Verb valence lexicons, and comparing them

Lars Hellan

DELPH-IN Summit, July, 2019

Cambridge

Valence

Verb valence is an essential aspect of what the mastery of a verb entails.

To the extent that HPSG grammars host information about valence, in a systematic manner, it is interesting to see how such information can be channeled into resources more accessible to the general language user.

More generally, one of the looming questions concerning valence is whether

- *If you know the meaning of a verb and the general lexico-grammar of the language, then you know the possible valence frames of the verb.*

In language teaching, this is perhaps more or less taken for granted; but numerous investigations show that it doesn’t hold, and so one general research question is to what extent meaning – however described – can be correlated with valence.

We outline an approach facing both of these perspectives.

First a look 5 years back.

2014: ***MultiVal*** – the Multilingual Valency database

The system hosts 4 languages, with altogether 40 000 verb entries, with valence frames classified in a uniform system. The languages hosted are:

Bulgarian *(lexicon import from BURGER, the Bulgarian Matrix grammar)*

Ga *(lexicon import from GaGram, the Ga Matrix grammar, whose lexicon is in turn imported from a ToolBox lexicon of Ga, created by M.E.Kropp Dakubu)*

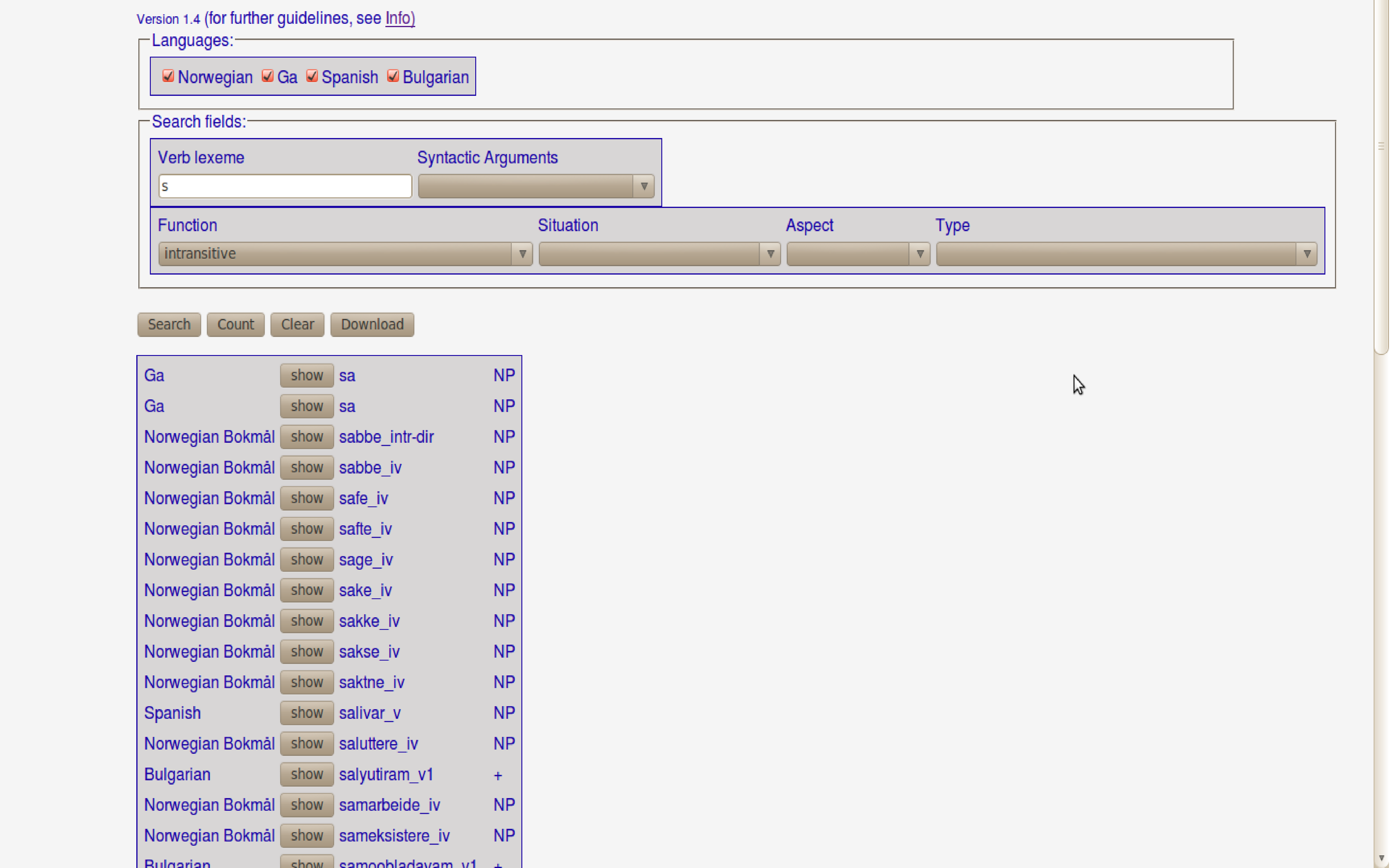
Norwegian *(lexicon import from NorSource , the Norwegian Matrix grammar)*

Spanish *(lexicon import from SRC , the Spanish Matrix grammar)*

For documentation of the system per February 2014 (before Bulgarian got added), see Hellan et al., LREC 2014.

