



method for generating c-structures from surface-indexed f-structures, as they are produced by cross-lingual f-structure projection. In particular, we investigate the possibility of inducing underlying c-structures from projected f-structures in a *linguistically perspicuous way*. In the spirit of general c-structure principles [1] we define principle-based c-structure and c-to-f-structure projection rules that permit a wide range of parametric variation to account for multiple languages.

In a cross-lingual projection scenario, the induction of c-structure from surface-indexed f-structures reduces to a controlled generation problem: Given a target language f-structure, its surface string, and a set of general c-structure rules with f-structure annotations, we can generate surface strings from the projected f-structures. The generated structures whose yields match the surface order of the aligned target sentence are candidate c-structures for a full-fledged c- and f-structure analysis for the target sentence. Alternatively, we can analyse the target strings using general c-structure rules and the f-structure-indexed words as lexicon entries, and record those analyses that result in the target f-structures that were created in the cross-lingual projection step.

We will present a small-scale experiment that is intended as a proof of concept for this grammar induction architecture. We are using two languages (English and German) and a restricted – but representative – set of core grammatical constructions (including complementation, modification, determination, extraction, coordination, etc.) as defined by a controlled test suite of 100 sentence pairs. In order to investigate the feasibility of the c-structure induction task proper, we restrict ourselves to translation pairs that conform to the DCH.

We make use of a core (multi-lingual) LFG grammar fragment, defined using the XLE platform, that utilises a range of abstraction and generalisation devices, such as macros for grammatical function names and constituents. The grammar defines general c-structure rules with functional projections that implement general c-/f-structure projection principles, using functional (C,I,D) and lexical categories, and intermediate projection levels for specification, complementation and adjunction structures, following [1]. A grammar excerpt is given below (with GF a variable over permissible grammatical functions, and XP a macro for maximal c-structure constituents). Note that we can use the shuffle notation (“,”) in rules  $M \rightarrow D_1, \dots, D_n$  that permits all possible orders for the daughters of a rule, in order to account for parametric variation of c-structure rules across languages. Alternatively, we may pre-define known serialisation properties of a target language in a specific language projection task.

<p>CP --&gt; (XP: { (^ SUBJ) = !    (^ TOPIC) = ! (^ GF+) = !  (^ FOCUS) = ! (^ GF+) = !}),  C': (^ PRED).  C' --&gt; (C),  (IP).</p>	<p>IP --&gt; (XP: { (^ SUBJ) = !    (^ TOPIC) = ! (^ GF+) = !    (^ FOCUS) = ! (^ GF+) = !}),  (I').  I' --&gt; (Io),  (VP).</p>
---	--

We pay particular attention to the amount of overgeneration in c-structure generation. Given that the analyses are constrained by a surface-indexed f-structures, we primarily observe spurious attachment ambiguities. These we intend to control by way of the principle of economy of expression ([1]), which can be implemented using XLE’s OT-based preference mechanism ([4]). We will present the grammar and results for the testsuite of 100 sentences, with detailed evaluation and discussion of specific types of syntactic phenomena.

## References

- [1] J. Bresnan. *Lexical-Functional Syntax*. Blackwell Publishers, Oxford, 2001.
- [2] M. Burke, A. Cahill, R. O’Donovan, J. van Genabith, and A. Way. Treebank-based acquisition of wide-coverage, probabilistic LFG resources: Project overview, results and evaluation. In *Proceedings of IJCNLP*, 2004.
- [3] A. Frank. Automatic F-structure Annotation of Treebank Trees. In M. Butt and T.H. King, editors, *Proceedings of the LFG 2000 Conference*, Berkeley, CSLI Publications, Stanford, 2000.
- [4] A. Frank. A (discourse) functional analysis of asymmetric coordination. In M. Butt and T.H. King, editors, *Proceedings of the LFG 2002 Conference*, pages 174–196, Athens, CSLI Publications, 2002.
- [5] Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003.
- [6] S. Padó. *Cross-lingual Annotation Projection Models for Role-Semantic Information*. PhD thesis, Saarland University, 2007.
- [7] P. Resnik R. Hwa, A. Weinberg, C. Cabezas, and O. Kolak. Bootstrapping parsers via syntactic projection across languages. *Natural Language Engineering*, 11(3):311–325., 2004.
- [8] L. Sadler, J. van Genabith, and A. Way. Automatic F-Structure Annotation from the AP Treebank. In M. Butt and T.H. King, editors, *Proceedings of the LFG 2000 Conference*, Berkeley, CSLI Publications, 2000.
- [9] K. Spreyer and A. Frank. Projection-based acquisition of a temporal labeller. In *Proceedings of IJCNLP*, 2008.
- [10] A. Tokarczyk. F-structure projection using word-aligned corpora. Softwareproject, University of Heidelberg, 2008.
- [11] J. van Genabith, A. Way, and L. Sadler. Semi-automatic generation of f-structures from tree banks. In M. Butt and T.H. King, editors, *Proceedings of the LFG99 Conference*, Manchester, CSLI Publications, 1999.
- [12] D. Yarowsky and G. Ngai. Inducing multilingual PoS taggers and NP bracketers via robust projection across aligned corpora. In *Proceedings of NAACL 2001*, pages 200–207, 2001.