On the organization of semantic data in passive bilingual dictionaries

Willy Martin

Introduction

In this paper I will proceed according to a very simple and well-known principle, viz. moving from the *more general* towards the *more particular.*

As a consequence I will,

— first of all, say a couple of words about the *lexicographical landscape* as it looks like now, in the nineties;

— secondly, I will try to give you an idea of what the *Van Dale-dictionaries* are like, more in particular what the Van Dale bilingual passive dictionaries are like;

— and thirdly and lastly, I will take up the *semantics* in the above mentioned dictionaries or at least elaborate upon some organizational aspects of the latter.

1. The changing lexicographical landscape

If one considers lexicography, in the traditional sense of the word, to be the description of (parts and/or aspects of) the lexicon of a language in a dictionary for human users, then it will be obvious that phenomena such as *Natural Language Processing* have, at least, entailed a change in focus, a widening of the field. In other words, with the breakthrough of the idea of the centrality of a lexical component in large, robust NLP-systems —a breakthrough which has been felt from the beginning of the 80's onwards— both the objects of interest for lexicography and the ways of describing the lexicon and aspects thereof changed.

Not being able in this context to deal with these changes in great detail[1] I will restrict myself to the enumeration of what I consider to be the most important ones:

a. Such as Fig. 1 below illustrates, the objects of interest for lexicography (both computational and otherwise) are much wider than they were before, implying next to dictionaries for human users (D's) such objects as:

— computer-based dictionaries (CBD's)
— machine-readable dictionaries (MRD's)
— lexical/termbanks (L/TB's)
— machine dictionaries (MD's)
— lexical databases (LDB's)
— and artificial intelligence lexicons (AIL's)

[1] For a more detailed report see Martin-Woltering 1989, from which Fig. 1 is taken.

b. Although the distinctions between the objects mentioned, unlike the linear representation may suggest, are gradual rather than disjunctive, one can discern within this set two subsets or families, each with a maximally different prototype, viz. dictionaries on the one hand, vs. lexical databases on the other.

c. Actually the main differences between D's and LDB's can be characterized as follows:

1. As an LDB can be defined as the lexical component of an NLP-system *in general*, the processes it is intended to perform will neither be *a priori* defined nor restricted. In other words such a component is not/should not be oriented towards one (or more) particular task(s). Rather, given certain tasks, an appropriate selection will be made from the knowledge the database contains. D's on the other hand, by the very fact that they, often for practical reasons, are obliged to orientate themselves towards certain users to perform certain tasks, are, as a rule, not neutral but user-oriented (see Martin-Al, 1990).
2. D's, in general, are less well-formalized than LDB's are, the latter demanding by definition a database form. Next to that LDB's do not only imply data, but retrieval procedures as well.
3. D's are not, or at least not explicitly, organized with regard to semantics, whereas this structural organization is/should be the most critical feature in LDB's. This does not mean that there is no linguistic/semantic structuring at all in D's (see e.g. Schnelle, 1990), only this feature is not 'foregrounded', or to quote Beubert: «Im Wörterbuch wird wohl nicht *vordergründlich* (my italics) ein System des Wortschatzes entworfen, doch widerspiegeln die Stichwörter und Stichwortnester zweifellos systemische Beziehungen zwischen den lexikalisch-semantischen Varianten» (Neubert, 1977). In Martin (1990) this structural feature is regarded to be an intelligence parameter.

d. Although I basically do agree with William Frawley where he writes: «Two things are readily said about lexicographic practice. First, it rarely changes. [...] Second, when lexicographic practice is criticized by lexicographers, it is examined almost entirely within the status quo [...] the very deeply foundational questions are rarely asked, and if such questions are asked, the answers infrequently conflict with established, conservative practice» (Frawley, 1988, 191-192), yet I would like to stress that during the last decade fundamental changes really have been taking place: Frawley's article itself *(New forms of specialized dictionaries)* is a case in point, such as is the Melcuk approach which he clearly illustrates and discusses in his article. Moreover, it is my contention that these changes have to do with the organization of syntax and semantics in dictionaries, the relationship between the two, and the impact the NLP-orientation has got on these matters. Consequently, the apparent gap between D's and LDB's has become, and is becoming, smaller and smaller.

