COLING 82, J. Horecký (ed.)
North-Holland Publishing Company

APPROACHES TO THESAURUS PRODUCTION

A. Michiels, J. Noël, English Department
University of Liège
Place Cockerill, 3,
B-4000 Liège
Belgium

We contrast two approaches to thesaurus production : the traditional and intuitive one versus the Amsler-type procedure, which interactively generates filiations among the genus words in a computerized dictionary. We discuss the application of such a procedure to our lexical data base (LONGMAN DICTIONARY OF CONTEMPORARY ENGLISH).

I INTRODUCTION

Since 1979 we have had available, by contract with LONGMAN Ltd, the computer tape of LDOCE (LONGMAN DICTIONARY OF CONTEMPORARY ENGLISH). Our main concern has been the development of a syntactico-semantic analyzer of general English making full use of all the formatted information contained in our dictionary file. (Michiels et al. 1980; Michiels 1982).

LDOCE is a medium-sized dictionary of core English containing some 60,000 entries which feature the following types of information :

a) fully formalized
   Part of speech (POS)
   Grammatical fields, i.e. sets of grammatical codes, which describe the environment that the code-bearing item can or must fit in.

What makes these grammatical fields particularly suitable for the purposes of machine disambiguation of natural language is that they are assigned to word-senses (definitions) as well as to whole lexical entries. An example is provided by the LDOCE entry CONSIDER (p. 233).

In the example string
I consider you a fool
the two-NP chain (YOU A FOOL) satisfies the $[X_1]$ code associated with the
NP_1 NP_2
second definition of the verb and enables the analyzer to select the appropriate definition in context ("scanning procedures" : cf. Michiels et al. 1980)

   Definition space, i.e.
      (i) semantic codes : inherent features for nouns, selectional restrictions for adjectives and verbs

Consider the entry HAMMER, verb. As the definition space does not appear in the printed version, we refer the reader to the computer file where, for the third definition, the semantic codes indicate that both the deep subject and the deep object must be [H] , i.e. [+HUMAN] .

      (ii) subject codes (field labels)

Ex : In the entry HAMMER, def. 3 is assigned SPXX (Sports) and def. 5 ECZS (EC : Economics, Z : subdivision indicator, S : Stock Exchange and Investment).

b) partly formalized

In most dictionaries, definitions are nothing else but strings of natural language, albeit of a special type (Smith and Maxwell 1973; Amsler 1980, p. 108). A first step towards formalizing definitions has been taken by the LDOCE lexicographers : all the LDOCE examples and definitions are written in a controlled defining vocabulary of some 2,100 items (lexemes - e.g. HISTORY - and morphemes - e.g. RE- and -IZATION - no morphological variants).

Our concern in this paper will be with how to produce thesauri from dictionary files. What prompts us to examine this problem is the existence of two contrasting approaches to thesaurus-production : the first is exemplified by LOLEX (LONGMAN LEXICON OF CONTEMPORARY ENGLISH, 1981), the second by Amsler 1980.

II THESAURUS PRODUCTION

Although LOLEX takes over a subset of the LDOCE definitions, both the choice of thesauric categories (e.g. J.212 verbs : DISMISSING AND RETIRING PEOPLE) and the assignment of a lexical item to one of several categories (e.g. DISBAND assigned to J. 212) are based on the lexicographer's intuition and knowledge of previous work in the field (cf. Roget's, etc.).

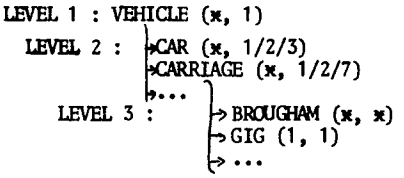
Amsler's approach is totally different (see Amsler 1980) : using as data base the computer files of the MPD (Merriam Pocket Dictionary) prepared by John Olney (Olney 1968), he develops an interactive procedure for thesaurus production. The first step is a manual selection and disambiguation of the GENUS TERMS in the definitions of nouns and verbs. By GENUS TERM is to be understood the first word of the definition which has the same POS as the definiendum and can serve as its superordinate. For example, in the first definition of HAMMER, the genus term is STRIKE, whereas in the fifth it is DECLARE.

