Rik Schutz
Van Dale Lexicografie Utrecht

VLIS: *V*an Dale *L*exicographic *I*nformation *S*ystem

Abstract

VLIS is a multilingual database, designed at Van Dale to support the production and maintenance of mono– and bi–lingual dictionaries of current usage. The storage of the data is product–independent and thus ready for any (electronic) product form.

The VLIS project is planned to run from 1992 until 1995 and it involves (1) design and construction of the database and (2) storage of the content of six existing dictionaries in which Dutch is L1.

The aims of this paper are to:

- sketch the project as such; the history, the steps from dictionaries to database
- describe the VLIS–system, the building blocks and the relations between these building blocks
- discuss certain features of VLIS in comparison with the dictionaries from which the data were extracted
- describe the path from VLIS to a new dictionary.

Introduction

VLIS is a natural follow–up to the Van Dale series of bilingual dictionaries of current usage, developed in the early eighties. A basic assumption for this series was a uniform Dutch basis for all dictionaries in which Dutch would be L1. The structure of this paper is as follows:

1. The project. The background of the project and some practical questions, like the way the existing dictionaries were used as a source of information.
2. The database. The structure of VLIS as a database; the building blocks and their relations.
3. From Dictionaries to VLIS. Adapting the data to the VLIS structure.
4. Using VLIS. The way it works: how does one get access to the data and how will a dictionary be derived from the database?

1. The project

When Van Dale was founded as a publishing house in the late seventies, the idea was to create one lexicographical description of current Dutch (the native tongue of the prospective users) as L1 in all the active bilingual dictionaries. This would mean it would be possible to prevent the description of L1 being heavily influenced by L2, which is a common characteristic of bilingual dictionaries. Furthermore, using one and the same L1 description in four dictionaries would be cheaper than doing the same job four times. The uniformity was considered a strong selling point of the new series.

Between 1984 and 1986 four dictionaries appeared with a (more or less) identical Dutch L1. First the monolingual explanatory dictionary of Dutch was completed. Then we turned to the left–half of the bilingual dictionaries: abbreviating the definitions to produce meaning distinctions. Then three identical copies were made and handed over to the editors of the French, German and English bilinguals.

```
D–D ->        D–              -> D–F
                              -> D–G
                              -> D–E
```

Figure 1

Now, after ten years of modifying and updating on an ad hoc basis, the original uniformity is still obvious on the L1–side of the dictionaries, but there is no formal control. New editions and derivations were edited and we added another title (Dutch–Spanish). Two more are in preparation. This resulted in a great number of files in which practically identical L1–information is stored.

As an example I will describe the situation for Dutch–French, for which we find now, in 1994, basically the same L1–information is stored in six distinct files. From the comprehensive first edition a concise version was made. This version was adapted for the French user (Le Robert 1). In the concise files selection codes for a paperback were added, but that did not result in a separate file. For each of these dictionaries an updated second edition has appeared. All the alterations and additions were made in all the distinct files. This is a costly procedure, and the inefficiency increases with the number of titles and editions in which L1 reappears. For Dutch–French the picture is as in Fig. 2:

```
D–F 1 (1985)  → concise 1 (1988)  → Le Robert D–F 1 (1988)  → Le Robert D–F 2 (1993)
                                  → paperback 1 (1991)
                                  → concise 2 (1994)
                                  → paperback 2 (1994?)
              → D–F 2 (1991)
```

Figure 2

So the original idea of editing and storing the source language only once goes back to prehistoric Van Dale times. After a decade of commercially successful exploitation of recycled data, the strength of the original concept was recognized more strongly than ever. And the need for control of the recycling process was felt more strongly than before. It was the further development of this idea that resulted in VLIS, a Van Dale Lexicographic Information System in which all the information will be stored centrally.

