The grammar *NorSource*

Lars Hellan

Norwegian University of Science and Technology (NTNU)
N-7491 Trondheim, Norway{lars.hellan@ntnu.no

Research Group in Digital Linguistics

PARSEME meeting, Athens 2014

NorSource (‘Norwegian HPSG Resource Grammar’) is a computational typed feature structure grammar of Norwegian built on the *LKB platform*, following the general format of the *HPSG Grammar Matrix*, the adopted format of the *DELPH-IN consortium*.
Parse outputs include semantic representations using the *Minimal Recursion Semantics* (MRS) format. The grammar both parses and generates (from the MRS representation).
The grammar currently uses three providing systems: LKB (for development), PET and (as of spring 2012) ACE for processing.

The grammar was started in 2001, by a group versed in Generative Grammar since the late 60ies, and formal semantics (‘Montague Grammar’) since the mid 70ies; from the mid 80ies the group developed a computational lexicon (under the acronym ‘TROLL’, see Hellan et al. 1989), mainly associated with research within ‘consolidated GB’. In the late 90ies the group reoriented itself towards HPSG, and started the grammar as part of the LinGO initiative with the LKB platform, the first grammar to be built on the Matrix, during the EU-project *DeepThought* (2002-4).

The grammar contains appx 5600 types, 250 syntactic rules, 40 inflectional rules, and 20 derivational rules. The lexicon is lemma-based, with about 83 000 lexical entries, thereof 12,500 verb entries, distributed over 350 verb lexeme types.

We can distinguish four main phases in the grammar's development:

- Phase 1, the *Grounding* phase (2001-04),
- Phase 2, the *Semantic Expansion* phase (2005-07),
- Phase 3, the *Cross-Linguistic Coding* phase (2008-10),
- Phase 4, the *Interoperability* phase (2010-14).

Phase 1 resided in the building of a basic core grammar around the Matrix skeleton (using the Matrix versions 0.1 – 0.6, as they developed; this included the MRS system). This stage included the accommodation of lexical entries lexicon adapted from the previously established resources TROLL and NorKompLex, where verb valence codes constituted major parts.

Phase 2 resided in the development of a fine-grained ontology and computing system of spatial and temporal relations, amenable to grammatical systems across languages and typologies, and a detailed semantics of comparative constructions.

Phase 3 was devoted to a revision of the valence code, to accommodate a cross-linguistically defined classification system of valence and construction types.

Phase 4, the current phase, can be divided into the following themes:

A. Deploying the grammar in ‘external’ applications: a ‘Grammar Sparrer’, as described in Hellan et al. 2013, accessed at http://typecraft.org/tc2wiki/Classroom:Norwegian_Grammar_Checking . This is a construct along the lines of Bender et al. 2004, and Suppes et al. 2014, falling within the overall initiatives described in Heift and Schultze 2007, where specific types of grammatical mistakes are accommodated by ‘mal-rules’ in an extended ‘mal’-version of the grammar, and parses involving such mal-phenomena are reported to the user as tutoring instructions. This system has been running as a webdemo since 2011. (The webdemo itself: http://regdili.idi.ntnu.no:8080/studentAce/parse .)

B. Exporting information from the grammar to independent resources:

1. A valence bank, which, with the same exporting strategy as for Norwegian, contains also two other languages, constituting the first instance of an in depth Multilingual Valence repository. In essence, the valence code used in verbal lexical types (cf. 3.2 below) is expanded to alternative and more easily inspectable formats, and the verb lexicons of the languages involved are imported into a database organized according to the newer codes, and searchable in terms of these codes. (See Hellan and Bruland 2013, and a web access at http://regdili.idi.ntnu.no:8080/multi_valence_web_demo/multivalence.)

B. Exporting information from the grammar to independent resources - contd:

2. A POS-tagger reflecting the lexical inventory of the grammar, useful for lexical acquisition from new text (http://regdili.idi.ntnu.no:8080/webtagger/tagger).

