Slovene-Croatian Treebank Transfer Using Bilingual Lexicon Improves Croatian Dependency Parsing

Željko Agić*, Danijela Merkler, Daša Berović****

*Department of Information and Communication Sciences, **Department of Linguistics
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10000 Zagreb, Croatia
zagic@ffzg.hr, dmerkler@ffzg.hr, dberovic@ffzg.hr

Abstract

A method is presented for transferring dependency treebanks between similar languages by using a bilingual lexicon, aiming to improve dependency parsing accuracy on the target language. It is illustrated by transferring the Slovene Dependency Treebank to Croatian by using a GIZA++ bilingual lexicon constructed from the Croatian-Slovene 1984 parallel corpus from the Multext East project. The transferred treebank is merged with the Croatian Dependency Treebank and the merged treebank is used to train and test two graph-based dependency parsers. MSTParser and CroDep accuracy on parsing the 1984 fictional text shows a statistically significant increase and a similar decrease on parsing the Croatian Dependency Treebank newspaper text.

Slovensko-hrvaški prenos drevesnic z uporabo dvojezičnega leksikona izboljša odvisnostno razčlenjevanje hrvaščine

Prispevek predstavi metodo za prenos skladenjskih oznak korpusov med podobnimi jeziki z uporabo dvojezičnega leksikona, katere namen je izboljšati točnost odvisnostnega razčlenjevanja na ciljnem jeziku. Metodo ilustriramo s prenosom Slovenske odvisnostne drevesnice na hrvaški jezik z uporabo dvojezičnega leksikona, ki smo ga s programom GIZA++ izluščili iz vzporednega hrvaško-slovenskega korpusa 1984 projekta MULTEXT-East. Prenešena drevesnica je združena s Hrvaško odvisnostno drevesnico, združena drevesnica pa je nato uporabljena za učenje in testiranje dveh odvisnostnih razčlenjevalnikov, ki temeljita na teoriji grafov. Natančnost razčlenjevalnikov MSTParser in CroDep na leposlovnem delu 1984 pokaže statistično signifikantno izboljšanje in podobno zmanjšanje na razčlenjevanju Hrvaške odvisnostne drevesnice.

Keywords: treebank transfer, bilingual lexicon, dependency parsing

1. Introduction

Dependency treebanks are considered to be a sparse language resource. As stated in (Ambati and Chen, 2010), only a few languages in the world enjoy the status of resource-rich languages from the viewpoint of dependency treebanking, while tools and language resources supporting syntactic analysis in the framework of dependency syntax are unavailable or inadequate for many other languages. An illustration is given in (Zhao et al., 2009) that the treebanks of ten different languages from the CoNLL 2007 shared task on multilingual dependency parsing (Nivre et al., 2009) – restricted to a maximum of 500 thousand tokens per treebank – summed up to approximately 2 million tokens, 0.5 million of those being allocated by the Prague Dependency Treebank and certain treebanks accounting for not more than 30 thousand tokens or approximately 1.5 percent of the sum.

From this specific viewpoint, syntactically annotated corpora of Croatian and Slovene are considered to be small, while still sufficient to perform meaningful dependency parsing experiments from the viewpoint of the CoNLL 2006 and 2007 shared tasks. Croatian Dependency Treebank (Tadić, 2007) currently contains approximately 90 kw, while two dependency treebanks implementing two different models of dependency syntax are available for Slovene – the 30 kw Slovene Dependency Treebank (Džeroski et al., 2006) and the 100 kw JOS corpus (Erjavec et al., 2010). The first data-driven dependency parsing experiments for Slovene were conducted with the former one within the CoNLL 2006 shared task and the overall parsing accuracy score (LAS) of approximately 74% was observed, also showing that graph-based parsing methods significantly outperformed the transition-based parsing methods used in the experiments. Dependency parsing of Croatian texts by using the Croatian Dependency Treebank was thoroughly investigated just recently (Agić, 2012; Berović et al., 2012), showing a similar preference for graph-based over transition-based parsing (ca 74% vs. 71% LAS). Parsing accuracy within the data-driven graph-based dependency parsing framework was further increased by utilizing a k-best spanning tree parsing approach (Hall, 2007) with valency lexicon reranking (Agić, 2012), reaching an overall accuracy of approximately 78% LAS.

