Adding Electronic Value.
The electronic version of the *Grote Van Dale*

Dirk GEERAERTS, Leuven, Belgium

Abstract
The paper describes the design features of the elektronische grote van dale or EGVD , i.e. the electronic version of the most authoritative dictionary of contemporary Dutch. A comparison between the EGVD and existing dictionaries on cd-rom leads to an identification of the innovative characteristics of the EGVD. Ranking high among these are a fully developed onomasiological search function, the layered representation of entries, and the use of generated speech for the multimedia representation of pronunciation. The design features of the EGVD are analyzed from a functional perspective: what functional advantages can electronic dictionaries achieve, in comparison with the paper dictionaries that they are based on?

1 Introduction

What value can electronic dictionaries add to their paper counterparts? Although there is a steadily growing market of dictionaries on cd-rom (and electronic dictionaries at large), the functional advantages that may be achieved for the dictionary user when existing dictionaries are converted into an electronic format are only sporadically analyzed in the lexicographical literature. In an attempt to do something about this relative neglect in the literature, I will present the design of one such dictionary on cd-rom that is currently being produced, the *Elektronische Grote Van Dale* or EGVD. I will try to transcend the level of a mere product presentation, by highlighting those features that are functional improvements arising from the use of an electronic medium, and by comparing the EGVD with existing dictionaries on cd-rom, concentrating on the features that are innovations, or that further elaborate innovative functions in existing electronic dictionaries.

The *Grote Van Dale* ("big Van Dale": GVD for short, *Van Dale Groot Woordenboek der Nederlandse Taal* in full) is the main dictionary of contemporary Dutch, universally acknowledged by the Dutch-speaking language community in The Netherlands and Belgium as the major reference point for linguistic (or at least lexicographical) matters concerning contemporary Dutch. In size, the GVD is comparable to the *New Shorter Oxford English Dictionary* (henceforth nsoed): a comparison of the paper versions, based on the number of characters per page and the number of pages, reveals that they contain roughly the same amount of information (or at least, the same amount of text). Their descriptive scope is different, however: whereas the nsoed is essentially a historical dictionary, covering a period of some twelve centuries, the GVD is basically a synchronic dictionary of 20th century Dutch, containing a fair amount of less frequent, technical, learned, specialized, or even obsolescent vocabulary. In this respect, the GVD rather resembles the *Oxford Dictionary of Contemporary English*, except for the difference in size and descriptive detail.

The first edition of the GVD appeared in 1864. The 13th edition appeared in a traditional paper version in September 1999 and is now being published in an electronic version. This is not the

first electronic product for the Van Dale company, but given the flagship status of the GVD, special attention is being given to the cd-rom version of this authoritative dictionary.

In order to bring out the features of the EGVD as clearly as possible, I will compare it to a number of existing dictionaries on cd-rom. Here is a list of the dictionaries serving as points of comparison, with the abbreviations that I will use to refer to them:

NSOED	*New Shorter Oxford English Dictionary*	[1996]
RHW	*Random House Webster's Unabridged Dictionary*	[1996]
WNT	*Woordenboek der Nederlandsche Taal op cd-rom*	[1995]
NE	*Van Dale Groot Woordenboek Nederlands-Engels*	[1997]
GR	*Le Grand Robert*	[1994]

The French and English dictionaries in this sample probably need no further introduction. With regard to the Dutch ones, a word of explanation may be useful. The *Woordenboek der Nederlandsche Taal* is, both in size and in purpose, the Dutch counterpart of the *Oxford English Dictionary.* The *Van Dale Groot Woordenboek Nederlands-Engels* is a Dutch-English desk dictionary for translation (it is used here as a representative of a set of dictionaries for translation, going from Dutch to English, French, and German, and vice versa).