3. A simple Reasoner over movement and spatial information exported from the MRS. (See Bruland 2013)

Valence specification in NorSource

Valence lists ‘SPR’ and ‘COMPS’ need to be emptied in the course of syntactic combination; their content is reflected in the GF specifications:

$$\begin{bmatrix} \text{HEAD } \textit{verb} & \\ \text{GF} & \begin{bmatrix} \text{SUBJ } \boxed{3}\begin{bmatrix} \text{HEAD}[\text{CASE } \textit{nom}] \\ \text{INDX } \boxed{1}[\text{ROLE } \textit{agent}] \end{bmatrix} \\ \text{OBJ } \boxed{4}\begin{bmatrix} \text{HEAD}[\text{CASE } \textit{acc}] \\ \text{INDX } \boxed{2}[\text{ROLE } \textit{patient}] \end{bmatrix} \end{bmatrix} \\ \text{SPR } \langle \boxed{3} \rangle & \\ \text{COMPS } \langle \boxed{4} \rangle & \\ \text{ACTNTS} \begin{bmatrix} \text{PRED } \textit{beissen - rel} \\ \text{ACT1 } \boxed{1} \\ \text{ACT2 } \boxed{2} \\ \text{AKTRT } \textit{achievement} \end{bmatrix} & \end{bmatrix}$$

Wiew of preposition represented with enriched semantics, in MRS of a Matrix grammar (Beermann and Hellan):

