Piek Vossen, University of Amsterdam

The automatic construction of a knowledge base from dictionaries: a combination of techniques

ABSTRACT: In this paper an approach is described to the construction of a knowledge base from dictionaries which combines an empirical point of view with a more theoretical framework. A distinction is made between extracting information for separate entries and building the lexical knowledge base in which this information is to be implemented. Without such an implementation the full impact of having information extracted is not explicit, while on the other hand the development of such a knowledge base cannot be done properly without an overview of the information, therefore presupposing some degree of analysis.

1.Introduction

In the Acquilex Esprit-project (BRA-3030) the feasibility and cost-effectiveness of (semi-) automatically extracting lexical knowledge from Machine Readable Dictionaries (MRDs), and representing this knowledge in a multilingual knowledge base is evaluated. The project is a joint enterprise of the Universities of Amsterdam, Barcelona, Cambridge, Dublin and Pisa. In Amsterdam the information from the Longman Dictionary of Contemporary English (henceforth LDOCE) and the Van Dale monolingual dictionary of Contemporary Dutch[1] is being extracted. Two central issues of the project are:

- to develop techniques for automatically extracting as much information (both syntactic and semantic) as possible from the individual entries.
- to store this information in a Lexical Knowledge Base (LKB) which exploits the hierarchical aspect of knowledge made explicit for each word sense by allowing inheritance of information from more general words to more specific words.

These two issues represent two perspectives in using MRDs: the empirical approach of looking at the data which is found and trying to represent its content in a systematic way, and the theoretical approach of setting up a consistent lexical representation system first and then trying to fill it with data extracted from the dictionaries. Various problems in extracting information from MRDs and representing these data in a formal and consistent way make clear that both approaches have serious limitations. For example, having definitions analysed in terms of their structure still does not make clear what the semantic impact is of having this information systematically available. Lexical knowledge just like any other kind of knowledge is hierarchically organized, i.e. concepts are based on the meaning of other more general concepts. The full impact of 'knowing' that e.g. "brandy" is a "drink" with the quality "containing alcohol" (the result of parsing its

definition) becomes clear when it is stored in a lexical knowledge base in which the concepts "drink" and "containing alcohol" are formally defined and the hierarchical relations can be exploited. Building such a knowledge base in a formal and consistent way is a non-trivial task. It is, however, impossible to represent the information actually found (to implement a realistic lexicon) without having a good notion of the kinds of information to be found. In the Acquilex project both approaches are combined in a complementary way described in this paper. In the next section the extraction process is described. In section 3 the limited usefulness of the results of parsing definitions is explained when it is not implemented in an LKB with a hierarchical element. Section 4 describes the LKB used in the Acquilex project and the limitations of building an LKB lexicon for a small domain. Section 5 describes an approach to use large scale rough material as an empirical resource for setting up an LKB lexicon.

2. Making the information stored in dictionaries explicit

The semantic information contained in dictionary definitions is stored in the form of expressions in natural language, compare the following examples from Van Dale, 1984:

bisschopwijn	=	gekruide en gesuikerde warme, rode wijn (literally "spiced and sugared warm red wine")
wijnvlek	=	door gemorste wijn veroorzaakte vlek (literally "by spilled wine caused spot")

Lexicographers thus rely on the fact that human users speak the language and know the meanings of the words. In a sense they build on the knowledge people already have. Speakers of Dutch know that "wijn" in the definition of "bisschopwijn" is the syntactic kernel and therefore the genus of the definition, whereas the same word is embedded as a differentia in the definition of "wijnvlek" and they thus infer that it is not a kind of "wijn". The function of the words in the structure of the definition as a whole determines the semantic effect. Computers can only have access to this information if this structure is made explicit and the meaning of each word is determined. Therefore syntactic parsers have been built (Vossen 1990, 1991b) which analyse the phrase structure of definitions not only in terms of their genus and differentiae, but also making the compositional structure of the differentiae explicit:

```
ENTRY(HWINFO(TNR(007593_00)
             EW(bisschopwijn)
             HSNR(00.01))
      NMD(NP(noun)
             (RE(PRM(co)(PRM(m)(VP(scnd)(PRDN(PRED(m)($VX(gekruide)))))
                        COORD($C0(en))
                        PRM(m)(VP(scnd)(PRDN(PRED(m)($VX(gesuikerde))))))
                 PRM(m)(STATE($A1(warme)))
                 ca(,)
                 PRM(m)(STATE($A1(rode)))
                 KE(m)($N0(wijn))))))
```

The brackets indicate the scope of the constituents, so that it is clear which word specifies which other word, while the constituent labels before the colons indicate the kind of

specification. In this example "wijn" is labelled "KE(m)", which in this case means that it is the syntactic kernel of the definition (and therefore also the genus). The other words "gekruide" (spiced), "gesuikerde" (sugared), "warme" (warm) and "rode" (red) are labelled PRM(m) (besides other labels) which means that they all specify some kernel KE, in this case "wijn". Further labels such as VP(scnd) indicate more precise relationships, i.e. that "wijn" is the (affected) object of the events "gekruide" and "gesuikerde" and not the subject (agent). The 'dollar codes' before the words, such "$A1" and "$N0", contain inflectional information. Within these trees the structural information of the definition is integrated in a single labelled bracketed structure. As a result this information is explicit but it still requires a lot of processing to scan the brackets and interpret the labels. The integrated structure, therefore, is converted into a much simpler list structure in which each piece of information is separately represented as a two-place relation between the specification (e.g. "rood", "warm") and the elements to which it is applied (e.g. "wijn"), each between separate brackets:

