Constructing a Multilingual Database of Verb Valence

Lars Hellan[1], *Tore Bruland*[2]

(1) Department of Language and Communication Studies, NTNU, N-7491 Trondheim
(2) Department of Informatics, NTNU, N-7491 Trondheim

lars.hellan@ntnu.no, torebrul@idi.ntnu.no

ABSTRACT

We show the initial stage of an incremental on-line multilingual valence pattern demo, presently populated with two languages, Norwegian and Ga. The procedure for establishing the Norwegian part of the valence database resides in reusing material available in the computational HPSG-grammar Norsource, which has a rich array of lexical information, in part developed from earlier existing lexical resources for Norwegian. The procedure used for Ga is based on a Toolbox lexicon for Ga, with a first stage of processing enabling its data to join the conversion strategy used for Norwegian. A common template is used for the valence information display, although neither source fills in the template completely, reflecting their original differences in content. Essential among these is the availability of example sentences illustrating each valence option for each verb – this is available for Ga, but not for Norwegian. The results are implemented but not yet widely published, serving at the moment partly for self-improvement through exhibiting weaknesses in the resources from which they were derived, and partly for development of the multilingual design.

KEYWORDS: Valence, Syntactic Argument Structure, Computational Grammar, Norwegian, Ga, LKB, HPSG, Toolbox.

1 Introduction

We present the initial version of an on-line multilingual database, so far with two languages represented – Norwegian and Ga (spoken in the Accra area of Ghana). Such a database is of potential value for machine-aided translation, and for language comparison from both practical and theoretical perspectives. No multilingual valence database exists yet, to our knowledge, thus neither standards nor 'good examples' are available for reference for the present enterprise. Monolingual valence databases do exist, including, for instance, the Erlangen Valence Patternbank, [1] VerbNet, [2] and E-VALBU, [3] and standards set by those could be carried over to the information provided by the partaking languages of a multilingual database. A requirement of a multilingual database, though, is that the encoding of information must be uniform across the languages involved, which means that (i) a 'mechanical' combination of existing monolingual bases would not suffice; and (ii) the design of a common categorization and classification will be a crucial challenge. Still, since assembling lexical material for a language, making a classification system, and making consistent use of it in the construction of the database for the language, are major tasks, the more one can make use of existing resources, the better.

In the present demonstration, we make use of two pre-existing lexical resources: the verb-lexicon of the Norwegian HPSG grammar *Norsource*, [4] based on the LKB platform, [5] and a Toolbox lexicon of Ga[6]. The former consists of about 10,000 entries (for verbs), the latter nearly 2000 (for verbs), both on what we may call a 'full-frame' basis, namely a design where alternative valence frames for a given lemma are represented by different entries.[7]

2 The Norwegian valence lexicon

The Norsource lexicon has adopted and adapted resources from the TROLL project[8] and the NorKompLex project, conducted previously at NTNU. The format of an LKB type entry is exemplified in (1) below from the Norwegian lexicon, with the lexical entry for regne 'rain', as in det regner 'it rains'; the lexical type of the entry - here *v-intrImpers* [9] - carries all aspects of syntactic and semantic information reflected in the grammar: [10]

```
(1)     regne_impers := v-intrImpers &
        [ INFLECTION nonfinstr,
        STEM < "regne" >,
        PRED "_regne_v_rel" ].
```

[1] http://www.patternbank.uni-erlangen.de/cgi-bin/patternbank.cgi?do=introtxt

[2] http://verbs.colorado.edu/~mpalmer/projects/verbnet.html

[3] http://hypermedia2.ids-mannheim.de/evalbu/

[4] http://typecraft.org/tc2wiki/Norwegian_HPSG_grammar_NorSource ; on HPSG, cf. Pollard and Sag 1994.

[5] Cf. (Copestake 2002).

[6] See Dakubu 2009, 2010

[7] This in contrast to approaches like (Levin 1993), and VerbNet, which rather identify as the main unit a verb together with its cluster of possible frames (belonging to what is called a 'verb class').

[8] Cf. (Hellan et al. 1989).

[9] For the system of type labels used, cf. (Hellan and Dakubu 2009, 2010).

[10] The expression in (1) reads : 'regne-impers' is a subtype of the type 'v-intrImpers', with the specifications for inflection class, stem and predicate values indicated, following the tdl code used in LKB grammars.

