Cross-Lingual Projection of LFG F-Structures: Resource Induction for Polish

Alina Tokarczyk and **Anette Frank**
Department of Computational Linguistics, University of Heidelberg, Germany
{tokarczy, frank}@cl.uni-heidelberg.de

Natural language processing has made rapid progress over the last decades. Yet, computational linguistic resources and tools are restricted to a handful of languages. It seems unrealistic to develop high-quality resources for all languages using traditional methods. Especially the creation of grammars and syntactic treebanks is an expensive process. Various methods aim at overcoming the shortage of NLP resources. One approach that is pursued in this paper targets the induction of linguistic annotations in a cross-linguistic setting: Using a bilingual corpus, existing analysis tools are applied to the resource-rich language side of the bitext. The resulting annotations are projected to the second, resource-poor language using automatic word alignments as a bridge. The projection approach for resource induction is built on the assumption that linguistic analysis of a sentence carries over to its translation in an aligned parallel corpus. While this assumption does not hold uniformly, the projected annotations can be used to train NLP tools for the target language. This has been shown for PoS tagging [1], NP-bracketing [1], dependency analysis [2, 3], word sense disambiguation [4], extraction of semantic roles [5] and temporal labelling [6].

Within the ParGram project [7], grammars for English, French, German, Norwegian, Japanese, Urdu and other languages are written according to the framework of Lexical Functional Grammar, using XLE as a processing platform [8]. Manual development of large-scale LFG grammars is an expensive process that may be sped up by automation techniques. One strand of work that targets the automatic induction of LFG grammars is the induction from existing syntactic treebanks [9]. However, this method rests on the availability of high-quality treebanks. To overcome the need of treebank creation, we investigate the cross-lingual projection approach to induce syntactically annotated corpora for new languages. Given the considerable divergence of constituent structures across languages, the grammar architecture of LFG, with its strong lexicon component and multiple levels of representations seems especially suited for a cross-lingual grammar induction task. F-structures are largely invariant across languages, and are thus especially suited to serve as a pivot for cross-lingual syntactic annotation projection. Following this insight, we pursue cross-lingual projection of grammatical functions (GFs) to induce an f-structure bank for Polish. We project English f-structures to aligned Polish texts, to yield an f-structure bank that may be used to train a dependency parser for Polish. A full-fledged LFG grammar for Polish may be obtained by mapping the induced f-structures to c-structures for Polish.

Our approach follows [2] and adapts their method of projecting dependencies (from English to Spanish/ Chinese) to the LFG framework. The experiment is conducted on the JRC-Acquis Multilingual Parallel Corpus [10], a large collection of European Union legislative texts that – unlike Europarl – includes texts in Polish. We create a subcorpus aligned on the sentence level containing 257,144 sentence pairs. Word alignment is provided by the statistical machine translation system MOSES [11], based on statistics captured from the entire corpus. In a pilot experiment we randomly selected a training corpus of 6000 sentence pairs, development and test corpora with 1000 sentences each. Filtering of duplicates and sentences with problematic tokenization yielded a development corpus (484 sents, 13.87 tokens/s), training corpus (1918 s, 14.13 t/s) and test corpus (486 s, 13.52 t/s). The English sides are parsed using the English Pargram LFG grammar, with selection of the most probable analysis. To a certain degree English is an isolating language that makes use of function and non-content words. Polish, by contrast, is a highly inflecting language and needs in general fewer or as many words as English to express the same content. In order to decide whether alignment of one Polish word with one or more English words conforms to general translational mappings, we use unidirectional Polish-English word alignment as a basis for projection. For projection, the GF *gf* encoded in the source f-structure fs_e is transferred to the target sentence using word alignment links *al* ($al \in AL: E \times F$). The set *GF* ($gf \in GF: E \times E$) consists of the GFs holding between two English words $gf(e_i, e_j)$ and are denoted by SUBJ, OBJ, OBL, ADJ etc. Following [2], the projection of GFs is defined as follows:

> (1) The GF $gf(e_i, e_j)$ in the source f-structure fs_e projects to the target f-structure fs_f as $gf(f_i, f_j)$ if and only if the source terminals te_i and te_j that project to the PRED values e_i, e_j included in the fs_e are aligned with the target terminals tf_i and tf_j of f_i and f_j, respectively.

