Exploring opportunities for comparability and enrichment by linking lexical databases

Isa Maks, Bob Boelhouwer

Instituut voor Nederlandse Lexicologie
TST-centrale
Postbus 9515
2300 RA Leiden
maks@inl.nl, boelhouwer@inl.nl

Abstract

Results are presented of an ongoing project of the Dutch TST-centre for language and speech technology aiming at linking of various lexical databases. The project involves four Dutch monolingual lexicons: WINT05, e-Lex, RBN and RBBN. These databases differ in organisational structure and content. To enable linkage between these lexicons, we developed a common feature value set and a common organisation structure. Both are based upon existing standards for the creation and reusability of lexicons: the Lexical Markup Framework and the EAGLES standards. Examples of the content and structure of each of the lexical databases are presented in their original form. Also, the structure and content is shown when mapped onto the common framework and feature value set. Thus, the commonalities and complementarity of the lexical databases is more readily apparent. Besides, this elaboration of the databases opens up the opportunity for mutual enrichment.

1. Introduction

The Dutch TST-centre for language and speech technology is responsible for the maintenance, distribution and accessibility of digital language resources, including lexical databases (LDBs). In our paper we will present results of an ongoing project of the TST-centre to increase the comparability of the various LDBs and to make these results online available to potential users. The project consists of the following stages (i) establishing a common organisation structure and creation of a common feature value set (ii) mapping of the different LDBs on this set (iii) design and implementation of a web interface. The methods and tools that will be developed (and/or chosen) to achieve this, should be sufficiently generic to accomadate LDBs that will be entrusted to the TST-centre in the future.

The need to perform this project is based upon the type of questions that are posed by potential users; they usually don't concern a specific lexicon but are more generic in nature like "in which lexicon will I find pronunciation information" (or irregular past participles, or syntactic complementation patterns)", etc. The aim of our project is that this type of questions can be answered at once, since the user is enabled to query all available LDBs simultaneously.

2. Four Dutch monolingual LDBs

Four LDBs that were recently assigned to the Dutch TST-centre are involved in this project. The LDBS are monolingual Dutch lexicons differing in size, intended use, lexicon organisation and (type of) content. For instance, phonology is exclusively covered by e-Lex, combinatorics is exclusively covered by RBN, and GB05 is the only lexicon with focus on orhography (see table 1).

	WINT	e-Lex	RBN	RBBN
lemma entries	110.000	200.000	46.000	4.000
wordform		600.000		

orthography	++	+	+	+
phonology	-	++	-	-
morphology	+	+	+	+
syntax	+	+	+	+
semantics	-	+	++	+
usage	-	-	+	++
combinatorics	-	-	+	-

Table 1

Clearly, these linguistic resources contain a wealth of information, precisely because of their mutual differences. At the same time, however, the fact that they are differently conceived, differ in lexicon organisation and use different feature value sets for expressing (often) the same kind information, makes it difficult for a user to obtain insight in these lexicons.

3. Method and Standards

The first stage of the project will be the subject of this paper. This stage is subdivided in
(1) the design of a common feature value set and the mapping of the data on this set and

(2) the design of a framework and the mapping of the data on this framework.
One precondition is that the native structures of the lexicons remain intact. Most LDBs are components of existing NLP applications and must therefore be kept in their original format. As a consequence, repeatable routines have to be developed for retrieving the data and transporting them to the target lexicons. Another precondition is the use of existing standards. In fact, two standards are used in this project: the Lexical Markup Framework and the Eagles standard.

3.1. Lexical Markup Framework

The Lexical Markup Framework (LMF) is a standard for the creation and reusability of NLP lexicons (Francopoulo et al. (2006), ISO-TC37 SC4 (2005)). The LMF consists of (i) a core model which describes the basic hierarchy of information in a lexical entry and (ii) the extensions of the core model which are expressed in a framework that describes the reuse of the core components in conjunction with the additional components required for a specific lexical resource.

The main structure of our common model will be based upon upon the LMF core model. As the extensions are still under development, we will use the native formats for the description of the specific parts of the lexicons.