((bisschopwijn) (TN 007593_00) (HN 0) (SN 1)
(DF (EV OG) ($VX kruiden EV:A1) ($N0 wijn OG:A3))
(DF (EV OG) ($VX suikeren EV:A1) ($N0 wijn OG:A3))
(DF (QA OG) ($A1 rood QA) ($N0 wijn OG))
(DF (QA OG) ($A1 warm QA) ($N0 wijn OG)))

After each word the typological status is indicated by a two letter code (i.e. "QA" is quality, "OG" is object which is also the genus, "EV" is event). Each relation as a whole is preceded by a general relation indicator also between brackets, containing the type-codes of the words that are related ("EV OG" means that a relation between an event and an object which is also the genus is expressed). Specific relations between verbs and arguments or PPs are expressed by specifiers after the type-code, i.e. "EV:A1" means that the predicate designates the event itself, "OG:A2" means that the genus-object is the first argument, "OG:A3" means that it is the second argument of the event, "OG:PP" means it is a PP-complement of the verb.

3. The lexical knowledge base

When indexed these differentiae lexicons can be loaded in a lexical database, called LDB, developed in the Acquilex-project at Cambridge University (Carroll 1990), in which very quick and easy access to the above semantic relations between words is possible (e.g. all words that have the quality (QA) "rood"). By formulating queries in which information from several dictionaries can be combined the LDB provides very fast access to the vocabulary of a language from the information side, e.g. all nouns which cannot be pluralized and refer to things which are liquid (have QA "vloeibaar") and can be drunk (have QA "drinkbaar"). Unfortunately a lexical database such as the LDB will not return the above example "bisschopwijn" (and also not all other kinds of wine) although all these properties hold for it. This is because the fact that it is "liquid" and "can be drunk" is not directly specified in its definition, but is inferred by human readers, since they know that "wijn" means a "drank" ('drink') and that "drank" is "vocht" ('liquid'):

bisschopwijn	=	gekruide en gesuikerde warme, rode wijn ("spiced and sugared warm, red wine")
wijn	=	alcoholische drank , uit gegist druivesap bereid" ("alcoholic drink, from fermented grape-juice made")
drank	=	drinkbaar vocht, al wat men drinkt ("drinkable liquid, all what someone drinks")
vocht	=	vloeibare stof ("liquid material")
stof	=	materie, substantie ("matter, substance")

Inheritance of features from more general levels to more specific levels is not an intrinsic property of the LDB (Boguraev et al 1991). This can only be achieved by an explicit inheritance mechanism that formally exploits the taxonomy relations between head-words and genuswords. In principle each genus is an entry in the dictionary and can therefore be looked up to find its own genus, etc. ("bisschopwijn", "wijn", "drank", "vocht", "stof"), thus revealing the hierarchical organisation of the vocabulary. Because of this special status of the genus terms, they have been seperated from the differentiae and stored in a separate genus lexicon in the LDB:

((bisschopwijn) (TN 007593_00) (HN 0) (SN 1) (GEN wijn::SG::CO::.::??.??))

A database which combines locally specified properties (differentiae) with inheritance via these hierarchical structures is a very powerful and efficient system. The taxonomies thus derived from MRDs are very large and complex structures in which thousands of words are interconnected via even many more relations (Amsler 1981, 133-138, Vossen and Serail 1990, Vossen and Copestake 1991). Once a property is expressed for a top node such as "person" it will be possible to derive it for all the ca. 6000-7000 words which are directly or indirectly classified as such. However, this also means that a wrongly stated property will be wrongly inherited for thousands of other words as well.

Another complicating aspect is the necessity of having exceptions at more specific levels. For instance not all words described as a "drink" refer to substances that can in fact be drunk or that are customarily drunk, still nobody would really want to deny the fact that "drinkable" is commonly inherited for drinks:

inmaakbrandewijn	=	brandewijn die men gebruikt om eetwaren in te maken ("brandy which people use to preserve food")
brandewijn	=	sterke drank met 35 a 80% alcohol, gestookt uit wijn, graan of andere grondstoffen ("brandy" = "strong drink with 35 a 80% alcohol, distilled from wine, corn or other ingredients")
drank	=	drinkbaar vocht, al wat men drinkt ("drink"
	=	"drinkable liquid, all what people drink")

The Dutch compound "inmaakbrandewijn" is usually not drunk but only used to preserve food. The typical 'function' or the 'telic role' inherited from "drank" ('drink') thus has to be over-written.