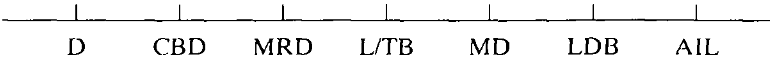

Fig. 1: Objects of interest for lexicography (computational and otherwise)

2. The Van Dale-dictionaries

2.1. Some history

As stated in the introduction, Van Dale-dictionaries are to be taken here as bilingual passive (Foreign Language (FL) → Mother Tongue (MT)) dictionaries. Before entering into details a short 'historical note' may be in order:

— In 1975, Van Dale Projectontwikkeling B.V. (Van Dale Project Development Ltd.) was founded as a new company within Kluwer publisher's. Its main aim was the publication of a new series of seven large dictionaries (six bilingual ones: English-Dutch, Dutch-English, French-Dutch, Dutch-French, German-Dutch, Dutch-German, and one monolingual volume: Dutch-Dutch). The series should be 'marked off both by its originality and by its usefulness' to use the publisher's words (compare with the quotation from Frawley in section 1).

— In 1976, after some preparatory investigations, a Mono-Bilingual Management Team was formed consisting of seven members, including the Publisher and the Dictionary Editors-in-Chief.

— In the ensuing three years (1977-1979) the common concept for the series was discussed and laid down, paying much attention to mutual comparability, implying that —as a rule— the formal grammar for the different volumes should be the same. Also during this period instruction manuals were written and sample entries tried out.

— The next period, from 1980 till 1986, was when the editing work was done. It fell apart into two parts: a first period (ending in 1984) in which the FL-MT-volumes were published —starting with the French-Dutch dictionary in 1983—, and a second one which began in 1984 with the publication of the Dutch-Dutch volume which served as an input for the MT-FL dictionaries and which ended with the publication of Dutch-English in 1986.

— Since then a series of medium-sized dictionaries has been derived from the larger ones (all seven smaller volumes being published in 1988), whereas also work on the 2nd edition has been started yet.

— For the layman-user both *size* and *structure* (the use of *structural indicators* such as double digit codes, signs/symbols such as ♦, ¶, ⇒, etc.) will be the overall features which are most likely to be mentioned first when comparing the Van Dale-dictionaries to other FL-Dutch ones. A glance at the 'bright'-examples given below will make clear what is meant.

2.2. Examples and comments

bright [brait] helder, schitterend, blinkend; illuster *(example)*; gelukkig; vlug, pienter, snugger, levendig, opgewekt; *the - side* lichtzijde.

Fig. 2. 'bright' according to Wolters-Noordhoff, English-Dutch, 1981

bright^2 ‹f3› ‹bn; -er: -ly; -ness› **0.1 hel(der)** ‹ook fig.› ⇒ *licht, stralend, glanzend, fleurig, klaar* **0.2. opgewekt** ⇒ *opgeruimd, levendig, kwiek* **0.3 schrander** ⇒ *snugger, vlug, pienter, intelligent.* ♦ **1.1** a - future *een mooie/rooskleurige toekomst*; one of the -est moments in the history of Europe *een v.d.meest glorieuze momenten in de geschiedenis v.Europa*; - as a new pin *zo helder als wat*; look on the side of things *de dingen van de zonzijde bezien, optimistisch blijven* **1.2** - eyes *heldere/stralende ogen* **1.3** a - idea *een slim idee* **1.4** the ~ lights *het uitgaanscentrum*; (BE: inf.; vaak iron.) **2.1** a - spark *een slimme kerel, een slimmerd, een groot licht* **2.2** - and breezy *levenslustig, opgeruimd.*

Fig. 3. 'bright' according to the English-Dutch Van Dale, 1989

A dictionary article in the English-Dutch Van Dale (and in the other FL-MT Van Dale-dictionaries for that matter as well) is to be regarded as a *framelike datastructure*.[2] Such as the frames used in AI it represents knowledge by means of *slots* (general conceptual categories) and *fillers* (specifications of the slots). Moreover, such as is the case for AI-frames also, it is not restricted to pure *declarative* structures only, but can contain procedural ones as well.

A typical, be it greatly simplified[3], *frame* in the FL-MT Van Dales contains the following *slots* (those mentioned between brackets being optional):

1 lemma (spelling variation)
2 (pronunciation)
3 (frequency)
4 grammatical data
5 (pragmatic data)
6 (supralexical (= proverbial) reference)
7 translation profile
8 (contextualized equivalence)

The simplification resides a.o. in the fact that only the top-most *slots* are given (a slot such as grammatical data, e.g. may get seven *subslots*), that no *modifiers* are mentioned (a slot such as 'spelling variation' e.g. gets such modifiers as ‹AE Sp.›, ‹AE Sp. also›, ‹esp. AE Sp.›, etc.) and, above all, that *different types of lemmas* go with different frames (the treatment of abbreviations differs from that of full forms such as lexical items differ from grammatical ones etc.). Another simplification has to do with the *fillers*: nothing is said here about their number, form, order, etc.