Croatian and Slovene are similar languages, i.e. they are both genetically and culturally close languages (Tadić, 2007) with small but usable dependency treebanks. Due to their similarity and resource availability, transferring a treebank from one language to another for purposes of improving dependency parsing accuracy is considered in this experiment. Treebank transfer is basically defined as "translating" a treebank from source language to target language while maintaining its syntactic annotation layer, effectively creating a syntactically annotated resource for the target language. Existing approaches mostly do not include

language similarity as a feature of significance for syntactic transfer and deal with generic approaches. These approaches include methods based on machine learning techniques (Jansche, 2005), word alignment and/or machine translation (Ambati and Chen, 2010) and parser delexicalization (Zeman and Resnik, 2008; Søgaard, 2011; McDonald et al., 2011). Lightweight machine-translation-related methods also exist and are mostly based on word-by-word transfer by using bilingual dictionaries of source and target languages (Zhao et al., 2009; Durrett et al., 2012). Relatedness of Croatian and Slovene in levels of linguistic description up to and including the syntactic level – both in terms of observed similarity and in terms of language resource compatibility – indicated that the computationaly inexpensive syntactic transfer method based on a Croatian-Slovene bilingual lexicon might improve dependency parsing scores. In the described experiment, Slovene was chosen as source language and Croatian as target language.
The following sections describe the resources and tools – parallel corpora, bilingual lexicon, treebanks and parsers – used in the experiment, the experiment preparation and its results in terms of observed dependency parsing accuracy in several test scenarios. Future work plans regarding treebank transfer and dependency parsing of Croatian are also briefly outlined.

2. Resources and tools

Treebank transfer from Slovene to Croatian by using a bilingual dictionary requires a dependency treebank of Slovene and a bilingual dictionary. Being that a dependency treebank for Croatian also exists, the transfer is implemented as a method for enlarging the Croatian treebank. To the best knowledge of the authors, a freely available Croatian-Slovene dictionary or bilingual lexicon is currently not available. Thus a bilingual lexicon was constructed for purposes of this experiment by using a freely available Croatian-Slovene parallel corpus. These resources are briefly described in the following section. Additionaly, the two graph-based parsers used in the experiment are also sketched.

2.1. Treebanks and other resources

The Croatian Dependency Treebank (HOBS) (Tadić, 2007) is a dependency treebank built along the principles of Functional Generative Description, as adapted in the Prague Dependency Treebank (Hajič et al., 2000). The ongoing construction of HOBS closely followed the guidelines set by the Prague Dependency Treebank, with their simultaneous adaptation to the specifics of the Croatian language. HOBS currently consists of 3,465 sentences in the form of dependency trees that were manually annotated with syntactic functions. These sentences, encompassing approximately 90,000 tokens, stem from the Croatia Weekly 100 kw corpus that is a part of the newspaper sub-corpus of the Croatian National Corpus. The Croatia Weekly sub-corpus was previously sentence-delimited, tokenized, lemmatized and MSD-annotated by linguists. Thus, each of the analyzed sentences contains the manually assigned information on part-of-speech, morphosyntactic category, lemma, dependency and syntactic function for each of the wordforms.

Sentences in HOBS are annotated according to the Prague Dependency Treebank syntactic annotation manual, with respect to differing properties of the Croatian language and consulting the Slovene Dependency Treebank (SDT) project (Džeroski et al., 2006). The syntactic functions utilized in HOBS are thus considered to be compatible with those used in SDT. The Slovene Dependency Treebank contains a part of the morphosyntactically annotated Slovene component of the parallel Multext East corpus (Erjavec, 2004), i.e. the first third of the Slovene translation of the novel 1984 by George Orwell, containing approximately 30,000 tokens in 2,000 sentences. Similar to HOBS, the SDT project was also based on the Prague Dependency Treebank – more precisely, development of HOBS stemmed from the experience of SDT in porting the Czech annotation rules to Slovene. With respect to this fact, the two treebanks can be considered to be highly compatible. The JOS syntactically annotated corpus (Erjavec et al., 2010) of aproximately 100,000 tokens in 6,100 sentences utilized a different syntactic annotation and was thus not used in this experiment. However, a mapping between the JOS annotation and the PDT-style annotation is possible.

