NOM DEF-big-NOM
'the big Karim', 'Karim the big' (كريمُ الكبيرُ)
'le grand Karim', 'Karim le grand'

b. * karīm-u-n kabīr-u-n
Karim-NOM-INDEF big-NOM-INDEF
'Karim is big' (كريمٌ كبيرٌ)
'Karim est grand'

(3) a. kitāb-ī
book-POSS.1.SING
'my book' (كتابي)
'mon livre'

[2]Un état construit (*état d'annexion*, *ḥālat al-'iḍāfah* ou *al-muḍāf wa al-muḍāf 'ilayh*) consiste en deux expressions nominales l'une à la suite de l'autre, formant ensemble un seul constituant au niveau syntagmatique.

Le paradigme complet des suffixes du possessif de l'arabe est synthétisé dans la table 1.

Singulier				
	1.	masc+fem	ـِي	-ī
	2.	masc	ـكَ	-ka
		fem	ـكِ	-ki
	3.	masc	ـهُ	-hu
		fem	ـهَا	-hā
Pluriel				
	1.	masc+fem	ـنَا	-nā
	2.	masc	ـكُمْ	-kum
		fem	ـكُنَّ	-kunna
	3.	masc	ـهُمْ	-hum
		fem	ـهُنَّ	-hunna
Duel				
	1.	masc+fem	ـنَا	-nā
	2.	masc+fem	ـكُمَا	-kumā
	3.	masc+fem	ـهُمَا	-humā

Tab. 1 : Le paradigme des suffixes du possessif en arabe

Le troisième type de SN défini sans l'utilisation de *al-* est l'état construit. Dans ce type de construction, ni *al-* ni *-n* ne peuvent être affixés au(x) nom(s) en état construit (cf. 4.c-d). Pourtant, les constructions construites sont marquées pour la définitude. Comment se matérialise alors cette définitude ? Dans l'état construit en (4.a), la tête construite *ṭullāb-u*, aussi connue en arabe sous la dénomination le *muḍāf* (i.e. ce qui est annexé, ou ajouté), ne peut être défini par préfixation de *al-*, et c'est le syntagme nominal complément *al-ǧāmi'at-i*, dit *muḍāf 'ilay-h* (i.e. auquel est annexé le *muḍāf*) qui porte la marque de définitude (*al-*, dans ce cas). Le syntagme en entier est alors défini si et seulement si le complément l'est déjà. En (4.b), il s'agit d'un état construit indéfini, et c'est le syntagme nominal complément qui porte la marque de l'indéfinitude.

(4) a. ṭullāb-u al-ǧāmi'at-i
students-Nom Def-university-Gen
'the students of the university', 'the university's students' (طلاّبُ الجامعةِ)
'les étudiants de l'université', 'les étudiants universitaires'

b. ṭullāb-u ǧāmi'at-i-n
students-Nom university-Gen-Indef
'students of university', 'university's students' (طلاّبُ جامعةٍ)

'des étudiants d'université', 'des étudiants universitaires'

c. * al-ṭullāb-u al-ǧāmi'at-i
DEF-students-NOM DEF-university-GEN

d. * ṭullāb-u-n ǧāmi'at-i-n
students-NOM-INDEF university-GEN-INDEF

Pour conclure, nous dirons que dans les états construits, la définitude du syntagme supérieur est récupérée par propagation (ou percolation) à partir des nœuds fils. Cette propagation peut être implémentée de deux manières différentes. Dans la première, le syntagme nominal père récupère le trait de définitude directement à partir du syntagme nominal complément. Dans la seconde, la propagation se fait de manière indirecte ; i.e., du complément vers la tête du syntagme nominal, ensuite, de la tête vers le syntagme supérieur. Dans cette dernière approche, la construction complète *ṭullāb-u al-ǧāmi'at-i* 'the students of the university', en (4.a), hérite tous les traits morpho-syntaxiques, y compris la définitude, à partir de la tête du syntagme. Dans l'analyse que nous développerons à la section 3, nous utiliserons des données concernant les noms et les syntagmes nominaux définis explicitement par l'article défini *al-*, ainsi que des données concernant les états construits.

1.2. La définitude dans le syntagme nominal étendu

Un syntagme nominal arabe type est constitué d'un nom représentant la tête du syntagme précédé et/ou suivi par un certain nombre de périphériques (ou modifieurs) : démonstratifs, quantificateurs, quantifieurs (cardinaux et ordinaux), adjectifs, adverbes intensificateurs, pronoms possessifs, compléments génitifs (possessifs) et relatives. Fassi (1999) (cité dans Kremers (2003, p :70)) résume l'ordre de ces modifieurs dans le syntagme nominal arabe comme suit :

1. Q-Dem-Ord-Card-Adj-(Det)-**N**-(Gen)
2. (Det)-**N**-(Gen)-Adj-Card-Ord-Dem-Q-Rel

Les deux schémas 1 et 2 ci-dessous donnent respectivement l'ordre des modifieurs prénominaux et l'ordre des modifieurs postnominaux. Le complexe *(Det)-N-(Gen)* est une combinaison fixe qui ne peut pas être brisée. *Det* et *Gen* sont mis entre parenthèse pour indiquer qu'ils sont en distributions complémentaires. Les modifieurs *Q*, *Dem*, *Ord*, *Card* et *Adj* peuvent apparaître des deux côtés : dans cet ordre, pré-nominalement, et dans l'ordre inverse (ou miroir), post-nominalement. En outre, les relatives peuvent seulement apparaître post-nominalement et suivent toujours tous les autres modifieurs. Comment interagissent alors ces différents composants au sein du syntagme nominal ? Et, qu'en est-il de la définitude au niveau de chaque composant ? Les périphériques que nous venons d'énumérer sont de deux genres ; ceux qui sont affixés à la tête du syntagme, et ceux qui ne le sont pas. Les premiers font partie intégrante de la tête nominale et on ne peut pas parler de définitude dans le cas de ces périphériques. Pour les seconds, certains d'entre eux peuvent être définis, d'autres doivent

être définis et d'autres encore ne sont jamais définis. Dans ce qui suit, nous n'allons pas examiner tous ces périphériques, nous nous contenterons des possessifs, des adjectifs, des intensificateurs et des démonstratifs ; les autres méritent plus d'espace. Les pronoms possessifs sont affixés aux noms indéfinis. Ils ne s'attachent jamais aux noms définis de quelque manière que ce soit. En fait, les pronoms possessifs servent déjà à définir et une double détermination n'est pas permise. En arabe, les adjectifs et les syntagmes adjectivaux suivent immédiatement les noms qu'ils modifient. Considérons les exemples suivants :

(5) a. kitāb-u-n kabīr-u-n
book-Nom-Indef big-Nom-Indef
'a big book' (كتابٌ كبيرٌ)
'un grand livre'

b. al-kitāb-u al-kabīr-u
Def-book-Nom Def-big-Nom
'the big book' (الكتابُ الكبيرُ)
'le grand livre'

c. * kitāb-u-n al-kabīr-u
book-Nom-Indef Def-big-Nom

d. * al-kitāb-u kabīr-u-n[3]
Def-book-Nom big-Nom-Indef
'the book is big' (الكتابُ كبيرٌ)
'le livre est grand'

e. kitāb-u-n kabīr-u-n ğiddan
book-Nom-Indef big-Nom-Indef very
'a very big book' (كتابٌ كبيرٌ جدًّا)
'un très grand livre'

f. al-kitāb-u al-kabīr-u ğiddan
Def-book-Nom Def-big-Nom very
'the very big book' (الكتابُ الكبيرُ جدًّا)
'le très grand livre'

g. * al-kitāb-u kabīr-u-n ğiddan
Def-book-Nom big-Nom-Indef very
'the book is very big' (الكتابُ كبيرٌ جدًّا)

[3]Dans ce papier, certaines constructions, telles que *al-kitāb-u kabīr-u-n* 'the book is big', sont des phrases grammaticalement correctes. Nous les avons marquées comme agrammaticales pour indiquer qu'elles ne sont pas acceptables comme syntagmes nominaux ; il s'agit, en fait, de phrases thématiques mais non de syntagmes nominaux ; dans *al-kitāb-u kabīr-u-n*, *al-kitāb-u* est le thème et *kabīr-u-n* le prédicat.

'le livre est très grand'

h. * kitāb-u-n al-kabīr-u ǧiddan
book-Nom-Indef Def-big-Nom very

Les exemples en (5.a-b) sont des exemples type du syntagme nominal arabe constitué d'un nom modifié par un adjectif. Dans ces deux exemples, nous remarquons que l'adjectif s'accorde avec le nom qu'il modifie en définitude ; ils s'accordent aussi en genre et en nombre, mais cet accord n'est pas important pour notre description actuelle. De plus, l'accord en définitude est obligatoire, et cela est confirmé par les agrammaticalités en (5.c-d). Dans les exemples en (5.e-h), l'adjectif est suivi par l'intensificateur *ǧiddan*. L'intensificateur est un spécifieur de l'adjectif modifiant le nom. L'adjectif doit toujours s'accorder en définitude avec le nom qu'il modifie, alors que son intensificateur ne peut recevoir l'article défini. En (6), on peut voir que le démonstratif *hāḏa*, qui précède le nom, n'est jamais un hôte pour *al-*. En plus, il exige du nom qui le suit d'être défini (6.a-b) et que cette définitude soit marquée par l'article défini *al-* et non par un quelconque autre moyen, ce qui est visible par les agrammaticalités en (6.c-d). En (6.c), le syntagme nominal *kitāb-u al-'ustāḏ-i* est défini par annexion mais il ne constitue pas avec *hāḏa* un syntagme nominal unique ; le pronom démonstratif *hāḏa* et le syntagme nominal *kitāb-u al-'ustāḏ-i* sont respectivement thème (ou *mubtada'*) et prédicat (ou *ẖabar*). De même en (6.d), le nom *kitāb-u-hu* est défini par suffixation du pronom possessif *-hu* et constitue le prédicat de la phrase.

(6) a. hāḏa al-kitāb-u
this Def-book-Nom
'this book' (هذا الكتابُ)
'ce livre'

b. * hāḏa kitāb-u-n
this book-Nom-Indef
'this is a book' (هذا كتابٌ)
'ceci est un livre'

c. * hāḏa kitāb-u al-'ustāḏ-i
this book-Nom Def-teacher-Gen
'this is the teacher's book', 'this is the book of the teacher' (هذا كتابُ الأستاذِ)
'ceci est le livre du professeur'

d. * hāḏa kitāb-u-hu
this book-Nom-his
'this is his book' (هذا كتابُهُ)
'ceci est son livre'

Il est à noter enfin que, *al-* ne peut se combiner avec les quantificateurs lorsqu'ils apparaissent pré-nominalement (7.a-b) ; si le quantificateur porte le *al-* (7.c) ou s'il est

suffixé d'un pronom possessif (7.d), il doit à lui seul former un syntagme nominal complet.

(7) a. ğamī'-u al-awlād-i ħaraǧū
all-Nom Def-boy.Pl-Gen leave.Perf-3Masc.Pl
'all the boys left' (جميعُ الأولادِ خرجُوا)
'tous les garçons sont sortis'

b. * al-ğamī'-u al-awlād-i ħaraǧū
Def-all-Nom Def-boy.Pl-Gen leave.Perf-3Masc.Pl

c. al-ğamī'-u ħaraǧū
Def-all-Nom leave.Perf-3masc.Pl
'the all left' (الجميعُ خرجُوا)
'tous, ils sont sortis'

d. ğamī'-u-hum ħaraǧū
all-Nom-they leave.Perf-3Masc.Pl
'all of them left' (جميعُهم خرجُوا)
'eux tous sont sortis'

2. La nature affixale de l'article défini en arabe

2.1. Affixes vs. clitiques

Dans la plupart des langues, il y a certains morphèmes qui sont problématiques parce qu'ils ont un statut non évident qui n'est ni celui d'un mot indépendant ni celui d'un affixe. De tels morphèmes semblent avoir un statut intermédiaire entre ces deux catégories bien établies. D'habitude, ils n'ont pas l'autonomie d'un mot normal et doivent s'appuyer sur un mot adjacent, l'hôte. Le statut spécial de tels items a été reconnu par des linguistes structuralistes et comparatistes qui les ont appelés *clitiques*. Les clitiques ont été définis essentiellement en termes de leur déficience d'un statut prosodique indépendant ; les clitiques sont phonologiquement dépendants d'un hôte sur lequel ils s'appuient, une propriété généralement rapportée à des caractéristiques prosodiques les empêchant de compter comme mots phonologiques. Mais cette déficience accentuelle est caractéristique aussi des affixes. En effet, les deux types de morphèmes apparaissent toujours attachés à d'autres constituants, d'où la nécessité d'établir des critères distinctifs explicites (Djebali, 2009, p :161). Ce qui a conduit Zwicky à proposer des critères qui servent à distinguer ces deux types d'unités. L'approche suivi par Zwicky (1977) est une approche lexicale qui rend compte des clitiques d'un point de vue morpho-phonologique. D'autres approches au sujet ont été proposées ; la première, purement syntaxique, revient à Kayne (1975) (cité dans Rossi (2007, p :22)), et la deuxième, due à Nespor (1986)(cité dans Rossi (2007, p :22)), est purement phonologique.

