Willy Martin, Free University of Amsterdam

On the parsing of definitions

Abstract: This paper describes a pilot project carried out in the Lexicology Subdepartment of the Free University of Amsterdam in which definitions are parsed by means of a concept-oriented parser. The basic features of the system are explained and their relevance shown both for practical purposes (such as information retrieval and 'term generation'), and for theoretical purposes (giving notions such as frames/templates, types, relations an explicit lexicological meaning).

1. Introduction

Computational lexicology/lexicography currently favours issues related to the *acquisition*, the *representation* and the *application* of lexical knowledge functioning within a NLP-environment. Especially the first issue, that of acquisition, is a central topic within so-called *re-usability* studies. At least if 're-use' is interpreted as "to exploit information implicitly or explicitly present in existing lexical resources" (see e.g. Calzolari 1991). This paper falls within the acquisition domain as it will explicitly deal with the extraction of (lexical) knowledge from (dictionary) definitions. However, it will be evident that acquisition without a representational framework does not make (much) sense. Furthermore we will also indicate how to use the knowledge obtained.

2. Definitions and Meaning Types

Our starting-point is the fact that words, with regard to their meaning, can be classified into *meaning types*. Words can have meanings that are predominantly conceptual, collocational, grammatical, figurative/associative, connotative, stylistic and contextual/discursive. The figure below makes clear that the meaning of a word moreover is not to be seen as one monolithic block, but as a conglomeration of meaning aspects (see also Neubert 1978 and Martin 1988 in this respect).

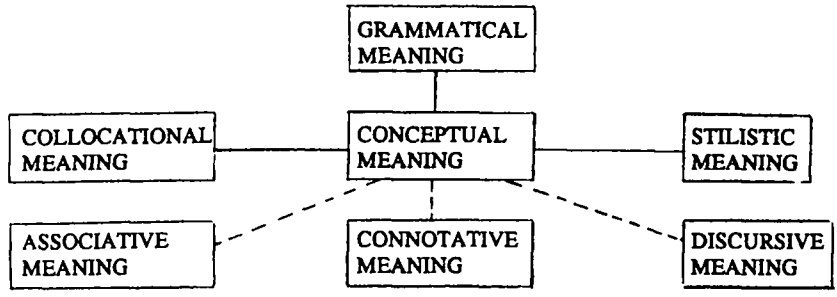

Fig. 1 Meaning aspects of words

Furthermore the above a.o. implies that

- although in many words the conceptual meaning is central (that is why it is put in the centre of the figure), this need not always be the case. So e.g. we find such words as 'thing' (which has primarily a discursive meaning), 'bloody' (as in 'you bloody fool', typically showing collocational meaning, i.c. intensification of the prototypical attribute of the noun it modifies), 'about' (which has a functional-grammatical meaning) etc.
- even in words where the conceptual meaning aspect is central, as a rule, one will find other meaning aspects as well. In this respect the full lines (——) indicate so-called 'obligatory' meaning aspects, whereas broken lines (- - - -) refer to 'optional' ones.

All in all however, one can state that the meaning of words, though not being monolithic, tends to be centered around one predominant aspect, thus giving rise to different lexical meaning types. As a corollary one expects different kinds of lexical meaning types to exhibit different descriptive treatments. So e.g. terms, showing 'par excellence' conceptual meaning, will require first and foremost conceptual meaning descriptions i.e. concept-oriented definitions. In what follows then we will concentrate on terms and their meaning as expressed in definitions, the typical locus for conceptual meaning information.

3. Terms and Concepts

For a start we will define terms as lexical expressions of concepts which are typically used within a particular knowledge domain (subject field) and by particular members of the linguistic community (experts in the domain). So e.g. one will find the concept "the part of the alimentary canal situated between the pharynx and the stomach" lexicalized in English as 'oesophagus' by experts in the field, whereas laymen will use here 'gullet' as an expression. Consequently the former will be considered a term, whereas the latter will not.

Although the situation is, by no means, always that clear-cut[1], it may suffice for our present purpose to state that terms

- are restricted as to their range of meaning aspects - as a rule they do not show any associative, connotative nor discursive meaning aspects, focussing primarily on conceptual meaning;
- are more concept-oriented in the sense that, in a first instance, they function within a conceptual system, only in a second instance, in a linguistic one;
- are linked to concepts[2] by means of their definitions.