As noted previously, no Croatian-Slovene dictionary-like resources were readily available and an approach with automatic construction of a bilingual lexicon from a parallel corpus was implemented here with respect to that fact. Two parallel corpora for the Croatian-Slovene language pair were usable when conducting the experiment – the fully completed 1984 parallel corpus from the Multext East project and the Croatian-Slovene Parallel Corpus in its early development state (Požgaj Hadži and Tadić, 2000). Being that the latter one is still in development, is not entirely document-, sentence- or word-aligned and differs in domain from the source treebank (i.e. SDT), the Croatian-Slovene subset from the 1984 parallel corpus was chosen for bilingual lexicon construction. The corpus was sentence-aligned using hunalign (Varga et al., 2005) in realignment mode, keeping only 1:1 sentence alignments. The resulting set of Croatian-Slovene sentences contained 6,337 sentence pairs and 210,948 tokens. Basic stats for this resource and the treebanks are given in table 1.

The dictionary was constructed from the sentence pairs using GIZA++ (Och and Ney, 2003). It contained 52,502 Slovene-Croatian word pairs for 16,432 different Slovene word forms that translated into 17,368 different Croatian word forms. The entries (i.e. word pairs) were sorted by translation probability obtained from the parallel corpus, respecting the GIZA++ format.

feature	HOBS	SDT	hr-1984	si-1984
sentences	3,465	1,997	6,337	6,337
tokens	88,045	36,554	101,774	102,837
MSD tags	828	789	802	1039
syntactic tags	69	69	N/A	N/A

Table 1: Basic stats for used corpora

2.2. Dependency parsers

Two graph-based dependency parsers were selected to be used in the experiment – MSTParser and CroDep. The selection was based on previous experiments with Croatian dependency parsing (Agić, 2012) that showed a strong preference towards graph-based, rather than transition-based dependency parsing of Croatian texts.

MSTParser (McDonald et al., 2006) is a state-of-the-art graph-based dependency parser generator with first and second order arc-factored language models, perceptron learning algorithm and both projective and non-projective parsing algorithms. It was used in this experiment to generate second order arc-factored non-projective parsers for Croatian. This MSTParser configuration was previously shown (Agić, 2012) to obtain the highest parsing accuracy on HOBS among all the tested standalone parsers (approximately 74.53% LAS).

CroDep is a novel k-best maximum spanning tree dependency parser with valency lexicon reranking (Agić, 2012), design specifically to increase the accuracy of parsing Croatian texts by utilizing a valency lexicon of Croatian verbs CROVALLEX (Mikelić Preradović, 2008; Mikelić Preradović et al., 2009). It is based on the k-MST parser (Hall, 2007) in that it produces a number of candidate dependency trees for an input sentence, sorted by confidence, and these candidate trees are then reranked by a rule-based reranking module that uses CROVALLEX as a knowledge base. Specifically, every dependency relation that attaches to a verb is assigned an additional weight, which is in turn decided by matching its properties with the constraints and requirements stated in CROVALLEX for that specific verb entry. Sums of these additional weights are assigned to the candidate trees and they are reranked by consulting both the ranking list of the k-best parser and the ranking list of the CROVALLEX-based module. The parser was tested on HOBS and achieved a parsing score of approximately 77.21% LAS, i.e. significantly better than the previously top-performing standalone MSTParser. As previously indicated, CroDep currently implements a first order arc-factored language model with perceptron learning, a k-best maximum spanning tree algorithm similar to the one implemented in k-MST and a valency lexicon reranker. It can be (made to be) used independently of the input language, as long as a verb valency lexicon is available for that language. The parser is currently a prototype and will be made publically available as soon as it leaves the early development stage and is tested on a certain number of languages meeting the stated requirements.