We expect the following advantages of central storage:

- Product independency; immediate re-usability of data for other products
- Guaranteed uniformity in D-information in any Van Dale product; only one adaptation needed in the case of an (expected) modification of official spelling rules
- Efficient use of editorial capacity
- Identical procedures for deriving products (books or software) from VLIS save time and money
- Increase of editorial quality by enabling and encouraging editors to approach the information in various ways. Not just alphabetical, but per subject field, in semantic clusters, morphologically related etc.
- Increasing independence of Dutch as metalanguage and anticipating the use of the lexical data in NLP-systems by more or less formalizing the semantics.

The basis for both the design and the content of VLIS was the semantic/hierarchical database from which the Van Dale synonym dictionary originated. The full title of this book is 'dictionary of synonyms and other semantically related words'. A characteristic of this database is that single and multi-word entities are on an equal footing. Each single or multi-word term in the synonym database is related to the central term in a semantic cluster. Each of these central terms is related to its hypernym. It contains 40,000 lexical entities.

2. The database

The building block in VLIS is a Lexical Entity (LE), a combination of form and meaning. It is relatively easy to formalize form and it is very hard to formalize meaning. From reports on formalizing meaning with semantic features in the literature we learned that much pioneering work would have to be done. We decided that our existing dictionaries – which are widely praised for their quality and which were composed without a formal descriptive system – would be the basis for the structure in VLIS. This is especially useful since VLIS is intended to support the production of dictionaries like the existing ones in the near future. We decided to stay close to the data and the structure of what we had, instead of risking losing our way in search of the ideal lexical description of language. We tried to minimize the limitations of this choice and to design the system in such a way that adaptations can be made without the need to revise all the data.

We accepted that we could not distinguish meaning according to a hard and fast set of rules. Meaning distinction in polysemous entries would be, as before, edited by controlled intuition, with the help of a handful of more or less formal criteria.

A comparison between VLIS and the source dictionaries may serve to illustrate the VLIS structure:

1. The central entity is not a traditional keyword, but any single or multi–word combination of meaning and a fixed form: a Lexical Entity or LE. We distinguish several types of LE's. In the lemma *groen* we find five types:

Single word	*groen 1* (colour green)
Single word	*groen 2* (unripe)
Formula	*een oude bok lust nog wel een groen blaadje* (there's life in the old dog yet)
Simile	*zo groen als gras* (as green as grass)
Collocation	*op groen springen* (turn green)
Idiom	*een groene weduwe* (a grass widow)

2. Multi–word LE's (former examples) are classified as an idiom, a collocation, or a formula with a fixed form and a 'part of speech' code. Relations to smaller lexical units (usually a single word) are fixed.

op groen springen is a 'verb', a collocation, composed of groen and springen.
een groene weduwe is a 'noun', an idiom composed of groen and weduwe.

Example sentences that cannot be treated as a lexical unit are identified as textual illustrations of or example sentences to an LE.

3. Every LE is related to at least one other LE in the database. Relations can be semantic (synonym, hypernym, translation), formal (derivation, inflectional morphology, male/female) or structural (component of).

During the current VLIS project we do not try to fix all the possible relations between LE's. We stress semantic relations and use the other types as a kind of gap filler. Eventually the relational network will be completed. Examples:

Semantic relations for the single word noun LE *groen*

groen 2	SYNONYM	onrijp (unripe)
groen 3	ANTONYM	rijp (experienced)
groen 4	SYNONYM	*milieuvriendelijk* (ecological, organic)

Formal relations

groen *3* <adj.>	NOMINALISATION	*groentje*<n.> (greenhorn)
directeur <m.>	MALE <–> FEMALE	*directrice*<f.> (manager(ess))

Structural relations

groene weduwe	COMPONENT	*groen 1* (green) *weduwe* (widow)

In a printed dictionary a component will act as an entry or headword under which the multi–word lexeme can be found. We intend to select the entry component automatically according to a set of rules. The rules may vary according to the type of dictionary. (See also 5.)