The following slide shows search results for ‘intransitive’, for verb starting with “s”.

The next slide in turn shows information as it looks for a given verb, and shows two features of interoperability with other applications – *TypeCraft* and *ImagAct*.



Version 1.4 (for further guidelines, see Info)

Languages:

☑ Norwegian ☑ Ga ☑ Spanish ☑ Bulgarian

Search fields:

Verb lexeme: s

Syntactic Arguments:

Function: intransitive

Situation:

Aspect:

Type:

Ga	show	sa	NP
Ga	show	sa	NP
Norwegian Bokmål	show	sabbe_intr-dir	NP
Norwegian Bokmål	show	sabbe_iv	NP
Norwegian Bokmål	show	safe_iv	NP
Norwegian Bokmål	show	safte_iv	NP
Norwegian Bokmål	show	sage_iv	NP
Norwegian Bokmål	show	sake_iv	NP
Norwegian Bokmål	show	sakke_iv	NP
Norwegian Bokmål	show	sakse_iv	NP
Norwegian Bokmål	show	saktne_iv	NP
Spanish	show	salivar_v	NP
Norwegian Bokmål	show	saluttere_iv	NP
Bulgarian	show	salyutiram_v1	+
Norwegian Bokmål	show	samarbeide_iv	NP
Norwegian Bokmål	show	sameksistere_iv	NP
Bulgarian	show	samoobladavam_v1	+

For the button SHOW:
Automatic import of urls for glossed examples from ***TypeCraft*** has been defined, and links to ***ImagAct*** scene videos are being added – here for *Marit tar seg på kinnet:*

Language	**Norwegian Bokmål**
Verb id	ta_tr-detachposs-refl
Syntactic Arguments	NP+NPrefl+PP
FCT	transReflxWithOblique
SIT	ternaryPossessorDetachment
Aspect	
Verb type	v-trObl-obRefl_oblPRTOFob
Example of type	Ola klør seg på ryggen
Orthography	ta
English gloss	[take] – only through TypeCraft link
Example	[Marit tar seg på kinnet] – only through TypeCraft link
Free translation	[Mary touches her cheek] – only through TypeCraft link
TypeCraft URL	http://typecraft.org/tc2/ntceditor.html#2790,45468
ImagAct URL	http://www.imagact.it/imagact/sceneMetadata.seam?sceneId=54&cid=9995

However -

One thing is missing in such a resource, namely what more than just formal frame identities may interconnect the verbs across the languages. Information about meaning similarity is missing.

Let us start over again, beginning with the notion of a *Verb Valence Lexicon*.

Verb Valence Lexicons (VVL)

In a Verb Valence Lexicon (VVL), each verb is specified with regard to the valence frames it can occur in. One possible format is to have one lexicon entry per frame per verb, so that for instance if each of x verbs has y frames, the number of entries in the VVL will be x times y (in practice, the number of frames varies significantly from verb to verb, and many verbs have just one frame).

By a 'lexval' we may understand such an entry - thus, such a VVL will have for instance x times y lexvals.

Another possible organization, more corresponding to standard lexicons and dictionaries, is to define each entry for a verb as such, with the various frames listed inside of this entry.

We can thus distinguish between *VVL_by_Lexval* and *VVL_by_Verb* with regard to the top organizational entry units. We here consider VVLs of the *type VVL_by_Lexval*. We sketch the organization of one such VVL, and consider possibilities of comparative VVLs

A Norwegian VVL

A VVL for Norwegian has been derived from the lexicon of the computational grammar Norsource, based on HPSG. It has nearly 13,000 lexvals. The 'by Lexval' format is illustrated by the following entries for "kaste“ ‘throw’:

(1)

a. kaste-opp__intrPrtcl

b. kaste__tr-obRefl

c. kaste__tr-obDir

The expression to the right of '__' is a frame type, and an item like '-opp' to the left of '__‘ in (a) is a specific word occurring with 'kaste' typical of a specific meaning, or being the sole case where a word of that category occurs with the verb.

’System sentences’ in lexvals

In a workflow for providing exemplification of lexvals one can assign 'system sentences' for all lexvals. There are 40-50 valence types recurring throughout the lexval lists, and one can construct system sentences such that one and the same sentence (apart from the verb) can be used for all lexvals of a given type. Thus, for instance, relative to example (1b) - kaste__tr-obRefl -, whose frame type 'tr-obRefl' is used in 660 lexvals, a common conversion schema could be written for this frame-type on the form (2b); likewise for (1a) the schema (2a) could be used, covering 60 lexvals, here reflecting the specific item "opp":

(2)

a. V-opp__intrPrtcl => <V-opp__intrPrtcl, "Hun V opp.">

b. V_tr-obRefl => <V_tr-obRefl, "Hun V seg.">

’System sentences’ in lexvals

A script applying from line to line can then instantiate the 'V' positions with the actual verb word of each line, for instance in present tense.