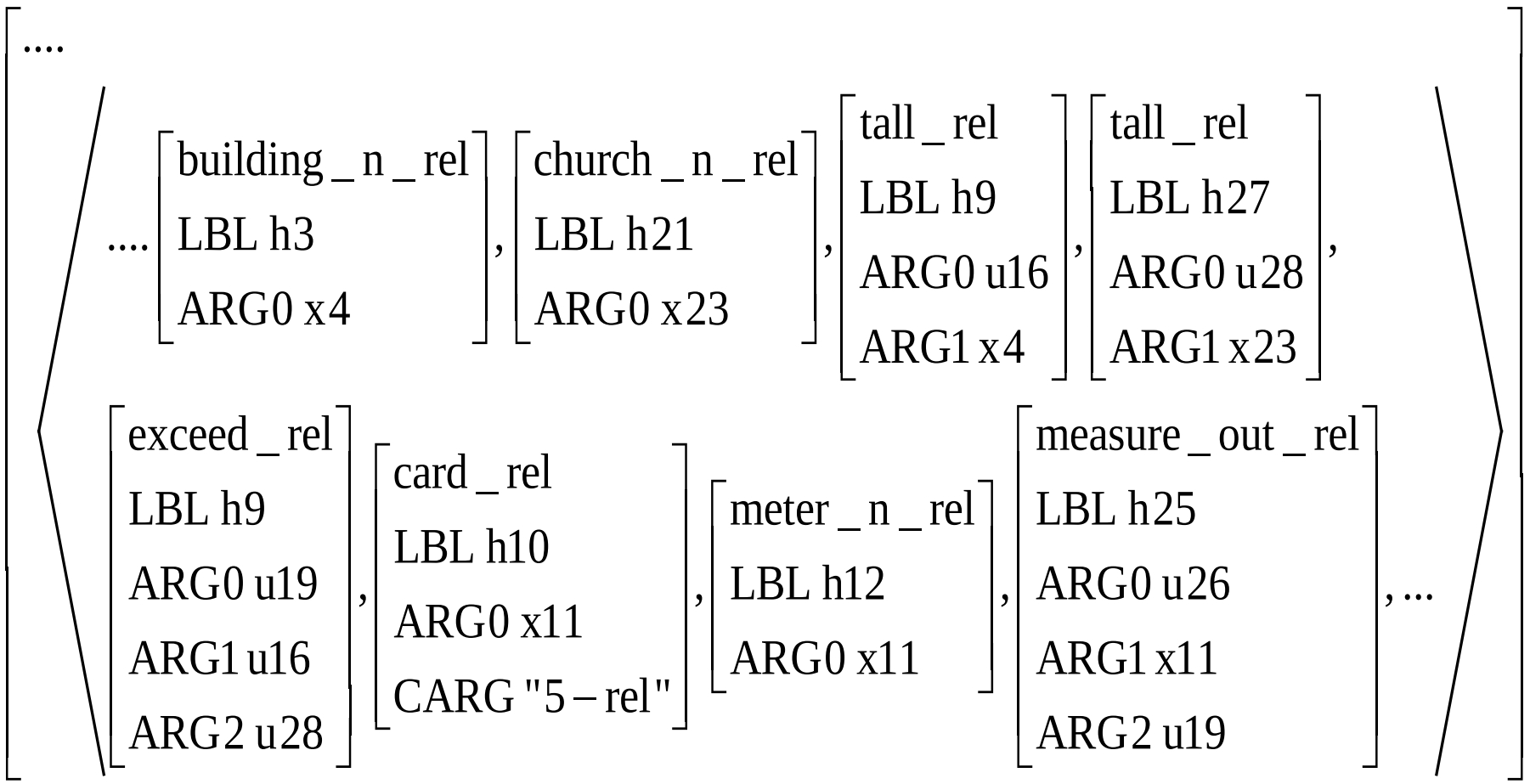
The boy runs through the forest

$$RELS\left\langle \ldots\left[\begin{array}{l} \_run\_v\_rel \\ ARG0\ e2 \\ ARG1\ x4\left[ROLE\ locomotor\right] \end{array}\right], \left[\begin{array}{l} \_through\_p\_rel \\ \\ ARG0\ e6 \\ ARG1\ x4\left[ROLE\ locomotor\right] \\ ARG2\ x9\left[ROLE\ viapnt-of-path\right] \\ TOPOL\left[\begin{array}{l} FIG\ x4 \\ GRND\ x9 \\ TOPFEAT\left[\begin{array}{l} FRONT- \\ BACK- \end{array}\right] \end{array}\right] \end{array}\right], \ldots \right\rangle$$

This building is 5 meters taller than the church:

"5-meters-taller-than (the building, the church)“

"For a degree *d1* and a degree *d2* such that *d1* is the height of the building and *d2* is the height of the church, *d1* exceeds *d2* to an extent *d3*, and *5 meters* measures out *d3*."

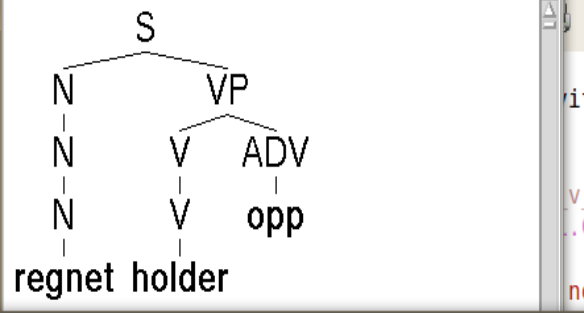

....
....
building _ n _ rel
LBL h3
ARG0 x4
church _ n _ rel
LBL h21
ARG0 x23
tall _ rel
LBL h9
ARG0 u16
ARG1 x4
tall _ rel
LBL h27
ARG0 u28
ARG1 x23
exceed _ rel
LBL h9
ARG0 u19
ARG1 u16
ARG2 u28
card _ rel
LBL h10
ARG0 x11
CARG "5 – rel"
meter _ n _ rel
LBL h12
ARG0 x11
measure _ out _ rel
LBL h25
ARG0 u26
ARG1 x11
ARG2 u19
, ...

Applications Places System

lhellan

Parse Tree #0

Close | Close All | Print



lex1.close.tdl

File Edit Options Buffers Tools LKB Help

```
    SYNSEM.LKEYS.KEYREL.PRED "_vare_v-intrPrtcl_rel" ].

holde_telicpart := v-intrPrtcl-COMPLETEDACTIVITY &
  [ STEM < "holde" >,
    INFLECTION nonfstr-strong,
    SYNSEM.LOCAL.KEY-SPEC opp-pcl,
    SYNSEM [LKEYS.KEYREL [ PRED "_holde_v-intrPrtcl_rel" ] ]].
```