A conceptual definition then will be one "in which a unique identification of the concept is provided for, with reference to the conceptual system of which it forms part and which classifies the concept within that system" (see Sager 1990, 39 in this respect).

From the above it should have become clear that 'terminological definitions' (here taken as definitions of terms) show a predominantly relational character: terms/concepts[3] can be defined by reference to all the terms/concepts surrounding them (in the field of knowledge they occur in) as represented in the figure below.

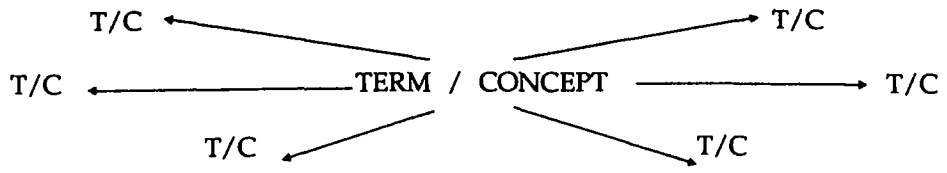

Fig. 2 Conceptual relations as definitions of terms

The next step for us will be to investigate whether this situation gives rise to a difference in parsing approaches.

4. 'Linguistic' vs. 'Concept-oriented' Parsing of Definitions

As stated before, the parsing of definitions has become rather popular recently. In most approaches, so e.g. Alshawi 1989 and Alshwede-Evens 1988, a kind of 'linguistic' approach is followed, meaning that, as a rule, most attention is spent on syntactic pattern matching before one starts building semantic structures.

A typical input/output in Alshawi 1989 e.g. reads

```
(roller coaster)

(a kind of small railway with sharp slopes and curves,
popular in amusement parks)

((CLASS RAILWAY) (COLLECTIVE KIND)
PROPERTIES (SMALL)))
```

Fig. 3 Analysis of 'roller coaster' according to Alshawi 1989

The output obtained typically is the result of applying phrasal patterns (e.g. det-adj-noun) to definition strings first[4], upon which, later on, semantic structures are built (e.g. in the above string noun = class, adj = properties etc.). Although we have to somewhat simplify Alshawi's approach, the basic idea should remain clear: this approach is syntax-driven, implying that the better the syntactic parse, the more one can semantically flesh out the result. The approach we are going to advocate starts from the opposite idea in that it tries to maximally profit from conceptual structuring in a particular subject-field so as to do a minimum of syntax. In the above case of 'roller coaster' e.g. this approach would first try to find out which concept type it was dealing with. Suppose it would find out that this was an ARTEFACT, then, making use of its knowledge about artefacts, such as

the fact that one expects artefacts to have a certain size, to consist of certain parts, to always have a certain function and/or use, and to possibly have a prototypical location, etc., would guide it in *specifying* the (unspecified) *expectation pattern* that has been triggered. In doing so, a minimum of syntax will be used and needed (see below).

5. Concept-oriented parsing of terms: Starting-point

Although we do not think there to be just one sacrosanct method for the parsing of definitions, it will be obvious that we consider the concept-oriented approach especially adequate in the case of terms. In what follows we briefly characterize the working method we used for the prototype we will present.

Being interested in *medical terms* we took what we considered to be the most central one, viz. 'disease', as a starting point from which to structure the knowledge domain. Thereafter

a) definitions of 'tokens' of this central 'type' were taken to generalize from;

b) specific combinations (conceptual collocations)[5] were taken to check and modify the model resulting from a).

As a first result a type definition was written represented here in BNF-notation[6]:

```
CONC_REL ::= isa CONC_TYPE(n)  FRAME(n)
CONC_TYPE(disease) ::= disease_lu_1
FRAME(disease) ::=  {SUBTYPE}? ({G_AFFECTS}* | {M_AFFECTS}*
                    {F_AFFECTS}*}+ {O_AFFECTS}+ {CAUSED_BY}*
                    {HAS_SYMPTOM}* {TRANSMITTED_BY}* {HAS_QUAL}*
SUBTYPE ::=  isa {word}+
G_AFFECTS ::= g_affects BODY_PART
BODY_PART ::= {word}+
etc.
```

From the above it should be clear that definitions of terms/concepts are regarded to be sets of relations between the term/concept in question and its neighbours (so e.g. 'hepatitis' is related to 'liver', the nature of the relation being 'g-affects'). The relations themselves are assembled in a frame and bound to a type (cf. FRAME(disease)). The slots of the frame call for fillers the domain of which is constrained (cf. g_affects BODY_PART).