2.2. Les critères de Zwicky

L'objectif principal de cette section est d'affirmer que l'article défini de l'arabe doit être vu comme un affixe et non un clitique. Pour cela, nous utiliserons des critères qui ont été établis par Zwicky (1977, 1983) et Miller (1992), et acceptés comme tests pour distinguer les affixes des clitiques (ou les affixes des non-affixes, ou encore les mots des non-mots). Selon Zwicky (1985a), ces tests sont utilisés pour diagnostiquer une situation linguistique particulière plutôt qu'une condition nécessaire et suffisante. À chaque fois qu'un certain élément se conforme avec certains de ces tests, il est plus probable qu'il soit dans une certaine situation particulière. Nous introduirons ces critères au fur et à mesure de leurs utilisations dans la sous-section suivante.

2.3. Les arguments pour une analyse affixale

Les morphèmes grammaticaux (ou séquence de morphèmes grammaticaux) reconnaissables comme formes liées ne manifestent pas toujours au même degré des caractéristiques d'intégration à un mot qui les engloberait. Certaines formes liées représentent des mots grammaticaux dont la position dans la phrase est déterminée par les règles de la syntaxe, mais qui sont susceptibles de se *cliticiser*, c'est-à-dire de perdre leur autonomie pour former une unité prosodique avec un mot plein auquel ils sont adjacents. C'est par exemple le cas en anglais pour la contraction *don't* forme réduite de *do not*. Un clitique est en principe une forme liée qui se caractérise par un faible degré d'interaction avec son hôte, tandis qu'un affixe se caractérise par une interaction forte avec la base à laquelle il s'attache. Pour l'article défini en arabe, son statut affixal est évident de plusieurs points de vue. D'un point de vue orthographique, *al-* apparaît toujours comme préfixe, nullement ailleurs et sous aucune autre forme. Ainsi, l'écrit reflète une intuition quant à cette nature affixale. D'un point de vue syntaxique, *al-* n'est pas actif syntaxiquement ; il ne peut être ni argument ni modifieur. De plus, cette intuition peut être vérifiée par les Critères de Zwicky. Les constructions avec les affixes, la portée de l'article défini sur les structures coordonnées, le degré de sélection du gouverneur, les idiosyncrasies morphophonologiques, et les règles phonologiques sont tous des arguments qui suggèrent un traitement morphologique de l'article défini en termes d'affixes. Avec ces critères en place, l'hypothèse proposée plus haut, concernant le statut affixal de l'article défini en arabe, sera explicitement examinée et rigoureusement testée dans les sous-sections subséquentes.

2.3.1. Le liage

Les affixes sont des morphèmes liés, et ceci est le cas avec *al-*. *Al-* ne peut jamais, et en aucune circonstance, apparaître comme forme isolée. Phonologiquement, *al-* ne porte pas d'accent.

2.3.2. Les constructions avec les affixes

L'article défini en arabe se combine avec des mots entiers, c'est-à-dire avec des mots qui sont déjà formés d'un point de vue morphologique. Ceci a pour effet que l'article défini peut entrer en combinaison avec des mots qui sont déjà fléchis et comportent tous leurs affixes et infixes. Zwicky (1977) considère qu'un affixe ne peut entrer en construction qu'avec une base ou un autre affixe. Ainsi, nous pouvons considérer que l'article défini en (1.b) supra est un affixe. Dans cet exemple, l'article défini est préfixé au mot *kitāb-u*, qui est déjà composé d'une base *kitāb* et d'un suffixe *-u*, pour former le mot *al-kitāb-u*.

2.3.3. Le déplacement

Les parties propres aux mots ne sont pas sujettes aux règles de déplacement : elles ne peuvent servir de trous dans des relations gap-filler. Les nominaux définis se plient totalement à cette règle : l'article défini ne peut jamais se déplacer seul, et à chaque fois qu'un nominal se déplace, il le fait avec son article attaché.

2.3.4. La portée de l'article défini

Miller (1992) redéfinit l'un des critères essentiels des affixes suggéré par Zwicky (1977), à savoir que ceux-ci ne peuvent avoir une large portée sur des éléments coordonnés, alors que les clitiques peuvent avoir ce genre de portée comme le montre (8) (Djebali, 2009, p :201), où les clitiques *'ll* et *'s* ont une portée sur les structures coordonnées placées entre parenthèses :

(8) a. (Pat and Leslie)'ll be there.
b. (Pat and Leslie)'s book.

Dans les exemples en (9), nous pouvons voir que l'article défini ne peut avoir une large portée sur des éléments coordonnés ; *al-* ne peut être factorisé à la manière des clitiques *'ll* et *'s*.

(9) a. al-walad-u wa al-bint-u ya-l'ab-ā-ni al-kurat-a
Def-boy.Sg-Nom and Def-girl-Nom Imperf-play.with-Nom-Dl Def-ball-Acc
'the boy and the girl play with the ball' (الولدُ والبنتُ يلعبانِ الكرةَ)
'le garçon et la fille jouent au ballon'
b. * al-walad-u wa bint-u-n ya-l'ab-ā-ni al-kurat-a[4]
Def-boy.Sg-Nom and girl-Nom-Indef Imperf-play.with-Nom-Dl Def-ball-Acc

[4]Certaines phrases, telles que (9.b-c), sont des phrases grammaticales. Nous les avons marquées comme agrammaticales pour indiquer que *al-*, en étant attaché au premier syntagme, ne peut définir (i.e. ne peut avoir une portée sur) le second via coordination (i.e. sans lui être affixé).

'the boy and a girl play with the ball' (الولدُ وبنتٌ يلعبانِ الكرةَ)
'le garçon et une fille jouent au ballon'

c. * walad-u-n wa al-bint-u ya-l'ab-ā-ni al-kurat-a
boy.Sg-Nom-Indef and Def-girl-Nom Imperf-play.with-Nom-Dl Def-ball-Acc
'a boy and the girl play with the ball' (ولدٌ والبنتُ يلعبانِ الكرةَ)
'un garçon et la fille jouent au ballon'

d. walad-u-n wa bint-u-n ya-l'ab-ā-ni al-kurat-a
boy.Sg-Nom-Indef and girl-Nom-Indef Imperf-play.with-Nom-Dl Def-ball-Acc
'a boy and a girl play with the ball' (ولدٌ وبنتٌ يلعبانِ الكرةَ)
'un garçon et une fille jouent au ballon'

Miller (1992, p :155) suggère alors ce critère sous la forme suivante :

Si un item doit être répété sur chaque conjoint dans une structure coordonnée, alors il doit être un affixe et ne peut être un clitique postlexical ; s'il échoue d'être répété, il doit être un clitique et non un affixe. Si la répétition est optionnelle, aucune confirmation n'est alors établie.

Ce test est facile à appliquer pour le cas de *al-*. Premièrement, notons que les éléments auxquels *al-* peut être attaché sont coordonnables, ceci est visible en (10).

(10) a. 'ištarayt-u kitāb-a-n wa ḥāsūb-a-n
buy.Perf-I book-Acc-Indef and notebook-Acc-Indef
'I bought a book and a notebook' (اشتريتُ كتابًا وحاسوبًا)
'j'ai acheté un livre et un ordinateur'

b. 'ayn-ā-ni kabīr-at-ā-ni wa ħaḍraw-at-ā-ni
eye-Nom-Dl big-Fem-Nom-Dl and green-Fem-Nom-Dl
'big green eyes' (عينانِ كبيرتانِ وخضروتانِ)
'de gros yeux verts'

Ensuite, quand ces éléments sont définis par *al-*, *al-* ne peut pas avoir une portée large sur la coordination, mais il doit, plutôt, être répété sur chaque conjoint (11).

(11) a. 'ištarayt-u al-kitāb-a wa al-ḥāsūb-a
buy.Perf-I Def-book-Acc and Def-notebook-Acc
'I bought the book and the notebook' (اشتريتُ الكتابَ والحاسوبَ)
'j'ai acheté le livre et l'ordinateur'

b. al-'ayn-ā-ni al-kabīr-at-ā-ni wa al-ħaḍraw-at-ā-ni
Def-eye-Nom-Dl Def-big-Fem-Nom-Dl and Def-green-Fem-Nom-Dl
'the big green eyes' (العينانِ الكبيرتانِ والخضروتانِ)
'les gros yeux verts'

L'omission de l'une des occurrences de *al-* sur l'un des constituants aboutit, dans le cas des syntagmes adjectivaux (12.a), à une agrammaticalité, et dans le cas des syntagmes nominaux (12.b), à une autre signification dans laquelle l'article a une portée restreinte.

(12) a. * al-ʿayn-ā-ni al-kabīr-at-ā-ni wa ħaḍraw-at-ā-ni
DEF-eye-NOM-DL DEF-big-FEM-NOM-DL and green-FEM-NOM-DL

b. ʾištarayt-u al-kitāb-a wa ḥāsūb-a-n
buy.PERF-I DEF-book-ACC and notebook-ACC-INDEF
'I bought the book and a notebook' (اشتريتُ الكتابَ وحاسوبًا)
'j'ai acheté le livre et un ordinateur'

2.3.5. La sélection du gouverneur

Zwicky (1983) admet que l'une des propriétés qui distinguent les affixes des clitiques, réside dans le fait que les affixes sont très sélectifs en ce qui concerne leurs bases, ce qui n'est pas généralement vrai pour les clitiques. Ce critère n'est en fait que l'expression d'une tendance générale dans les affixes, surtout dérivationnels, à sélectionner la base à laquelle ils s'attachent. Cette tendance est vraie pour l'article défini en arabe. En fait, *al-* peut se combiner avec la plupart des nominaux, en l'occurrence les noms (communs et propres), les adjectifs, les numéraux (ordinaux et cardinaux) et les quantificateurs, à l'exception des démonstratifs. Bien que cela paraisse comme un faible degré de sélectivité, il faut noter que : d'une part, tous ces éléments forment une même classe naturelle, et ils peuvent être utilisés, comme le montre l'exemple (13), pour se substituer à une même entité (dans l'exemple, au syntagme nominal *al-ṭālib-u* en (13.a)) ; et d'autre part, *al-* ne s'attache jamais aux prépositions et que rarement aux adverbes.

(13) a. naǧaḥa al-ṭālib-u fī al-ʿimtiḥān-i
pass.PERF.MASC DEF-student.SG-NOM in DEF-exam-GEN
'the student passed the exam' (نجحَ الطالبُ في الامتحانِ)
'l'étudiant a réussi à l'examen'

b. naǧaḥa al-moqdād-u fī al-ʿimtiḥān-i
pass.PERF.MASC DEF-moqdad-NOM in DEF-exam-GEN
'Almokdad passed the exam' (نجحَ المقدادُ في الامتحانِ)
'Almokdad a réussi à l'examen'

c. naǧaḥa al-mušawwiš-u fī al-ʿimtiḥān-i
pass.PERF.MASC DEF-troublemaker-NOM in DEF-exam-GEN
'the troublemaker passed the exam' (نجحَ المشوِّشُ في الامتحانِ)
'le perturbateur a réussi à l'examen'

d. nağaḥa al-ʼarbaʻat-u fī al-ʼimtiḥān-i
pass.Perf.Masc Def-four-Nom in Def-exam-Gen
ʼthe four passed the exam' (نجحَ الأربعةُ في الامتحانِ)
ʼles quatre ont réussi à l'examen'

e. nağaḥa al-rābiʻ-u fī al-ʼimtiḥān-i
pass.Perf.Masc Def-fourth-Nom in Def-exam-Gen
ʼthe fourth passed the exam' (نجحَ الرّابعُ في الامتحانِ)
ʼle quatrième a réussi à l'examen'

f. nağaḥa al-ğamīʻ-u fī al-ʼimtiḥān-i
pass.Perf.Masc Def-all-Nom in Def-exam-Gen
ʼthe all passed the exam' (نجحَ الجميعُ في الامتحانِ)
ʼtous ont réussi à l'examen'

Il est à noter enfin que, *al-* ne peut se combiner avec les quantificateurs lorsqu'ils apparaissent pré-nominalement et fonctionnent comme déterminants aux nominaux qui les suivent, comme dans les exemples (14.a-b). En effet, si le quantificateur porte le *al-*, comme en (14.c), il doit former à lui seul un syntagme nominal.