(1) states that if two English words are related by a GF, the same GF will relate their word-counterparts in Polish. The GFs obtained from LFG parsing that satisfy the conditions in (1) will thus be transferred to the corresponding Polish sentence via word alignment links. The projected GFs may be incorrect, due to (i) errors in the source annotations obtained from automatic LFG parsing, (ii) poor accuracy of word alignment, or (iii) true mismatches of functional structure between English and Polish. The annotation and alignment errors radically impair the quality of the projected GFs. Those shortages can be overcome by applying correction rules similar to [2] that locally transform the induced Polish f-structures. We defined two post-projection correction rules that are motivated by general linguistic properties of Polish such as (i) absence of articles that could correspond to specifiers (SPEC-DET) 'the' or 'a/an' in English, and (ii) possessive relations expressed by genitive marking of NP as opposed to an *of*-PP in English. Further correction rules may be formulated, taking into account morpho-syntactic information concerning case, number, tense etc.

We evaluate the quality of the automatically induced f-structures against a gold standard of 50 Polish f-structures from the test corpus which were manually corrected (11.98 t/s for English, 9.76 t/s for Polish). The f-structures were transformed to the SALSA/TIGER XML format [12], which enables efficient modification by adding, deleting or correcting GFs, using the SALTO annotation tool. We calculated (i) precision, recall and f-score for exact

match of projected GFs, distinguishing direct projection and projection with post-modification using the two correction rules mentioned above, and (ii) accuracy of the projected dependencies taking into account automatically derived (noisy) in contrast to hand-corrected (optimal) word alignments, to establish an upper bound (Table 1, left). As expected, direct projection of GFs is noisy (49.98% f-score). Application of language-specific transformation rules considerably improves the accuracy of projected GFs (63.5% f-score). The quality of word alignment is a crucial factor for the quality of projection: projection based on corrected word alignments enhances the quality of the induced f-structures by 12 percentage points (pp) f-score for direct projection and by 19.45 pp f-score for projection with transformation rules. In line with [2] (cf. Table 1, second left), these results clearly indicate that direct projection of GFs is significantly outperformed by projection using post-projection transformations, both for automatic alignment (13.52 pp f-score improvement), and for manually corrected word alignments, yielding an upper bound (20.97 pp f-score improvement). The upper bound (projection based on perfect word alignment) in conjunction with two correction rules achieves an accuracy of 82.95% f-score. According to [10], the quality of word alignment can still be improved using a factored alignment model using part-of-speech, lemma and morphological information. Thus, if both language sides can be enriched with word level information, we can expect increases of word alignment, and thus projection quality, within the ranges of the upper bound, possibly enhanced by further post-transformation rules. Compared to the results of unlabeled dependency projection for Spanish/Chinese in [2], we gain higher f-score (on average by 16.4 pp). Compared to [3] (Table 1, 2nd right), precision of direct projection is lower by 18/32 pp. [3] relies on one-to-one alignment links (intersection) only, which increases precision but decreases recall. As [3] does not report recall figures, we cannot compare the results. As our aim is to induce f-structures as complete as possible, it is worth noting that we obtain balanced precision and recall values. [13] represents an LFG-based approach, like ours (Table 1, right). We observe that combining argument information from two languages (English and German) enhances precision but degrades recall. Considering f-score, our projection of GFs via automatic word alignment outperforms the projection by [13] by 16.7 pp. However, due to the different languages and corpora considered, full comparison of the approaches is difficult.

In future work, we will explore extended training sets, improved word alignments, and use of morpho-syntactic information to further improve the projection quality. The resulting Polish f-structure treebank may be used to train a dependency parser (in [2], a dependency parser for Spanish/Chinese could be trained to yield 72.1%/52.4% labeling f-score, using projected dependency f-scores of 72.1%/53.9% as input for training). We will further explore induction of a full-fledged LFG grammar for Polish, by adding a module that learns f-to-c-structure mappings for Polish.