In a first instance then no syntactic patterns occurring in definitions are looked for, instead of that, a *conceptual frame* is constructed, which acts as a kind of expectation pattern[7] steering the actual parsing.

6. The parsing algorithm

The basic algorithm, used in the prototype, can be roughly characterized as consisting of the following steps:

a. read definition
b. segment definition
c. look for head of definition

d. check clues
e. look for subhead(s) of definition
f. fill frame subhead(s) taking into account (checks on)
   - coordination
   - clues
   - postmodification
g. fill frame head
h. write sense frame

A typical input[8] reads like this:

> **rheumatoid arthritis**: a chronic disease of the musculo-skeletal system, characterized by inflammation and swelling of the joints, muscle weakness, and fatigue.

The corresponding output looks like

```
rheumatoid_arthritis:
[disease        g_affects       musc_skel_syst]
[disease        has_qual        chronic]
[disease        has_symptom     fatigue]
[disease        has_symptom     weakness]
[disease        has_symptom     inflammatic..]
[disease        has_symptom     swelling]
                        [weakness       g_affects    muscle]
                        [swelling       g_affects    joints]
                        [inflammation   g_affects    joints]
```

In what follows we will try to make clear the main features (a system leading to) such a result implies.

7. The system

7.1. Overall architecture

The parser under review is set up to analyze definitions of medical terms in English. As such it is but one of the components of a system consisting of

- a preprocessor
- a segmentor
- a lexicon
- a set of conceptual relations
- a parser proper

7.2. Input specifications

Up till now we have only dealt with definitions for diseases (terms for nosology concepts). These definitions can be taken from all kinds of sources, e.g. from termbanks or from (terminological) dictionaries. The example given above should make clear that we

work with *analytical definitions* exhibiting all kinds of *difficulties* in both *lexis* and *syntax* (such as structural ambiguities cf. inflammation and swelling of the joints; muscle weakness, and fatigue).
In a later stage the parser also should handle definitions such as the following[9]:

> **cystinosis**
>
> A disease characterized by lysosomal accumulation of free cystine and its crystallization in reticuloendothelial cells in many tissues, including bone narrow, liver, spleen, lymph node, kidneys, retina, uvea and conjunctiva. The disease probably is transmitted as an autosomal recessive. It is often associated with the Fancomi syndrome. Three clinical forms with different prognoses have been recognized: benign, intermediate, and nephrogenic. (definition taken from CHURCHILL'S ILLUSTRATED MEDICAL DICTIONARY).

7.3. Lemmatizer-tagger as Front-end

It goes without saying that a *lemmatizer-tagger* is a basic requirement for the efficient operation of the parser. This way text words (= word forms occurring in the definitional text) can be linked up with the items occurring in the lexicon (see below). For that purpose we use an adapted version of *Dilemma* (see Martin e.a. 1988 and Paulussen-Martin 1992).

7.4. Minimal syntax

After having been lemmatized and tagged, the definition gets split up into smaller parts (segments) by the segmentor. This module is a *minimal syntactic processor* which, on the basis of categorial information (such as Boolean values for NP compatibility and NP delimitation), delimits word groups in the input string. Unlike other approaches (cf. supra) which make use of syntactic pattern matching techniques, syntax is kept to a strict minimum as we assume that much of what is done (by others) syntactically, can be left out when one disposes of more powerful, i.c. conceptual, knowledge. As a result our input definition now looks as follows (| indicating delimiters, | | indicating boundaries):

> a chronic disease | of the musculo-skeletal system |, (characterized) | by inflammation | and swelling | of the joint(s) |, muscle weakness |, and fatigue | |.