2 Overview

A systematic analysis of the design features of a dictionary on cd-rom involves two different levels: the *technical* level, involving the background programming techniques and the technological characteristics of the user interface, and the *functional* level, involving the purposes that the dictionary intends to serve for its users. The former level takes into account questions such as whether the dictionary uses multimedia, to what extent it contains hyperlinks, or how it adheres to Windows standards. The latter level explores the functions that are supported by the technical design of the dictionary. Focussing on the technological characteristics is appealing but misleading: technological fireworks that are not motivated by functional considerations may contribute to the immediate appeal of the dictionary, but ultimately, they are superfluous as much as they are superficial.

Therefore, rather than merely listing the technical characteristics of the EGVD , I will organize my analysis on the basis of the functions that the EGVD purports to serve, and show how the technological innovations contribute to the way in which these functions are fulfilled. There are three basic functions: a semasiological one, an onomasiological one, and an edutainment function. The fact itself that these three functions are combined in a single dictionary is not an entire novelty, but it does illustrate the *enhanced multifunctionality* that can be achieved in an electronic dictionary. In what follows, I will deal with each of the three functions in more detail, with special emphasis on the way in which the EGVD adds new features to the overall possibilities of dictionaries on cd-rom. Saving the best wine (in our case, the most innovative feature) for the last, I will save the onomasiological function for the end, and start with a discussion of the semasiological function. I will say next to nothing about features that are more or less standard in software applications, such as the availability of on-line help or the possibility of customizing the display of the application.

3 Semasiology

The technologically supported innovations in the semasiological part of the dictionary fall into three classes: the *addition* of new information (3.1), the macrostructural *retrieval* of lexical entities (3.2), and the microstructural *presentation* of data (3.3).

3.1

In the EGVD , two types of information are added in comparison with the paper dictionary. First, the EGVD introduces multimedia in the form of an *audio representation of the pronunciation* of the headwords. In contrast with those dictionaries on cd-rom that include actual recordings of pronunciation (a technique that is expensive and that requires quite a lot of disk space), the multimedia representation of the pronunciation of all 250,000 EGVD headwords takes the form of diphone-based generated speech, taking as its input the phonetic transcription of the words.

Second, adding information is facilitated by the pop-up functionalities of the electronic product. Information that would take up too much space in the classical dictionary, can now be made accessible from within an entry without disturbing the layout of the article. The EGVD intends to use this possibility for the addition of *inflectional paradigms*: the inflectional expansions of the headwords are shown in a separate pop-up window. This addition is useful not just for the inflections themselves (which often take an irregular form), but also for difficulties of language use associated with these forms, like hyphenation. The morphological pop-ups are enriched with this type of data, together with an additional type of enrichment: where necessary, explanatory text specifies the usage characteristics of the inflected forms. For instance, even for an abstract word like *boosheid* “anger”, the plural form is included in the morphological pop-up window; because *boosheid* does not conventionally occur in a concrete and countable sense, the pop-up window explains under which circumstances abstract words may be used in the plural (e.g. when they are used to refer to a specific type of anger, or metonymically to refer to an instance of angry behaviour).

In this way, the differences between dictionary and grammar begin to diminish: the dictionary entries are linked to a grammatical description of the language that offers more detail than the grammatical compendium that is sometimes included with paper dictionaries. It should be added, though, that this feature (the grammatical pop-up windows) will not yet be implemented in the first release of the EGVD. It is projected as one of the features to be added for release 1.1 or 1.2.

3.2

The retrieval of relevant information is electronically optimized by enhancing the accessibility of entries. One aspect of this type of improvement is the *internal cross-referencing* of the dictionary: any word within an entry may be immediately looked up in the dictionary, so that difficulties of understanding relating to the definitions of words may be minimized.

Another form of improved retrieval is the inclusion of a ‘*suggestive search*’ function, i.e. a func function that generates suggestions when the search string as typed by the dictionary user does yield direct hits. In the EGVD , this function will basically take into account three different types

of deviation from the potential targets: spelling corrections, near-homophones, and inflectional forms.