To account for these phenomena a Lexical Knowledge Base (LKB) has been developed at Cambridge University (Copestake 1991) which makes use of typed feature-structures

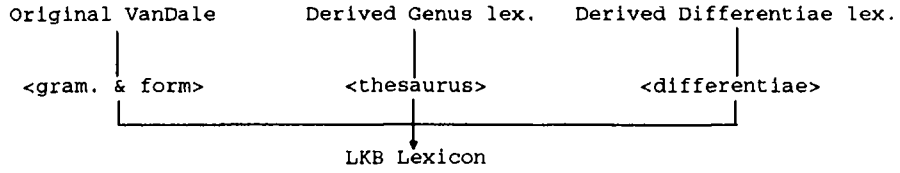
(similar to those described in Carpenter (1990)) to store the information from the differentiae and which uses PSORTs based on the genus terms for controlling default inheritance of this information. In the feature structures (FS) a distinction is made between the different dimensions of differentiae (the features) and the values which can be filled in for each feature, thus constituting feature-value pairs. In the case of "bisschopwijn" "rood" (red) is a value for the feature "colour" and "warm" (warm) is a value for the feature "temperature". Since the number of features that are used in the differentiae is relatively small, whereas the number of values can be rather large it is thus possible to predefine the lists of possible features in advance while the values are left unspecified until they can be filled in by the differentiae in the actual definitions. These pre-fab lists of features or FSs are stored in the LKB as TYPEs, which can be seen as abstract concepts consisting of clusters of differentia-types or features. Examples of such TYPEs in the current LKB system in the Acquilex project are concrete, abstract, substance, object, place, artifact, natural. Each cluster represented by a TYPE contains only those features which are relevant for a specific semantic field. The following two FSs: artifact_substance and natural_object for example represent different clusters of features (animacy is false vs true, shape is non-individuated vs individuated,[2] agentive process from which it originates is man-made vs natural, etc.):

```
[artifact_substance
 ORIGIN_AREA:string                 QUANT:quantity
 TELIC:verb-sem                     SIMILAR:string
 PHYSICAL:true                      APPEARANCE:appearance
 ANIMACY:false                      FORM:[phys_form VOLUME:scalar
 PHYSICAL_STATE:state_a                             WEIGHT:scalar
 QUAL:[phys_qual                                    SHAPE:non-individuated]
        COLOUR:colour               CONSTITUENCY:constituency
        SMELL:smell                 AGENTIVE:man-made]
        TASTE:taste
        TEMPERATURE:temperature
        TEXTURE:texture
        SIZE:size]
[animate_natural_object
 ORIGIN_AREA:string                 QUANT:quantity
 TELIC:verb-sem                     SIMILAR:string
 PHYSICAL:true                      APPEARANCE:appearance
 ANIMACY:true                       FORM:[phys_form VOLUME:scalar
 PHYSICAL_STATE:state_a                             WEIGHT:scalar
 QUAL:[phys_qual                                    SHAPE:individuated]
        COLOUR:colour               CONSTITUENCY:constituency
        SMELL:smell                 AGENTIVE:natural
        TASTE:taste                 ORIGIN:string
        TEMPERATURE:temperature     AGE:age
        TEXTURE:texture             SEX:gender]
        SIZE:size]
```

In these examples of FSs features are in capital letters before the colon, whereas the values follow the colon in small bold letters. At the TYPE-level most values are still open since almost all values stand for value-classes, which means that they only restrict the class of possible values which can occur. These value-classes are defined elsewhere in the system, e.g. colour is defined by the set of all colour terms, although in some cases they are still underspecified, e.g. *string* which occurs at several features and which allows any LISP string as value. Nevertheless, each value has to have some kind of definition in order to

mean anything to the system (i.e. to make the right semantic inferences). The TYPEs in the lexicon can thus be seen as the concepts in which the information found in the lexicon has to be expressed so that its content can be made fully explicit. The notion that the information in dictionary definitions builds on other knowledge is as such to some extent formalised by specifying this knowledge as types.

By combining the information from the different dictionaries loaded in the LDB, an LKB lexicon can be built up in which the headword-genus relations constitute the taxonomies (or PSORTs) via which properties are inherited, the interpreted differentiae constitute these properties and the orginial Van Dale dictionary[3] is used to extract form-information and grammatical properties:



The role of the taxonomies consists in relating all the words from the vocabulary to the correct differentiae TYPE (e.g. "artifact_substance") via its top word (e.g. "drank") so that these FSs become available for all subtypes. The TYPE is thus only specified once for a whole taxonomy. The differentiae which are found at each specific word then have to be interpreted as values for only those features which are relevant for the word according to the taxonomically determined TYPE (Rodriguez et al 1991). In case a differentia cannot be interpreted as such a value it is not represented. It is obvious that the LKB system thus operates as a very strong filter on the data being extracted from the MRDs. The resulting LKB lexicon can be loaded into the LKB system which further controls all inheritance processes (default and non-default) giving a formal representation for the information contained.