	Projection of GFs – current experiment				Hwa et al. (2005)				Ozdowska	Bouma et al. (2008)		
Languages	en-pl				en-sp		en-ch		en-pl	ge-du	en-du	(ge,en)-du
sents (corpus)	50 (JRC-Acquis)				100 (UN/FBIS/Bible)		88 (FBIS)		50 (AC-corpus)	222 (Europarl)		
	direct	+corr	direct	+corr	direct	+corr	direct	+corr	direct proj	direct projection		
Alignment	automatic		manual		automatic		manual		automatic	automatic		
LP/ULP	49/50	64/64	61/62	85/85					67/82	52.2	54.3	74.6
LR/ULR	51/52	63/63	63/63	81/82						52.9	48.8	34.1
F-score/ UF-score	50/51	63.5/63.5	62/62.5	83/83.5	/33.9 /26.3	/65.7 /52.4	/36.8 /38.1	/70.3 /67.3		52.6	51.4	46.8

Table 1: LP/ULP: precision for labeled/unlabeled GFs; LR/ULR: recall for labeled/unlabeled GFs; +corr: direct proj. plus correction

References

[1] Yarowsky, D. and Ngai, G. (2001). Inducing Multilingual POS Taggers and NP Bracketers via Robust Projection across Aligned Corpora. In: *Preceedings of NAACL-2001*, pp. 200-207.

[2] Hwa, R., Resnik, P., Weinberg, A., Cabezas, C. and Kolak, O. (2005). Bootstrapping parsers via syntactic projection across parallel texts. In: *Natural Language Engineering* 11 (3). Cambridge, pp. 311-325.

[3] Ozdowska, S. (2006). Projecting POS tags and syntactic dependencies from English and French to Polish in aligned corpora. In: *EACL 2006 Workshop on Cross- Language Knowledge Induction.*

[4] Diab, M. and Resnik, P. (2002). An Unsupervised Method for Word Sense Tagging using Parallel Corpora. In: *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics*, Philadelphia.

[5] Padó, S. and Lapata, M. (2005)**.** Cross-linguistic projection of role-semantic information**.** In: *Proc. of HLT/EMNLP 2005*.

[6] Spreyer, K. (2007). Projecting Temporal Annotations Across Languages. Diploma thesis. Saarbrücken.

[7] Butt, M., Dyvik, H., King, T.H., Masuichi, H., and Rohrer Ch. (2002). The Parallel Grammar Project. The parallel grammar project. In: *Proceedings of COLING 2002. Workshop on Grammar Engineering and Evaluation*, pp. 1-7.

[8] Crouch, D., Dalrymple, M., Kaplan, R., King, T., Maxwell, J. and Newman, P. (2007). XLE Documentation. PARC.

[9] Cahill, A., Forst, M., Burke, M., McCarthy, M., O'Donovan, R., Rohrer, C., van Genabith, J., Way, A. (2005): Treebank-Based Acquisition of Multilingual Unification Grammar Resources. In: *J. of Research on Language and Computation 3(2)*, 247-279.

[10] Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D. and Varga, D. (2006). The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In: *Proceedings of the LREC 2006.* Genoa.

[11] Koehn, P., Federico, M., Shen, W., Bertoldi, N., Bojar, O., Callison-Burch, Ch., Cowan, B., Dyer, Ch., Hoang, H., Zens, R., Constantin, A., Moran, Ch.C.,Herbst, E. (2007). Moses: Open Source Toolkit for Statistical Machine Translation**.** In: *ACL 2007*.

[12] Erk, K. and Padó, S. (2004). A powerful and versatile XML format for representing role-semantic annotation. In: *LREC 2004.*

[13] Bouma, G., Kuhn, J., Schrader, B., and Spreyer, K. (2008): Parallel LFG grammars on parallel corpora: A base for practical triangulation. In Butt, M. and King, T.-H. (eds): *Proceedings of LFG 2008*, Sydney.