A final example of improved retrieval involves the introduction of a *phraseological search dialogue*. Phraseological units like idioms and proverbs constitute a typical difficulty for the traditional dictionary: finding adequate criteria for deciding where to include them is not easy, and communicating those principles to the average dictionary user may be even more difficult. In the electronic dictionary, by contrast, accessing phraseologisms is made easy by the introduction of a phraseological dialogue box, allowing the user to directly access the phraseological units in the dictionary, on the basis of the form of the expressions, or on the basis of the type of expression.

3.3

Finally, optimized microstructural representation is achieved through the *layered representation of entries*. In a dictionary of the size of the GVD, entries may be quite long, with a complex internal structure. In order to facilitate finding one's way through the entries, the entries may be accessed at different levels. There are four levels:

- the headword level, containing information about spelling, pronunciation, hyphenation, grammatical and morphological characteristics, and etymology
- the level of senses, i.e. the definitions constituting the semantic backbone of the dictionary
- the level of nuances and phraseological units (collocations, idioms, proverbs etc.) that belong with a given sense
- the quotations and examples that illustrate senses, or nuances, or phraseological entities.

The layered representation implies that the dictionary user may choose at which level he wishes to see an entry. If he is only interested in etymological information, for instance, he may restrict the representation to the headword level. Or if he is interested in a quick overview of the semantic potentialities of a word, he may decide to see only the sense level. The layered representation is a contextualized one, to the extent that the switch from one level to another may not just occur globally (i.e. for the article as a whole), but may also be achieved locally (i.e. from within a single structural entity). This means, for instance, that the user may successively 'peel off' the internal structure of a specific sense, going from the main definition to the nuances and phraseological units listed under that meaning, and from there exploring what specific illustrative quotes or examples are given for, say, a specific expression.

	NSOED	RHW	WNT	NE	GR	EGVD
multimedia pronunciation	-	+	-	-	-	+
inflectional pop-ups	-	-	-	+	+	+
internal cross-referencing	+	+	+	+	+	+
suggestive search	-	-	-	+	+	+
phraseological search dialogue	+	-	-	-	-	+
layered representation of entries	+	+	-	+	+	+

Table 1: Available functions in different dictionaries

Table 1 charts to what extent functions similar to the ones described here are present in the dictionaries on cd-rom included in the sample serving as point of comparison. It does not show to what extent the EGVD improves on the way in which these functions are realized in the other dictionaries. Here is an indication of the restrictions on the way in which the functions are implemented in the dictionaries included in Table 1:

- multimedia representation of pronunciation: RHW contains recorded speech only
- pop-up windows with inflectional paradigms: although GR and NE contain separately accessible inflectional paradigms, they do not develop it with the type of added explanatory text (the mini-grammar) described earlier
- suggestive search: the existing functions make use of orthographic (and to a limited extent, morphological) information, but do not enrich the search function by making use of pronunciation information; in this way, for instance, NE is not able to relate the (frequent but unofficial) spelling *kado* to the correct form *cadeau*, in spite of identical pronunciation
- phraseological search dialogue: the search function in NSOED is basically an alphabetical list of phrases and compound words, with no added search functions, like Boolean operators or the possibility of querying specific types of expressions; looking for *barrel*, for instance, will yield *barrel-chested, barrel-fish, barrel-house, barrel-organ, barrelvault* but not *a barrel full of fun*, which is to be found under *a*
- layered representation: many of the dictionaries in the sample allow the dictionary user to determine which elements from the entries he wishes to see, for instance by selecting or deselecting definition fields, etymology, pronunciation, or quotes and examples; but they do not organize this function in such a way that the user can locally probe ('peel off') an article with increasing depth of detail.