Simple MRS

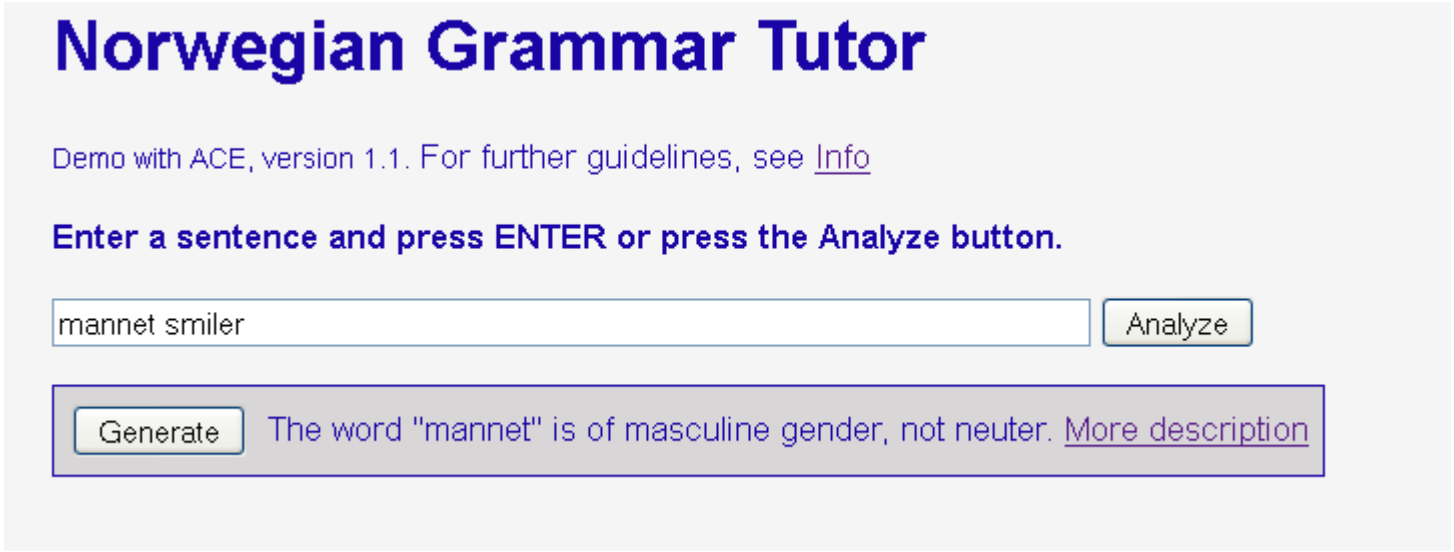
Close | Close All | Print | Save as XML | Generate

```
[ LTOP: h1 [ h ROLE: ROLE ]
 INDEX: e2 [ e E.TENSE: PRES E.MOOD: INDICATIVE E.ASPECT: SEMSORT E.DELIMITED: BOOL PATH-TELIC: BOOL SIT-TYPE.COMPLETED: + SIT-TYPE.DYNAMIC: BOOL SIT-TYPE.F
 RELS: <
      [ "_regn_n_rel"<0:6>
        LBL: h3 [ h ROLE: ROLE ]
        ARG0: x4 [ x ROLE: ROLE WH: - BOUNDED: + PNG.NG.NUM: SING PNG.NG.GEN: N PNG.PERS: THIRDPERS ] ]
      [ "_def_q_rel"<0:6>
        LBL: h5 [ h ROLE: ROLE ]
        ARG0: x4
        RSTR: h6 [ h ROLE: ROLE ]
        BODY: h7 [ h ROLE: ROLE ] ]
      [ "_holde_v-intrPrtcl_rel"<7:13>
        LBL: h8 [ h ROLE: ROLE ]
        ARG0: e2
        ARG1: x4
        ARGX: h9 [ h ROLE: ROLE ] ]
      [ "_opp_pcl_rel"<14:17>
        LBL: h9
        ARG0: u10 [ u WH: BOOL ROLE: ROLE ]
        ARG1: x4 ] >
 HCONS: < h6 qeq h3 > ]
```