The following figure shows the LMF core model consisting of a database which contains one of more lexicons. The lexicon consists of one of more lexical entries and each lexical entry consists of one or more lemmatised forms. The sense and the inflected form components are not obligatory.

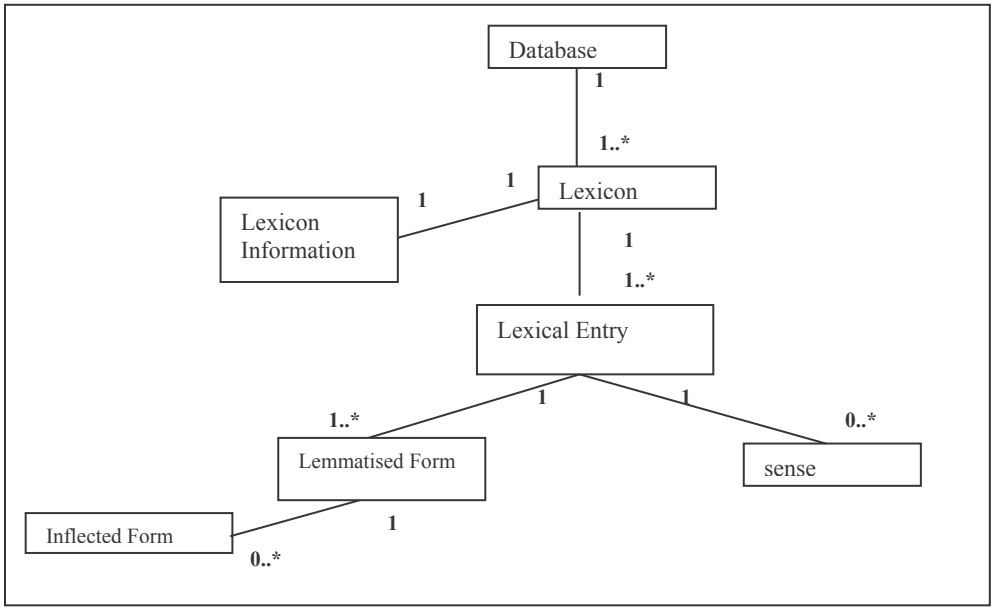

Figure 1 LMF core model

3.2. The Eagles standard

Not only the structure of the lexicon needs to be a uniform structure for all involved LDBs, the feature value set needs to be standardised as well. For the morphosyntactic information we use the standards proposed by EAGLES (Monachini et al.(2003)). As an example, we here show the feature value set for nouns. Only the features that are relevant for the description for Dutch language have been selected from this standard.

feature	value
cat	noun
number	singular/plural
gender	masculine(m)/feminine(f)/neuter(n)
type	common/proper
case	dative/accusative/genitive
(count)[1]	

The mapping of the data of the native lexicons on the unified feature value set, involves the decomposition of values into minimal units and the explicitation of implicit values. Example (1) is an illustration of decomposition; example (2) is an illustration of explicitation (since the article de implicitly means that the word category is a noun).

LDB	feature	value
e-Lex (original)		n(soort,ev,zijdig)
e-Lex (new)	cat	N
	type	common
	number	singular
	gender	non neuter

Example 1: decomposition

[1] We chose to consider countability as a semantic feature.

LDB	feature	value
WlNT (original)	lidwoord	de [v.]
WlNT (new)	cat	n
	gender	feminine (f)
	article	*de*

Example 2: explicitation

4. Mapping the data on the LMF

In the following sections we describe how, for each lexicon separately, the data are restructured in order to be able to map them on the common framework.

4.1. WlNT

The WlNT2005 (Dutch orthographic dictionary) is a lemma lexicon presenting for each lemma the standard spelling, word category, the inflected forms and hyphenation patterns. The following table shows three WlNT noun entries:

Id	5341	6226	6227
lemma	aardappel *(potatoe)*	bal *(ball)*	bal*(ball)*
hyph1-lemma	aard/appel/	bal/	bal/
cat	noun	noun	noun
article	de	de	het
plural form	aardappel, aardappels	ballen	bals
hyph-plural form	aard/ap/pel/en aard/ap/pels/	bal/len/	bals/
meaning indicator		voorwerps naam (*object)*	danspartij (*dance party)*

Table 2: WlNT entries

The noun aardappel (potatoe) has two interchangeable plural forms. The results of the mapping of this entry on the LMF core model are shown by figure 2.