4 Edutainment

Dictionaries are not just used for professional purposes: they are not just consulted, but they may also simply be read for the sheer pleasure of learning about the language. Apart from such browsing through dictionaries for the mere fun of it, the edutainment function of the dictionary may involve more specific goals, like searching for a word that fits a crossword pattern defining a word with ten letters of which the last one happens to be *y*. The formal possibilities of the electronic dictionary have greatly enhanced the support offered for the latter type of function. Typical features of electronic dictionaries with specific appeal for scrabble players, lovers of language games, and aficionados of linguistic curiosa, include the following: search functions featuring wildcards, reverse indexes, and anagrams. (The first two could, of course, also be included in the group of mechanisms enhancing the accessibility of lexical entries, as mentioned above.) In the sample, these functions are fairly well represented, as shown by Table 2.

In the EGVD , these functions will be supplemented in a number of ways. I will skip, for the moment, the possibilities offered by the onomasiological part of the dictionary as described in the following paragraph. We will see there how the dictionary supports, for instance, a structured search for etymological information, or a multi-indexed search for expressions from a specific register. Apart from such functionalities, which fall out automatically from the onomasiological search engine to be described below, the EGVD tries to improve on existing functions.

	NSOED	RHW	WNT	NE	GR	EGVD
reverse index	-	-	-	-	-	+
wildcards	+	+	+	+	+	+
anagrams	+	+	-	+	-	+

Table 2: Additional Features of electronic dictionaries, e.g. for scrabble players

Specifically, the wildcard search function goes beyond the standard possibilities involving any string (*) or a single character (?), by offering a wildcard for consonants versus vowels and the possibility of searching for the string *ij* considered as a single character. The latter function particularly illustrates the edutainment purposes of the dictionary. The string *ij*, representing the diphthong [EŐ], is normally treated on a par with other digraphic signs like *eu* [è] or *ui* [¿y]. In Dutch scrabble games, however, *ij* is a single chip, representing a single character. As a gadget for scrabble players (a group of heavy users as far as dictionaries are concerned), the EGVD introduces the option of considering *ij* a single character in wildcard searches.

As a major addition, the EGVD contains a full-fledged *rhyming dictionary*. In the cultural context of The Netherlands, this is not a function with limited popular appeal, as one might possibly think: making verses to accompany presents given at the occasion of the feast of St Nicholas (Santa Claus, if you wish), is a widespread tribal rite of the Dutch. Now, how is the rhyming dictionary realized? The phonetic representation that is part of the headword information, is input for a rhyming function that yields a structured set of rhyming words. The notion 'structured set' is not without importance: the rhyming words are ordered in a way reflecting their rhyming value. For instance, they are ordered according to the number of syllables they contain, because versification theory, as based on the classical definition of meter, distinguishes between a monosyllabic set like *bot, pot, rot* and the bisyllabic *kalot, kapot*. Similarly, the electronic rhyming dictionary has to take into account the accentuation pattern of the words: *bóycot* is not a true rhyme of *kapót, kalót*, and the dictionary user has to be told so.

Further, the rhyming dictionary is not just based on the headwords of the EGVD , but it also takes into account the inflected forms of these words. With the inflected form *zotte* of the adjective *zot* as input, for instance, one will get *botte, rotte, kapotte* (but not **potte* or **kalotte*, for these are non-existing forms of the nouns *pot, kalot*). There are two ways, then, in which the EGVD rhyming dictionary extends the possibilities of the rhyming function in the NSOED. When asking for rhyming words for *treat*, for instance, one gets a series like *treat, undertreat, metrete, retreat, street, by-street*, in which neither the number of syllables nor the accentuation pattern is taken into account. And when asking for an inflected form like *treats*, the NSOED does not give the relevant forms, like *retreats*.