4. Building LKB lexicons for small domains

4.1. Advantages of working with a formally predefined TYPE system

One of the major advantages of the LKB system over the lexical database are the control possibilities mentioned above. In addition to the fact that the lexical information represented must be consistent with the system, consistency is also achieved by building up the TYPEs by hand. These TYPEs can be used as a filter to guide the interpretation of differentiae and to warn for underspecifications. Because of the far-reaching impact of the information specified at the highest level setting up these TYPEs manually also seems desirable and since the number of different features involved is relatively small this will not involve too much work. Another major advantage already indicated above is the possibility of deriving massive data via the thesaurus, thus revealing all indirectly implied properties by inheritance which will then automatically be checked for consistency. As an LKB entry the above "inmaakbrandewijn" example with only locally specified information looks as follows:

```
[lex-noun-sign
 ORTH: inmaakbrandwijn
 SENSE-ID:[sense-id
           FS-ID: inmaakbrandewijn_v_0_1
           LANGUAGE: dutch
           DICTIONARY: vand
           LDB-ENTRY-NO: 29597
           SENSE-NO:1]
 CAT:[nominal-mfeats
      NUM:singular
      GENDER:male
      COUNTABILITY:false]
 RQS:[artifact_substance: TELIC:[ARG1:[PRED:inmaken_v_0_1]]]]
<lex-noun-sign rqs> < BRANDEWIJN_V_0_1 <lex-noun-sign rqs>
```

In this specification the actual FS is given between square brackets containing information on the orthography (ORTH:), on the source dictionary from which it is derived (SENSE-ID:), syntactic information stored at CAT and interpreted differentiae from its definition (RQS) which in this case is a specific telic-role or function "inmaken" (preserve). The bottom line contains the PSORT relation or genus "brandewijn". When this entry is fully expanded by the system the result is the complete template given for artifact_substance above with all values inherited from more general words:

```
[lex-noun-sign
 ORTH: inmaakbrandwijn
 SENSE-ID:[sense-id
           FS-ID: inmaakbrandewijn_v_0_1
           LANGUAGE: dutch
           DICTIONARY: vand
           LDB-ENTRY-NO: 29597
           SENSE-NO:1]
 CAT:[noun-cat
      CAT-TYPE:n
      M-FEATS:[nominal-m-feats
               REG-MORPH:true
               AGR:[nominal-agr
                    PERS:3
                    NUM:singular
                    GENDER:male]
               COUNTABILITY:false]]
 RQS:[artifact_substance:
      TELIC:[ARG1:[PRED:inmaken_v_0_1]]]
      ORIGIN_AREA:string
      PHYSICAL:true
      ANIMACY:false
      PHYSICAL_STATE:liquid
      QUAL:[phys_qual
            COLOUR:colour
            SMELL:smell
            TASTE:taste
            TEMPERATURE:temperature
            TEXTURE:texture
            SIZE:size]
      QUANT:quantity
      SIMILAR:string
      CONSTITUENCY:[constituents PRED: "alcohol"]
      APPEARANCE:appearance
      FORM:[phys_form
            VOLUME:scalar
            WEIGHT:scalar
            SHAPE:non-indivuated]
      AGENTIVE:[ARG1:[PRED:stoken_v_0_1]]]]]
```

Only the property "inmaken_v_0_1" (preserve) is directly specified for "inmaakbrandewijn"; values such as "stoken_v_0_1" (distilled) as value for agentive-process and constituents such as "alcohol" will be inherited from "brandewijn_V_0_1" (brandy), to which it is related as a hyponym in the lexical specification (see specification at the bottom-line of the entry). Other values may be inherited again from "drank" (drink), etc. (however in this specific example the telic role "drink" which is inherited by default is overwritten). Although these values are specific for this entry the FS as a whole is available for all words which are related to the type artifact_substance. This also means that it will be easy to automatically compare all entries of this type and find out which words have identical values. The system could then mark these words as being (near-)synonymous. Future lexicographical work could then be guided to discriminate between underspecified entries, making use of the templates to check and trigger further enrichments. Finally, within ACQUILEX the data from several monolingual dictionaries ranging over four languages (English, Dutch, Itialian and Spanish) are stored in the same type system. At Amsterdam University the analytic procedure described for the Van Dale dictionary has also been implemented for the Longman Dictionary of Contemporary English (Vossen 1990, 1991b). As a result the English LKB lexicon is highly compatible with the Dutch lexicon, thus making automatic cross-linguistic comparison possible (Copestake and Jones 1991).

4.2. Problems with building LKB lexicons

In the Acquilex project we are currently developing lexicons for small subsets of the vocabulary, i.e. "food", "drinks", "persons with occupations", "instruments" and "places". These selections are made by using the taxonomies. In building up the relevant types and LKB lexicons for these domains some problems are encountered:

- What are the criteria for distinguishing different TYPEs, and for deciding which features are relevant for what TYPEs? In order to build up a FS-representation a theory is needed which predicts what is required and which explains the distribution of TYPEs and features. - Domains of the vocabulary cannot be seen in isolation. The words of a language are strongly interrelated and intermingled posing serious theoretical and methodological problems for anyone trying to set up a TYPE system. How to infer, for instance, the correct features for all the differentiae involving mainly verbs and adjectives if only restricted domain information for nouns is available?

4.2.1. On what basis are the different TYPEs distinguished?

When manually building a TYPE system decisions have to be made about which TYPEs should be distinguished and which features should be included where. A starting point for these clusters of features could be the distinctions between classes of words in linguistic theories, such as e.g. mass, count, group and plural nouns. Evidence for distinguishing such classes is often based on different grammatical (syntactic and semantic) behaviour or different implied inferences of words belonging to such classes. From the fact that e.g. "one water" and "two water" are unacceptable and "some water" is acceptable we can infer that "water" is a mass noun. The possibility or impossibility for items to occur in such 'test-phrases' constitutes a form of empirical evidence (e.g. to be looked for

in corpora). Once a TYPE system for such notions has been set up, a possible way of proceeding could be to scan the definitions of words in MRDs for clues to include or exclude words in terms of these classes. A TYPE system set up in this way could be made to accept only that knowledge from the dictionary which fits the distinctions made.