It should be realized that genus term and syntactic head do not always coincide, and this mismatch is a major obstacle in the development of automatic procedures for genus term selection. Contrast in this respect the first and the second homographs of the LDOCE headword BOA (page 105). The second poses no problem : syntactic head and genus term are identical (GARMENT). In the first, however, the genus term is lodged inside the second OF-phrase, itself embedded in the first, which in its turn depends on the syntactic head ANY.

Once they have been selected, the genus terms are disambiguated with reference to the data base itself by selecting the appropriate homograph and definition numbers. A convenient example, drawn from LDOCE, is the disambiguation of the genus term CONSIDER in the definitions of LOOK ON[2] ([X 9 esp. as, with]: to consider; regard) CONSIDER here will be disambiguated as CONSIDER (*, 2) (* = non-homographic, 2 = second definition - cf. LDOCE entry CONSIDER, p. 233)

The next step is the use of a tree-growing algorithm, which Amsler has programmed and applied to his MPD data base. It is based on a filiation technique between lexical entries and genus terms. We shall illustrate it with respect to the item VEHICLE (*, 1) in our own data base. Descending the filiation path, the procedure will select all the items which use the word VEHICLE (*, 1) as genus term in their definitions. Among these are CAR (*, 1/2/3) and CARRIAGE (*, 1/2/7). CARRIAGE in turn functions as a genus term and yields its own sub-class, which contains, among others, the items BROUGHAM (*, * - non-homographic + a single definition) and GIG (1,1) - which are themselves defined by means of the genus term CARRIAGE. In our example, the procedure stops at BROUGHAM and GIG because these lexical items are nowhere in the dictionary used as genus terms. It results in a partial

taxonomy headed by the item VEHICLE :



Going up the filiation path from the word-sense VEHICLE (x, 1) one finds as syntactic head the pro-form SOMETHING - there is no genus term. Even if one is prepared to consider SOMETHING as the genus term (relaxing the POS identity condition), the thesauric link that is obtained does not yield more information than the semantic codes associated with the relevant definition.

A clear advantage of Amsler's procedure over intuitive thesaurus-production (as exemplified in LOLEX) is that it can lead to an improvement of the dictionary data base that is used as source. To take only one example : suppose that one is convinced that there should be a thesauric link (hyponym - superordinate) between VEHICLE and INSTRUMENT. If LDOCE is used as source data base for thesaurus - production, the link in question will not be retrieved (INSTRUMENT is not used as genus term in the LDOCE definition of VEHICLE (x, 1)), which inevitably raises the question of whether or not to revise the definition of VEHICLE.

III EXPLOITING LDOCE DEFINITIONS

When applied to the LDOCE definitions, Amsler's technique reveals an interesting consequence of a controlled defining vocabulary : the thesauric hierarchies are more shallow in LDOCE than in MPO (which does not feature a controlled defining vocabulary). To give an example, MPO defines LIMOUSINE by means of the genus term SEDAN.

Level one : VEHICLE
Level two : AUTOMOBILE
Level three : SEDAN
Level four : LIMOUSINE : ..... sedan

SEDAN is not available as genus term in LDOCE because it is not in the defining vocabulary. LIMOUSINE, defined by means of the genus term CAR, is level 3, not 4 in LDOCE :

Level one : VEHICLE
Level two : CAR
Level three : LIMOUSINE : ..... car .....

The shallow hierarchies based on LDOCE definitions are no doubt less revealing for the purpose of thesauric organisation. But the use of a controlled defining vocabulary makes it easier to process dictionary definitions in terms of both :

1) automatizing genus term selection and disambiguation and
2) parsing whole definition strings (as opposed to 1)

This is because the lexicon that the parser must have access to can be determined in advance. It is NOT open-ended (open-ended means, practically, as extensive as the defined vocabulary, i.e. the whole list of dictionary entries - cf. Amsler 1980, p. 109).