5 Onomasiology

Combining a semasiological with an onomasiological dictionary is an old dream of lexicographers. Usually, this is taken to mean the combination in a single product of two separate dictionaries – basically, an alphabetically ordered semasiological dictionary and a hierarchically, or

at least thematically ordered synonym dictionary or thesaurus. A prime example of a paper dictionary achieving this aim is the *Oxford Dictionary and Thesaurus*, in which each alphabetical headword may be accompanied by a hierarchically ordered set of synonyms and near-synonyms. In the electronic domain, the *Van Dale Groot Elektronisch Woordenboek Hedendaags Nederlands* (an electronic dictionary in the same series as NE) combines the *Van Dale Hedendaags Nederlands* (a single volume alphabetical dictionary of contemporary Dutch) with the *Van Dale Synoniemenwoordenboek* (a hierarchically ordered synonym dictionary): any word in the alphabetical dictionary is cross-referenced to the synonym dictionary, and vice versa. In both of these cases, the combination of onomasiological and semasiological functions in a single dictionary is achieved through linking of separate dictionaries.

There is, however, an entirely different perspective that may be taken with regard to the question of how to combine a semasiological and an onomasiological dictionary: why not develop a function that allows one to search the semasiological dictionary in an onomasiological direction? Basically, this involves searching from meanings to forms, or more generally and more practically defined, from within an entry to the headword. This new perspective naturally leads to further refinements. On the one hand, the onomasiological target may be defined more broadly than just 'headwords': what a dictionary user may be looking for is any type of lexical entity, regardless of whether it takes the form of a single word or not. Basically, this includes all types of idiomatic expressions next to single words, but one may also think of quotations as a searchable type of lexical entity. On the other hand, the kind of features that form the source of the onomasiological search, should be systematically identified. These features involve the different types of information that characterize lexical items (either words or expressions), like definitions, labels, synonyms and antonyms, or etymological data. Each of these features can be input for an onomasiological query, i.e. for a search for a lexical means of expression exhibiting the desired features.

From a slightly more technical perspective, this function is but a further step in the development towards *structured full text searches*. The rather coarse function represented by a full text search has been more and more refined in recent dictionaries on cd-rom. A few examples may suffice to illustrate the point. In the NSOED, the full-text search function can be restricted to searches within definitions, and/or within quotations, and/or within the etymological part of the dictionary. This type of design insufficiently distinguishes between what one is looking for, and on the basis of what one is looking for it (searching within definitions is searching for words conforming to that definition, but searching within 'quotation text' is probably searching for the quotation itself, and not for the word under which it is filed). Still, it illustrates what is meant by a structured search: the search operation is limited to (combinations of) specific parts of the dictionary. The structuration goes, in fact, one step further: when searching the etymological information, specific queries based on the names of languages are possible, and quotations may be identified by author or work quoted. Another type of structural restriction can be found in the Van Dale dictionaries for translation on cd-rom (like the NE), which include the possibility of restricting the search to a specific word class, or to words that are either British English or American English.

These examples show that the idea of a structured search, while somehow being 'in the air', has not yet been developed to its logical conclusion, viz. that of making a principled distinction beinction between the lexical entities one could be looking for, and the features on the basis

aan de hand van:	Zoek naar WOORDEN	VERBINDINGEN	CITATEN
FORM			
SOORT			
VERSPREIDING/STIJLWAARDE			
VERKLARING			
ETYMOLOGIE/HERKOMST			

Table 3: Onomasiological Search Matrix

of which one may start looking for them; and at the same time, that of systematizing the different types of characteristics that may be input to the onomasiological search. In an attempt to carry through this logical step, the EGVD will contain an *onomasiological search matrix* as schematically represented in Table 3. Technically speaking, this type of search matrix is based on a mechanism of multi-indexing, allowing Boolean searches across different indexes (like an index of words with a specific label, or an index on the definitions of idiomatic expressions).

At the interface level, the matrix distinguishes horizontally between three types of onomasiological targets: headwords, idiomatic expressions, and quotations. (The latter are primarily literary quotations.) Along the vertical dimension, different types of query features are specified.