(14) a. ğamīʻ-u al-awlād-i ħarağū
all-Nom Def-boy.Pl-Gen leave.Perf-3Masc.Pl
ʼall the boys left' (جميعُ الأولادِ خرجُوا)
ʼtous les garçons sont sortis'

b. * al-ğamīʻ-u al-awlād-i ħarağū
Def-all-Nom Def-boy.Pl-Gen leave.Perf-3Masc.Pl

c. al-ğamīʻ-u ħarağū
Def-all-Nom leave.Perf-3masc.Pl
ʼthe all left' (الجميعُ خرجُوا)
ʼtous, ils sont sortis'

2.3.6. Les idiosyncrasies morpho-phonologiques

L'une des propriétés essentielles des affixes selon Zwicky (1983) est qu'ils connaissent plus d'idiosyncrasies morphophonologiques que les clitiques. Miller (1992) propose même que ce soit l'une des caractéristiques indéniables des affixes que de connaître de telles idiosyncrasies qu'on ne peut expliquer sur les seules bases des règles phonologiques productives. Les règles de la syntaxe, étant régulières, ne peuvent non plus rendre compte de ces idiosyncrasies, qu'il faut lister dans le lexique et traiter avec les règles de la morphologie (Djebali, 2005). L'article défini en arabe connaît plusieurs idiosyncrasies morphophonologiques quand il est combiné avec son hôte. En effet,

comme beaucoup d'affixes flexionnels, *al-* est sujet à des alternances particu-lières dans la forme, déclenchées par la phonologie des mots adjacents au sein d'une expression. En arabe, l'article défini expose une alternance idiosyncrasique *al-/a-/l-/Ø*, déclenchée par la phonologie du mot auquel il se rattache (i.e. son hôte) et celle du mot qui le précède. Lorsque l'article est l'élément initial d'une phrase ou d'un syntagme (i.e. lorsqu'il est prononcé après une pause), il est prononcé (i) */al-/*, s'il est suivi par l'une des consonnes non assimilées[5] (15.a), et (ii) */a-/*, s'il est suivi par l'une des consonnes assimilées et dans ce cas le *-l-* est assimilé à la consonne qui est alors doublée en prononciation (15.b).

(15) a. al-manzil-u wāsi'-u-n
DEF-house-NOM large-NOM-INDEF
'the house is large' (اَلْمنزلُ واسعٌ)
'la maison est grande'

b. ad-dār-u wāsi'-at-u-n
DEF-house-NOM large-FEM-NOM-INDEF
'the house is large' (اَلدَّارُ واسعةٌ)
'la maison est grande'

c. al-manzil-u -l-wāsi'-u
DEF-house-NOM DEF-large-NOM
'the large house' (اَلْمنزلُ الْواسعُ)
'la grande maison'

d. al-manzil-u -ṣ-ṣaġīr-u
DEF-house-NOM DEF-small-NOM
'the small house' (اَلْمنزلُ الصَّغيرُ)
'la petite maison'

Dans toutes les autres positions, l'article doit être nécessairement précédé par une voyelle, et il est prononcé */l/* lorsqu'il est suivi par l'une des consonnes non assimilées (15.c) ; s'il est suivi par l'une des consonnes assimilées, le *-l-* est assimilé à la consonne qui est alors doublée en prononciation (15.d). Cette assimilation est indiquée dans les textes vocalisés en laissant le *-l-* sans aucune marque et en plaçant une *šadda*, qui représente le signe graphique de gémination en arabe, sur la consonne initiale du mot (i.e. celle suivant l'article défini).

[5]Les lettres dites assimilées (ou encore, lettres solaires ; *al-ḥurūfu aṣ-ṣamsiyyah*), au nombre de quatorze, sont : (i) toutes les dentales (t, d, ṯ, ḏ, ṭ), (ii) toutes les sifflantes (s, š, z, ṣ, ḍ, ḏ) et (iii) toutes les liquides (r, l, n). Toutes les autres lettres, (', b, ǧ, ḥ, ẖ, ', ġ, f, q, k, m, h, w, y), aussi au nombre de quatorze, sont dites non assimilées (ou encore, lettres lunaires ; *al-ḥurūfu l-qamariyyah)*.

2.4. Conclusion

En conclusion de cette section, nous pouvons affirmer que la majorité des tests présentés sur la nature de l'article défini en arabe pointent dans la même direction et nous indiquent que sa distribution et ses propriétés sont déterminées par des règles morphologiques plutôt que par des règles syntaxiques. L'article défini n'est pas attaché à son hôte dans la syntaxe, mais se comporte comme un morphème et est engendré dans la morphologie avant son insertion dans la syntaxe. Son comportement ne peut pas être expliqué d'une manière simple et satisfaisante dans une analyse purement syntaxique. Contrairement à la morphologie, la syntaxe ne dispose pas de règles particulières pour chacune des langues et ne possède pas de principes pour des constructions spécifiques. Elle consiste en un système de règles qui s'appliquent d'une façon très générale.

3. Analyse

Nous avons soutenu dans la section précédente que *al-* est un affixe et que, comme tous les affixes, il s'attache aux éléments d'une classe bien déterminée ; dans le cas de l'arabe, aux éléments qui ont des traits nominaux. Dans ce qui suit nous adoptons une approche purement lexicaliste. Dans une telle conception, les règles syntaxiques combinent des mots complètement bien formés issus de la composante morphologique, y compris des mots formés d'une base et d'un ou de plusieurs affixes pronominaux. Ainsi, nous considèrerons l'article défini *al-* comme un affixe flexionnel non sujet aux principes gouvernant le composant syntaxique de la grammaire. Dans tous les cas, il se combine avec son hôte comme résultat d'un processus lexical, non syntaxique ; la composante morphologique de la grammaire serait donc responsable de sa génération et de son insertion dans la structure. Nous suggérons dans cette section une analyse HPSG de la définitude au sein du syntagme nominal en arabe, expliquant les données présentées dans la section 1. Cette analyse est basée sur une deuxième hypothèse qui stipule que le nom constitue la tête syntaxique et sémantique du syntagme nominal arabe. En effet, en arabe, tous les traits morphosyntaxiques qui doivent être transférés à la projection maximale du syntagme nominal (i.e. le syntagme nominal final), pour des raisons d'accord ou de gouvernement, sont manifestés sur le nom. En plus, l'article défini arabe, étant un affixe flexionnel, peut être aisément préfixé au radical nominal par le biais d'une règle lexicale. Cette solution a été proposée, pour la première fois, par Wintner (2000) pour l'hébreu. Dans la sous-section 3.2, nous exposerons l'approche de Wintner (2000) qui rend compte du phénomène de la définitude en hébreu. Dans cet exposé, nous décrirons la règle lexicale utilisée pour établir la relation entre les formes définie et indéfinie, et concrétisant ainsi l'affirmation que l'article défini est un affixe. Par la suite, dans la sous-section 3.3, nous montrerons les limites de la solution proposée et son inadéquation pour l'arabe. Enfin, dans la sous-section 3.4, nous présenterons la solution que nous avons proposée et retenue pour

rendre compte de la définitude et de l'accord en définitude dans le syntagme nominal arabe. Dans la sous-section subséquente, nous donnerons un survol de l'analyse HPSG pour les spécifieurs et nous dirons quelque chose quant à son adéquation à l'analyse du syntagme nominal de l'arabe.

3.1. Les spécifieurs et leur analyse en HPSG

L'analyse HPSG standard pour l'anglais, présentée dans Pollard (1994, section 9.4), considère les articles, qui sont des mots indépendants précédant les noms, comme des compléments sous-catégorisés des noms. L'article se combine avec le nom, qui est la tête de la construction, par le biais du Schéma Spécifieur-Tête. Et comme HPSG exige que les syntagmes soient saturés (i.e. ayant leurs traits de valence (Subj, Comps, Spr) tous vides), un nom nu (i.e. sans article) ne sera pas accepté comme syntagme nominal valide. Bien que cela soit légitime pour des langues telle que l'anglais, il n'en est pas le cas pour l'arabe, pour la simple raison que les noms nus en arabe fonctionnent parfaitement comme des syntagmes nominaux complets ; il suffit pour s'en convaincre de voir les exemples de la section 1 supra.

3.2. La définitude comme processus lexical : l'approche de Wintner (2000)

Avec l'extension de HPSG aux langues où l'expression de la définitude fait partie de la composante morphologique de la langue, Wintner (2000) proposa, pour l'hébreu, de rendre compte du phénomène de la définitude en renforçant l'inventaire des règles lexicales de la langue par un ensemble de règles lexicales traitant tous les phénomènes liés à la définitude. Wintner (2000) suggéra alors de traiter l'article défini en hébreu comme un affixe qui se combine, par le biais d'une règle du type *word-to-word* qu'il nomme la Règle Lexicale du Défini (RLD) (cf. Figure 1), avec les nominaux dans le lexique, et ce n'est qu'après que cette combinaison eu lieu que les nominaux définis entrent dans la syntaxe. Wintner (2000) fait recours à un trait (booléen) additionnel, DEFiniteness (dorénavant Def), qu'il utilise pour encoder la valeur de la définitude dans les nominaux. Comme l'accord en définitude en hébreu n'est pas un processus sémantique, il ajoute ce trait dans le trait CATegory des nominaux (plutôt que dans leurs traits CONTent). En plus, étant donné que la définitude est un trait des syntagmes qui est hérité à partir de la tête lexicale, il considère Def comme un trait de tête approprié pour tous les nominaux. La RLD opère sur tous les mots nominaux, à condition que la valeur de leur trait Def soit '-'. Pour toutes les catégories, son effet sur la phonologie est déterminé par les mêmes règles phonologiques ; il utilise alors une fonction *definite* pour faire abstraction sur l'ensemble de ces règles. Cette fonction change la valeur du chemin SynSem | Loc | Cat | Head | Def de '-' à '+'.

Pour la propagation de la définitude vers la projection maximale, Wintner (2000) propose que cette propagation se fasse par le biais du nom tête du syntagme. Dans le cas de l'état construit, cela se fait en deux pas : le nom tête en état construit hérite,

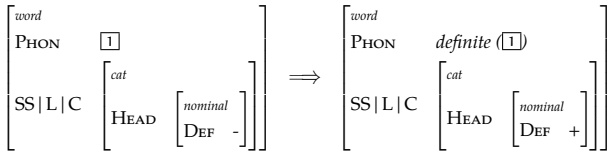

FIG. 1 : La RÈGLE LEXICALE DU DÉFINI pour l'hébreu (Wintner, 2000, p :349)

dans un premier temps, le trait DEF du complément (soit DEF -, soit DEF +) ; par la suite, l'état construit en entier hérite, à son tour, ce trait, avec le reste des autres traits morphosyntaxiques, à partir du nom tête.

3.3. L'approche de Wintner (2000) et le traitement de la définitude en arabe

La règle de Wintner (2000) pourrait-elle être reconduite dans une analyse de la définitude au sein du syntagme nominal arabe ? Ou, plutôt, serait-elle suffisante pour rendre compte de la définitude en arabe ? Il faut noter que la règle de Wintner (2000) ne s'applique qu'aux noms indéfinis (DEF -), auxquels on préfixe l'article défini pour les rendre définis (DEF +) ; bien sûr, avec un certain nombre de changements phonologiques. Dans la tradition grammaticale arabe, les noms sont par définition indéfinis dans le lexique, et pour les rendre définis il suffit de leur préfixer le *al-* et d'éliminer le *tanwīn* de leurs terminaisons. Cette idée renforce encore l'approche de Wintner (2000). Mais cela n'est pas pratique dans le cas d'une grammaire computationnelle ; dans une telle grammaire, le lexique est formé à partir de lexèmes. En plus, en arabe, il n'y a pas que la préfixation de l'article défini ; il y a aussi la suffixation du *tanwīn*. Ce qui veut dire, qu'au départ, le nom arabe ne devrait pas être spécifié pour la définitude (i.e. ni DEF -, ni DEF +), et ce n'est que par la suite qu'il peut (i) soit devenir DEF - (ii) soit devenir DEF + (iii) soit encore resté définitivement sous-spécifié pour la définitude (dans le cas d'un nom en état construit ou auquel est suffixé un pronom possessif). De plus, dans le cas des états construits, la solution de Wintner (2000) propose de propager la définitude du complément vers la tête, la rendant ainsi définie. Mais, la question qui se pose : pourquoi rendre une tête définie alors qu'elle ne l'est pas par essence ! En conclusion, une seule règle comme celle proposée dans Wintner (2000) n'est pas suffisante pour rendre compte de la définitude au sein des nominaux arabes.