If nevertheless we take up *bright* as a case in point it is because it can make the overall structure more clear: next to the obligatory slots, viz. lemma, grammatical data and translation profile/reference —representing a *minimal frame* or *expectation pattern*— some optional slots are filled as well, viz. frequency and contextualized equivalence. In other words in the case of *bright* the minimal frame, containing a mini-

[2] For a comparable statement see Meyer e.a. 1990, 5.

[3] The frames actually are differentiated according to types. A fairly simple type of lexical items such as *abbreviations* e.g. already demands a rather sophisticated structure. A representation of it (in the form of a predominantly CF-grammar) required some twenty categories (preterminal symbols).

mum— (but as the case may be, sufficient) amount of both formal and semantic data, is expanded in both directions.

Last but not least, *bright* is a typical entry because of its *non-linear structure*. As is well known, the semantics of a translation dictionary are to be found both in the translation equivalents (the *context-free* part) and in the translation of what we will call for convenience sake the examples (the *context-bound* part). Both parts, contrary to what is mostly done in translation dictionaries (see e.g. Fig. 4: *kick* according to Collins-Robert) are, on the one hand, separated in the Van Dale-dictionaries (cf. the black diamond which is used a separator), on the other hand, they are linked together by means of a *numerical code*.

Whereas in the translation profile the numbers enumerate different meanings (*bright* 0.1 = first meaning *hel(der)*, 0.2 = second meaning *opgewekt*, etc.), the double digit code in the contextualized equivalence section has quite another meaning: as this section contains combinations with the entry word, the second digit refers to the meaning number of the entry word (e.g. *bright eyes* (1.2) is a combination of *eyes* with *bright* in its second meaning). The first digit refers to the grammatical category (with 1 standing for noun, 2 for adjective, 3 for verb, etc.) of the word the entry word is combined with (e.g. in the expression *bright and breezy* (2.2) *bright* is combined with another adjective).

Why this is done this way will be explained in the next part. For the moment it may suffice to notice that the way the Van Dale passive bilingual dictionaries are structured offers the possibility to separate out data directly (e.g. the expression *bright and breezy*) and —comparable to the *if-needed procedures* of the AI-frames— flesh those data out, or check them, or shade them, etc., by using the second digit as a pointer to information to be found in another slot (in this case in the translation profile, in other cases, as for example when dealing with grammatical words and/or with grammatical usage, by referring to the appropriate passage(s) in the grammar which serves as a companion to the dictionary).

kick [kik] **1** *n* **(a)** *(action)* coup *m* de pied. **to give the door a** - donner un coup de pied dans la porte; **to aim** *or* **take a - at sb/sth** lancer un coup de pied à qn/qch *or* dans la direction de qn/qch; (...). **(b)** (*_fig: thrill etc._) **she got quite a - out seeing Paris** elle a été tout émoustillée *or* excitée de voir Paris; (...). **(c)** *(gun)* recul *m. (Aut)* **a - of the starting handle** un retour de manivelle. **(d)** *(Ftbl etc.)* **he's a good** -* il a un bon dégagement.

Fig. 4. 'kick' according to Collins-Robert, English-French, 1978

3. Semantics in the Van Dale-dictionaries

For the time being I will take up three dimensions in the discussion of the *organization of the semantics in the passive bilingual Van Dale-dictionaries*, viz.:

- the paradigmatic dimension
- the contextual dimension
- and the computational dimension

3.1. The paradigmatic dimension

With paradigmatic dimension here is meant the way equivalence is rendered on the context-free word level.

As one will observe in the example *bright,* e.g. the meaning (represented by the translation equivalents here) of this item is subdivided into three *clusters:* one representing the perceptual, another the emotional and a third one the cognitive meaning.

Not just one translation equivalent is given in each case, instead of this, a series of variants follows a main translation. *Bright* in its emotional sense e.g. gets *opgewekt* as main translation and *opgeruimd, vrolijk, levendig* and *kwiek* as variants (separated from the main translation by a double arrow (⇒)).

Equivalences then are not regarded as one-to-one relationships between discrete items but as relations between *continuous, variable elements.* There are of course exceptions (think of terms and other standardized items), and depending on the kind of word (lexical, grammatical, collocational, pragmatic, etc.) different paradigmatic types, different *equivalence models,* should be worked with. As to words belonging to a general core vocabulary and having a primarily conceptual meaning, we think that continuity not discreteness should be the rule.