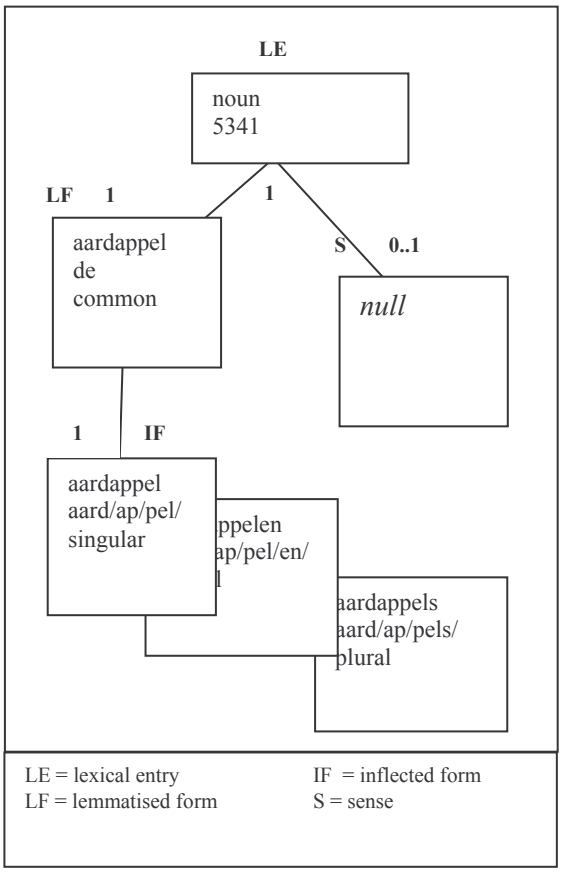

Figure 2: WlNT-LMF aardappel (*potatoe)*

In general the entries don't have a sense component; as a consequence the relationship between the lexical entry and the sense component is a 1 to 0 relationship. Only occasionaly, in order to identify homonyms with different inflection patterns, also a short definition or meaning indication is given in WlNT. An example is the lemma bal where different plural forms coincide with different meanings of the word (*ball* vs. *dance party*). For these entries two lexical entries are specified each one with a sense component linked to it.

4.2. E-Lex

E-Lex is a rather large lexicon meant for tagging and lemmatising of Dutch corpora of written and spoken text. It is a word form lexicon with for each entry the lemma, word category, gender, article, various pronunciation patterns, syntacic complementation patterns, etc. Like in WlNT, a meaning indication is given only in case of homonymy. The semantic type was added after completion implying that no syntactic-semantic linking could be accomplished. The following tables show the singular noun entry *jaloezie* (table 3) and the plural noun entry *jaloezieën* (table 4). In Dutch, *jaloezie* is a polysemous word which is shown by the two semantic types. A syntactic complementation pattern is given which indicates the use of the fixed preposition *jegens (towards*).

Lemma	jaloezie (*jealousy, venetian blind)*
morph	[jaloezie]N
cat	noun
art	de
type	common
syntactic complementation pattern	[pp:[hd:<jegens>] [obj1:np]] (*preposition: towards)*
semantic type	abstract; concrete
wordform	jaloezie
number	sing
pron1 (standard Dutch_)	jaluzi
pron2 (standard Flemish)	Jaluzi
pronsyll	ja-lu-‘zi

Table 3

lemma	jaloezie
morph	[jaloezie]n
cat	noun
art	de
type	Common

syntactic complementation pattern	[pp:[hd:<jegens>] [obj1:np]]
semantic type	abstract; concrete
wordform	jaloezieën
number	plur
pron1	jaluzi@
pron2	jaluzij@
pronsyll	ja-lu-‘zi-j@

Table 4

The organisation of the semantic features is problematic: in cases of polysemy there is no linking between the syntactic and semantic component. This can be illustrated by the example of *jaloezie*: the use of the preposition jegens is only correct with *jaloezie* in the meaning of *envy* and not in the meaning of *venetian blinds*. In other words: the semantic type is specified at the lemma level and not at the meaning (or sense) level. Figure 3 shows an extra component linked to the lexical entry. However, the LMF model allows only for components linked to the sense or (lemmatised) form components. Probably, we are going to solve this problem by splitting up e-Lex in two separate LDBs.

