ACTA UNIVERSITATIS LODZIENSIS
FOLIA LINGUISTICA 36, 1997

PART I. LEXICON

Barbara Lewandowska Tomaszczyk

CORPUS LINGUISTICS AND THE LEXICON*

1. INTRODUCTION

An attempt is made in this study to present the place and function of computers in the lexicological analysis of naturai language and its lexicographic applications. Issues examined in this paper are connected with the acquisition of lexical knowledge from linguistic corpus data, reusability of the lexical knowledge in monolingual and multilingual lexicographic tasks as well as possible implications of such methodologies for the analysis of human language lexis.

2. CORPUS LINGUISTICS

Corpus Linguistics and, more precisely, Computational Corpus Linguistics is a relatively new development in the study of language, rapidly developing in the eighties (cf. the first corpus and its description by Kučera and Francis 1967, cf. also Makkai 1980, Meijs 1987, Sinclair 1991). The primary task of Corpus Linguistics is gathering and storing (originally in a book format, at present – in electronic form) of large quantities of authentic language data, spoken and written. The concept of **corpus** does not entail the sense that would cover any arbitrary collection of language data. A corpus, in the sense used here, is, as G. Leech [1991: 11] put

* This research has been partly covered by the European Community Program of Cooperation in Science and Technology between Western and Eastern European Countries Nr. 8453 as well as the TEMPUS grant Nr. 2087.

it, a collection of machine-readable linguistic data "designed or required for a particular 'representative' function". These 'databanks', as they are sometimes called, provide linguists with the real materials against which they can test their hypotheses.

Corpora of written language are more numerous. However, the work on speech output points to the need for corpora of spoken language. The authentic spoken data are more difficult to collect and an additional problem is an orthographic or phonetic transcription which is a very time consuming enterprise. Attempts at automatic speech analysis and transcription, are much less developed than methods of written text handling. At the same time, such large quantities of linguistic data can generate questions and issues which could have never been asked had such quantities of material not been collected and analysed. The management of such databanks, i. e. the access and retrieval of lexical information in the digital form, is made possible by computer software of tagging, parsing and concordancing type. Different computer programs are used to generate lexica, dictionaries and thesauri of the acquired lexical knowledge.

The linguistic corpus can be treated as a significant lexical resource, an embodiment of a token 'native speaker' with a cumulative competence of all and each native speaker-member of a given linguistic community. And yet, it should be borne in mind that such 'surface' phenomena, classically attributed to 'performance' rather than 'competence', as false starts, clumsy syntax, abductive lexical uses, often patently wrong, etc., are also there in the corpus and even though they may be treated as symptomatic of the synchronic language variability or of future linguistic developments, it is precisely for the analyst to decide what their current status is.

3. THE LEXICON

The lexicon used to be treated either as "an appendix of the grammar" [Bloomfield 1933] or as a depository of syntactic irregularities [Chomsky 1965]. With the rise of first semantically based models (e.g. Generative Semantics) and cognitively oriented approaches to language at present [*Frame Semantics* – Fillmore 1977, *Cognitive Grammar* – Lakoff and Johnson 1980, Langacker 1991, *Conceptual Semantics* – Jackendoff 1983, 1992], the place and role of the lexicon in linguistic models have radically changed. In the place of modular components incorporating syntax, phonology, semantics in autonomous compartments, cognitive grammatical models at present view those 'levels' as a continuum rather than modules, uniting lexicon, morphology and syntax, each associated with

a phonological and semantic structure. Semantics is treated as a separate component feeding syntax in Chomsky's models, while in the cognitive models it is equated with conceptualizations and encompasses different kinds of human experience immersed in the recognized social, physical, and linguistic context [cf. Langacker 1991: 2]. The semantic structure of a linguistic unit (lexical, phrasal, sentential, etc.) is characterized relative to *cognitive domains*, [*frames*] understood as structural conceptualizations of experience. This fact alone eliminates the feasibility of the inter-level linguistic distinctions as well as that of a strict dichotomy between the linguistic and the encyclopaedic knowledge. I would be prepared to defend a hypothesis that it is precisely the large language corpora that provide a tool to extract the knowledge of the lexis in its entirety in the context of the lexical frames. Such corpus-based lexical knowledge parallels the concept of the lexicon in its cognitive linguistic format.