What we mean is further illustrated by the example *key,* where the label ‹ben.(aming) voor› (= designation for) actually indicates a *concept,* not a translation equivalent. In doing so, we move away from a more enumerative, finite approach towards one in which prototypicality, relationship and extensible dynamism are of primary importance.[4]

key[1] [ki:] ‹f3› ‹telb.zn› ‹-›sprw.231› **0.1** ‹ben.voor› **sleutel** ‹v. slot: om iets vast te draaien› ⇒ ‹fig.› *toegang; (strategische) sleutel, strategische plaats; oplossing, verklaring, lijst met antwoorden; letterlijke vertaling; sleutelwoord* ‹v. geheim- of cijferschrift›; ‹biol.› *determineertabel;* ‹schaken› *sleutelzet; opwindknop* ‹v. horloge› **0.2** ...

Fig. 5. ‘key’ according to the English-Dutch Van Dale, 1989

3.2. The contextual dimension

As stated earlier, the semantics of a bilingual dictionary are not only to be found in the translation equivalents but in the translation of examples as well. As a matter of fact this part has been called the *contextualized equivalents* in the English-Dutch Van Dale.

One of the problems of these contextualized equivalents is their organization, especially when dealing with large entries: how to give them a place which is both maximally accessible and maximally informative. Hausmann 1988 discusses this matter and discerns three ordering principles, viz. the semantic, the categorial and the alphabetical principle. Sometimes these principles are combined.

An example of a semantic ordering is e.g. Collins-Robert (see *kick*). Its disad-

[4] A similar point-of.view is to be found in Neubert (this volume) who speaks about translation equivalents as cognitive prototypes.

vantages are obvious. Suppose you do not know what the expression «more kicks than halfpence» means, then you have to go through the whole of the article only to find out that the expression is not there. In other words, you can not use semantic orderings as such in an FL-MT dictionary because one can not use *non-existent* knowledge (knowledge one does not have at one's disposal (i.e. semantics)) as an organizing principle.

Wahrig is a typical example of a categorial alphabetical ordering: examples come after the meaning profile and are ordered according to the word class category of the main combination word. So one finds combinations with nouns, with adjectives, with verbs etc., grouped together, and within these groups arranged alphabetically.[5]

The Van Dale-dictionaries group their contextualized equivalents, to use Hausmann's words, according to a: 'kategoriell-semantisch-alphabetisches Ordnungsprinzip' (Hausmann, 1988: 146). Indeed, when taking a look at *bright* e.g. one can observe that examples are ordered according to those three principles applied in that order. However, inserting the semantic structure —the so-called mediostructure— in between the categorial and alphabetical ordering is, according to Hausmann, a serious methodological mistake —basically because of the fact that it contradicts the form → meaning principle[6] and so can only be a hindrance.[7]

At this point I would like to stress that organizing the semantics in a dictionary is both a matter of accessibility and of information content.[8] The more formal a criterion is, the higher it may score on the level of access, not necessarily however on that of information content: the alphabetical ordering of e.g. the macrostructure being a case in point. The very fact that the categorial alphabetical ordering preferred by Hausmann does not make it possible to extend or enrich the translation profile by making use of the context, and vice versa, made us decide to lower somewhat priority of access in favour of a richer, more dynamic, a more related representation format.

In other words, context is considered to be too important so as not to be loosened up completely from what is usually called 'isolated meaning'. By making use of the categorial-semantic odering (expressed in the double digit code) the advantages of the semantic approach (direct linkage between the general (translations) and the specific (examples) and so giving rise to both generalization and specification if needed or wanted) is preserved, without loosing completely sight of the looking-up facilities (and so of the form → meaning principle).

[5] Wiegand 1989 uses the term 'integrated' (e.g. Collins-Robert) vs. 'unintegrated' (e.g. Wahrig) microstructures.

[6] As meaning cannot be used as an organizing principle (cf. supra) we decided to use form as a means of organization of contextual equivalence.

[7] Hausmann 1988 also mentions two other 'hindrances' in the active Van Dale-dictionaries, viz. the use of the semantic mediostructure for the *looking up of idioms* and the *division* of many dictionary articles into different parts corresponding to different grammatical behaviour. As to the latter remark we will (partly) deal with it in section 3.3, as to the former may it suffice here to point at the fact that it is rather extraordinary for dictionaries to differentiate formally such as the Van Dale-dictionaries do between idioms and non-idioms, and so make idioms recognizable, which can hardly be taken as a hindrance for looking them up!