A conversion list with members like (i), (ii) and (iii) in TABLE 1 below is then used to populate the database; to the left of the arrow is the lexical type name, and the lines to the right provide 'expansions' of this type name, in the format chosen for the database:

(i)	v-intrImpers	=>	SAS: "EXPL" SFP: impersonal-weatherProcess Example of type: "det regner"
(ii)	v-intrImpersPrtcl	=>	SAS: "EXPL+adpos" SFP: impersonal-weatherProcess Example of type: "det klarner opp"
	…		
(iii)	v-ditr	=>	SAS: "NP+NP+NP" SFP: ternaryRel
	…		

TABLE 1: Sample from the conversion of valence-types to specifications used in the database view.

'SAS' stands for 'syntactic argument structure'. For Norwegian, the set of possible SAS specifications is currently 158, which is close to being exhaustive at this level of specification, based on so-called 'formal' categories – (2) is a snippet of the list:

(2) ….
NP+INF
NP+INF:equiSBJ
NP+INF:raisingSBJ
NP+NP
NP+NP+APpred
…

'SFP' stands for 'semantic and functional properties', and is so far a more tentative and restricted assembly of categories, but an area of further development. The total number of conversion rules including those in TABLE 1 is currently 350.

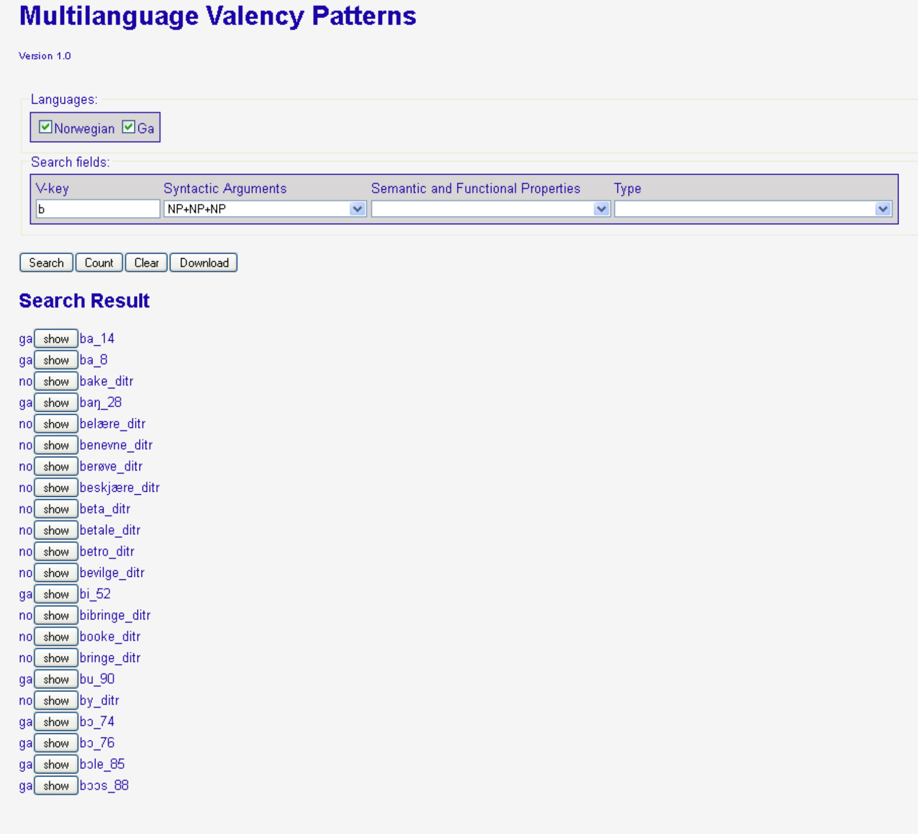
Every entry of the grammar's lexicon (i.e., entries like (1) above) is run through the conversion list, and the expansions for each verb populate the database. In the user interface, the search for any specific verb lemma can be combined with a menu of lexical types, and with a SAS or SFP specification; see below.