7.5. Conceptual knowledge and calculation

The knowledge banks which form the core of the system are the lexicon and the set of *conceptual relations*. A lexical entry, e.g. 'aids', is a three-place predicate consisting of the actual lexeme, its concept type and its word category. So:

> (aids, concept (nosology-concept, aids, [u, u, u, u, u, u]), n).

As one will observe, the second argument, the concept type, consists of six unspecificied slots. The parsing of definitions is precisely aimed at *filling or specifying these slots*. It is the set of conceptual relations that a concept type may have that determines this specifica-

tion. At the moment such a *relational template* for diseases (nosology concepts), somewhat simplified, looks as follows (also see sub 5):

```
nos-concept
        g_affects (nos, (macro, micro, funct, embryo))
        o_affects (nos, organism)
        caused_by (nos, etiology)
        has_symptom (nos, finding)
        transmitted_by (nos, trans)
        has_qual (nos, qual)
```

Of course next to nosology or disease types we make use of other ones such as the figure underneath partially shows (at this moment we make use of 21 concept types).

(For the construction and definition of the underlying concept system we refer to Mars e.a. 1991 and Martin e.a. 1991.) At this point it is important to see that the implicit aids concept (and so the conceptual meaning of the lexeme 'aids') can a.o. be defined/specified by concepts taken from the domain of macro- and micro-anatomy, and that, in the given case, the relation between both arguments will be established. In this respect it is crucial for the parser to find the *head concept* of the definitional phrase. It does so by setting up a syntax-based hypothesis (taking the rightmost noun occurring in front of the first delimiter) and checking it with conceptual knowledge. In case of a definition of 'aids' as

> "a group of diseases secondary to a defect in cell-mediated immunity associated with a single newly discovered virus" (taken from Eurodicautom)

in a first instance 'group' will be taken up as head. Afterwards it will be rejected on *conceptual grounds*, a.o. because of the fact that 'group' is not considered a medical concept. In other cases *head shifting* will take place because of the fact that the head candidate can not be conceptually specified by its subheads (conceptual incompatibility between the assumed head and its subhead(s)). In the same vein, when being confronted with "classes of phenomena that present great difficulties for all syntactic formalisms (...) [One of], the most important of these being conjunction (...)" (Winograd 1983, 257-258), the parser again will solve (or try to solve) these cases by making use of conceptual information. That, in the case of rheumatoid arthritis (see definition in section 6), it does not yield parses such as 'swelling of muscle (weakness)' and that it manages to combine 'joints' both with 'swelling' and 'inflammation' proves it to be fairly successful in this respect. Other examples of conceptual calculation imply the establishment of new concept types out of old ones ('throat' e.g., being a macro-anatomical concept, becomes a finding concept when in combination with a qual concept such as in 'sore', this way 'sore throat' can 'fill' a symptom relation with a nosology concept), or rules for PP-attachment (compare: "a disease characterized by a sense of constriction *in the chest*" vs. ... constriction *in children* and young adults").

7.6. Frames

Given a definition of which the head or conceptual type has been established, the parser tries to fill its *conceptual template* or frame as much as possible. It does so by looking recursively for pre- and postmodifiers (the latter are called subheads), which 'fit' the