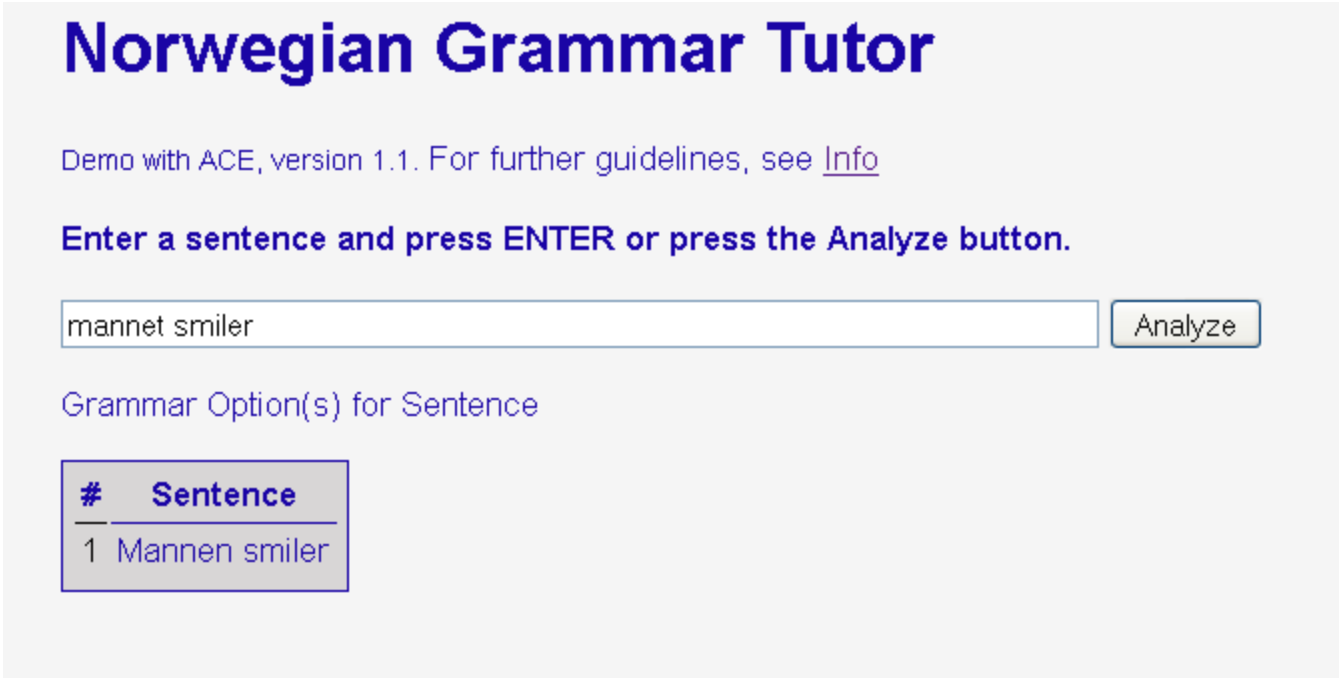
lhellan@HFF3... | lex1.close.tdl | Lkb Top | tsdb() | norsk.tdl | lex1.close.tdl | Simple MRS | Parse Tree #0

Grammar Tutor – error message

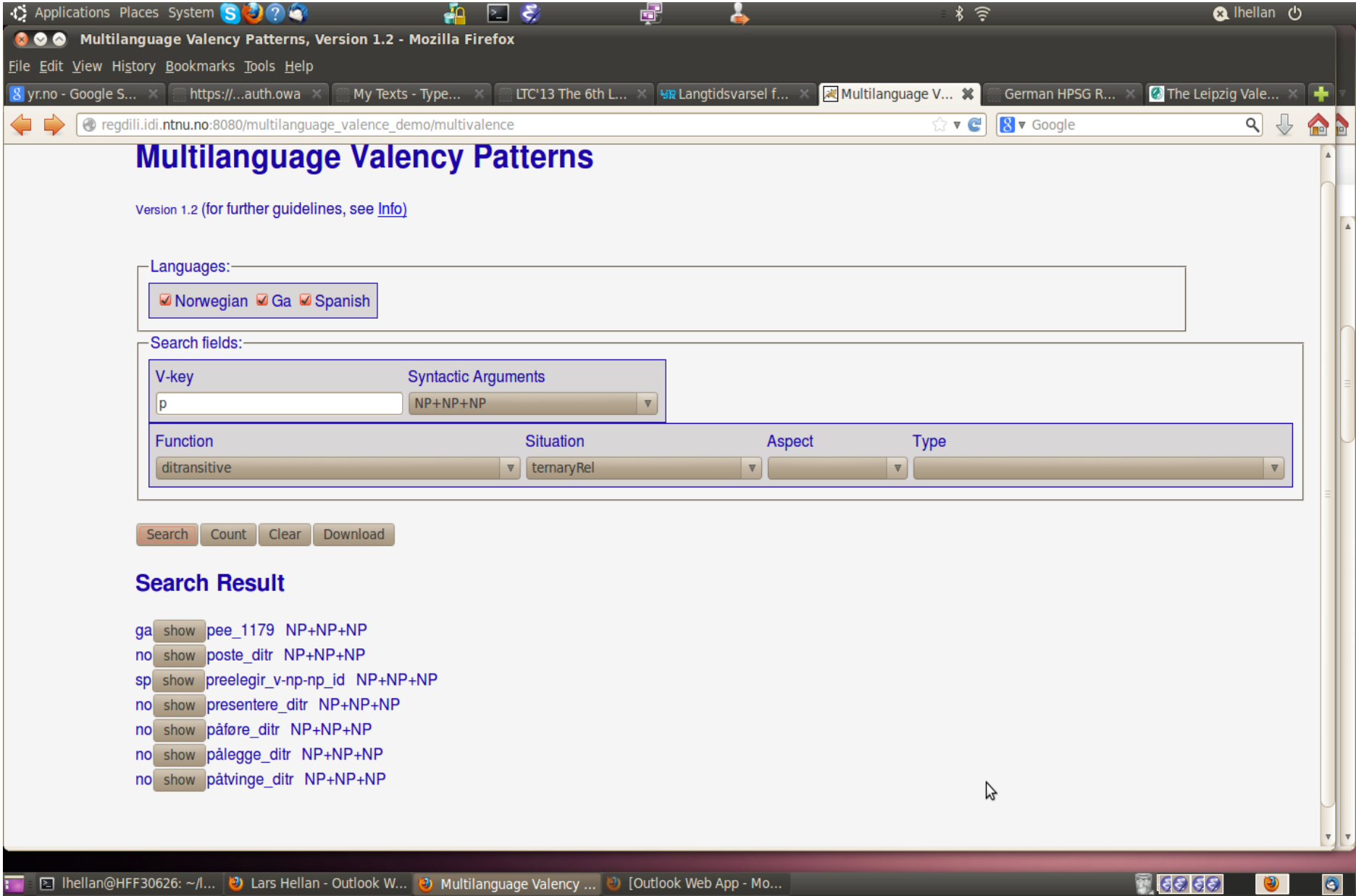
- Enter an ungrammatical sentence
- Receive an error message
- Select the first MRS and classify it with Utool
- If the MRS is accepted, a button to generate is displayed