Schematically, the decision to use a controlled vocabulary to write dictionary definitions can have three undesirable consequences :

1).- reduction of the amount of information conveyed by the definition : OVERUSE of implicitly or explicitly partial definitions (in the sense of Bierwisch & Kiefer 1969, p. 66-68) - the latter are incomplete definitions which wear

their incompleteness on their sleeve, for example :

TARANTULA : spider of a certain kind.

2).- semantic overloading of all-purpose items such as GET, HAVE, MAKE, TAKE, etc.

E.g. KEEP (1, 8) : to have for some time or for more time (LDOCE, p. 603)

3).- uncontrolled increase in syntactic complexity in the differentia (non-genus part of the definition) :
   a) degree of embedding - not only in clauses, but also - and perhaps more importantly - in complex nominal groups (cf. Amsler 1980, p. 108 on ANT-EATING in the definition of AARDVARK)
   b) anaphoric relations
   c) scope relations (conjunction plays a prominent part here)

Compare the following two definitions of INSULIN

i).- OALDOCE (Hornby 1980) - 18 words
substance (a hormone*) prepared from the pancreas* of sheep used in the medical* treatment of sufferers from diabetes*
(* = does not belong to the LDOCE defining vocabulary).

ii).- LDOCE - 37 words
a substance produced naturally in the body which allows sugar to be used for ENERGY, esp. such a substance taken from sheep to be given to sufferers from a disease (DIABETES) which makes them lack this substance.
(ENERGY and DIABETES in capital letters because not in LDOCE defining vocabulary).

This third consequence stems from the avoidance of non-defining vocabulary items by means of PARAPHRASE, which displaces the burden towards syntactic elaboration, a point cogently made in Ralph 1980 (p. 117).

This "grammaticalization" of much of the information conveyed by LDOCE dictionary definitions points to the need to analyse whole definition strings rather than just the genus terms (see the process of ANNOTATING dictionary definitions in Noël et al. 1981).

Before we consider how to tackle the problem of disambiguating definition strings, we must examine a much easier way of retrieving at least some thesauric links from the LDOCE dictionary file. The LDOCE lexicographers sometimes provide ready-made thesauric links :

1).- cross-reference to an item belonging to the defining vocabulary :
CAPTAIN (2, *) : to be captain of; command; lead
↑ ↑ ↑
synonyms

2).- cross-reference to a non-defining vocabulary item :
ABBEY (*, 1) : .....; MONASTERY or CONVENT
↑ ↑ ↑
synonyms

3).- cross-reference to a non-defining vocabulary item inside an LDOCE definition, with a paraphrase in the defining vocabulary. An example is to be found in the LDOCE definition of INSULIN quoted above :
disease (DIABETES) which ....
↓ ↙ hyponym
genus term
↓
superordinate

In Noël et al. 1981 and Michiels et al. 1981 we have shown the power of the LDOCE grammatical codes to disambiguate items in context, more specifically in the context provided by the definition strings themselves. For instance, in the LDOCE definition of SERPENT (*, 2) :

- a wicked person who leads people to do wrong or harms those who are kind to him

the annotating process will select the V3 code for LEADS, because it occurs in the syntactic environment NP + TO + VP (NP = people, VP = do wrong) defined by V3 . This assignment enables the system to reject all the word senses for LEAD in LDOCE except the appropriate one (one out of nine; cf. entry LEAD[1] page 622).

We would like here to put forward a further possible exploitation of the LDOCE grammatical codes for the purpose of disambiguating dictionary definitions. It applies to genus terms and consists in the selection of a preferred word-sense for the genus term on the basis of a similarity in grammatical code between definiens and genus term. Let us turn back to our fourth example, the entry LOOK ON (2, *). The first genus term is CONSIDER. LOOK ON is assigned the grammatical code X9 . The second definition of CONSIDER is assigned the X (to be) 1, 7 code. The similarity in grammatical code X serves as criterion to disambiguate CONSIDER in the definition of LOOK ON as CONSIDER (*, 2).