3.4. Une approche pour la définitude dans le syntagme nominal arabe

Puisque l'analyse que nous cherchons concerne l'aspect morphosyntaxique de la grammaire, nous présentons notre approche selon les deux volets : morphologie et syntaxe. En morphologie, quatre règles lexicales sont nécessaires pour rendre compte de la définitude des nominaux (pour être plus précis, il s'agit de deux règles et deux schémas de règles). Ces règles sont toutes flexionnelles. Les noms, par exemple, se

trouvant dans le lexique sous la forme de lexèmes (i.e. noms nus), subissent, tous, comme première opération d'affixation, la flexion du cas. La table 2 donne un résumé des opérations d'affixation possibles que pourrait subir un *lexème nom* dans le lexique par le biais de règles morphologiques. La première règle, que nous appellerons Règle Lexicale de Marquage de Cas, sert, comme son nom l'indique, à marquer le nom pour le cas (i.e. nominatif, accusatif ou génitif) ; tous les noms, qu'ils soient définis ou indéfinis, suffixés d'un pronom possessif ou non suffixés, ou encore composés ou non composés, subissent préalablement un marquage de cas. Cette règle est du type *lexeme-to-lexeme* ; elle reçoit en entrée un *lexème nom* et génère en sortie un *lexème nom marqué pour le cas*. En plus, ces lexèmes marqués pour le cas peuvent par la suite être utilisés comme noms en état construit. Par conséquent, ils peuvent se combiner directement avec un complément. Pour cela, cette règle spécifie le complément dont le nom pourrait en avoir besoin par la suite. Ce complément sera retiré par la suite de la liste Comps si le nom n'est pas en état construit ou n'est pas suffixé d'un pronom possessif au génitif ; ce retrait sera assuré par les trois autres règles lexicales proposées ci-après.

Lexème de base	Mot cible à générer		Opération(s) nécessaire(s)	Impact sur la définitude	Impact sur l'entrée lexicale
kitāb-(book)	kitāb-u-n (book-Nom-Indef)	1.	suffixation du cas (-u)	-	
		2.	suffixation de l'indéfinitude (-n)	Def -	
	kitāb-u (book-Nom)		suffixation du cas (-u)	Def ±	Spécification du complément dont le nom en a besoin (on est dans le cas d'un état construit). Ce complément peut être, à son tour, soit Def + soit Def -.
	kitāb-u-hu (book-Nom-his)	1.	suffixation du cas (-u)	-	La suffixation de *hu-* est matérialisée par le retrait du complément de nom de la liste Comps (liste des compléments attendus).
		2.	suffixation du possessif (-hu)	Def +	
	al-kitāb-u (Def-book-Nom)	1.	suffixation du cas (-u)	-	
		2.	préfixation de la définitude (al-)	Def +	

Tab. 2 : Affixation des nominaux dans le lexique

La deuxième et la troisième règles (respectivement, Règle Lexicale de Marquage de la Définitude et Règle Lexicale de Marquage de l'Indéfinitude) seront utilisées respectivement pour le marquage de la définitude et l'indéfinitude ; il y a soit la préfixation de *al-* et le marquage du trait de définitude à Def +, soit la suffixation de *-n* et le marquage du trait de définitude à Def -. Ces deux règles sont du type *lexeme-to-word* ; elles reçoivent en entrée un *lexème nom marqué pour le cas* et génèrent en sortie un *nom* (i.e. un mot) *marqué pour la définitude*, donc soit défini soit indéfini. Ces deux

règles assurent, en outre, le retrait du complément attendu de la liste Comps du nom (i.e. Comps < >). Enfin, la quatrième règle (Règle Lexicale du Possessif), qui est du type *lexeme-to-word*, permet de suffixer un pronom possessif (e.g. *-hu*) au *lexème nom*, marquer le trait de définitude à Def + et retirer le complément figurant dans la liste Comps du nom. L'introduction du trait Def facilite un encodage de la définitude qui soit indépendant de ses manifestations morpho-phonologiques réelles. Ce processus de définitude ayant pris place dans le lexique, les processus syntaxiques peuvent dès maintenant opérer sur ce trait, sans qu'ils aient accès à sa manifestation réelle. En syntaxe, le problème fondamental auquel nous faisons face est le suivant : Comment faire propager le trait Def à la projection maximale ? Pour répondre à cette question, nous devons répondre à une autre question qui en constitue une prémisse : quel est le constituant qui porte le trait Def qui doit percoler vers la projection maximale ? De cette question découle une autre question très importante : si le trait Def est approprié pour le trait Cat, quelle est sa position exacte dans la hiérarchie de la structure de traits des nominaux ? L'approche que nous adopterons en syntaxe et qui répondra à l'ensemble de ces interrogations se base sur des faits empiriques que nous avons constatés et conclus à travers les exemples (16) à (19) ci-dessous.

(16) a. al-maktab-u al-kabīr-u
Def-office.Masc.Sg-Nom Def-big.Masc.Sg-Nom
'the big office' (المكتبُ الكبيرُ)
'le grand bureau'

b. maktab-u-n kabīr-u-n
office.Masc.Sg-Nom-Indef big.Masc.Sg-Nom-Indef
'a big office' (مكتبٌ كبيرٌ)
'un grand bureau'

Ces faits confirment deux hypothèses traditionnelles de la grammaire arabe : (a) les adjectifs s'accordent totalement avec les noms ou les syntagmes nominaux qu'ils modifient (cf. (16.a-b)), et (b) les noms en états construits ne peuvent en aucun cas être définis par *al-* ; on peut facilement vérifier, dans les exemples (17.a-e), que, dans un état construit défini, seul le constituant le plus à droite (i.e. le complément auquel tous les autres compléments sont annexés) est défini (soit par *al-*, soit par annexion d'un pronom possessif), et ce quel que soit le nombre de constituants qui le composent.

(17) a. maktab-u al-mudīr-i
office-Nom Def-principal-Gen
'the principal's office','the office of the principal' (مكتبُ المديرِ)
'le bureau du directeur'

b. maktab-u mudīr-i al-madrasat-i
office-Nom principal-Gen Def-school-Gen

'the school principal's office'(مكتبُ مديرِ المدرسةِ)
'le bureau du directeur de l'école'

c. maktab-u mudīr-i madrasat-i al-ḥayy-i
office-Nom principal-Gen school-Gen Def-quarter-Gen
'the quarter school principal's office' (مكتبُ مديرِ مدرسةِ الحيِّ)
'le bureau du directeur de l'école du quartier'

d. maktab-u kātibat-i mudīr-i madrasat-i al-ḥayy-i
office-Nom secretary-Gen principal-Gen school-Gen Def-quarter-Gen
'the quarter school principal secretary's office' (مكتبُ كاتبةِ مديرِ مدرسةِ الحيِّ)
'le bureau de la secrétaire du directeur de l'école du quartier'

e. maktab-u mudīr-i madrasat-i-nā
office-Nom principal-Gen school-Gen-our
'our school principal's office'(مكتبُ مديرِ مدرستِنا)
'le bureau du directeur de notre école'

Par conséquent, nous pouvons dire que dans un état construit, c'est le complément le plus à droite qui décide de la définitude de l'état construit en entier ; en d'autres termes, si le constituant le plus à droite est (in)défini, l'état construit en entier est alors (in)défini. Ceci est visible dans l'exemple (18).

(18) a. maktab-u mudīr-i-n
office-Nom principal-Gen-Indef
'a principal's office','an office of a principal' (مكتبُ مديرٍ)
'un bureau d'un directeur','un bureau de directeur'

b. maktab-u mudīr-i madrasat-i-n
office-Nom principal-Gen school-Gen-Indef
'a school principal's office' (مكتبُ مديرِ مدرسةٍ)
'un bureau d'un directeur d'une école'

Pour synthétiser, nous dirons que : (a) La première hypothèse nous permet d'avancer que la valeur de la définitude au sein d'un syntagme nominal Nom-Adjectif est la même dans les deux constituants : nom et adjectif. Par conséquent, le trait Def du SN peut être récupéré indifféremment du nom tête ou de l'adjectif. (b) La deuxième hypothèse nous permet d'avancer que la définitude au niveau de la projection maximale d'un état construit ne peut provenir de la tête construite ; la définitude de cette dernière est d'ailleurs floue (i.e. sous-spécifié, ou Def ±). Donc, la définitude doit provenir plutôt du complément du nom. Mais, le complément du nom peut avoir à son tour une définitude floue, et ce flou peut s'étaler théoriquement (par récursivité) à plusieurs constituants. Parmi tous ces constituants, un seul devrait porter une valeur

claire pour la définitude ; c'est le complément le plus à droite. Par conséquent, c'est le trait Def du complément le plus à droite qui doit percoler vers la projection maximale. En plus, si l'état construit est modifié (19.a), ou encore, si le complément du nom est modifié (19.b), c'est au modifieur de porter le trait Def de l'état construit.

(19) a. maktab-u al-mudīr-i al-ğadīd-u
office-Nom Def-principal-Gen Def-new-Nom
'the principal's new office','the new office of the principal' (مكتبُ المديرِ الجديدُ)
'le nouveau bureau du directeur'

b. maktab-u al-mudīr-i al-ğadīd-i
office-Nom Def-principal-Gen Def-new-Gen
'the new principal's office', 'the office of the new principal' (مكتبُ المديرِ الجديدِ)
'le bureau du nouveau directeur'

La table 3 résume tous ces constats et donne pour chacun des exemples (16, 17, 18 et 19) les valeurs du trait Def associées au syntagme nominal en entier, à sa tête et à son complément le plus à droite. Ce qui nous permet de conclure le principe fixant le calcul de la définitude du syntagme nominal arabe, que nous énonçons comme suit :

Principe d'Héritage de la Définitude dans le SN Arabe :
La définitude d'un syntagme nominal non-unaire, construit ou non construit, doit être héritée à partir de l'adjoint (i.e. modifieur) sinon le complément le plus à droite.

	Syntagme Nominal	Def du SN	Def de l'adjoint (modifieur) ou le complément le plus à droite[6]	Def de la tête du SN
1.	al-maktab-u al-kabīr-u	Def +	Def +	Def +
2.	maktab-u-n kabīr-u-n	Def -	Def -	Def -
3.	maktab-u al-mudīr-i	Def +	Def +	Def -
4.	maktab-u mudīr-i al-madrasat-i	Def +	Def +	Def -
5.	maktab-u mudīr-i madrasat-i al-hayy-i	Def +	Def +	Def -
6.	maktab-u kātibat-i mudīr-i madrasat-i al-hayy-i	Def +	Def +	Def -
7.	maktab-u mudīr-i madrasat-i-nā	Def +	Def +	Def -
8.	maktab-u mudīr-i-n	Def -	Def -	Def -
9.	maktab-u mudīr-i madrasat-i-n	Def -	Def -	Def -
10.	maktab-u al-mudīr-i al-ğadīd-u	Def +	Def +	Def -
11.	maktab-u al-mudīr-i al-ğadīd-i	Def +	Def +	Def -

Tab. 3 : Les valeurs possibles pour le trait Def portées par les éléments concernés par la définitude au sein du syntagme nominal

Revenons, maintenant, à notre question première. Comment se fait la propagation du trait Def ? Si la propagation devait se faire à partir de l'adjoint ou du complément le plus à droite, nous arriverions au niveau de la projection maximale à une contradiction. Cette contradiction est due au fait qu'en HPSG, le principe des traits de têtes (HFP) permet de propager les traits Head du constituant fils-tête au syntagme parent (Sag, 1999, p :63). Donc, si le trait Def fait partie des traits de tête, comme le conçoit Wintner (2000), nous aurons deux valeurs pour ce même trait ; l'une provenant de la tête du syntagme, transportée par le HFP, et l'autre provenant de l'adjoint ou le complément le plus à droite comme nous le souhaitons. Par conséquent, le trait Def ne doit pas faire partie des traits de têtes ; en d'autre termes, il ne sera plus concerné par le HFP. Cette idée a été déjà introduite par Beermann & Ephrem (Beermann, 2007, p :29) qui ont été en face de la même situation en amharique et ont proposé d'exprimer la définitude comme un trait indépendant situé au même niveau que le trait Head dans la structure de traits des nominaux. Maintenant, par quel moyen se fait la propagation du trait Def ? Le HFP en est un. Nous procédons donc à sa révision pour qu'il puisse rendre compte de ces dernières propositions ; c'est-à-dire ne pas s'occuper seulement du trait Head mais aussi du trait Def dans le cas des nominaux. Nous le reformulons donc selon les termes suivants :

Le Principe des Traits de Tête (HFP) *(version modifiée)* :

Dans tout syntagme avec tête, les valeurs des traits Head et Def du père doivent être unifiées avec respectivement les valeurs Head du fils tête et Def du fils non tête le plus à droite.