4. As many Lexical Entities as possible are labelled with a UDC (universal decimal code) subject code. Thus terms that share a domain can easily be clustered. The numbers can of course easily be represented as an abbreviation of the subject. The code also helps in editing and recognizing polysemy:

mol	<zool.>	(mole)
	<muz.>	(flat)
	<chem.>	(mol(e))

5. Lexicographical decisions for a certain dictionary derived from VLIS will be taken on the level of the whole dictionary. For example: in many existing dictionaries the user is uncertain about the entry under which a multi–word lexeme is to be found. In a general bilingual dictionary derived from VLIS all idioms will be treated under the headword corresponding with the first noun in the idiom; Adj+N collocations will be treated under both the adjective and the noun (or treated under the noun with a cross reference from the adjective).

6. Both an LE and a relation can be specified by additional information. Part of speech, inflectional forms and the labels from the dictionary will often end up as a piece of additional information, specifying either an LE or a relation between two LE's.

7. During the first phase of the project, translations are treated as lexical forms, exactly in the form in which they occur as L2 translations in the

Dutch–Foreign language dictionaries. So the four Dutch LE's that share the form *groen* all relate to the English form *green*. But this does not provide a total picture of the polysemy of *green* in English, of course. It is likely that *green* is just as polysemous as its Dutch counterpart. By interweaving the L1 information from the English–Dutch dictionary, the semantic profile of the English form *green* becomes explicit. Only then can the distinct combinations of form and meaning of English Lexical Entities be distinguished.

Once these are included in VLIS a maximal equivalence between Dutch–English and English–Dutch will be achieved. A translation relation will then be identified as uni– or bidirectional. It will be possible for example to suggest *immature* as a translation of *groen*, without giving *groen* as a translation of *immature*.

It is likely that in the future more information will be added that is not (explicitly) found in the existing dictionaries. Argument structures, selectional restrictions and formalized semantic features belong to this category. As pointed out above, for the time being VLIS is essentially a system to support the production of traditional and electronic dictionaries.

3. From dictionaries to VLIS

Once the structure of the database was laid out, the synonym database was modified accordingly and the data in the dictionary files were adapted to that structure.

From the dictionaries Dutch–Dutch and the four bilingual dictionaries together approximately 60,000 word meanings and 100,000 example sentences were available and not yet included in the synonym database.

The route from the dictionaries to VLIS will be illustrated with abbreviated entries of dictionary articles from Dutch–English (Fig. 3)

Dutch-English
groen <adj.>
0.1 [kleur] ***green***
0.2 [onrijp] <+ fig.> ***green***
0.3 [milieuvriendelijk] ***green***
♦
1.1 een oude bok lust nog wel een groen blaadje
there's life in the old dog yet
1.1 (iem.) het groene licht geven (om . . .)
give (s.o.) the g. light/go-ahead (to . . .)
2.1 groen en geel worden van nijd
→ nijd
8.2 zo groen als gras
as g. as grass
1.¶ een groene weduwe
a grass widow

weduwe <f>
0.1 ***widow***
♦
2.1 groene weduwe
frustrated housewife

Figure 3

Each item (single or multi–word) from each article was exposed to the critical eye of an editor before it was allowed to enter the database. The editing involved:

- Clearing the data. Items that were obsolete or otherwise undesirable were coded for deletion. The acceptable ones became a Lexical Entity (LE) in VLIS.
- Coding the part of speech (for multi–word LE's; for single–word LE's this information was already available in the dictionary)
- Coding the kind of LE (single word, idiom, collocation, simile etc.).
- Choosing a standard form for variation in the sources (for the single word LE *wc* the following forms were available *W.C./WC/w.c./wc*; a multi–word example is *het groene licht geven/iem. het groene licht geven om*).
- Relating each LE semantically to at least one other LE.
- Relating multi–word LE's to single word components.
- Adding subject codes (in UDC).

For each potentially new LE the form of the entry word plus the definition from the monolingual dictionary was automatically added to the synonym database. The editing took place in the database via specially designed screens. For the entry *groen* this editing resulted in four single word Lexical Entities:

groen 1			(colour)
groen 2	SYNONYM	of onrijp	(unripe)
groen 3	SYNONYM	of onervaren	(unexperienced)
groen 4	SYNONYM	of milieuvriendelijk	(ecological; organic)

Note that some text strings that indicate the meaning distinction in D–E became formal semantic relations in VLIS and that meaning 0.2 was divided, according to different synonyms and distinct selectional restrictions: fruit vs. people.