3. Experiment and results

3.1. Experiment setup

The experiment was basically envisioned in three stages. Firstly, SDT is "translated" to Croatian by simple word-to-word mapping with the Croatian-Slovene bilingual lexicon. Secondly, the translated resource (henceforth called hr-SDT) is assigned the Croatian metadata required for training the parsers, i.e. lemmas and morphosyntactic tags. Thirdly, parsers are trained and tested on manually dependency parsed Croatian texts. The translation of SDT was done at unigram level, i.e. by mapping each of the Slovene tokens to a respective Croatian token from the bilingual dictionary. Only the pairs with highest probabilities attached by GIZA++ were chosen from the dictionary. The resulting hr-SDT treebank therefore contained the same number of sentences and tokens as the original SDT (1,997 sentences and 36,554 tokens, see table 1) and the same syntactic features. A total of 29,344 token and 25,786 lemma replacements occured by consulting the lexicon. The transfer process is illustrated by Figure 1.

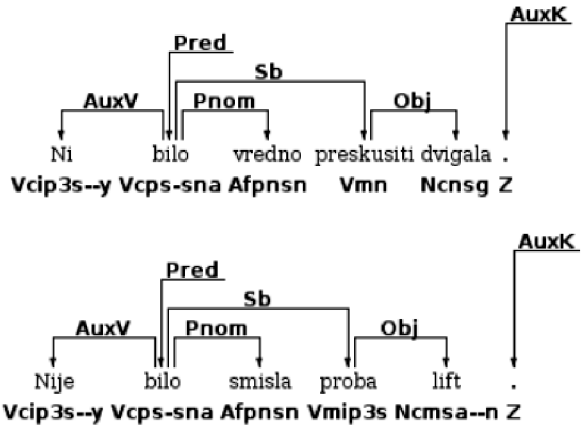

Figure 1: Dependency tree transfer example

For assessing the translation quality, 100 sentences were randomly selected from hr-SDT, paired with respective sentences from SDT and Croatian 1984 corpus and manually evaluated for adequacy and fluency on 1-5 scale. Translation adequacy was scored at approximately 3.64 and fluency at 2.99.

Two courses of action were taken with respect to metadata assignment. In the first one, Slovene lemmas were also translated to Croatian via bilingual lexicon and Slovene MSD tags were held. In the second one, hr-SDT was lemmatized and MSD-tagged using the CroTag HMM tagger and lemmatizer (Agić et al., 2008; Agić et al., 2009) trained on the Croatian 1984 corpus. Croatian MSD tags frequently differed from Slovene (14,216 occurences), but very infrequently in part-of-speech information. The tagging accuracy can be estimated at approximately 85% on basis of previous experiments (Agić et al., 2008).

Training sets for the parsers were created by attaching hr-SDT to training sets created from HOBS. As described in detail in (Agić, 2012) and according to CoNLL 2006 and 2007 shared task rules, 10 disjoint testing sets of approximately 5,000 tokens were extracted from HOBS, leaving 10 disjoint training sets for creating language models, approximately 83,000 tokens each. Each of these training sets was merged with both versions of hr-SDT, i.e. the one with Slovene MSDs and the one with lemmas and MSDs assigned by the CroTag tagger. This resulted in two batches of 10 training sets to be used in training MSTParser and CroDep. Tenfold cross-validated testing was to be done on both HOBS and the Croatian 1984 corpus. As the latter one is not syntactically annotated, we created a test set by man-

ually annotating 345 sentences and 5,226 tokens from the corpus respecting the HOBS and SDT standard. Results of a previous experiment (Agić, 2012) with parsing on HOBS were used here as a reference point. Labelled attachment score (LAS) was observed.

3.2. Results

The obtained results are displayed in table 2. They can be observed from several viewpoints.

Firstly and most importantly, the results indicate that the usefulness of treebank transfer is domain-dependent in this specific experiment. More specifically, introducing variants of hr-SDT – respecting the fact that SDT is a corpus of fictional text – to the HOBS training set decreases the overall parsing accuracy on parsing newspaper texts from HOBS by 0.53 and 0.57% LAS for MSTParser and by 0.32 and 0.44% LAS for CroDep, using MSD-tagged and untagged hr-SDT, respectively. The negative influence of introducing hr-SDT to HOBS changes to positive when parsing the hr-1984 test set of fictional text. The observed improvements over the baseline are 0.93 and 1.18% LAS for MSTParser and 0.89 and 1.11% LAS for CroDep. Parser language models obviously benefit from the quantity of data in HOBS when parsing fictional text, while the introduction of hr-SDT diverts the models from the properties of sentences in newspaper text.