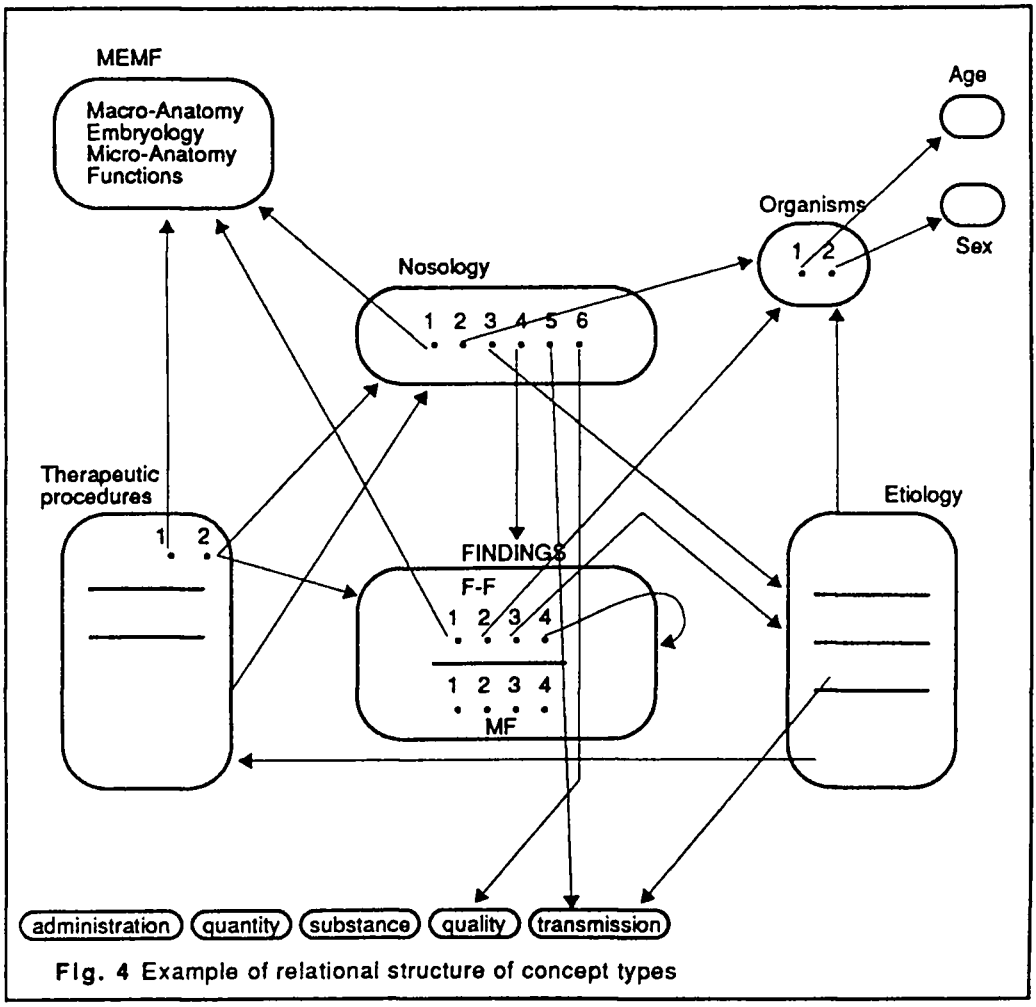

Fig. 4 Example of relational structure of concept types

Fig 4. Part of concept system 'diseases'

head (or its modifiers). Fitting here means that the concept type of the governed lexeme corresponds with the concept type of one of the arguments of the template of the governing lexeme. In the 'rheumatoid arthritis' example above e.g. the functional concept type of which 'musculo-skeletal system' is an instantiation, 'fits' or 'fills' the first argument or slot of the concept type 'rheumatoid arthritis' belongs to. M.m. the same can be said for all the other slot-fillers.

From the above it will have become clear that for the representation of conceptual meaning we have chosen for a *frame-based system* (see e.g. Habel 1985): concept types are defined by frames, i.e. sets of conceptual slots, attributes or features. By parsing defini-

tions concepts get instantiated, slots get concrete fillers as one more example can make clear.

> asthma: "a respiratory disorder, often of allergic origin, characterized by difficulty in breathing, wheezing, and a sense of constriction in the chest"

```
asthma:
[disease      caused_by      allergy]
[disease      g_affects      respiration]
[disease      has_symptom    constriction]
[disease      has_symptom    wheezing]
[disease      has_symptom    difficulty]
                  [constriction      g_affects   chest]
                  [difficulty        g_affects   breathing]
```

8. Usefulness

The parser described here tries to serve a twofold aim. In the first place its aim is *practical*. By making definitions conceptually explicit it is possible on the one hand to enhance the access to data bases (by making search items available in a systematic way), on the other hand, because of the fact that definitional knowledge becomes available in a systematic way, it also becomes possible now to generate from partial conceptual knowledge (answering such questions as: what is the term for the disease caused by HIV?, affecting the immune system? etc.) In the second place the system-cum-parser was set up as a pilot project in order to shed some light on the *lexicon structure*, more in particular on the *structure of the definitions* and the *organization of conceptual meaning* as – prototypically – expressed in terms. We hope that the framework of types, relations and frames as advocated in this approach, will be further elaborated upon not only by ourselves but by others as well.