3 Expansion to a bi- and multi-lingual valence lexicon

For the multilingual demo, the material also for Ga comes from verb lexicon files available in the form of a TDL-formalism lexicon, based on a separate conversion from the Toolbox format to the

LKB lexicon format. [11] In the TDL-entries, exemplified in (3), each attribute corresponds to a 'field' in the Toolbox file:

A conversion list as exemplified in TABLE 1 will convert the lexical types used in the Ga file into the same general array of SAS and SFP labels as for Norwegian (with some labels specific to Norwegian, and some to Ga, reflecting typological differences). The database for Ga will have specification slots corresponding to all of the attributesexemplified in (3), along with those induced by the conversion rules, just as the Norwegian database will have specifications for the relevant parameters in (1) along with those induced by the conversion rules. Thus, the database is created on a 'pot-luck' basis, each language contributing its own resources, on a common form.



Multilanguage Valency Patterns

Version 1.0

Languages: ☑ Norwegian ☑ Ga

Search fields:

V-key	Syntactic Arguments	Semantic and Functional Properties	Type
b	NP+NP+NP		

Search Result

```
ga show ba_14
ga show ba_8
no show bake_ditr
ga show baŋ_28
no show belære_ditr
no show benevne_ditr
no show berøve_ditr
no show beskjære_ditr
no show beta_ditr
no show betale_ditr
no show betro_ditr
no show bevilge_ditr
ga show bi_52
no show bibringe_ditr
no show booke_ditr
no show bringe_ditr
ga show bu_90
no show by_ditr
ga show bɔ_74
ga show bɔ_76
ga show bɔle_85
ga show bɔɔs_88
```

FIGURE 1: Search result for ditransitive frame.

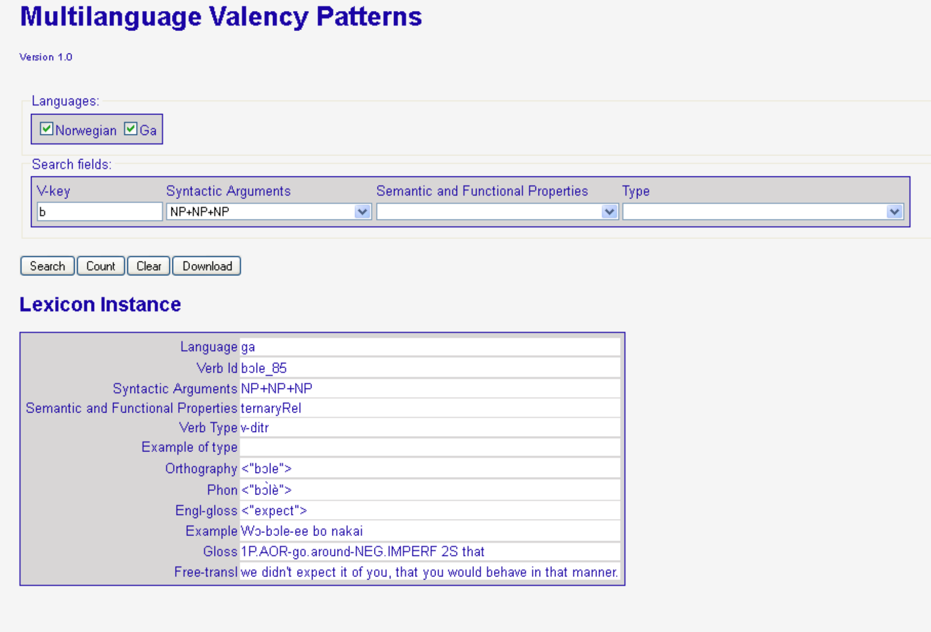
With these prerequisites, a search in the combined database will be able to define its targets by SAS, or SFP, or Lexical Type, or any combination of these - see FIGURE 1 above and FIGURE 2 below for screen-shot examples of the two types of views available. [12] The view in FIGURE 1

[11] See (Hirzel 2006, 2012) for earlier explorations of such conversions.

[12] The demo website is http://regdili.idi.ntnu.no:8080/multilanguage_valence_demo/multivalence.

brings out, for the language(s) selected, the set of entry-IDs of verbs satisfying the criteria indicated, viz. here the SAS frame “NP+NP+NP” (the formal pattern of a d-transitive or double-object construction).