- VORM: involves the form of the search targets, like looking for words ending in *-heid,* or looking for expressions containing the words *appel* "apple" or *citroen* "lemon".
- SOORT: relates to the restricted set of types that are customarily defined for words and idiomatic expressions: word classes on the one hand, types of idioms (like proverbs or formulaic expressions) on the other.
- VERSPREIDING/STIJLWAARDE: is based on the different kinds of labels that specify the variational distribution (*verspreiding*) or the stylistic value (*stijlwaarde*) of the search targets. These include labels like (the counterparts of) *formal, vulgar, jocular, euphemistic, offensive, Belgian Dutch, army slang,* and so on.
- VERKLARING: supports searches in the definitions and explanatory remarks that accompany words and phrases. It allows for queries of the type 'Give me all the words that contain, in their definition, the string *hond* "dog" or *honden* "dogs"'. Synonyms are also included.
- ETYMOLOGIE/HERKOMST: takes into account the etymology of the words, and the origin (*herkomst*) of the quotations if the search is for quotations rather than words. An etymology-based query may be further structured by references to the languages mentioned in the etymologies or to the period of the oldest attestation of the word. The origin of the quotations, needless to say, refers to the author or the source from which they are taken.

Each entry field in the matrix allows for Boolean expressions. The entry fields may be filled in directly, or through the intermediary of a dialogue window specifying the possibilities. (The dialogue box for searches based on the type of word, for instance, contains a list of word classes.)

It is not without importance to note that the basic search result is not just a word, or an expression, but a word or an expression in a specific meaning. If you define a query for a word containing the words *hond* “dog” or *honden* “dogs” in its definition, you only want to see the sense(s) whose definitions refer to dogs; any other meanings of the words in question are basin are basically irrelevant.

The onomasiological search matrix is an innovative and powerful tool for probing the dictionary. It serves functions that rather belong in the edutainment sphere (like etymology-based queries), but above all, it enables the dictionary user to actively look for linguistic means of expression. Given a certain flexibility and inventiveness in formulating queries on the part of the user, highly specific questions like the following may now be easily answered.

- What is the specific name of those heavy towers that one finds in medieval castles and fortresses? Looking for definitions containing *toren* “tower” and *burcht* “fortress” or *kasteel* “castle” leads to the word *donjon.*
- What is the specific name of those heavy towers that one finds in medieval castles and fortresses? Looking for definitions containing *toren* “tower” and *burcht* “fortress” or *kasteel* “castle” leads to the word *donjon.*
- Does Belgian Dutch have a specific variant for the word *overgordijnen* ? Looking for words with the label *Belg.* and with *overgordijnen* as part of the definition may lead to *draperie*.
- Are there any euphemisms for talking about cancer? Looking for the label *euf.* and the string *kanker* on the definitional level yields the abbreviated, euphemistic form *K*.

6 Conclusion

The design features that I have highlighted in my presentation of the EGVD all illustrate the functional progress that can be made by electronic dictionaries. These features include

- enhanced accessibility and retrieval
- a maximal exploitation of the representational flexibility of the electronic format, viz. in the form of a layered representation of the dictionary entries and the introduction of grammatical information in the form of pop-up windows
- the inclusion of multimedia in the form of audible pronunciations
- the incorporation of a rhyming dictionary based on the phonetic transcriptions of the pronunciations
- a structured onomasiological query function
- overall, an enhanced multifunctionality of the lexicographical product.

Many of these features are further developments of functions that are already present in other electronic dictionaries; others constitute more radical innovations. The onomasiological search matrix in particular constitutes an interesting new perspective on the concept of a full text search. By highlighting that the dictionary is ultimately a relational database that can be accessed from different directions – from words to meanings, and from characteristics of lexical

entities to those entities themselves, the onomasiological search matrix may well turn out to be a major step forward in our conception of what digital dictionaries could be.

Overall, the development of electronic dictionaries may now be characterized as a major process of *integration*: the explanatory dictionary is integrated with special purpose dictionaries like a rhyming dictionary or a reverse dictionary, the dictionary as such is integrated with the grammar, and the semasiological dictionary is integrated with the onomasiological one. And with further releases of the electronic dictionaries ahead, this process is not likely to end here...