4. LEXICAL ACQUISITION

The extraction of lexical information, termed also the *acquisition of the lexicon*, is based on the extracting of the lexical knowledge from the corpora of texts as well as from machine readable dictionaries (MRDs). Extracting of full lexical information from a large corpus manually could turn out to be a life time job. Therefore there is an urgent, and continually growing need to handle the search automatically. Corpora of English texts are quite numerous and grow rapidly (ICAME, Helsinki, Longman, Lancaster-Lund, Lund-London, Oslo-Bergen, etc.). The situation concerning other languages, including Polish, is much worse. In Poland, some newspaper publishers (e.g. *Gazeta Wyborcza*) are ready to share with researches their linguistic resources in electronic form. Other possibilities include scanning techniques – the OMNIPAGE packet at our disposal is being used for building *monolingual corpora* (Polish and English) as well as *bilingual* (translated texts) and *parallel corpora* (authentic texts covering the same domain in both languages). On the other hand, there exist in Poland a few centers which contribute to the domain of Language Technology. Activities represented there pertain to different topics in Language Technology, such as computational lexicography, speech generation and recognition, text understanding as well as expert systems for knowledge representation (for a more exhaustive list cf. Vetulani 1994).

The automatic acquisition of bilingual lexical knowledge (cf. examples below) from bilingual corpora is only in the *statu nascendi* at present. The available software, even though quite effective at the sentence-alignment

level, uses very little linguistic sophistication. The programs are based mostly on character lengths and/or item distribution in the sentence, assisted by the 'anchoring' techniques via proper names, numbers, or other fixed features in the texts. Lexical alignment is the next step in this process, currently under investigation (Lancaster UCREL team). The acquisition of lexical knowledge from non-translated parallel corpora, centered around similar domains, on the other hand, is only a matter of theorizing at present. Software for bilingual sentence-alignment (tested at different centers at present) combined with concordancing programs, may prove very useful not only for lexicography but also for CALL as well as in particular for the training of translators/interpreters and for the translation practice. Below are presented two pairs of sentences from the English-Polish bilingual corpus in the Department of English, at the University of Łódź, aligned at UCREL, University of Lancaster (A. McEnery and M. Oakes):

(1) **sub** d = 2 ----------g
Properly read and interpreted these statements give the reader a complete, synoptical picture of the firm's operations and results in quantified form.
----------g
Sprawozdania, właściwie czytane i interpretowane dają czytelnikowi kompletny, poglądowy obraz działalności firmy i jej wyników w wyrażeniu liczbowym.
con d + 232 ----------g
But I don't think there's anything wrong with the school, particulary, I've seen better and I've seen worse.
----------g
Ale nie wydaje mi się, żeby ze szkołą było coś szczególnie nie w porządku. Widziałem w życiu lepsze i gorsze.

Symbols used:
sub – one-to-one-sentence substitution
con – contraction [two sentences in one language corresponding to one sentence a in the other one]
d – distance

Interpretation:
a lower **d-score** signifies a more confident alignment;
d – depends on:
a) difference in length in characters
b) likelihood of alignment type.

5. REUSABILITY OF LEXICAL RESOURCES

As has been mentioned before, Lexical Databases (LDBs) and Lexical Knowledge Bases (LKBs) are products of lexical extraction from machine-readable corpora (i.e. texts and dictionaries) and can serve, in turn, a number of functions for both human as well as machine natural language processing tasks such as: verb frame acquisition, virtual lexica building, etc. This can improve the lexical acquisition process again and further enhance the LDB/LKB in the reusability cycle. To meet the requirements of reusability of lexical resources, there have to be assigned standardized mark-up, more specifically in *lemmatization*, part-of-speech (grammatical) *tagging*, syntactic, semantic, and discourse *parsing*. To meet these conditions and facilicate the interchange of corpus data a team of specialists grouped around the Text Encoding Initiative (TEI) originated in 1987 and sponsored by the *Computers in Humanities, Association for Literary & Linguistic Computing and Association for Computational Linguistics* is working on the production of a uniform system of guidelines for text encoding standards called SGML (*Standard General Markup Language*). Terms such as 'tagging' or 'parsing' are partly a misnomer. What they involve is in fact all that is pertinent to linguistic analysis, from sound to meaning. The approaches to these tasks center around two different methods, the first one based on the conceptual analysis, the other one–utilizing statistical methodology.