Each verb-ID in the list in FIGURE 1 is provided with a button ‘Show’, and FIGURE 2 is an instance of what can be called up with this button, a view for the second lowest verb on the list in FIGURE 1, stating all of the information directly or indirectly encoded in the entry in (3). This information thus includes the attribute ‘Verb Type’, with a value *v-ditr* being the expression explicitly entered behind the symbol ‘:=’, and a value for ‘Semantic and Functional Properties’, which is not explicitly entered in (3), but induced from the Verb Type *v-ditr* in the conversion list illustrated in TABLE 1, here the by item (iii) in that list.



Languages:

☑Norwegian ☑Ga

Search fields:

V-key	Syntactic Arguments	Semantic and Functional Properties	Type
b	NP+NP+NP		

Lexicon Instance

Language	ga
Verb Id	bɔle_85
Syntactic Arguments	NP+NP+NP
Semantic and Functional Properties	ternaryRel
Verb Type	v-ditr
Example of type	
Orthography	<"bɔle">
Phon	<"bɔ̀lè">
Engl-gloss	<"expect">
Example	Wɔ-bɔle-ee bo nakai
Gloss	1P.AOR-go.around-NEG.IMPERF 2S that
Free-transl	we didn't expect it of you, that you would behave in that manner.

FIGURE 2: Result for ‘Show’ for a verb displayed on the search result in FIGURE 1.

Having mentioned machine translation as a possible application of such a database, it is clear that valence information by itself is not a carrier of translation, but together with information about semantic equivalence of putative verb pairs in two languages, valence equivalence or similarity will serve as an additional parameter to induce quality of translation. For the marking of semantic equivalence there will be in principle two routes – ‘bridging’ through the availability of shared English (or other language) glosses, and sameness of semantic representation in some formalism rich and tractable enough to be readily searchable. Per this day, no system of the latter kind is available, [13] and the information located in the present slot 'semantic and functional properties' is too general to induce translation. For the option of going via English glosses, the Ga database has the necessary information, but the Norwegian database not, hence the database at present could not contribute substantively towards translation by itself, only offer valence information supplementing translation hypotheses generated elsewhere.

4 Logistic perspectives

The database architecture is open for the addition of further languages. The database and its interface are independent of the data provenance – the data, when already acquired and systematized, could come from lexically based grammars (of which HPSG grammars are only one type), or from digital lexical resources generally. The classification systems can vary, as long as correspondence rules like those suggested in TABLE 1 can map them to the system of SAS and SFP here adopted. The conversion pipeline can be of any form whatsoever - in the present case, they come for both languages from LKB files, but this is because one of the provenance sources actually is formalized in LKB, and for Ga, there is indeed an LKB grammar also for this language,[14] where the lexicon file is also used.

The scalability of the approach nevertheless will depend on for how many languages a process as here suggested can be fairly directly replicated. Since there are many languages for which LKB grammars have been defined, the present process can in principle be used for many of these languages;[15] the same holds for other frameworks of 'deep grammars'. Replicability of the process from Toolbox projects is likewise conceivable, but since these are developed in much less mutual concordance than the grammar types mentioned, they have to be considered more on a one-by-one basis.

Another issue is that of incremental improvement of the database – what we have so far described is a 'once-and-for-all' line of action, which is obviously insufficient for tasks of such complexity as a valence database. LKB grammars, among others, typically lack specifications like the lower lines in (3), and a question will be how to add such specifications, i.e., example sentences with instructive glossing. Incremental improvement of any aspect of the database can in general be done through the grammar/lexical resource, with uploading of the system to the database from time to time; acquisition of examples from corpora, or by individual contribution, will constitute a different path. Conceivably there could be a direct contribution interface to the database for relevant examples; or one could use corpus and annotation tools to feed the relevant information into the 'source' grammars,[16] from which the information could be further propagated to the database using the established pipeline. This issue will be in focus in further design developments, with Norwegian being first in line for testing.

[13] Initiatives that may lead towards such formats include FrameNet ((cf. https://framenet.icsi.berkeley.edu/fndrupal/), and the Leipzig Valency Classes Project (http://www.eva.mpg.de/lingua/valency/index.php ; cf. Comrie and Malchukov, to appear).

[14] Cf. Dakubu et al. 2007.

[15] Aside from a conversion strategy from verb types as here illustrated, one can also induce SAS information from the actual feature structure of the lexical entries, when run through the grammar unification mechanism; this has been done for Norsource, but is not included in the current system.