Regardons, maintenant, ce que cela va donner avec les adjoints. Pour les adjoints, Pollard (1994) propose le Schéma Tête-Adjoint (connu dans la littérature sous le nom : Schéma Tête-Modifieur). Les Adjoints spécifient les têtes qu'ils sélectionnent comme valeur du trait Mod dans leurs entrées lexicales. Comme tous les autres nominaux (à l'exception des démonstratifs), ils ont un trait Def, dont la valeur est partagée avec la valeur du chemin Mod|Loc|Cat|Def. Quand la Règle Lexicale de Marquage de la Définitude (respectivement la Règle Lexicale de Marquage de l'Indéfinitude), que nous avons introduits supra, est appliquée aux adjoints, les deux chemins résultent dans une même spécification de la valeur '+' (respectivement '-'). Ainsi, il est garanti que les adjectifs (in)définis, par exemple, ne sont pas seulement spécifiés comme (in)définis, mais en plus, sélectionnent des têtes (in)définies. L'effet des deux règles, la Règle Lexicale de Marquage de Cas (RLMC) et la Règle Lexicale de Marquage de la Définitude (RLMD), appliquées au nom et à l'adjectif, est illustré dans les figures 2 et 3.

Une fois que le processus de l'ajout de l'article défini a pris place dans le lexique, les schémas Head-complément et Tête-Modifieur peuvent rester intacts.

[6]C'est le constituant le plus à droite ; il s'agit soit d'un adjectif soit d'un complément. De plus, nous pouvons voir, dans ces exemples, que la tête est toujours le constituant le plus à gauche.

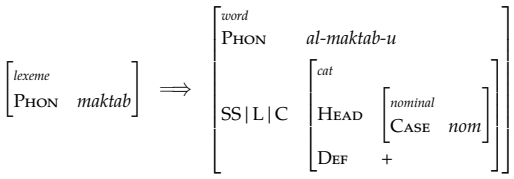

FIG. 2 : L'effet des règles RLMC et RLMD sur les noms

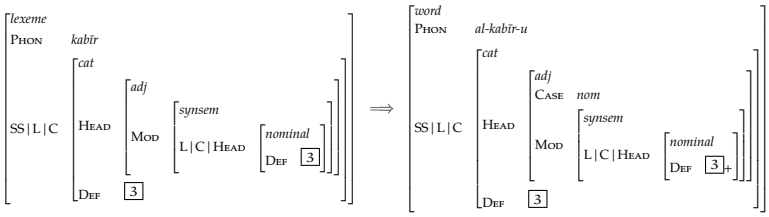

FIG. 3 : L'effet des règles RLMC et RLMD sur les adjectifs

La figure 4 représente la structure du syntagme nominal *al-maktab-u al-kabīr-u* 'the big office' (cf. (16.a)). Dans cet exemple, nous utilisons le SCHÉMA TÊTE-MODIFIEUR pour combiner un nom avec un adjectif. Sur cet exemple, nous pouvons voir l'effet du PRINCIPE DES TRAITS DE TÊTES MODIFIÉ au niveau du syntagme parent où la valeur du trait HEAD du fils-tête est coindexée (i.e. réentrante) avec la valeur HEAD du syntagme, instanciant ainsi le HFP modifié, ce qui est indiqué par l'indice 4, alors que la valeur du trait DEF est récupérée à partir du fils-adjoint, i.e. l'adjectif *al-kabīr-u* 'the-big'. Le lecteur averti notera que la contrainte exprimant l'accord en définitude entre l'adjectif et le nom modifié est assurée par l'indice 5. En effet, l'accord en définitude entre un nominal et ses adjoints est exprimé dans les entrées lexicales des adjoints, exactement de la même manière que l'est l'accord en nombre et en genre. Cependant, il y a une différence mineure entre les deux processus d'accord : puisque l'accord en définitude n'est pas un processus sémantique en arabe, le trait DEF ne fait pas partie du CONT des nominaux (contrairement au nombre et au genre). Mais puisque les modifieurs ont accès aux catégories des têtes qu'ils modifient (c'est une partie de la valeur de leurs traits MOD), les adjectifs peuvent sélectionner pour modification des noms définis ou indéfinis, selon leur propre valeur de définitude.

Notons enfin que les adjectifs, qu'ils soient définis ou indéfinis, peuvent modifier aussi bien les noms que les syntagmes nominaux. Ils n'imposent aucune contrainte sur les valeurs des traits de valence des nominaux qu'ils modifient. Pour analyser un état construit modifié tel que *maktab-u al-mudīr-i al-ǧadīd-u* 'the new office of the prin-

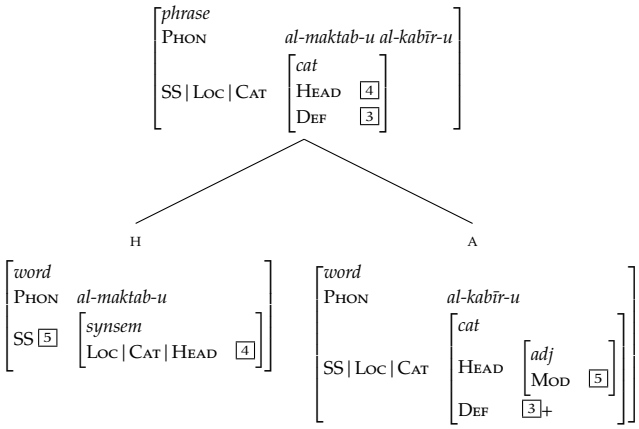

FIG. 4 : L'accord en définitude dans le syntagme nominal Nom-Adjectif

cipal' (cf. (19.a)), nous procédons comme suit. Le SCHÉMA TÊTE-COMPLÉMENT combine le nom en état construit *maktab-u* 'office' avec le complément qu'il attend *al-mudīr-i* 'the principal'. Le SCHÉMA TÊTE-MODIFIEUR intervient, ensuite, pour combiner l'état construit résultant *maktab-u al-mudīr-i* 'the office of the principal' avec l'adjectif modifieur *al-ǧadīd-u* 'the-new'. L'analyse de cet exemple (cf. Figure 5), nous permet de voir l'effet du PRINCIPE DES TRAITS DE TÊTES MODIFIÉ : (i) au niveau du syntagme intermédiaire, i.e. l'état construit *maktab-u al-mudīr-i* 'the office of the principal', la valeur du trait HEAD du fils-tête (indice 4) est coindexée avec la valeur HEAD du nœud parent, instanciant ainsi le HFP modifié, alors que la valeur du trait DEF est récupérée à partir du complément *al-mudīr-i* 'the principal' (indice 2) ; (ii) au niveau de la projection maximale, les valeurs des traits HEAD et DEF sont toutes les deux récupérées à partir du fils-tête *maktab-u al-mudīr-i* 'the office of the principal'.

Conclusion

Nous avons fourni dans ce papier une analyse de la définitude au sein du syntagme nominal arabe. Cette analyse nous a permis de proposer une nouvelle approche pour rendre compte de la définitude au sein du SN arabe argumentée sur des données empiriques. Notre approche consiste à promouvoir la définitude du syntagme nominal à partir non pas de sa tête lexicale mais plutôt du complément ou l'adjoint le plus à droite. L'analyse que nous avons proposée, et implémentée en HPSG, est basée sur deux hypothèses principales. La première, consiste à voir l'article défini de l'arabe (*al-*) comme étant un affixe flexionnel qui se combine avec les nominaux au niveau du lexique ; cette hypothèse a été confirmée par une variété d'arguments en utilisant

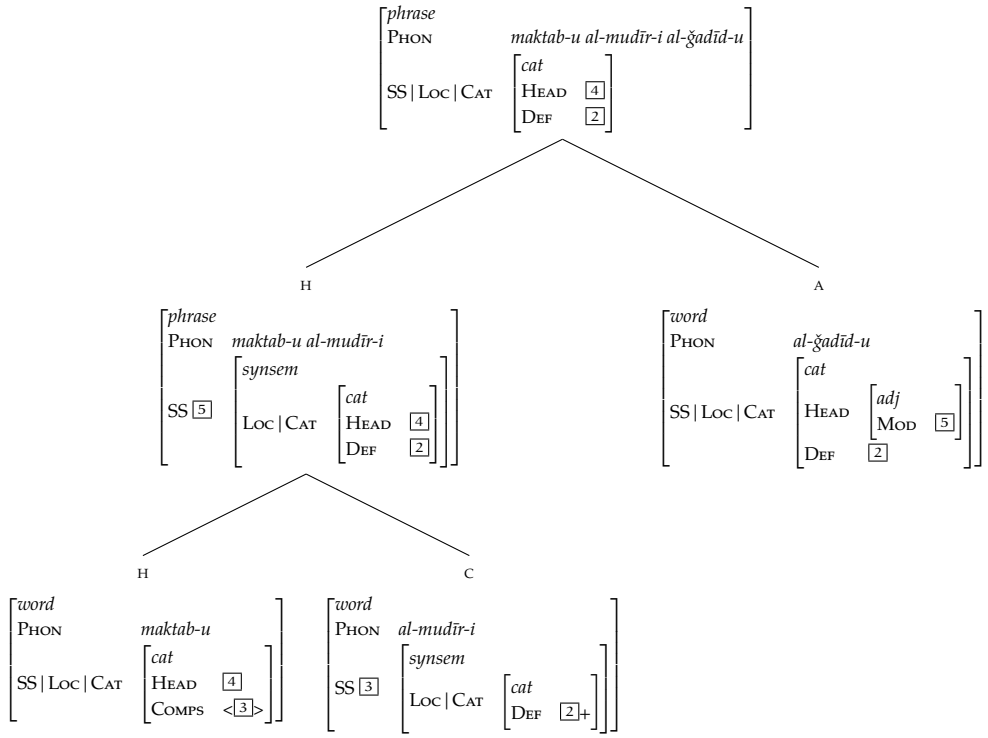

FIG. 5 : L'accord en définitude dans l'état construit modifié

les critères dits de Zwicky. La seconde stipule que le nom constitue la tête syntaxique du syntagme nominal. Cette analyse morphosyntaxique est scindée en deux parties : une première composante, morphologique, composée d'un certain nombre de règles lexicales, responsable de la génération des noms (in)définis et leur insertion dans la syntaxe ; une deuxième composante, syntaxique, basée essentiellement sur le Principe des Traits de Têtes Modifié, qui permet de rendre compte de l'accord en définitude entre les différents constituants du syntagme nominal ainsi que de l'héritage de la définitude au niveau de la projection maximale du syntagme nominal.

Summary

This paper proposes an HPSG analysis of the Modern Standard Arabic noun phrase. The focus will be on the definiteness phenomenon. For this purpose, we examine a varieties of related phenomena, including the definiteness marking within the nominal phrase, in particular within the genitive phrase, the definiteness agreement between the different nominal phrase elements, the affixal status of the definite article, the combining of the definite article with the elements of the construct state, and the definiteness at nominal phrase maximal projection. The analysis which we develop lies on two essential assumptions. The first one stipulates that the definite article *al-* in Arabic is an inflectional affix, and therefore, takes its place in morphology (or lexicon) rather than in syntax. To claim this assumption, we give several arguments based on some criteria commonly known as Zwicky Criteria. Furthermore, our analysis, which is implemented in a lexicalist framework, is based on a second assumption which identifies the noun as the head of the noun phrase ; therefore, we consider noun phrases as NPs, rather than as DPs.

Références

Beermann, D. & Ephrem, B. The Definite Article and Possessive Marking in Amharic. *CSLI Publications*, pages 21–32, 2007.

Djebali, A. Les pronoms liés en arabe classique sont-ils des clitiques ? *RÉLQ*, 1(1) :20–40, 2005.

Djebali, A. *La Modélisation des Marqueurs d'Arguments de l'Arabe Standard dans le Cadre des Grammaires à Base de Contraintes*. PhD thesis, Université du Québec à Montréal, Montréal, Qc, Canada, 2009.

Fassi, F. Arabic modifying adjectives and DP structures. *Studia Linguistica*, 53(2) :105–154, 1999.

Kayne, R. M. *Syntaxe du français : le cycle transformationnel*. Paris : Seuil, 1975.

Kouloughli, D. E. La Langue Arabe : Esquisse d'un Profil Historique et Linguistique. *LALIES 13*, 1994c.

Kremers, J. *The Arabic noun phrase : A minimalist approach*. PhD thesis, University of Nijmegen, Nijmegen, Netherlands, 2003.

Miller, Philip H. *Clitics and Constituents in Phrase Structure Grammar*. New York : Garland, 1992.

Nespor, M. & Vogel, I. *Prosodic phonology*. Dordrecht : Foris Publications, 1986.

Pollard, C. J. & Sag, I. A. *Head-Driven Phrase Structure Grammar*. University of Chicago Press, Chicago, 1994.

Rossi, E. *Clitic Production in Italian*. PhD thesis, University of Groningen, Groningen, Netherlands, 2007.

Sag, I.A. & Wasow, T. *Syntactic Theory : A Formal Introduction*. CSLI Publication, Stanford University, 1999.

Wintner, S. Definiteness in the Hebrew Noun Phrase. *Journal of Linguistics*, 36 :319–363, 2000.

Zwicky, A. M. *On Clitics*. Reproduced by Indiana University Linguistics Club, Bloomington, IN, 1977.

Zwicky, A. M. Clitics and Particles. *Language*, 61 :283–305, 1985a.