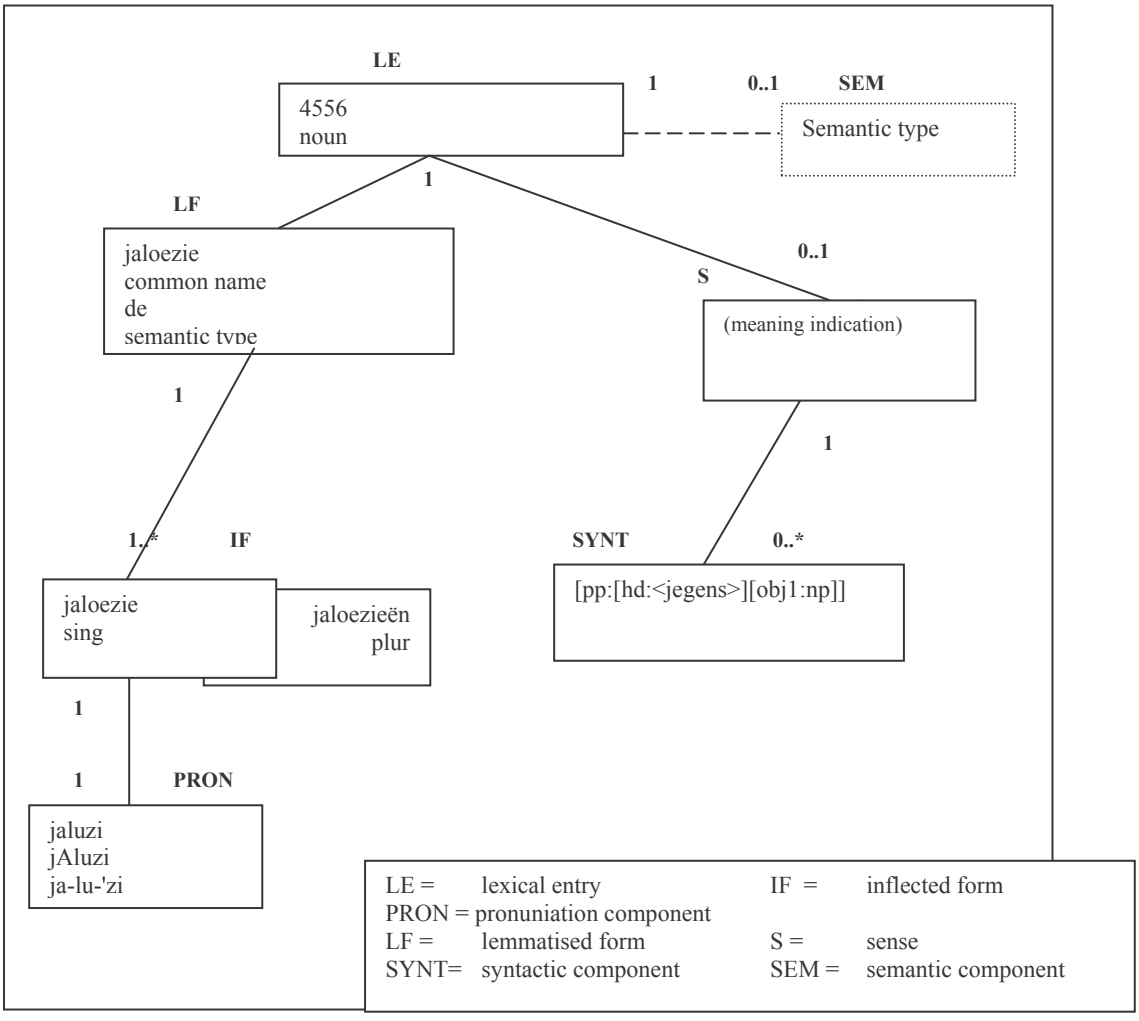

Figure 3

4.3. RBN

The entries of the RBN (Reference Lexicon for Dutch) are semantic entries, i.e. all information refers to a semantic unit. The following table (5) shows two entries which represent two meanings of the lemma entry jaloezie. The meaning of the word is expressed by a short description of the meaning (meaning indication). For each entry a semantic type, semantic shift, syntactic complementation patterns, usage, etc. are given. The semantic entries are illustrated by examples (the combinatorics) which show the word in context. The description of the example unit consists again of a syntactic, pragmatic and semantic component which – because of lack of space – is not represented in the following examples.

meaning nr	**18415**	**18416**
lemma	jaloezie	jaloezie
article	de	de
type	common	common
gender	f	f
morph. type	simplex	simplex
number	sing	sing
plural form	-	jaloezieën
complementation1	fixprep 'jegens'	-
meaning indication	afgunst (*envy)*	zonnescherm (*venetian blind)*