Generate correct option(s)



A multilingual valence database



- The **Languages**:
A selection of languages L1,..., Ln;
- The **Parameters**:
A set of specification parameters defined across all the languages (i.e., *common* parameters, in the sense of being independent of any particular language, although not in the sense of necessarily being relevant for all of the languages);
- The **Valence-profiles**:
For each language, an inventory of its valence types characterized in terms of the parameters available, called its *valence-profile*;
- The **Valence-type suites**:
For each language, a list of sentences instantiating each of its valence types, indexed according to the types;
- The **Valence Lexicons**:
For each language, a verb lexicon where each verb entry is classified according to its valence type (in addition to other lexical information);
- The **Valence Corpora**:
For each language, a sentence corpus instantiating each verb in each of the valence frames it can support.

With a multilingual main focus, what comes closest is probably the *Leipzig Valency Classes Project* (ValPaL), where 80 verb meanings have been analyzed across 30 languages, and may be said to provide more or less directly the *Languages*, *Parameters*, *Valence-profiles* and *Valence-type suites* in terms of the list above, relative to the 80 verbs. This project has just been published.

http://www.eva.mpg.de/lingua/valency/

How we get the data

The three current repositories are all derived from HPSG-based computational grammars of the languages in question.

Information comes in part from lexical entries for the individual verbs, in part from the grammatical type of each verb. In the previous example, the expression under ‘Verb type’, i.e., *v-intr-suDir*, is such a type, among the appx. 250 verb lexeme types based on argument frame properties.

For each such type, a *correspondence* rule is defined mapping the type onto some of the specification parameters of MultiVal, such as the following:

v-ditr	=>	SAS:	“NP+NP+NP”
		FCT:	ditrans
		SIT:	ternaryRel

The field ’Syntactic Arguments’ – ‘SAS’, is illustrated in the list snippet below; the symbol ‘+’ stands for linear order.

NP+INF
NP+INF:equiSBJ
NP+INF:raisingSBJ
NP+NP
NP+NP+APpred

For Norwegian, the set of possible SAS specifications is currently 158, which is close to being exhaustive at this level of specification.

The field ’Function’ – ‘FCT’ - relates to a more traditional type of descriptive terms, such as ‘intransitive’, ‘transitive’, ‘transitive with oblique’, etc.. They provide less detail in differentiation than the SAS field, thus, for Norwegian, there are currently only 88 FCT term. In contrast to the SAS list, the FCT terms say nothing about linear order.

The fields ‘Situation’ and ‘Aspect’ contain situation type and aspectual properties of situations expressed, thus both representing semantic information.

‘Verb type’ represents a classificatory index of the frame relative to the lexical type system of the grammar of origin.