Zwicky, A. M. & Pullum, G. K. Cliticization vs. Inflection : English n't. *Language*, 59 :502–513, 1983.

Address for correspondence :
Mahmoud Fawzi Mammeri
Mahmud.mammeri@gmail.com
École Supérieure de Commerce d'Alger,
2, Rampe Salah Gherbi, Agha,
Alger, Algérie

PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 97 APRIL 2012 83–98

Kriya – An end-to-end Hierarchical Phrase-based MT System

Baskaran Sankaran, Majid Razmara, Anoop Sarkar

Simon Fraser University

Abstract

This paper describes *Kriya* – a new statistical machine translation (SMT) system that uses hierarchical phrases, which were first introduced in the Hiero machine translation system (Chiang, 2007). Kriya supports both a grammar extraction module for synchronous context-free grammars (SCFGs) and a CKY-based decoder. There are several re-implementations of Hiero in the machine translation community, but Kriya offers the following novel contributions: (a) Grammar extraction in Kriya supports extraction of the full set of Hiero-style SCFG rules but also supports the extraction of several types of compact rule sets which leads to faster decoding for different language pairs without compromising the BLEU scores. Kriya currently supports extraction of compact SCFGs such as grammars with one non-terminal and grammar pruning based on certain rule patterns, and (b) The Kriya decoder offers some unique improvements in the implementation of cube-pruning, such as increasing diversity in the target language n-best output and novel methods for language model (LM) integration. The Kriya decoder can take advantage of parallelization using a networked cluster. Kriya supports both KENLM and SRILM for language model queries. This paper also provides several experimental results which demonstrate that the translation quality of Kriya compares favourably to the Moses (Koehn et al., 2007) phrase-based system in several language pairs while showing a substantial improvement for Chinese-English similar to Chiang (2007). We also quantify the model sizes for phrase-based and Hiero-style systems and also present experiments comparing variants of Hiero models.

1. Introduction

Hierarchical Phrase-based Machine Translation (Chiang, 2005, 2007) is a prominent approach for Statistical Machine Translation (SMT). It is usually comparable to or better than conventional phrase-based systems for several language pairs.

Corresponding author: baskaran@cs.sfu.ca
Cite as: Baskaran Sankaran, Majid Razmara, Anoop Sarkar. Kriya – An end-to-end Hierarchical Phrase-based MT System. The Prague Bulletin of Mathematical Linguistics No. 97, 2012, pp. 83–98.
doi: 10.2478/v10108-012-0004-y.

In this paper, we present Kriya which implements a hierarchical phrase-based machine translation system which includes a grammar extraction module and a decoder. The name Kriya is the Sanskrit word for *verb* to signify that syntactic parsing techniques can be useful for machine translation.

Kriya is similar to some of the existing hierarchical phrase-based systems, but has some distinguishing features. For example, Kriya uses a unique approach for computing the language model (LM) heuristic in cube-pruning (Chiang, 2007) and it can also optionally support better diversity in the cube-pruning step. Kriya supports extraction of different types of more compact grammars as an alternative to full grammars typically extracted using the synchronous CFG (SCFG) extraction heuristics described in the original Hiero paper. The full grammar is typically associated with issues such as over-generation and search errors (de Gispert et al., 2010) and the use of compact grammars can achieve BLEU scores comparable to full grammar. Kriya also supports shallow-n decoding that leads to faster decoding while maintaining the same BLEU scores as the full decoding.

The rest of the paper is structured as follows. First we review some of the existing Machine Translation systems focusing on Hiero-style systems (Section 2) highlighting specific features. We then give a brief definition of synchronous context-free grammar (SCFG) in Section 3 to set the stage. In Section 4 we describe both grammar extractor and decoder modules interspersed with the features in Kriya. We finally present some experiments (Section 5) comparing Kriya with the well-known phrase-based system Moses which is used to benchmark Kriya's performance for several language pairs.

2. Related Works

`Moses`[1] (Koehn et al., 2007) is an open source toolkit that supports three types of state-of-the-art statistical machine translation systems: phrase-based, hierarchical phrase-based and syntax-based SMT. The toolkit is written in C++ and supports SRILM (Stolcke, 2002), KENLM (Heafield, 2011), randLM (Talbot and Osborne, 2007) and irstLM (Federico et al., 2008) for language model queries. To speed up training, tuning and test steps, Moses supports Oracle Grid Engine[2] (formerly Sun Grid Engine) and Amazon EC2 cloud and implements several memory/speed optimization algorithms. Chart decoding is done by the CKY+ algorithm which enables it to process arbitrary context free grammars with no limitations on the number of terminals or non-terminals in a rule. It also implements Chiang (2007)'s cube-pruning algorithm. Advanced methods such as Factored Models (Koehn and Hoang, 2007), Minimum Bayes Risk (MBR) decoding, Lattice MBR, Consensus Decoding and multiple translation table decoding (to name a few) have been implemented in Moses.

[1] `http://www.statmt.org/moses`

[2] `http://www.oracle.com/us/sun`

Joshua[3] (Li et al., 2009, 2010; Weese et al., 2011) developed at the Center for Language and Speech Processing at the Johns Hopkins University, is an open source machine translation toolkit written in Java that implements most critical algorithms required for hierarchical decoding such as chart-parsing, n-gram language model integration, beam and cube-pruning and k-best extraction. An advantage of this toolkit is that each component in the machine translation pipeline can be run with other components or separately such as Z-MERT (Zaidan, 2009) which is a stand-alone implementation of Och (2002)'s algorithm written in Java. The toolkit implements training corpus sub-sampling by which the most representative subset of the training corpus is used to extract rules from resulting in a faster training phase. In addition, Minimum Bayes Risk, Deterministic Annealing and Variational Decoding algorithms are implemented in this toolkit.

cdec[4] (Dyer et al., 2010) is another translation toolkit written in C++ which allows training and decoding a number of statistical machine translation models, including word-based models, phrase-based models and hierarchical phrase-based models. cdec provides support for Hadoop (an implementation of a distributed filesystem and MapReduce) for parallelization. Input to this system can be a sentence, lattice or context-free forest, which is then transformed to a unified translation forest. Secondly, language model re-scoring, pruning, inference algorithms and k-best derivation extraction are uniformly applied to the generated translation forest. cdec supports a number of optimization algorithms, including Minimum Error Rate Training (MERT) (Och, 2003), LBFGS (Liu and Nocedal, 1989), RPROP (Riedmiller and Braun, 1993) and Stochastic Gradient Descent. Compared to Joshua, cdec uses a smaller memory footprint with the same running time (Dyer et al., 2010).

Jane[5] (Vilar et al., 2010; Stein et al., 2011), RWTH's hierarchical phrase-based translation system, is a more recent open source toolkit which offers similar features. It is written in C++ and includes tools for phrase extraction and translation. Most of the operations can be parallelized by supporting grid engine clusters. The implementation of Jane allows for augmenting the feature set with arbitrary number of additional features as described in Stein et al. (2011). It also offers two ways to support additional models: combination in log-linear fashion and a mechanism to get the model to score a derivation to be incorporated in the main model's score. For tuning, Jane supports three different optimization methods: Minimum Error Rate Training (MERT) (Och, 2003), Margin Infused Relaxed Algorithm (MIRA) (Crammer et al., 2006) and the Downhill Simplex method (Nelder and Mead, 1965). Stein (2011) shows that Jane is 50% faster than Joshua on identical settings.

[3] http://www.sourceforge.net/projects/joshua

[4] https://github.com/redpony/cdec

[5] http://www.hltpr.rwth-aachen.de/jane

3. Synchronous Context-Free Grammars

This section provides a formal definition of a synchronous context-free grammar (SCFG) as a precursor to the discussion of the implementation in Kriya.

Formally a grammar G in a hierarchical phrase-based model is a special case of probabilistic synchronous context-free grammar (PSCFG) that is defined as a 4-tuple: $G = (T, NT, R, R_g)$, where, T and NT are the sets of terminals and non-terminals in G. Hiero grammars typically use two non-terminals X and S where S is the start symbol. R is a set of production rules of the form:

$$X \rightarrow \langle \gamma,\ \alpha,\ \sim \rangle,\ \gamma, \alpha \in \{X \cup T^{+}\} \tag{1}$$

The $\sim$ in the hierarchical rule indicates the alignment indices for the non-terminals in the production rule such that the co-indexed non-terminal pairs are rewritten synchronously. These production rules are combined in the top by the *glue* rules R_g leading to the start symbol S:

$$S \rightarrow \langle X_1,\ X_1 \rangle \tag{2}$$

$$S \rightarrow \langle S_1 X_2,\ S_1 X_2 \rangle \tag{3}$$

where the non-terminal indices indicate synchronous rewriting of the source and target non-terminals having the same index.

4. Kriya

Our implementation of Kriya closely follows the original exposition in Chiang (2007) with extensions that provide several additional features. Broadly, Kriya consists of two independent modules: a *grammar extractor* and a *CKY-based decoder*. Traditionally, grammar extraction has been a bottleneck in Hiero-style translation, due to the massive size of Hiero SCFG grammars and also due to the increasing availability of parallel data. The grammar extractor in Kriya has been designed to efficiently learn translation model even for a very large data set and this is achieved by way of parallelization and optimization. Thus our approach does not resort to sub-sampling to choose a smaller representative training set. Alternately Kriya also supports extraction of several variants of more compact grammars, for example extracting a 1 non-terminal grammar or filtering the full grammar based on certain greedily selected rule patterns (Iglesias et al., 2009).

Kriya decoder currently supports SCFG models for string inputs and features cube-pruning (Chiang, 2007) for integrating the language model scores with the decoder. We introduce a novel approach for improving the heuristic language model scores for the left and right contexts in CKY-based decoder by taking into account the potential position of the target hypothesis fragment in the final candidate. Kriya also supports

shallow-n decoding (de Gispert et al., 2010) for fast decoding without impacting the translation quality for certain close language pairs.

Kriya has been written primarily in Python (versions 2.6 and 2.7). This allows us to test new ideas by quickly implementing them in short duration at the same time keeping the code-base small, manageable and easy to read. On the negative side, Kriya is bit slower mainly due to the well-known speed issues in Python, which we alleviate using several engineering optimizations. These optimizations have resulted in practically acceptable training and decoding speeds in Kriya as we later quantify in Section 5.

4.1. Kriya Grammar Extractor

The Hiero grammar extraction algorithm (Chiang, 2007) starts from the set of initial phrases that are identified by growing the word alignments into longer phrases. Given the initial phrases corresponding to a sentence pair, the heuristic algorithm first designates the smaller initial phrases (such as phrase pairs having non-decomposable alignments) as terminal rules, expanded from a non-terminal X. The algorithm then extracts hierarchical rules by substituting the smaller spans within the larger phrases by the non-terminal X if the phrase pair corresponding to the smaller span has already been identified as a rule. It extracts all possible rules from the initial phrases subject to a maximum of two X non-terminals in a rule such that they do not rewrite adjacent spans in the source side. The Hiero extraction assumes unit count for each initial phrase and distributes this uniformly to all the rules extracted from the phrase. The parameter estimation then proceeds by relative frequency estimation.

Chiang proposed the total number of source side (terminals and non-terminals) terms and a maximum rule length to be 5 and 10 respectively. We found improvements with longer source side rules. In comparison, phrase-based models typically use a maximum phrase length of 7 and sometimes even longer phrases for certain language pairs. The source side length and maximum rule length are customizable parameters in the Kriya rule extraction, to facilitate experiments with different lengths.

A major issue in the extraction of Hiero grammar is the exponential size of the resulting grammar such that the full grammar cannot be held in memory for parameter estimation. Some of the existing Hiero systems use sub-sampling (Weese et al., 2011) to reduce the size of the training corpus and run the grammar extraction on the sub-sampled corpus, resulting in approximate probability estimates. In contrast Kriya uses the entire training data and we use memory optimizations and parallelization to achieve this.

The grammar extractor in Kriya is modularized to run in three phases in order to efficiently extract grammars even for large training corpora. In the first phase the extractor splits the training corpora into smaller chunks and extracts the rules for each chunk by trivial parallelization over the cluster. The second step scans the rules from the individual chunks and filters them based on the source side texts of a tuning or

test set; at the same time collecting the accurate counts for the target phrases in the filtered rules. The final step estimates the forward and reverse probabilities using relative frequency estimation.

The grammar extractor has been customized to the cluster environment (Kriya will soon support the Hadoop framework for extraction and decoding) and thus the extraction can be massively parallelized to efficiently extract Hiero grammar for large corpora. For a smaller data set, it is however possible to estimate the parameters for the full grammar by way of changing the configuration file.

4.1.1. Extracting Compact Grammars

Apart from the original Hiero-style model, Kriya grammar extractor supports the extraction of some variants that are smaller than the full grammar using different pruning[6] strategies. The main motivation for such pruned grammars is twofold, i) to reduce the grammar size and ii) to speed up the decoding enabling faster experiments. In some cases, the resulting compact grammars are suggested to improve the translation by way of reducing search errors (Iglesias et al., 2009; He et al., 2009), which has been contradicted elsewhere (Zollmann et al., 2008; Sankaran et al., 2011).