(3)

kaste-opp__intrPrtcl & de kaster opp &

kaste__tr-obRefl & hun kaster seg &

kaste__tr-obDir & du kaster ballen/du kaster ballen hit/du kaster ballen til meg/du kaster ballen hit til meg &

’System sentences’ in lexvals

In assigning sentences across the board for a given frame, one wants to be as independent as possible of restrictions imposed by the verb on its arguments, and thus such sentences can be made as ‘anonymous’ as possible, say with just plural 3p pronouns. Thus “ballen” in (3c) may be natural for “kaste” but not for other verbs of the type. Selection of suitable replacements of such pronouns then have to be made by hand.

However, even from totally anonymous sentences, one can derive POS-signatures (with ‘pron’ or ‘N’) and, with a suitable environment, run the signatures with the verb through text corpora and assemble sentences actually instantiating the signatures (as, e.g., seen in the ‘Kurze Sätze’ initiative in the Leipzig Corpus Collection project group). That could then be an automatized procedure for establishing not just examples for each lexval but corpus links as well.

A comparative verb lexicon extended with valence

A comparative verb lexicon will align a verb, with its synset, in one language with its set of synonyms in the other language, thus creating a bi-lingual synset (in the Wordnet sense).

Here, what we aim to compare is rather a combination of a verb and a valence frame in which it can occur in one language with corresponding combinations in the other language; such bilingual sets of lexvals we may call bilingual "valsynsets".

A comparative verb lexicon extended with valence

What does it mean that two lexvals are semantically equivalent? It can be construed as equivalence relative to selected parameters of situations such as number of participants, semantic roles, causation, etc. It can also be construed as translational equivalence between 'system sentences' instantiating the lexvals in question.

The latter runs a danger of arbitrariness in not exposing factors that may condition equivalence or lack of equivalence in each case. But for a comparative verb valence lexicon to serve as an initial observational basis for such a study, the translational strategy may be sufficient, and obviously the more easily executable.

Of course, the use of such translational equivalents has to be hooked up to a more elementary bilingual synset of the type one could derive from any bilingual dictionary or wordlist.

A comparative verb lexicon

For instance*, 'kaufen_tr'* and *'kjøpe_tr'* constitute a bilingual valsynset, which can be represented as '{kaufen_tr, kjøpe_tr}; adding a system sentence illustration for each member of the pair, we get what may be called a *translation-valset* as in (4):

(4) {<kaufen_tr, "Ich kaufe eine Zeitung.">, <kjøpe_tr, "Jeg kjøper en avis.">}

The common instantiated 'meaning' represented by this translation-valset is not separately indicated in this representation, but each of the two instantiations express it already.

A comparative verb lexicon

A German VVL might be constructed in a similar manner as the Norwegian VVL (although for German there are already so many valence resources that one should be able to use among what there already is).

Connecting lexvals across the languages will be to a large extent a manual task.

A typical issue in such a connection is when a lexval in one language corresponds to two lexvals in the other, as when German "holen" corresponds to both "hente" and "kjøpe" in Norwegian. In the present approach we will represent such a situation with two distinct translationvalsets – here with examples attuned to the contrast:

(5)

a. {<holen_tr, "Ich hole eine Zeitung.">, <kjøpe_tr, "Jeg kjøper en avis.">}

b. {<holen_tr, "Ich hole den Ball.">, <hente_tr, "Jeg henter ballen.">}.

A related example, with clearer selectional restrictions, is when “spise” corresponds to “essen” and to “fressen”, according to ontological status of who/what is eating.

Meaning – valence correspondence

Imagine that for the two languages we have 10,000 translation-valsets as in (4); relative to a given lexval-type on the left side we can then investigate the consistency of the lexval-types on the right side.

That would be a strategy for exploring the meaning-valence correspondence issue mentioned at the beginning on a bilingual basis.

Of course, also within a given language, one can explore pattern similarities relative to the sets of lexvals that given verbs can take – so, verbs a, b and c may all take the same set of lexvals, which opens for investigating what a, b and c may have in common otherwise, and whether, in particular, there is a semantic similarity (that will be a counterpart to the 'VerbNet'/Levin 1993 approach to 'verb classes').

A comparative verb lexicon

User-relevant valence resources can be derived from either of the approaches.

In general user interfaces one will probably omit the more detailed frame-type labels that any HPSG grammar or its derivatives would use, but for establishing lexval sets in the first place, such labels are essential.

LKB-grammars are quite relevant in such a project, since they do have articulated valence specifications for large sets of verbs, on a format which is easily operated on.

The initiative here outlined has so far has produced the mentioned 13,000 lexvals for Norwegian with anonymous example sentences. The bouquet of following possible steps remains to be started on.