This would be perfectly all right if we had a full-blown linguistic theory of what distinctions play a role and are necessary to describe the whole vocabulary. The problem is that we do not have such a complete theory. Not only is there discussion on the definition of basic categories (compare the ongoing discussions on differences between the above noun classes) but when it comes to lexical semantics there is not even the beginning of a consensus on what properties it should capture. In addition to this, linguistic theories have traditionally concentrated on the generalisations that could be made about language, regarding the lexicon as a repository of idiosyncratic properties that could not be predicted. However, how much of the information necessary to use words properly is idiosyncratic and how much is captured by these generalisations? Furthermore, these claims about distinctions in subclasses have never been tested against real size vocabularies so that on the one hand it is not clear to what extent the distinction cover the whole vocabulary (perhaps there are classes of noun that cannot be described as either count, mass, group or plural nouns), and on the other hand it is possible that idiosyncratic properties are still to some extent regular but have not yet been captured in a generalised class. We have as yet no idea to what extent the linguistic behaviour of the words of the vocabulary of a language is covered by the general categories and how many words can be captured. In this respect the question to what extent grammars cover all expressions in corpora is similar, and perhaps these issues are two sides of the same coin.

Lack of theory is most saliently felt for lexical semantics. The semantics of a word should describe the typical conceptualization associated with it. In this respect not knowledge of the object to which the word normally refers should be captured but the way in which the vocabulary and in particular this word cuts up the conceptual space. That is why we speak of "sunset" and "sunrise" and not of "earthturn", and that is why language can differ considerably in the way in which the vocabulary is related to its potential reference. Furthermore the vocabulary is not just the output of a common cognitive system that neatly divides the conceptual space into clearly distinguished parts having single separate words attached to each part. Various other aspects (such as culture, history, social aspects, formal linguistic aspects) trigger lexicalisation processes in languages and thus influence the structure of the vocabulary (although probably less strongly than conceptual aspects). The following words, for example, all refer to the same concept "policeman" but they differ in the attitudinal, diachronical and regional information carried along:

bobby	infml BrE a policeman	flatfoot	sl a policeman
bull	sl, esp. AmE a policeman	peeler	BrE old sl a policeman
copper	infml a policeman	pig	sl policeman
cop	infml policeman		(Examples from LDOCE, 1978)

The specification of a FS that should cover a particular semantic domain or field should therefore be based on extensive linguistic and cognitive knowledge on how the vocabu-

lary of the language is structured as the result of its diachronic development. Thus predicting what is relevant for a domain and what, therefore, has to be included in a FS seems to be a very ambitious undertaking, and it is unlikely that, given the current state of linguistic and cognitive theory, a TYPE system can be developed in a top-down fashion which also fits realistic sets of words. In this respect the empirically observable diversity of the vocabulary is in no proportion to the shallow semantic classificational apparatus of linguistic theory. Instead of adapting the meaning of words to pre-defined concepts, therefore, a case could be made for the opposite approach: fitting concepts to the words: i.e. an empirical approach starting out from the actual words might be far more appropriate.

Although dictionaries do not necessarily provide adequate and consistent information, they are nevertheless rich resources on a more general level enabling one to at least test a theory of the lexicon in terms of its coverage, and to some extent find out how the vocabulary is structured as a whole (e.g. how diverse conceptualizations are given dictionary definitions as a rough indication). As far as they do not provide that information, additionally, the claims of linguistic theories on the lexicon should be extended so that corpora can be searched (semi-automatically) for support. The MRDs can still be of some help to extend these claims and to enlarge their coverage.

4.2.2. The inappropriateness of building a knowledge base for a small domain

LKB lexicons in the Acquilex project are set up for restricted domains of lexical knowledge, e.g. nouns denoting food and drinks. One of the reasons for doing this is to be able to set up a TYPE system for a domain by hand and to evaluate the usefulness and consistency of the data which is directly and indirectly derivable. Another motivation for such a domain specific approach could be that it makes it possible to keep track of the diversity of the lexicon. However, such a restriction, which was necessary given the limited resources in the project, leads to various problems, both practical and theoretical.

4.2.2.1. Practical problem: attaching differentiae as values to the right features

By making a distinction between features (to be specified at the TYPE-level) and values (to be filled in at the entry level) a lot of problems are shifted from the representation to the extraction process. Although the number of features may be rather restricted and can be specified by hand, the range of values can be very large, and it is often not possible to infer from the form of the differentiae what feature it is a value of. Given the following table of most frequent modifiers in food-definitions, how can the system 'know' that "small" and "klein" refer to size and "flat" and "plat" to shape?