Kriya supports different approaches to prune the hierarchical phrase-based grammars and here we restrict ourselves to two methods that have been proposed earlier. Other approaches involving Bayesian methods for inducing a compact grammar (Sankaran et al., 2011) or a left-to-right decoding model similar to Watanabe et al. (2006) are under active research as mentioned later in the future directions.

First, the Hiero grammar can be simplified to have just 1 non-terminal, instead of 2 as proposed by Chiang. This grammar eliminates large number of rules, many of which turn out to be composed rules (He et al., 2009) that can be constructed by combining two or more smaller rules leading to spurious ambiguity (Chiang, 2007) during translation. Such 1 non-terminal grammar has been shown to have BLEU scores similar to the full Hiero grammar (Sankaran et al., 2011) for some language pairs such as English-Spanish, but suffer a reduction of about 1 BLEU point for Chinese-English and Urdu-English (Zollmann et al., 2008).

Another alternative is pruning based on rule patterns (Iglesias et al., 2009) which can reduce the size of the grammar. Kriya supports pattern-based filtering, and this is optionally triggered by using a separate configuration file specifying the patterns in the training process. The list of rule patterns to be included/excluded in the extraction process can be specified in the configuration file, in a notation similar to the one used by Iglesias et al. (2009).

[6]Some earlier works use the word *filtering* for this. We prefer *pruning* (or simplification) to indicate the case where some rules are removed that are otherwise applicable in decoding a given tuning/test set, while reserving the word *filtering* to the process of removing rules that will never be applicable for the tuning/test set. The latter is thus risk-free while the former is lossy.

4.2. Kriya Decoder

Kriya currently supports decoding with hierarchical phrase-based models employing CKY-style chart parsing. Given a source sentence f, the decoder finds the target side yield $\hat{e}$ of the best scoring derivation obtained by applying rules in the synchronous context-free grammar.

$$\hat{e} = \mathscr{Y}_e\left(\arg\max_{d \in D(f)} P(d)\right) \tag{4}$$

where, $D(f)$ is the set of derivations attainable from the learned grammar for the source sentence f. The model over derivations $P(d)$ is formulated as a log-linear model (Och and Ney, 2002) employing a set of features $\{\phi_1, \ldots, \phi_M\}$ apart from a language model feature that scores the target yield as $P_{lm}(e)$. The model can be written by factorizing derivation d into its component rules R_d as below.

$$P(d) \propto \left(\prod_{i=1}^{M} \prod_{r \in R_d} \phi_i(r)^{\lambda_i}\right) P_{lm}(e)^{\lambda_{lm}} \tag{5}$$

where, λ_i is the corresponding weight of the feature ϕ_i. The feature weights λ_i are optimized against some evaluation metric (Och, 2003), typically BLEU (Papineni et al., 2002). The default settings in Kriya support the standard features as will be mentioned later.

The decoder parses the source sentence with a modified version of CKY parser with the target side of corresponding derivations simultaneously yielding the candidate translations. The rule parameters and other features are used to score the derivations along with the language model score of the target translation as in Equation 5.

The derivation starts from the leaf cells of the CKY chart corresponding to the source side tokens and proceeds bottom-up. For each cell in the CKY chart, the decoder identifies the applicable rules and analogously to monolingual parsing, the non-terminals in these rules should have corresponding entries in the respective antecedent cells. The target side of the production rules yield the translation for the source span and the translations in the top-most cell correspond to the entire sentence. We encourage readers to refer to Chiang (2007) for more details.

Similar to Chiang, we use cube-pruning, specifically its lazier version (Huang and Chiang, 2007) to integrate the language model scoring in the decoding process. We introduce a novel approach in improving the heuristic language model score by taking into account the likely position of the target hypothesis fragment in the final translation which we explain in detail in the next section.

4.2.1. Novel Enhancements in Cube-pruning

The traditional phrase-based decoders using beam search generate the target hypotheses in the left-to-right order. In contrast, CKY decoders in Hiero-style systems

can freely combine target hypotheses generated in intermediate cells with hierarchical rules in the higher cells. Thus the generation of the target hypotheses are fragmented and out of order in Hiero, compared to the left to right order preferred by n-gram language models.

This leads to challenges in the estimation of language model scores for partial target hypothesis, which is being addressed in different ways in the existing Hiero-style systems. Some systems add a sentence initial marker (<s>) to the beginning of each path and some other systems have this implicitly in the derivation through the translation models. Thus the language model scores for the hypothesis in the intermediate cell are approximated, with the true language model score (taking into account sentence boundaries) being computed in the last cell that spans the entire source sentence.

We introduce a novel improvement in computing the language model scores: for each of the target hypothesis fragments, our approach finds the best position for the fragment in the final sentence and uses the corresponding score. We compute three different scores corresponding to the three positions where the fragment can end up in the final sentence, viz. sentence initial, middle and final: and choose the best score. As an example for fragment t_f consisting of a sequence of target tokens, we compute LM scores for i) <s> t_f, ii) t_f and iii) t_f </s> and use the best score for pruning alone[7].

This improvement significantly reduces the search errors while performing *cube-pruning* (Chiang, 2007) at the cost of additional language model queries. For example, a partial candidate covering a non-final source span might be reordered to the final position in the target translation. If we just compute the LM score for the target fragment as is done normally, this might get pruned early on before being reordered by a production rule. Our approach instead computes the three LM scores and it would correctly use the last LM score which is likely to be the best, for pruning.

Our experiments indicated a small but consistent increase in the BLEU scores and also reduction in the search errors due to this improvement in the computation of LM scores and we also found this to be especially helpful in the shallow-n decoding setting (Section 4.2.2) as we later discuss in experiments (Section 5).

However, additional queries to the language model result in a slight reduction in the decoding speed. This could be partly addressed by saving the three LM scores for both left and right edges with the hypothesis and reusing them appropriately when either or both edges remain unchanged. Secondly following the general strategy, we exploit n-gram state information in KENLM (Heafield, 2011) to query the language model for incremental target fragment following a stored state.

As a second enhancement, Kriya optionally supports improved diversity in cube-pruning by allowing a fixed (typically 3) number of candidates for each cube that are not represented in the cell. These hypotheses are included in the cell in addition to

[7]This ensures the the LM score estimates are never underestimated for pruning. We retain the LM score for fragment (case ii) for estimating the score for the full candidate sentence later.

the hypotheses pushed into the stack through cube-pruning. We found cube-pruning diversity to be marginally effective in different settings as we discuss later.

4.2.2. Shallow-n Decoding

Shallow-n grammars (de Gispert et al., 2010) are a class of grammars that restrict the number of successive hierarchical rules in a derivation in order to reduce the over-generation caused by large Hiero grammars. While this has restricted reordering capability compared to full Hiero, the degree of reordering can be customized to the requirements of specific language pairs by way of changing n. For example, Shallow-1 grammar might be sufficient for language pairs such as English-French and Arabic-English, whereas higher order shallow grammars are required for Chinese-English because of their large syntactic divergence. As a direct consequence of the reduction in the search space, shallow-n decoding results in substantially faster decoding.

Formally, a Shallow-n grammar G is defined as a 5-tuple: $G = (N, T, R, R_g, S)$, such that T is a set of finite terminals and N a set of finite non-terminals $\{X^0, \ldots, X^N\}$. As earlier R_g refers to the glue rules that rewrite the start symbol S:

$$S \rightarrow \langle X,\ X \rangle \tag{6}$$

$$S \rightarrow \langle SX,\ SX \rangle \tag{7}$$

R is the set of finite production rules in G and has two types, viz. hierarchical (8) and terminal (9). The hierarchical rules at each level n are additionally conditioned to have *at least* one X^{n-1} non-terminal in them. The $\sim$ in the hierarchical rule serves as the index for aligning the non-terminals such that the co-indexed non-terminal pair can be rewritten synchronously.

$$X^n \rightarrow \langle \gamma,\ \alpha,\ \sim \rangle,\ \gamma, \alpha \in \{\{X^{n-1}\} \cup T^+\} \tag{8}$$

$$X^0 \rightarrow \langle \gamma,\ \alpha \rangle, \qquad \gamma, \alpha \in T^+ \tag{9}$$

Kriya supports Shallow-n decoding, without requiring additional non-terminals to be explicitly created in the Hiero grammar (this is similar to other implementations of this idea). We simply keep track of the number of hierarchical nestings in the partial hypotheses stored in the decoder as part of the hypothesis state. We find the shallow grammars to be comparable to closer language pairs such as English-French and English-Spanish, but the translation performance suffers for Arabic-English and Chinese-English without additional hacks.

4.3. Optimizing Feature Weights

Kriya uses the well-known MERT algorithm (Och, 2003) for optimizing feature weights and it has been integrated with both MERT implementation in Moses (Koehn

et al., 2007) and zMERT (Zaidan, 2009). Recently we have also added an implementation of Pairwise Ranking Optimization (Hopkins and May, 2011) (PRO) within Kriya for tuning the feature weights.

5. Experiments

In this section we present experiments to evaluate Kriya on several language pairs. We use five different language pairs in our experiments - representing a wide range of diversities, such as close languages (English-French), translating into a slightly more inflected language than English (English-Spanish) and languages with high syntactic divergence (Chinese-English). Table 1 shows some statistics about the corpora used for our experiments.

Language pair	Corpus	Train/ Tune/ Test	Language Model
English-Spanish	WMT10 (Europarl + News commentary)	1.7 M/ 5078/ 2489	WMT10 *train* + UN
English-French		1.7 M/ 5078/ 2489	WMT10 *train*
French-English		1.7 M/ 5078/ 2489	Gigaword
Chinese-English	*Train*: HK + GALE Phase-1 *Tune*: MTC parts 1 & 3; *Test*: MTC part 4	2.3 M/ 1928/ 919	Gigaword
Arabic-English	ISI automatically extracted Parallel text	1.1 M/ 1982/ 987	Gigaword

Table 1. Corpus Statistics - English-French uses a 4-gram LM and other pairs use 5-gram LMs. Chinese-English experiments use four references for tuning and testing.

First we present experimental results to benchmark Kriya's performance in all these language pairs by comparing it with the well-known Moses phrase-based system. We used standard settings for Moses in all these experiments except for the maximum phrase length, which we set to 7. For Kriya models, we set the total source side terms to be 7 for Chinese-English and Arabic-English and 5 for others. We ran Kriya in standard setting which includes 8 features: inverse and direct phrase translation probabilities $p(f|e)$ and $p(e|f)$; inverse and direct lexical weights $p_l(f|e)$ and $p_l(e|f)$; phrase penalty; word penalty; glue rule penalty and language model.

For both Moses and Kriya, we trained lower-cased models for Chinese-English and Arabic-English, while training true-cased models for the rest. All the experiments described here use MERT (Och, 2003) for optimizing the weights of the features. The BLEU (Papineni et al., 2002) scores are computed using the official NIST evaluation script.[8]

[8] ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-v11b.pl

Language Pair	Moses		Kriya	
	Model size	*BLEU*	*Model size*	*BLEU*
English-Spanish	154.6	28.12	632.8	**28.19**
English-French	81.8	23.48	519.9	**23.54**
French-English	81.9	26.15	439.9	***26.63***
Arabic-English	68.0	37.31	331.5	***37.74***
Chinese-English	83.6	24.48	286.1	***25.96***

Table 2. Model sizes and BLEU Scores - Model sizes are in millions of rules. Bold face indicates best BLEU score for each language pair and italicized figures point to statistically significant improvements assuming significance level $\alpha = 0.1$.

Table 2 lays out the BLEU scores as well as the model sizes of the Moses and Kriya phrase tables. As shown, the hierarchical phrase-based system always has larger models (unfiltered phrase table size), ranging between 342.2% and 635.5% of their phrase-based counterparts. In terms of BLEU scores, Kriya results in higher BLEU scores in all the language pairs with the best improvement coming for Chinese-English, confirming the results of Chiang (2007). Further, Kriya achieves statistically significant improvements for Arabic-English and English-French experiments.

As mentioned earlier, the huge model size of the Hiero systems slows down decoding and earlier research has proposed two different approaches for this: Shallow-n decoding as opposed to the full decoding restricts the depth of non-glue hierarchical rules in the derivation. Orthogonal to this, more compact models that are substantially smaller than the full Hiero models can be used with full decoding. In this experiment we compare the basic variants of these two approaches in terms of BLEU scores and model size.

In the Shallow decoding setting, we use shallow-1 thus restricting the hierarchical rules in the grammar to directly rewrite into terminal rules with the glue rules freely combining the hierarchical rules. We compare this with a simpler Hiero grammar consisting of one non-terminal and this generally results in a compact model compared to the original Hiero model. Note that these two ideas are orthogonal and hence can be combined; however we generally find them to result in poor performance and so we ignore the combination experiments here.