Endnotes

1 Next to field-internal terms (terms used by experts, not by laymen), there are also field-external terms (terms used by both) e.g.. Compare in this respect 'hybrid computer' (field-internal) to personal computer (field-external).

2 Concepts are taken here to be the abstract categories representing the individual objects of our sensation, perception and imagination, or to quote Sager: "constructs of human cognition processes which assist in the classification of objects by way of systematic or arbitrary abstraction" (Sager 1990, 22).

3 As a rule there is a n:1-relationship between terms and concepts. Some schools of terminology even aim at a 1:1 relationship. When we write term/concept we actually mean term+/concept where + indicates one or more.

4 Actually Alshawi uses a hierarchy of phrasal patterns, see Alshawi 1989.

5 For a definition of conceptual collocations see Martin (to appear).

6 The following conventions are being used:
- words in lower case are terminal
- words in upper case are non-terminals
- constraints are given between ( ) so as to generalize over context-dependent values

- disjunctions are separated by |
- { }* = 0 or more
- { }? = 0 ore one
- { }+ = 1 or more

7 This does not imply that all features mentioned in the frame have equal status. In the 'disease'-frame e.g. some such as 'isa' and 'g/m_affects' have criterial status, meaning that they are necessary (not sufficient) criteria. Others such as has_symptom and the like are but expected (also see Cruse 1986, 20 ss. in this respect).

8 As a rule, up till now, only analytical definitions have been dealt with, which are, by the way, the only definitions which are recognized by terminology theory. This does not mean that practice always obeys theory here (see Sager 1990, 42 in this respect).

9 As stated in the preceding endnote up till now analytical definitions exhibiting NP-structures, have been dealt with. To cope with exemplars such as 'cystinosis' the parser should be both syntactically and conceptually adapted. However already some experience has been gained in parsing 'full text'-definitions (see Martin e.a. 1991).

10. Bibliography

ALSHAWI, H. (1989): "Analysing the dictionary definitions". In: Computational Lexicography for Natural Language Processing. Ed. by B.Boguraev and T.Briscoe. Longman, London/New York.

ALSHWEDE, T. and EVENS, M. (1988): "Generating a Relational Lexicon from a Machine-Readable Dictionary". In: IJL 1:3. OUP.

CALZOLARI, N. (1991): "Representation of semantic information in Acquilex". (Document 9/6, chapter 3 of Eurotra-7 study).

CRUSE, D.A. (1986): Lexical Semantics. OUP, Oxford.

HABEL, C. (1985): "Das Lexikon in der Forschung der künstlichen Intelligenz". In: Handbuch der Lexikologie. Ed. by C. Schwarze and D. Wunderlich, Königstein.

MARS, K. e.a. (1991) Eindrapportage Sapiens-project, UT-gedeelte. Universiteit Twente.

MARTIN, W. (1988) Een kwestie van woorden. VU, Amsterdam.

MARTIN,W, e.a. (1989): "Dilemma, an automatic lemmatizer". In: Colingua. Antwerp Papers in linguistics 56. Antwerp.

MARTIN, W. e.a. (1991): Over Atlex, Relset, Conceptor e.a., (VU-bijdrage tot het Sapiens-prototype). VU, Amsterdam.

MARTIN,W. (to appear): "Remarks on collocations in sublanguages". In: Proceedings ETI-conference Geneva, 2-4 October 1991. To be published in Terminologie et Traduction.

NEUBERT, A. (1987): "Kinds of lexical meaning". In: ZAA 26. Leipzig.

H.PAULUSSEN and W.MARTIN (1992): "Dilemma-2: a lemmatizer-tagger for medical abstracts". In: Proceedings 3rd conference on Applied Natural Language Processing. Trento. ACL.

REEDIJK, M. (1992), "Een conceptuele parser voor definities van medische termen". Scriptie. Amsterdam.

SAGER, J. (1990): A practical course in terminology processing. John Benjamins, Amsterdam/Philadelphia.

WINOGRAD, T. (1983): Language as a cognitive process. Vol 1: Syntax. Addison-Wesley, 1983.

KEYWORDS: computational lexicology, computational lexicography, parsing of definitions, concept-oriented parsing, terms, frames, relations