Most Frequent pre-modifiers in the subset of food in LDOCE and Van Dale:

white	4	hard	15	vlezig (fleshy)	7	eetbaar (eatable)	10
breadlike	4	thin	17	zacht (soft)	8	rond (round)	11
liquid	5	soft	17	wit (white)	9	langwerpig (long)	12
dry	6	round	21	zoet (sweet)	9	plat (flat)	13
large	10	flat	25	droog (dry)	9	fijn (fine)	14
light	10	sweet	30	dun (thin)	9	groot (large)	19
thick	14	small	46	rood (red)	10	klein (small)	28

A solution could be to manually type in all possible values for the relevant features of a specific domain in the TYPE system and let the system recognize the value, but this can be rather time-consuming for larger domains. Some very significant features are in principle (almost) unrestricted as to their possible values, e.g. anything can be a constituent of anything else, or any process ("cut", "fry", "boil", "crush", "twist", "fill", "press", etc.) can be used to produce or prepare some kind of food. Although a formal definition for the value of constituency is given (and must be given), namely any string, this is rather meaningless in a conceptual sense. Furthermore, some of the values (especially if they are expressed by verbs) have far-reaching consequences with respect to several features. All the processing verbs imply properties on the result which can be very different from the object before processing ("frozen water", "mashed potatoes"). And a multilingual database is even faced with the extra problem of interlinking the values of these 'open class' differentiae for the different languages, although a bilingual dictionary could be of help in that. By restricting oneself to nouns it is obvious that concepts that are typically associated with adjectives and verbs are not available to the system and cannot automatically be exploited to represent the knowledge expressed in the differentiae.

4.2.2.2.Theoretical problem: domain restriction leads to unrealistic semantics

Although the availability of lexicons in the LDB and LKB opens up new possibilities of studying systematic classes of words (by selecting (sets of) items belonging to the same part of the vocabulary) and building up very sophisticated semantic representations for them, there is also a very real danger in doing this. Various phenomena in language which cause problems in building Natural Language Processing (NLP) programs suggest that meaning in language is an enormously complex phenomenon. By isolating taxonomic parts of the vocabulary the phenomenon as a whole might be reduced to a seemingly manageable issue as well. For one thing the problem of polysemy and homonymy is no longer relevant in a lexicon that is restricted to a single conceptual and syntactic class. To account for the polysemy of an entry all its senses have to be considered and not just those senses belonging to two specific domains (e.g. "animal" and "food", Copestake and Briscoe 1991, 88-101, Briscoe and Copestake 1991). It will only be possible to find regular classes of sense extensions and look at regular morphological derivation after a wide range of domains covering several parts of speech have been represented in the system. Many abstract nouns, for example, have senses that can be described as derivations from verbs. However, the rules predicting these senses can only be formulated after one has specified these verbs. Another problem is that within a Saussurean structuralist view on lexical semantics meaning is essentially relations between words, and certainly not less than that. The lexicon thus functions as an enormous grid in which each word, in function and meaning, fully depends on the other words it is related to. The syntagmatic aspects of words are highly intermingled not only with the semantic aspects but also in the form of collocational restrictions between words. Ambiguity and language generation problems in various NLP applications suggest that information on what words tend to combine with which other words from a semantic, syntactic and collocational perspective is indispensable. This means that studying the semantics of particular nouns necessarily means also taking into account the possible verbs and adjectives they can combine with. The lexical semantic grid should therefore

not be built starting from one part of speech and one perspective (which is restricted to a particular domain such as food), but it should gradually be woven from all parts of speech and from many different fields towards each other. In terms of selection restrictions for example this means that we should not only look at which nouns can occur in which slots but also the other way round: in which events does the entity designated by a noun typically tend to be involved (Pustejovsky 1989, 17-25).

5. Combination of techniques

To overcome some of the problems mentioned above a combination of techniques is used in Amsterdam. First of all the process of parsing the definitions is clearly distinguished from the process of getting at representations in terms of feature-value pairs in the LKB.

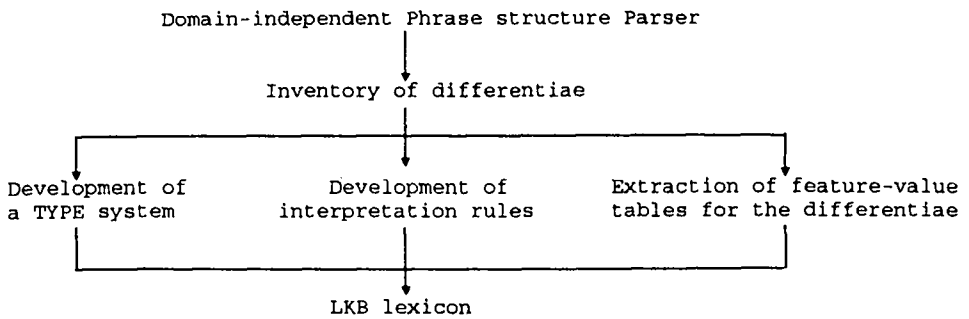
The main reasons for this are:

- the parser-grammar would need a lot of semantic information to get at very specific interpretations. This information could become partly available if one knew in advance what TYPE applies to a specific domain that is being parsed. However, in order to get at the features of a TYPE one has to know what is relevant for the semantic field that the TYPE represents, and one thus needs some access to the data in advance.
- to set up a representative TYPE system it is necessary to fully understand what kind of inheritance mechanisms are necessary and how to implement them. Having the intermediary parsing results in the LDB with its extensive query possibilities can be helpful in setting up a knowledge representation system for complex but not yet fully understood relations such as part-whole or group-member.
- whole-sale parsing of definitions will enable one to isolate those words which make up the basic concepts of a language having specific functions in the definitions, for which one can then start to build some LKB representation that will help future work in building up LKB lexicons in small domains