The experimental results are summarized in Tables 3 and 4. We find the simpler model consisting of one non-terminal employing full decoding to be competitive to the full model for closer language pairs such as French-English and Arabic-English at the same time clocking higher decoding speed. However, we see a reduction in the BLEU score for Chinese-English as has also been found by Zollmann et al. (2008). We thus hypothesize that 1 NT models have the same expressive power as the regular Hiero models (with 2 non-terminals), at least for languages with little syntactic diver-

Language Pair	Original (2 NT)/ Shallow-1	Compact (1 NT)/ Full	
	BLEU	*BLEU*	*Model size*
English-Spanish	27.70	**28.15**	351.3 (55.5%)
English-French	23.22	**23.48**	290.3 (55.8%)
French-English	**26.67**	**26.66**	248.2 (56.4%)
Arabic-English	37.15	**37.71**	161.4 (49.0%)
Chinese-English	24.04	25.25	154.2 (53.9%)

Table 3. Shallow-1 decoding vs. Compact (1 NT) model - Bold face indicates BLEU scores comparable to the original Hiero model in Table 2. Size of the compact 1 NT model as a % of original Hiero model is given within the brackets.

Model	Decoding level	Decoding time
Original (2 NT)	Full	0.71
Original (2 NT)	Shallow-1	0.24
Compact (1 NT)	Full	0.50

Table 4. Kriya Decoding time (in secs/word) for Chinese-English translation

gence. They also reduce the model size almost by half achieving a highest reduction of 51% for Arabic-English.

Shallow-1 decoding achieves highest decoding speed among the three but suffers a small reduction in the BLEU score except for French-English and incurs a larger reduction of 1.9 BLEU points for Chinese-English. It is three times faster than full decoding and twice faster than the 1 NT model. Higher order shallow decoding (not shown here), for example shallow-2 for Arabic-English and shallow-3 for Chinese-English achieve competitive performance but shallow-3 case suffers substantial reduction in decoding speed and is only marginally faster than full decoding.

In analyzing the effect of our novel approach of LM integration, we compare our approach to the naive approach of computing heuristic LM score that prefixes a beginning of sentence marker (<s>) to the candidate hypothesis. We believe that our approach will reduce search errors by finding better scoring candidates than the usual ones and we look at two parameters to test this: translation quality as measured by BLEU scores and search errors. In measuring search errors, we compared the model scores of the candidates (with one-to-one correspondence) in the N-best lists obtained by the two approaches and computed the percent of high scoring candidates in each N-best list.

In the shallow setting, our method improved the BLEU scores by 0.4 for both Arabic-English and Chinese-English. Our approach also obtained a much better N-

best list, with 94.6% and 77.3% of candidates in the N-best list having better scores than the naive approach for Chinese-English and Arabic-English respectively. In the full decoding setting the improvements, were lower with 69% and 57% of candidates obtaining better model scores for the two language pairs in the same order and the BLEU score increasing by 0.3.

We also found cube-pruning diversity to be useful in our experiments in Arabic-English and Chinese-English. We set cube-pruning diversity to be 3 (where, top-3 hypotheses from each unrepresented cube being added to the hypotheses in the corresponding CKY-chart cell) and in different settings involving full decoding and shallow-n decoding, the translation quality improved by a small margin of 0.25 BLEU points.

6. Future Directions

Kriya continues to be in active development and we are planning to add several new features. Currently, it supports the TORQUE cluster environment for parallelizing training and optimization processes. We are currently working towards supporting the MapReduce framework, specifically using a Hadoop setup. In recent developments in the Kriya decoder, we are exploring a new left-to-right decoder similar to Watanabe et al. (2006) in order to take advantage of its straight-forward language model integration in order to achieve a faster decoding time. Furthermore, we plan to extend the novel *ensemble* framework (Razmara et al., 2012) for decoding, which has already been implemented in Kriya. The ensemble decoder dynamically combines information from multiple translation and/or language models and can better exploit training data from different domains. This can particularly be useful in scenarios such as domain adaptation, multi-source translation. In inducing better SCFG grammars, we are working on efficient alternatives to the usual heuristic rule extraction approach, extending our earlier work using a Bayesian model (Sankaran et al., 2011). Finally, we are interested in incorporating syntax in our research with Kriya, in order to exploit the divergence between different languages.

7. Conclusion

We presented *Kriya* – a new implementation of hierarchical phrase-based machine translation which has novel features and which achieves competitive performance in several language pairs. Kriya's grammar extractor can efficiently extract Hiero grammars from large training sets and supports extraction of several compact Hiero grammar variants. The decoder currently uses CKY-based decoding, and we are currently working on left-to-right decoding to speed up the decoder. Kriya is under active development and several new features are being planned with specific focus on Bayesian models for extracting compact grammars, ensemble decoding, support for MapReduce framework and so on. We also presented experimental results on five language

pairs using the Kriya system comparing different variants of the basic Hiero-style decoder and different Hiero-style grammars.

Bibliography

Chiang, David. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of Annual Meeting of Association of Computational Linguistics*, pages 263–270, 2005.

Chiang, David. Hierarchical phrase-based translation. *Computational Linguistics*, 33, 2007.

Crammer, Koby, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. Online passive-aggressive algorithms. *Journal of Machine Learning Research*, 7:551–585, 2006.

de Gispert, Adrià, Gonzalo Iglesias, Graeme Blackwood, Eduardo R. Banga, and William Byrne. Hierarchical phrase-based translation with weighted finite-state transducers and Shallow-n grammars. *Computational Linguistics*, 36, 2010.

Dyer, Chris, Adam Lopez, Juri Ganitkevitch, Johnathan Weese, Ferhan Ture, Phil Blunsom, Hendra Setiawan, Vladimir Eidelman, and Philip Resnik. cdec: A decoder, alignment, and learning framework for finite-state and context-free translation models. In *Proceedings of the Annual Meeting of Association of Computational Linguistics 2010 System Demonstrations track*, ACLDemos '10, pages 7–12, 2010.

Federico, Marcello, Nicola Bertoldi, and Mauro Cettolo. IRSTLM: an open source toolkit for handling large scale language models. In *INTERSPEECH 2008, 9th Annual Conference of the International Speech Communication Association*, pages 1618–1621. ISCA, 2008.

He, Zhongjun, Yao Meng, and Hao Yu. Discarding monotone composed rule for hierarchical phrase-based statistical machine translation. In *Proceedings of the 3rd International Universal Communication Symposium*, pages 25–29, 2009.

Heafield, Kenneth. KenLM: faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, 2011.

Hopkins, Mark and Jonathan May. Tuning as ranking. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1352–1362. Association for Computational Linguistics, 2011.

Huang, Liang and David Chiang. Forest rescoring: Faster decoding with integrated language models. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 144–151. Association for Computational Linguistics, 2007.

Iglesias, Gonzalo, Adrià de Gispert, Eduardo R. Banga, and William Byrne. Rule filtering by pattern for efficient hierarchical translation. In *Proceedings of the 12th Conference of the European Chapter of the ACL*, pages 380–388, 2009.

Koehn, Philipp and Hieu Hoang. Factored translation models. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 868–876, Prague, Czech Republic, 2007. Association for Computational Linguistics.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej

Bojar, Alexandra Constantin, and Evan Herbst. Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 177–180. Association for Computational Linguistics, 2007.

Li, Zhifei, Chris Callison-Burch, Chris Dyer, Sanjeev Khudanpur, Lane Schwartz, Wren Thornton, Jonathan Weese, and Omar Zaidan. Joshua: An open source toolkit for parsing-based machine translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 135–139. Association for Computational Linguistics, 2009.

Li, Zhifei, Chris Callison-Burch, Chris Dyer, Juri Ganitkevitch, Ann Irvine, Sanjeev Khudanpur, Lane Schwartz, Wren Thornton, Ziyuan Wang, Jonathan Weese, and Omar Zaidan. Joshua 2.0: A toolkit for parsing-based machine translation with syntax, semirings, discriminative training and other goodies. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 133–137. Association for Computational Linguistics, 2010.

Liu, Dong C. and Jorge Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45:503–528, 1989.

Nelder, John A. and Roger Mead. A simplex method for function minimization. *Computer Journal*, 7:308–313, 1965.

Och, Franz Josef. Minimum error rate training in statistical machine translation. In *Proceedings of the Annual Meeting of Association of Computational Linguistics*, pages 160–167, 2003.

Och, Franz Josef and Hermann Ney. Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of the Annual Meeting of Association of Computational Linguistics*, pages 295–302, 2002.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the Annual Meeting of Association of Computational Linguistics*, pages 311–318, 2002.

Razmara, Majid, George Foster, Baskaran Sankaran, and Anoop Sarkar. Mixing multiple translation models in statistical machine translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, Jeju, Republic of Korea, July 2012. Association for Computational Linguistics.

Riedmiller, Martin and Heinrich Braun. A direct adaptive method for faster backpropagation learning: The RPROP algorithm. In *IEEE International Conference on Neural Networks*, pages 586–591, 1993.

Sankaran, Baskaran, Gholamreza Haffari, and Anoop Sarkar. Bayesian extraction of minimal SCFG rules for hierarchical phrase-based translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 533–541, 2011.

Stein, Daniel, David Vilar, Stephan Peitz, Markus Freitag, Matthias Huck, and Hermann Ney. A guide to Jane, an open source hierarchical translation toolkit. In *The Prague Bulletin of Mathematical Linguistics, No. 95*, pages 5–18, 2011.

Stolcke, Andreas. SRILM – an extensible language modeling toolkit. In *Proceedings International Conference on Spoken Language Processing*, pages 257–286, 2002.

Talbot, David and Miles Osborne. Randomised language modelling for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 512–519, Prague, Czech Republic, 2007. Association for Computational Linguistics.

Vilar, David, Daniel Stein, Matthias Huck, and Hermann Ney. Jane: open source hierarchical translation, extended with reordering and lexicon models. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 262–270. Association for Computational Linguistics, 2010.

Watanabe, Taro, Hajime Tsukada, and Hideki Isozaki. Left-to-right target generation for hierarchical phrase-based translation. In *Proceedings of the 21st International Conference on Computational Linguistics (COLING)*, pages 777–784. Association for Computational Linguistics, 2006.

Weese, Jonathan, Juri Ganitkevitch, Chris Callison-Burch, Matt Post, and Adam Lopez. Joshua 3.0: Syntax-based machine translation with the Thrax grammar extractor. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 478–484. Association for Computational Linguistics, 2011.

Zaidan, Omar F. Z-mert: A fully configurable open source tool for minimum error rate training of machine translation systems. *Prague Bulletin of Mathematical Linguistics*, 2009.

Zollmann, Andreas, Ashish Venugopal, Franz Och, and Jay Ponte. A systematic comparison of phrase-based, hierarchical and syntax-augmented statistical machine translation. In *Proceedings of the 22nd International Conference on Computational Linguistics*, pages 1145–1152, 2008.

Address for correspondence:
Baskaran Sankaran
baskaran@cs.sfu.ca
School of Computing Science
Simon Fraser University
8888 University Dr, Burnaby
BC V5A 1S6, Canada

PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 97 APRIL 2012

INSTRUCTIONS FOR AUTHORS

Manuscripts are welcome provided that they have not yet been published elsewhere and that they bring some interesting and new insights contributing to the broad field of computational linguistics in any of its aspects, or of linguistic theory. The submitted articles may be:

- long articles with completed, wide-impact research results both theoretical and practical, and/or new formalisms for linguistic analysis and their implementation and application on linguistic data sets, or
- short or long articles that are abstracts or extracts of Master's and PhD thesis, with the most interesting and/or promising results described. Also
- short or long articles looking forward that base their views on proper and deep analysis of the current situation in various subjects within the field are invited, as well as
- short articles about current advanced research of both theoretical and applied nature, with very specific (and perhaps narrow, but well-defined) target goal in all areas of language and speech processing, to give the opportunity to junior researchers to publish as soon as possible;
- short articles that contain contraversing, polemic or otherwise unusual views, supported by some experimental evidence but not necessarily evaluated in the usual sense are also welcome.

The recommended length of long article is 12–30 pages and of short paper is 6-15 pages.

The copyright of papers accepted for publication remains with the author. The editors reserve the right to make editorial revisions but these revisions and changes have to be approved by the author(s). Book reviews and short book notices are also appreciated.

The manuscripts are reviewed by 2 independent reviewers, at least one of them being a member of the international Editorial Board.

Authors receive two copies of the relevant issue of the PBML together with the original pdf files.

The guidelines for the technical shape of the contributions are found on the web site `http://ufal.mff.cuni.cz/pbml.html`. If there are any technical problems, please contact the editorial staff at `pbml@ufal.mff.cuni.cz`.