[8] In this respect compare Wiegand 1989, where he writes «... the microstructure of dictionary articles is not the only (partial) structure within the complete article structure. The other important structure is the addressing structure.»

3.3. The computational dimension

This dimension can be taken both in its original, etymological sense —Latin: *computare* in the sense of to calculate more in particular to calculate meaning— and in its more recent meaning of being deliverable by *computer.*

One of the basic principles underlying the FL-MT-Van Dale-dictionaries was that they should function as *semantic problem solvers* on word level. Processing language from the unknown (the FL) to the known (the MT) they basically would rely on formal aspects for the FL (cf. the preceding paragraph) whereas the representation language would not get any further explicitation, being the mother tongue of the user. The semantic calculus therefore is fairly simple in the case of monosemous words where the possible choice between variants is left over to the human user. In the case of polysemous words however the English-Dutch Van Dale e.g. scores relatively high as an *automatic lexical problem solver.* This has to do with the fact that, as to disambiguation or meaning discrimination, this dictionary (and the other Van Dales as well) does not leave the decision completely to the human user as such, but instead, uses quite a wide range of means such as part-of-speech categories, subcategorization, pragmatic data, examples, collocations, idioms, and last but not least, combination words.

In a prototype system developed a year algo called *Lexpert* (see Martin-Mortier, 1989) consisting mainly of a lemmatizer-tagger as preprocessor, the English-Dutch Van Dale as a lexical knowledge base and an inference engine, 'unrestricted' English text could fairly well be disamiguated.

In other words, the fact that a) meaning distinctions are organized on the basis of categorial, subcategorial, pragmatic-contextual and/or combinatorial *constraints* and b) that often these constraints are made explicit or explicitable, gives an extra dimension to the semantics of the Van Dale-dictionaries thus not only serving human, but computer-aided, translation as well. Moreover, with hindsight, it is especially the *organization of the combinational constraints* and of the *subcategorization-features* which makes the Van Dale-dictionaries a powerful and, to use a buzz word, reusable, tool both for humans and computers. The framework is there and has become (well) known by now. It is to be hoped that subsequent editions will further explore and enhance its possibilites.

References

ATKINS, B. T. *et al.* (Eds.), *Collins-Robert French-English, English-French Dictionary*, Collins, London, Glasgow and Toronto. 1978.

FRAWLEY, W., «New Forms of Specialized Dictionaries», in *IJL*, 1, 1988, 189-213.

GERRITSEN, J. - OSSELTON, N. (Eds.), *Engels Woordenboek deel Engels-Nederlands*, Wolters-Noordhoff, Groningen, 1981.

HAUSMANN, F. J., «Grundprobleme des zweisprachigen Wörterbuchs», in K. Hylgaard Jensen and A. Zettersten (Eds.), *Symposium on Lexicography III*, 1988, 137-154.

HAUSMANN, F. J. - WIEGAND, H. E., «Component parts and structures of general monolingual dictionaries: a survey», in: F. J. Haussmann e. a. (Eds.) *Dictionaries, An international encyclopedia of lexicography*, Walter de Gruyter, Berlin, New York, 1989, 328-360.

MARTIN, W., «Towards the construction of intelligent lexical databases», paper read at the Complex-conference, Balatonfüred, 8-11 September 1990.

Martin, W. - Al, B., «User-orientation in dictionaries: 9 propositions», in: T. Magay, J. Zigany (Eds.), *BUDALEX '88* proceedings, Budapest, Akademiai Kiado, 1990.

Martin, W. - Mortier, L., *Lexpert*, Internal Report, Universitaire Instelling Antwerpen 1989.

Martin, W. - Tops, G. (Eds.), *Van Dale Groot Woordenboek Engels-Nederlands*, 2nd edition, Van Dale Lexicografie, Utrecht-Antwerpen, 1989.

Martin, W. - Woltering, M., *Basic issues in computational lexicography*, report written for the EC-Commission, Utrecht, 1989.

Meyer, I. e.a., *Lexicographic Principles and Design for Knowledge-based Machine Translation*, Carnegie Mellon University, 1990.

Neubert, A., «Fact and fiction of the bilingual dictionary», in this volume.

Schnelle, H., «A formal view of the logic of the dictionary», paper read at the Complex-conference, Balatonfüred, 8-11 September 1990.

Wahrig, G. (Ed.), *Deutsches Wörterbuch*, Bertelsmann, Gütersloh, 1977.