The LDOCE semantic and subject codes can be exploited in a similar way. It can be hypothesized that the combined use of all the formalized information types in LDOCE will prove to have a high disambiguating power and turn out to be a useful tool for the setting up of thesauric classes.

A last point that we wish to touch on concerns the nature of the genus terms in a dictionary data base which makes use of a controlled defining vocabulary. The grammaticalization of information due to paraphrase in LDOCE gives rise to a special distribution of genus terms along a FULL WORD PROFORM gradient.

FULL WORD		PROFORM
LIQUID	SUBSTANCE	SOMETHING
		ANYTHING
		cf. LDOCE def. of VEHICLE (*, 1)
	ANALYSIS	PROCESS
		ACTION

(hyponym superordinate)

As compared with MPD, for example, LDOCE genus terms tend to cluster toward the proform end of the gradient. When the point is reached where the genus term does not provide more specific information than the semantic codes assigned to the definiendum, two conclusions can be drawn :

1).- the lexicographers of the source dictionary must consider whether their definition is appropriate, as it does not show the thesauric links perspicuously;
2).- the whole definition string must be processed and disambiguated, so as to retrieve the information that a dictionary which does not use a controlled defining vocabulary would have included in the genus term.

At the same time, the analysis of whole definition strings will reveal a number of thesauric links (such as that between INSTRUMENT and ACTION discussed in Michiels et al. 1980) that the study of genus terms, limited to the HYPONYM-SUPERORDINATE relation, is unable to retrieve.

REFERENCES

OALDOCE = Hornby, A.S., (editor-in-chief) OXFORD ADVANCED LEARNER'S DICTIONARY OF CURRENT ENGLISH, OUP London, 1980

LDOCE = LONGMAN DICTIONARY OF CONTEMPORARY ENGLISH, editor-in-chief : P. Procter, 1978

LOLEX = LONGMAN LEXICON OF CONTEMPORARY ENGLISH, Tom McArthur, 1981

Roget's = Roget's THESAURUS OF ENGLISH WORDS AND PHRASES, Penguin ed, 1966

Amsler 1980 = Amsler, R.A., THE STRUCTURE OF THE MERRIAM-WEBSTER POCKET DICTIONARY, TR-164, University of Texas at Austin Ph D., Dec. 1980

Bierwisch and Kiefer 1969 = Bierwisch, M. and Kiefer, F., Remarks on Definitions in Natural Language, in Kiefer, F. (ed), STUDIES IN SYNTAX AND SEMANTICS, D. Reidel, Dordrecht, Holland, 1969

Michiels 1982 = Michiels, A., EXPLOITING A LARGE DICTIONARY DATA BASE, Ph D thesis, University of Liège, 1982 (mimeographed)

Michiels et al. 1980 = Michiels, A., Mullenders, J., Noël, J., Exploiting a large data base by Longman, in COLING 80, 1980, p. 373-382

Michiels et al. 1981 = Michiels, A., Noël, J., Hayward, T., LE PROJET LONGMAN-LIEGE, DEVELOPPEMENTS THESAURIQUES, Congrès du LASLA, Liège, Novembre 1981

Noël et al. 1981 = Noël, J., Michiels, A. Mullenders, J.,LE PROJET LONGMAN-LIEGE, Congrès sur la lexicographie à l'âge électronique, Luxembourg, 1981

Olney 1968 = Olney, J., To all interested in the Merriam-Webster transcripts and data derived from them. Systems Development Corporation Document L-13579

Ralph 1980 = Ralph, B., Relative Semantic Complexity in Lexical Units, in COLING 80, 1980, p. 115-121

Smith and Maxwell 1973 = Smith, R. and Maxwell, E., An English dictionary for computerized syntactic and semantic processing, International Conference on Computational Linguistics, Pisa, 1973.