Creating a Norwegian valence corpus from a deep grammar
DEMO

Lars Hellan Dorothee Beermann Tore Bruland Tormod Haugland Elias Aamot

NTNU, 7491 Trondheim, Norway
lars.hellan@ntnu.no dorothee.beermann@ntnu.no t-brul@online.no tormod.haugland@gmail.com
elias.aamot@gmail.com

Keywords: interoperability of resources, valence, interlinear glossed text, automatic generation of valence corpus, TypeCraft, Leipzig Corpus Collection, Norsource.

The paper underlying this DEMO presents a procedure for generating a valence corpus of Norwegian (Bokmål) from a deep grammar using the Leipzig Corpus Collection (LCC). The corpus is presented in the form of IGT (interlinear glossed text) augmented by valence information. As our deep parser we use the computational grammar Norsource (Hellan and Bruland 2015) while our online IGT repository is TypeCraft (Beermann and Mihaylov (2014)). The HPSG-based grammar *NorSource*[1] has about 85,000 lexical entries, 250 syntactic rules, 40 lexical rules for derivation and inflection (Norsource having a lemma-based lexicon), and about 400 valence frame types, covering nearly all morpho-syntactic construction types in the language, with lexical specifications and selection defined for about 12000 verb entries, with a well developed system of argument structure. Most of these aspects of information are carried over to the IGT corpus, as illustrated in Figure 1,[2] which shows a sample schema of augmented IGT in the TypeCraft[3] editing interface. Here the valence frame instantiated by each occurring verb is shown underneath the standard IGT, using three different types of code for marking valence frame ((‘SAS’ for ‘Syntactic Argument Structure’, ‘FCT’ for ‘functional label’, using alternative codes for the same content):

String: Jeg vet at hun forbauset Ola
Free translation: I know that she surprised Ola

Jeg	|vet vite	|at	|hun	|forbause forbause	|t	|Ola
1.SG.NOM	|PRES	|DECL	|3.SG.FEM	|	|PAST	
PN	|V	|COMP	|PN	|V		|Np

vet:
SAS: NP+Sdecl
FCT: transWithSentCompl
ConstructionLabel: v-tr-obDECL

forbauset:
SAS: NP+NP
FCT: transitive
ConstructionLabel: v-tr

Figure 1 Sample schema of augmented IGT in the TypeCraft editing interface

The purpose of the project is twofold: (i) to create a human- and machine readable IGT valence corpus with working access to a larger group of users, including linguists and language experts, and more generally to increase the usability of each of our tools by aligning them in a useful way; (ii) to create datasets pairing valence codes with in-depth annotation of morpho-syntactic information as is typical for IGT data. This will allow for studies into general correspondences between word level and sentence level grammatical information on a large scale. The focus of the present paper is on the data import from Norsource to TypeCraft. The work we report on here covers 22 000 random sentences from the Norwegian corpus hosted at the *Leipzig Corpus Collection,*[4] where the main task is to develop a conversion algorithm that creates TypeCraft’s IGT-XML from Norsource XML, being able to identify the relevant grammatical information which, while systematic, is nevertheless widely dispersed over the Norsource XML.
The DEMO will show various aspects of the algorithm and stages in the conversion, as well as the search interface and examples of investigations which can be supported by the corpus.

[1] Cf. Hellan and Bruland (2015). Online access, for description: http://typecraft.org/tc2wiki/Norwegian_HPSG_grammar_NorSource . Demo: http://regdili.hf.ntnu.no:8081/linguisticAce/parse.
[2] See https://typecraft.org/tc2wiki/Norwegian_Valency_Corpus.
[3] https://typecraft.org
[4] http://asvdoku.informatik.uni-leipzig.de/corpora/