The second half of the lemma (in a Van Dale dictionary the part where the phraseology is to be found) was treated in a different way. To be able to bring together similar examples from the various source dictionaries, we used a trick. Each example sentence from each of the six sources was automatically reduced to a temporary 'kernel' by removing very frequent words. For example all articles, propositions, pronouns and auxiliary verbs were removed from the text. By clustering identical kernels automatically, examples from different dictionaries and/or different entries could be brought together, even if they were not literally identical. The kernel only served to make the clustering possible. It played no role after the clustering took place.

One could say that the phraseology in the existing dictionaries was used as a corpus.

groene weduwe	[KERNEL]
een groene weduwe	[example from D–D under the entry groen]
een groene weduwe	[example from D–E under the entry groen]
groene weduwe	[example from D–E under the entry weduwe]

After editing this raw material in the form of examples for VLIS it resulted in:

FORM	*een groene weduwe* (a grass widow)
POS	noun
SORT-LE	IDIOM
HYPERNYM	*echtgenote* (wife)
SPECIFICATION	<in buitenwijk> (in suburb)

After this editorial job was carried out – on text files in an ordinary word processor – we had a neat collection of various kinds of lexical units to be loaded in the VLIS database. Because the original form of the example was preserved, together with an identification tag, the translations could be collected from the source dictionary file and related to the Dutch LE.

een groene weduwe	E: a grass widow
	E: a frustrated housewife

If the standard form was not exactly identical to the original example, the translation had to be adapted to the standard form. If the necessary alterations were unacceptable for the translating editor, for instance because of some resulting contradiction or inconsistency,, the original form was preserved as an intermediate form between the LE and the translation.

One troublesome job left to do was decoding the typographically compressed information in L2. For example:

give the g. light/go–ahead to . . .	*give the* | *green light to . . .*
	give the | *go–ahead to . . .*

In the end, though, we managed to load all the Dutch multi–word lexical entities plus their translation equivalents from the four bilingual dictionaries into VLIS.

4. Using VLIS

The stereotypical user of VLIS is an editor of a Van Dale dictionary. Passwords ensure that only authorized people can change the data. VLIS automatically documents the date and the name of the editor. He/she will select information for further editing in one of the following ways:

- Separately. Every single LE can be selected and edited.
- Alphabetically/traditionally. A lemma window represents the information in the order of a traditional dictionary. Per headword meanings are distinguished and per meaning the related multi–word LE's (collocations, idioms) are accessible. A filter enables the editor to choose between all the available VLIS information or the items earlier selected for a specific L2 or dictionary.
- Semantically. Per cluster of related LE's: *green, ecologic(al), organic,* etc. can be accessed and edited simultaneously via a shared synonym or hypernym.
- Per subject. The UDC coding enables clustering of LE's that have a subject field in common. All medical words, or even surgical terminology, can then be edited together.
- According to part of speech (all pronouns or numerals).
- According to labels. These may be stylistic, like <formal> and <vulg.>, or regional.

- In fact any kind of information that is available in VLIS can be used as a criterion for selecting and sorting data. Thus if someone wishes to work on monosyllabic adjectives ending in *-s*, he/she can do so, in principle.

To finish this brief introduction to a never complete, but already operational database, I will describe how we hope to derive dictionaries from it.

As an example I will take a hypothetical Dutch–Turkish dictionary. First we would define the characteristics of the users. We assume that native speakers of Turkish and Dutch, both living in the Netherlands and Belgium, will use it. The size is limited to 25,000 entries, for commercial reasons.

So the preselection of potential LE's is our basic vocabulary with the addition of terms from the social security system, names and abbreviations of official bodies, and frequently used legal terms from The Netherlands and Belgium. No vulgarisms, no obsolete or archaic terms. All this can be automatically selected by using VLIS codes and labels.