At this point, it might be argued that using the 1984 Croatian-Slovene parallel corpus to construct a bilingual lexicon and to facilitate syntactic transfer introduces a bias with respect to the obtained results, being that the parsed text originates from the same source. However, the bias is here considered to be accounted for by the small size of the parallel corpus and the resulting bilingual lexicon and the resulting adequacy and fluency of translated text. Moreover, as discussed previously, other Croatian-Slovene parallel resources were not available at the time of conducting the experiment and thus using the 1984 parallel corpus was not a matter of choice.

Secondly, the observed decrease in parsing accuracy when shifting from newspaper to fictional text is substantial. Top scores for CroDep on these domains differ by 4.73% LAS in favor of newspaper text, while this difference amounts to 4.84% LAS for MSTParser. Treebank transfer benefits from lemmatization and MSD-tagging in all test scenarios. However, the observed difference between parsing accuracy when using tagger-assigned as opposed to transferred morphosyntactic tags is not shown to be statistically significant here.

Finally, The top-scoring parser on both fictional and newspaper text is CroDep. Its average difference over MSTParser is approximately 2.71% LAS across domains. It scored 72.48% LAS on hr-1984 and 77.21% LS on HOBS, topping MSTParser by 2.78% LAS and 2.68% LAS on these two test samples. UAS and LA metric were also used in the experiment and were shown to closely follow the pattern displayed by the LAS metric and were therefore excluded as they are less informative.

Test set	Model	MST	CroDep
hr-1984	HOBS	68.51	71.37
	HOBS + hr-SDT	69.44	72.26
	HOBS + hr-SDT tagged	69.69	72.48
HOBS	HOBS	74.53	77.21
	HOBS + hr-SDT	73.96	76.77
	HOBS + hr-SDT tagged	74.00	76.89

Table 2: LAS for MSTParser and CroDep hr-SDT language models on hr-1984 and HOBS

4. Conclusions and future work

Using Croatian Dependency Treebank, Slovene Dependency Treebank, Croatian-Slovene parallel resources and existing dependency parser generators, this experiment has shown that treebank transfer between similar languages by using a bilingual lexicon improves dependency parsing accuracy for the target language. It was also experimentally shown that the observed improvement is domain-dependent. Parsing accuracy peaked for the hybrid dependency parser CroDep at 72.48% LAS on fictional text and 77.21% LAS on newspaper text.

Future work in Slovene-Croatian treebank transfer will be targeted to several directions. Domain-specific bilingual lexica might introduce a positive bias for in-domain parsing. One such lexicon could be constructed from the Croatian-Slovene parallel corpus even in its early development stage, e.g. by using parallel sentence extractors such as LEXACC (Stefanescu et al., 2012), that operate on comparable corpora. The issue of bilingual lexica might also be addressed by using English as interlingua. Regarding bilingual lexica and transfer, a more elaborate approach to machine translation could be implemented along the lines of (Zhao et al., 2009) by using probabilistic word-by-word decoding to obtain translations of higher quality. The experiment presented here could also be repeated by setting Croatian as source and Slovene as target language. The syntactic annotation of the JOS corpus could also be mapped to SDT style and vice versa, as well as HOBS, providing an even larger resource for syntactic transfer. Moreover, the method could also be tested on other language pairs with compatible treebanks, e.g. Czech-Slovene and Czech-Croatian, even though syntactic transfer via bilingual lexica or statistical machine translation methods might pose a challenge with respect to availability of parallel corpora for these language pairs.

5. Acknowledgements

The authors would like to thank the anonymous reviewers for their valuable comments.

The results presented here were partially obtained from research within project CESAR (ICT-PSP, grant 271022) funded by the European Commission, and partially from projects 130-1300646-0645 and 130-1300646-1776 funded by the Ministry of Science, Education and Sports of the Republic of Croatia.