[16] For instance using TypeCraft (http://typecraft.org – cf. (Beermann and Mihaylov 2011)), as described in (Hellan and Beermann 2011),

5 Linguistic perspectives

There is general consensus that parameters like the following ought to be represented in an account of valence types:[17]

(4) a. syntactic argument structure, i.e., whether there is a subject, an object, a second/indirect object, etc., referred to as grammatical functions, and the formal categories carrying them;

b. semantic argument structure, that is, how many participants are present in the situation depicted, and which roles they play (such as ‘agent’, ‘patient’, etc.);

c. linkage between syntactic and semantic argument structure, i.e., which grammatical functions express which roles; - identity relations, part-whole relations, etc., between arguments;

d. aspect and Aktionsart, that is, properties of a situation expressed by a sentence with the valence in question in terms of whether it is dynamic/stative, continuous/instantaneous, completed/ongoing, etc.;

e. type of the situation expressed, in terms of some classificatory system.

The slot ‘SAS’ used presently may be said to represent the formal part of (4a), while the functional part of (4a), the ‘–arity’ part of (4b), (4d), and to a small extent (4e), are included under the current slot ‘semantic and functional properties’, a slot which may well become split up according to these parameters in the future. ‘Type’ in the current demo represents the lexical type label used in the input grammar; the array of such types relevant for Norwegian is described in (Hellan 2008), and the general system of type labels in question is described in (Hellan and Dakubu 2009, 2010). Other grammars or sources providing valence data may well use other type or type label inventories, and we leave open at this point whether each provenance system should be introduced as it is under ‘Type’, or some approach of standardization be attempted – counting against the latter are the circumstances that the other slots will be fully standardized anyhow, and that the ‘pot luck’ approach to acquisition will be easier the less conversion operations are called for. What one may still strive for is that similar ranges of discriminants are included in the classification systems used - thus, the type labels used presently cover the parameters already mentioned, whereas inclusion of participant roles, for instance, or more fine-grained situation type specifications, would increase the number of types considerably.

In general, it is given that valence frames will differ across languages and across language typologies, Norwegian and Ga illustrating both. We may refer to a language’s valence type inventory as its *v(alence)-profile*, and it will be a natural desideratum that a valence demo as here constructed should display not only valence information for particular verbs of each language involved, but also its v-profile generally. For the purpose of exposing such profiles, the online database has a ‘select’ functionality whereby for a specific language chosen, its v-profile is shown through the circumstance that only the Types and SAS options of relevance for the language are displayed on the roll-down menu. The task of establishing v-profiles, on the other hand, is a purely linguistic enterprise, aided through general descriptive work, corpus acquisition, or any combination of strategies. As an instance of decisions to be made in the linguistic

[17] See, for instance (Fillmore 2007), and other articles in (Herbst et al. 2007).

classification, a point brought out in the representation of the present pair of languages is that Ga, as a typical Kwa language making extensive use of verb serialization, encodes through serial verb constructions many of the contents for which Norwegian uses extended valence frames like those involving secondary predicates[18] and other complex relations.[19] The question then arises whether v-profiles should include constellations based on two verbs or more, or one should create a related notion of *c(onstruction)-profile* where also constructs like the relevant serial verb constructions are included. In the Ga lexical database, we have chosen the latter option, thereby including labels like 'ev-' for 'extended verb construction', visible in the Ga Type inventory, and 'pv-' for 'preverb', and 'SVC' (as a placeholder for possible more detailed specifications) for 'serial verb construction', in the Ga SAS inventory.[20]

Both the Norwegian and the Ga underlying resources are in active development. In the case of Ga, the Toolbox version indeed has a layer of fine-grained semantic information added to the discriminants shown currently in the Type inventory of Ga, not included in the current LKB file but included in another file not used yet. The inclusion of this information will be done at a later point, and the SFP list for Ga will then be considerably richer than it is currently. For Norsource there is also a background inventory of not-yet-employed semantic notions. In addition, some valence frame specifications are employed on a try-out basis for a few verbs, which means that in the Type inventory, whereas types like 'v-intr' and 'v-tr' have a large number of entries (2195 and 4668, respectively), some types have just 2 or 3 members. As the current import to the valence database is based on the actual state of the lexicon, some instances of mal-proportions thereby ensue, which might be eliminated at later points; on the other hand, the views offered by the current demo are instructive to the grammar developers in these respects, and so are being maintained for the present.