The parsers developed for LDOCE and Van Dale are domain-independent phrase structure parsers specific for the definitions of different parts of speech. The results of the parser are stored as separate derived lexicons (a genus and differentiae lexicon) in the LDB as described in section 2. Using the LDB, first an inventory is made for all differentiae for a particular domain for both languages English and Dutch. The frequency lists are distinguished in terms of adjectives specifying the genus, verbs designating events in which the genus is involved, and preposition phrases relating other nouns to the genus. In the current Acquilex project, which has a restricted set-up, these inventories are used for:

- setting up a common TYPE system for a domain for both English and Dutch, and to chart out cross-linguistic differences that might occur (in fact the FSs for artifact_substance and animate_natural_object in section 3 result from such inventory),
- formulating interpretation rules for those differentiae whose structure clearly suggests a feature-value interpretation,
- extracting interpretation tables for open-class differentiae for which no rules can be formulated so that all the possible values have an explicit formal representation (and thus a 'meaning').

Finally the TYPE system, the rules and the feature-value tables are used together to derive an LKB lexicon for the particular domain:



In case of the food & drink domain in the Acquilex project the inventory showed that both Van Dale and LDOCE make a clear conceptual distinction between on the one hand things directly classified as food and which are without exception artifacts, and, on the other hand, natural things that can be eaten, but which are not classified as food but as plants, parts of plants, fruit, animals, etc. Although there are some exceptions ("milk" and "meat" are to some extent processed and still natural) this distinction has led to the complementary TYPEs artifact and natural in the TYPE system. The relation between these two classes is further indicated by the very frequent use of verbs in both dictionaries which designate the process by which the artifact food is made, and by various specifications of the edible natural things that have been used as ingredients in these processes. The other differentiae mainly refer to general features which are relevant to all physical things, such as shape, colour, taste, temperature, etc. Furthermore, within a specific domain some frequent differentiae structures can be interpreted directly. For instance food, being non-animate passive matter, hardly fills first argument slots in differentiae. In those cases where it does, however, often a special construction is used which also has a special interpretation: i.e. most frequent are "are" followed by a property designating adjective, "have" followed by a property-designating NP (e.g. "a bitter taste"), and "contain" or "consist of" followed by an ingredient. Similar rather fixed interpretation can be made for very frequent verbs such as "used" and "made". In much the same way PPs which normally are rather ambiguous have fairly straightforward interpretations within a restricted domain. PPs with the preposition "with" either refer to ingredients or constituents of food, or in case of "taste" or "colour" refer to properties with a special status in the TYPE system. A PP with "for" almost without exception refers to the class of animates for which the food is intended. In this way the fact that nothing is known about the words that occur in the differentiae can be partially overcome. However, to deal with this problem properly, in the end, it will be necessary to provide a formal semantic representation for all these adjectives, verbs and nouns that occur in the differentiae of the subset that is being extracted, and to relate this semantic representation to the values of the features that are relevant. In case lexicons of different

languages are loaded the values of these languages also have to be linked. To provide a semantics for these words again other words will probably have to be defined as well. The only way to get around this is to start at some point where values are atoms (this must be a small manageable set of concepts) and to proceed top-down from there. In order to get at this set all words that are used to define others have to be collected in a bottom-up way (which is another reason for a whole-sale parsing approach on a more superficial level clearly separated from the interpretation process). It may well turn out to be unavoidable to define the semantics of any part of the lexicon without having defined these words first (Meijs and Vossen 1991, 113-126, Dik, Meijs and Vossen 1991).

6. Conclusions

A complete analysis of the content of definitions in one run is neither possible nor desirable, since the full semantic impact of their content can only be expressed in a database in which the hierarchical relations can be exploited and this database, in its turn, can only be developed on the basis of knowledge about the information to be contained in it. In the same way as a lexicographer builds on the knowledge which he or she assumes available for the words that are used to define another word, so also in an LKB the knowledge of words can only truly be represented after these words have been defined. Once the data in the MRDs have been roughly analysed they are therefore first stored in the lexical database LDB. Being systematically accessible in the LDB these data will form the empirical starting point to set up a TYPE system, which can then be used to guide further interpretation of the values represented by the differentiae in the LKB. Such LKB lexicons can be initially set up in a TOP-DOWN fashion since from these inventories the set of "core" words that is used to define that lexicon can be isolated and represented first, and explained in terms of TYPEs based on notions and distinctions envisaged in linguistic theory. The result will be a 'controlled vocabulary' not in a didactic and educational sense but in a technical knowledge engineering sense, which is expressive enough to capture the information in the dictionaries (given its 'data-driven' basis) and is still fully formalised. This 'core' LKB and TYPE system (capturing both linguistically based classes and the most elementary words) can be seen as a general hypothesis about the structure and content of the lexicon. The LKB then forces one to finally implement the overall rough data in an explicit and consistent knowledge representation language. As such the model is continuously tested against the data extracted for specific domains. In this way an empirical (LDB) and deductive (LKB) set up can be combined so that the one compensates for the restrictions of the other. By extending the system to other parts of speech the coverage of the model is hopefully improved. The problem of interpreting the 'open-class' differentiae will then be minimalised, since the same frequent and general verbs and adjectives will re-occur all the time, and after a while most of these words will have got a formal representation. It is obvious that the current TYPE system has to be changed when other domains are included as well. In this respect having parsed all the definitions may also open up other more overall strategies such as automatic clustering of differentiae on a large scale. Where such clusters correspond with the taxonomic categories that arise from the entry word – genus relations they can be used to form the basis of a more general TYPE system.