6. References

Agić Ž, Tadić M, Dovedan Z. 2008. Improving Part-of-Speech Tagging Accuracy for Croatian by Morphological Analysis. *Informatica*, 32(4), 2008.

Agić Ž, Tadić M, Dovedan Z. 2009. Evaluating Full Lemmatization of Croatian Texts. *Recent Advances in Intelligent Information Systems*, Warsaw, Academic Publishing House EXIT, 2009, pp. 175–184.

Agić Ž. 2012. *Pristupi ovisnosnom parsanju hrvatskih tekstova*. PhD thesis. University of Zagreb, Faculty of Humanities and Social Sciences, 2012.

Ambati V, Chen W. 2010. Cross Lingual Syntax Projection for Resource-Poor Languages.

Berović D, Agić Ž, Tadić M. 2012. Croatian Dependency Treebank: Recent Development and Initial Experiments. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, Istanbul, 2012.

Durrett G, Pauls A, Klein D. 2012. Syntactic Transfer Using a Bilingual Lexicon. In *Proceedings of the 2012 EMNLP-CoNLL*.

Džeroski S, Erjavec T, Ledinek N, Pajas P, Žabokrtský Z, Žele A. 2006. Towards a Slovene Dependency Treebank. In *Proceedings of Fifth International Conference on Language Resources and Evaluation*.

Erjavec T. 2004. Multext-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*.

Erjavec T, Fišer D, Krek S, Ledinek N. 2010. The JOS Linguistically Tagged Corpus of Slovene. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*.

Hajič J, Böhmová A, Hajičová E, Vidová Hladká B. 2000. The Prague Dependency Treebank: A Three-Level Annotation Scenario. *Treebanks: Building and Using Parsed Corpora*, Kluwer, 2000.

Hall K. 2007. k-Best Spanning Tree Parsing. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*.

Jansche M. 2005. Treebank Transfer. In *Proceedings of the IWPT 2009*.

McDonald R, Lerman K, Pereira F. 2006. Multilingual Dependency Parsing with a Two-Stage Discriminative Parser. In *Proceedings of CoNLL-X*.

McDonald R, Petrov S, Hall K. 2011. Multi-Source Transfer of Delexicalized Dependency Parsers. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*.

Mikelić Preradović N. 2008. *Pristupi izradi strojnog tezaurusa za hrvatski jezik*. PhD thesis, University of Zagreb, Faculty of Humanities and Social Sciences, 2008.

Mikelić Preradović N, Boras D, Kišiček S. 2009. CROVALLEX: Croatian Verb Valence Lexicon. In *Proceedings of the ITI 2009 — 31st International Conference on Information Technology Interfaces*, SRCE, Zagreb, 2009, pp. 533–538.

Nivre J, Hall J, Kübler S, McDonald R, Nilsson J, Riedel S, Yuret D. 2007. The CoNLL 2007 Shared Task on Dependency Parsing. In *Proceedings of the CoNLL Shared Task of EMNLP-CoNLL 2007*.

Och F J, Ney H. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1), pp. 19–51.

Požgaj Hadži V, Tadić M. 2000. Hrvatsko-slovenski paralelni korpus. In *Proceedings of the Language Technologies Conference*, Jožef Stefan Institute, Ljubljana, 2000.

Søgaard A. 2008. Data Point Selection for Cross-Language Adaptation of Dependency Parsers. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*.

Stefanescu D, Ion R, Hunsicker S. 2012. Hybrid Parallel Sentence Mining from Comparable Corpora. *Proceedings of the 16th EAMT Conference*, pp. 137–144.

Tadić M. 2007. Building the Croatian Dependency Treebank: the initial stages. *Suvremena lingvistika*, 63.

Varga D, Németh L, Halácsy P, Kornai A, Trón V, Nagy V. 2005. Parallel Corpora for Medium Density Languages. *Proceedings of the RANLP 2005*, pp. 590–596.

Zeman D, Resnik P. 2008. Cross-Language Parser Adaptation between Related Languages. In *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*, pp. 35–42.

Zhao H, Song Y, Kit C, Zhou G. 2009. Cross Language Dependency Parsing using a Bilingual Lexicon. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*.