semantic type	nondynamic (abstract)	artefact
countability	uncount	count
combination	**55768**	**55790**
canonical form	iemands jaloezie opwekken (*arouse someone's envy)*	de jaloezieën neerlaten (*let the blinds down)*
syntactic type	fixed	free
syntactic subtype	lexical collocation	

Table 5: RBN entry

Figure 4 shows the mapping of the RBN-data on the LMF core structure. Each lexical entry may be linked to one or more sense components thus allowing for polysemous entries.

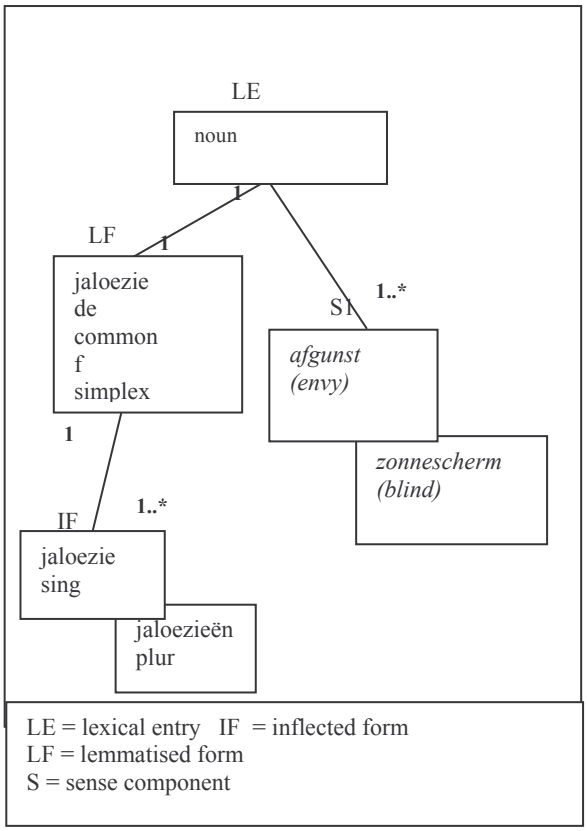

Figure 4: RBN – LMF (1)

Figure 5 shows further extensions of the core structure. The syntactic, semantic and pragmatic data are organised in components which are linked to the core model. For the extensions we use the feature value set of the native lexicon. The model is able to represent syntactic-semantic linking as is illustrated by figure 5 and figure 6. They show that the preposition *jegens* combines with *jaloezie* in the meaning of *envy* (figure 5) and not in the meaning of *blind* (figure 6).