In general, the linguistic notions encoded in such a database have to be generally understood, sufficiently that contributors can agree in the ways notions apply or correspond across frameworks, and so that users can understand them fairly directly. When a large range of such notions are put together in a comprehensive database, and with an in principle unlimited span of typological linguistic variation, the enterprise of course faces profound challenges relating both to the analytic understanding of phenomena and the terminologies and analytic systems applied to them. Rather than perceiving such a situation as just a 'challenge', however, one could, from a linguistic point of view, appreciate it for providing a space of investigation which is central to the concerns of linguistics, namely arriving at cross-typological, cross-framework understanding of phenomena and their analyses.

6 Concluding remarks

The architecture of a valence demo here described is fairly straightforward, computationally speaking; the basic work sits in the previous development of resources like the underlying grammar or lexicon, and in maintaining theoretical command of the linguistic issues involved. The latter task has been facilitated through the prior close linguistic cooperation behind the

[18] In the v-profile encoded in Norsource, no less than 40 valence frames include a secondary predicate ('small clause predicate').

[19] We here have in mind constructions referred to as 'integrated' serial verb constructions, as opposed to 'chaining' constructions, i.e., those linking together indefinite sequences of consecutive events.

[20] For explanation of these notions, see (Hellan and Dakubu 2010), and (Dakubu et al. 2007).

contributing resources; bringing together resources with less of such a background is likely to be more challenging, linguistically as well as computationally.

We hope to have identified an interesting avenue for re-utilization of linguistically rich resources, to be confirmed through further developments of the current design, including the population of it with more languages, and through further developments of the demo, both in its monolingual and its bi-lingual capacities.

References

Beermann, D. and Mihaylov, P. (2011). e-Research for Linguists. *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*

Comrie, B. and Malchukov, A. (eds) (to appear) *Handbook of Valency classes*.

Copestake, A. (2002). *Implementing Typed Feature Structure Grammars*. CSLI.

Dakubu, M.E.K. (2009). *Ga-English Dictionary with English-Ga Index*. Accra: Black Mask Publishers.

Dakubu, M.E.K. (2010) Toolbox project Ga, University of Ghana.

Dakubu, M.E.K., L. Hellan., and D. Beermann. (2007). Verb Sequencing Constraints in Ga: Serial Verb Constructions and the Extended Verb Complex. In St. Müller (ed) *Proceedings of the 14th International Conference on Head-Driven Phrase Structure Grammar*. CSLI Publications, Stanford. (/http://csli-publications.stanford.edu/)

Fillmore, C. (2007): Valency issues in FrameNet. In: Herbst and Götz-Votteler (eds.).

Hellan, L. 2008. Enumerating Verb Constructions Cross-linguistically. COLING Workshop on Grammar Engineering Across frameworks. Manchester. http://www.aclweb.org/anthology-new/W/W08/#1700

Hellan, L., Johnsen, L. and A. Pitz. (1989) The TROLL Project. Ms., NTNU.

Hellan, L. and M.E.K. Dakubu. 2009. A methodology for enhancing argument structure specification. *Proceedings from the 4th Language Technology Conference (LTC 2009)*, Poznan.

Hellan, L. and Dakubu, M.E.K. (2010) *Identifying Verb Constructions Cross-linguistically*. SLAVOB series 6.3, Univ. of Ghana.

Herbst, T and K. Götz-Votteler (eds.) (2007): *Valency: Theoretical, Descriptive and Cognitive Issues*, Berlin/New York: Mouton de Gruyter.

Hirzel, H.. 2006. Deriving LKB lexicons from Toolbox. Talk given at Workshop on Grammar Engineering, NTNU, June 2006.

Hirzel, H. 2012. Converting a Toolbox lexical database to LKB format. http://typecraft.org/tc2wiki/Converting_a_Toolbox_lexical_database_to_LKB_format.

Levin, B. (1993) *English Verb Classes and Alternations*. Univ. of Chicago Press, Chicago, IL.

Pollard, C. and Sag, I. (1994). *Head-Driven Phrase Structure Grammar*. Univ. of Chicago Press, Chicago, IL.