Endnotes

1 This research was made possible by the publishers Longman and Van Dale who have been willing to let us use their MRDs for research purposes.

2 Non-individuated nouns are mass nouns, individuated nouns are count nouns. Individuation means that something is conceived as a distinguishable seperate entity and is therefore also countable.

3 The original Van Dale tape was enriched with the syntactic information made explicit in the Philips Rosetta project (Smit & Medema 1987).

Bibliography

AMSLER, R. (1981): "A taxonomy for English nouns and verbs". In: Proceedings of the 19th ACL, Stanford, pp. 133-138.

BOGURAEV, B., T. BRISCOE, J. CARROLL, and A. COPESTAKE (1991): Database Models for Computational Lexicography. Research Report RC 17120, IBM Research Center, Yorktown Heights, New York.

BRISCOE, T., and A. COPESTAKE (1991): "Sense extensions as Lexical Rules". In: Proceedings of the IJCAI Workshop on Computational Approaches to Non-Literal Language, Sydney Australia.

BRISCOE, T. (1991): "Lexical issues in Natural Language Processing". In E. Klein and F. Veltman (eds.) Natural Language and Speech, Springer-Verlag.

CALZOLARI, N. (1991): "Acquiring and representing semantic information in a lexical knowledge base". In: Proceedings of the ACL Siglex Workshop on Lexical Semantics and Knowledge Representation, Berkeley, California, pp. 188-197

CARPENTER, R. (1990): "Typed feature structures: Inheritance, (In)equality and Extensionality". In: Proceedings of the Workshop on Inheritance in Natural Language Processing, Tilburg, pp. 9-18.

CARROLL, J. (1990): Lexical Database System: User Manual. Esprit BRA-3030 Acquilex deliverable no. 2.3.3(c), Computer Laboratory, Cambridge University.

COPESTAKE, A. (1991): "The LKB:a system for representing lexical information extracted from a machine-readable dictionary". In: Proceedings of the ACQUILEX workshop on Default Inheritance in the Lexicon, Cambridge University.

COPESTAKE, A. and T. BRISCOE (1991): "Lexical operations in a Unification-based Framework". In: Proceedings of the ACL Siglex Workshop on Lexical Semantics and Knowledge Representation, Berkeley, California, pp. 88-101.

COPESTAKE, A. and JONES B. (1991): Support for multi-lingual lexicons in the LKB system. Computer Laboratory, Cambridge University.

DIK, S.C., W. MEIJS and P. VOSSEN (1991): "Lexigram: A functional lexicon for knowledge engineering". In: Proceedings of the LIKE workshop, Tilburg, 17-18 January 1991.

MEIJS, W. and P. VOSSEN (1991): "In so many words: Knowledge as a lexical phenomenon". In: Proceedings of the ACL Siglex Workshop on Lexical Semantics and Knowledge Representation, Berkeley, California, pp. 113-126.

PROCTOR, P. (ed) (1978): The Longman dictionary of contemporary English. London: Longman.

PUSTEJOVSKY, J. (1989): "Current issues in computational lexical semantics". In: Proceedings of the 4th European ACL, Manchester, pp. 17-25.

RODRIQUEZ, H. et al (1991): Guide to the extraction and conversion of taxonomies. Acquilex draft user manual, Universita Politechnica de Catalunya, Barcelona.

SMIT, H. and J. MEDEMA (1987): Description Van Dale Dictionary N-N. Internal Report Rosetta Translation Project, Philips Research Laboratories, Eindhoven.

STERKENBURG, J. van, and W. J. J. PIJNENBURG (1984): Groot woordenboek van hedendaags Nederlands. Van Dale Lexicografie, Utrecht.

VOSSEN, P. (1990): A Parser-grammar for the Meaning Descriptions of the Longman Dictionary of Contemporary English. Technical Report NWO, project no. 300-169-007, University of Amsterdam, Amsterdam.

VOSSEN, P. (1991a): Comparing noun-taxonomies cross-linguistically. Acquilex Working Paper no 014, Esprit BRA-3030, January 1991, Amsterdam.

VOSSEN, P. (1991b): Converting data from a lexical database to a knowledge base. Acquilex Working Paper no 027, Esprit BRA-3030, November 1991, Amsterdam.

VOSSEN, P. and A. COPESTAKE (1991): "Untangling definition structure into knowledge representation". In: Proceedings of the ACQUILEX workshop on Default Inheritance in the Lexicon, Cambridge University.

KEYWORDS: Lexical knowledge base, computational lexicology and lexicography, natural language processing, definition-parsing, concept-building, lexical-semantics.