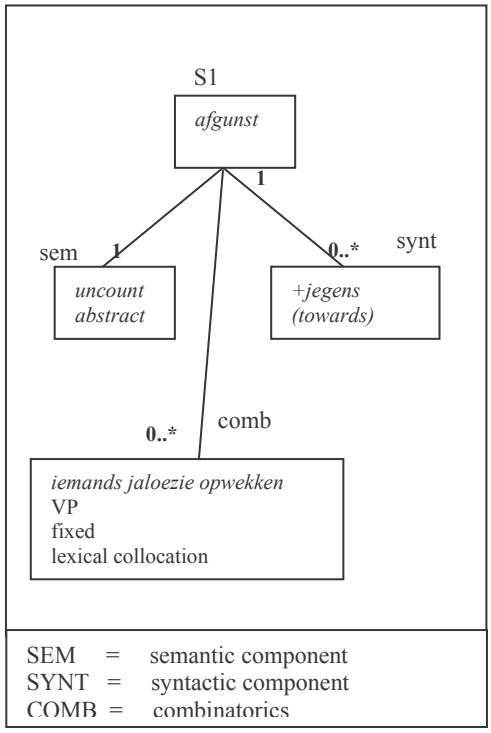

Figure 5: RBN – LMF (2)

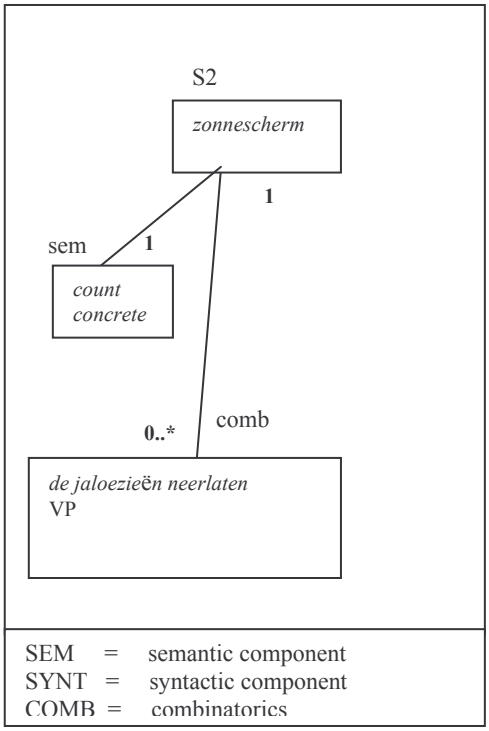

Figure 6: RBN – LMF (3)

When mapping the data, a problem arises due to the fact that in the RBN the inflected forms and their use are coupled to the semantic description. Figure 6 says that the plural form *jaloezieën* is not or not frequently used in the meaning of *envy*. To maintain this information, we decided to consider the feature countability as a semantic feature with overruling power.

4.4. RBBN

The organisation structure of the RBBN (reference lexicon Belgian Dutch) is similar to that of the RBN. The feature value set is less elaborated – it doesn't give, for example, any inflected forn - but the basic information can be represented as in figure 7. The RBBN has a very rich pragmatic description which will be represented with an extra pragmatic component linked to the sense component.

5. Conclusions

We have shown that the lexicons WlNT, e-Lex, RBN and RBBN have quite different structures. In spite of this, we largely succeeded in mapping all the data contained in these lexicons to the LMF core model. The major problem is posed by e-Lex which has a 'loose' semantic component.

Although our major aim was to accomplish and facilitate mutual comparisons between lexicons, this approach clearly opens up the opportunity for mutual enrichment. Also, the creation of new lexical resources by combining data from different lexical databases is far more feasible as the result of the uniformation of the lexicons.

The final stage that needs to be made in this project is the online representation of the data. We believe that the uniform structure and the uniform feature value set will facilitate this. Also, the categorisation of the data into meaningful groups, i.e. the LMF components, will contribute to this final objective.

6. References

Francopoulu, G., M. George, N. Calzolari, M. Monachini N. Bel, M. Pet, C. Soria, (forthcoming 2006). Lexical Markup Framework. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, Genova.

ISO TC37 SC 4: 2005. *Lexical Markup Framework*..

Monachini, M. , F.Bertagna, N. Calzolari, N. Underwood, Navarretta (2003). *Towards a standard for the creation of lexica*.

Monachini, M., F. Calzolari, M. Mammini, S. Rossi, M. Uliveri (2004) Unifying lexicons in view of a phonological and morphological Lexical DB. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, Lisbon