For each of these terms the additional information to appear in the dictionary can be defined and selected: pronunciation, regular or irregular inflection, grammatical collocations, (very frequently used) idioms.

We assume that this basic selection, which can be accessed directly in VLIS through a specially designed D–T filter, provides the starting point for the manual part of the editorial work. The final selection will be made, contextual illustrations (example sentences) will be added and Turkish equivalents will be added to VLIS for the selected LE's.

The entry under which multi–word LE's will be found in the dictionary is a matter of rules. The decision whether to give a multi–word LE under each component will depend on considerations of space. In a small dictionary for inexperienced dictionary users we will give and translate each multi–word LE under only one headword and give many cross references. Whether an idiom like *groene weduwe* belongs under the adjective or the noun is a decision that is not taken for this one fixed phrase, but for every similar LE. If we choose the first noun for any type of LE, the entry *groen* would look like fig. 4

groen [xrun] <adj.; ~er, ~st>		
0.1	[kleur]	***yeşil***
0.2	[onrijp]	. . .
0.3	[onervaren]	. . .
0.4	[milieuvriendelijk]	. . .
♦		
2.1	groen en geel zien van . . .	*Turkish translation*
8.2	zo groen als gras	*Turkish translation*
→	**bok, weduwe**	

Figure 4

By using the selection codes for D–T the derivation of the dictionary results in an (SGML–like) structured text file. The structure ensures a simple phototypesetting procedure. Automatic spatial reduction can be achieved, for example by compressing the four meanings into one, if they would share my unauthorized translation *yesil.*

groen [xrun] <adj.; ~er, ~st> 1 ***yeşil*** <kleur; onrijp; onervaren; milieuvriendelijk> ♦ ~ *en geel zien van . . .; zo ~ als gras . . .* → **bok, weduwe.**

Figure 5

Of course a simulator of the result in print will have to be available within VLIS, so that the selection does not need any editorial intervention, after derivation from VLIS. Only when the result is given the green light, does the actual derivation take place. The selection then goes rapidly via phototypesetting, printers and bookbinders to the bookshops.
For the near future we foresee the following VLIS–Products:

D–D explanatory, combinatorial, idioms, thesaurus, specialized (medical, finance), neologisms

D–Foreign bilingual dictionaries, idioms, specialized (medical, finance)

All of these can be produced in any size and in print or in any electronic format, according to the needs of the market.

Conceivable VLIS–Products in the long term could be:

Foreign–D The Foreign language–Dutch dictionaries, complementary to the Dutch–Foreign language volumes that are being incorporated into VLIS during the current project exist as structured data files. The addition of these is foreseen and will be a matter of time (and budget).

Foreignl–F2 Deriving bilingual dictionaries that do not include Dutch, for example Spanish–German, is a more complicated matter. Of course, this could never be done by just pushing the right button. It will involve, like in any other decent dictionary project, a carefully executed editorial programme of selection and addition. But we believe that VLIS could be a valuable source of up–to–date and well–structured information.

References

Bogaards, Paul 1991. *Groot woordenboek Nederlands–Frans.* Utrecht: Van Dale Lexicografie.

Cox, Heinz L. 1991. *Groot woordenboek Nederlands–Duits.* Utrecht: Van Dale Lexicografie.

Martin, Willy. and Tops, Guy A.J. 1991. *Groot woordenboek Nederlands–Engels.* Utrecht: Van Dale Lexicografie.

Slagter, Peter J. 1992. *HandwoordenboekNederlands–Spaans.* Utrecht: Van Dale Lexicografie.

Sterkenburg, Piet G.J. van 1991. *Groot woordenboek van synoniemen en andere betekenisverwante woorden.* Utrecht: Van Dale Lexicografie.

Sterkenburg, Piet G.J. van 1991. *Groot woordenboek van hedendaags Nederlands.* Utrecht: Van Dale Lexicografie.

Sterkenburg, Piet G.J. van 1992. "Electronic onomasiology" in H. Tommola (eds.), *Euralex Proceedings '92.* Tampere: Tampereen Yliopisto.