5.1. Cognitive models in NL processing

Probabilistic approaches in NL processing have been preceded by methods based on cognitive models of knowledge representation. Rooted in psychological findings of *spreading activation networks* [Anderson 1977], they too aim at capturing syntactic, semantic and discourse structures of natural language by means of graph diagrams of Augmented Transition Networks. The networks are composed of nodes representing states and arcs representing relations. The problem with cognitive modelling is that such parsers (frequently written in PROLOG) incorporate predicates based on truth conditional semantics. While useful in certain computer tasks, truth conditional semantics does not cover all aspects of natural language meaning and, as the cognitively oriented linguists would argue, is not what the natural language semantics is about at all. No wonder then that for the classical truth-conditional frameworks cognitive semantics is a notorious problem. McEnery [1993: 109] notices in his *Computational Linguistics*

in connection with that: "One of the problems with prototype semantics is that it is not always easy to specify what attributes an object is composed of, let alone enumerate the range values that attribute may take with respect to the object". The 'problem' McEnery points at here is exactly what prototype semantics is about. No wonder then that new tools have to be looked for in order to progress in natural language processing.

One of them, in the implementation stage, is an attempt for a natural language processing system to be based entirely on cognitive grammar principles. Its author, K. Holmqvist [1993], proposes a valence accommodation methodology to capture natural language comprehension. This approach, ambitious as it is, is in the prototype phase for the time being.

The ideal, hardly attainable at present, which would guarantee unproblematic reusability of data, would be an entirely 'theory neutral' acquisition of lexicon, however. The first approximation to the ideal might be approaches based on statistical probability techniques.

5.2. Statistical methods in parsing

Collecting corpora is only one side of the coin. Another, equally important one, as we have shown above, is to build computer programs that, first of all, tag the corpus sentences with the parts-of-speech labels, then syntactically analyse (parse) these sentences. The problem here is that practically each sentence in a corpus, if analysed in a content-free environment, can be proved ambiguous not only with respect to strict syntactic marking, but also with respect to reference fixing deictic elements. An automatic parser is not only to cover every possible structure of the sentence, but also to be able to choose from among them the most probable parsing in the particular context. In fact, then, the computer program is expected to be able to perform tasks left unaccounted for in many linguistic theories.

For such practical applications of computer in the domain where the analysis provided by the experts is not, or cannot be perhaps, fully axiomatized, it is the statistical methods that prove to be most promising. Fully automatic methods of statistically based parsing are underway in a few computer centers in Europe and the United States, but the results have not been published yet. Other methods, involving human-assisted parsing involve linguistic rules proposed by the analysts and their application based on the statistical algorithm [cf. McEnery 1993]. The grammar provided by the linguist is tested against the computer data and corrected ('debugged') accordingly. As a result of processing bilingual corpora in future, one could aim at building a *Computational Contrastive Grammar* of

the languages concerned, which could be reused in the tasks of text generation, e.g. for the machine translation. In order to apply this method, the program has to be trained on a set of manually-parsed sentences (usually around one million words), referred to as a *treebank* in the computational linguistics terminology. This treebank or *skeleton parsing* (cf. IBM/Lancaster group) is usually complemented by grammatical tagging, the corpus annotation technique of primary use in lexicographic practice. There are numerous tagsets available reported in the computational linguistic literature, the one, however, relativelly widely spread is the CLAWS Tagset (Consistent-Likelihood Automatic Word-Tagging System, versions one, two, and four cf. Black et al. 1993) referred to also as the Lancaster Tagset. The reported success rate of the CLAWS System reaches 94%. The examples of tagged and parsed sentences are drawn from the corpora of the UCREL group [Eyes and Leech 1993: 55]:

An example of a treebank text with appropriate *grammatical tags* (linked, by underlined symbols, to each word and punctuation mark) is drawn from the Canadian Hansard Corpus:

(2) May_VM I_PPIS1 say_VVI ,-, Mr._NNSB1 Speaker_NNS1 ,-, that_CST I_PPIS1 have_VHO sent_VVN a_AT1 copy NN1 of_IO this_DD1 to_II the_AT chairman_NNS1 of_IO the_AT committee_NNJ and_CC to_II the_AT two_MC ministers_NNS2 involved_VVN.-.

Tag symbols explained:

VM – modal auxiliary verb
PPIS1 – personal pronoun, first person, subjunctive, singular
VVI – general lexical verb infinitive
NNSB1 – noun, preceding singular noun of style or title, abbreviatory
NNS1 – noun of style, singular
CST – *that* as a conjunctor
VHO – base form *have*
VVN – past participal of lexical verb
AT1 – singular article
NN1 – singular common noun
IO – *of* as preposition
DD1 – singular determiner
II – general preposition
AT – article, neutral for number
NNJ – organization noun, neutral for number
CC – coordinating conjunction
MC – cardinal number, neutral for number
NNS2 – plural noun of style

VVN – past participle of lexical verb
– punctuation tag – full stop
[There are over 150 tags used in the CLAWS 2a Tagset by the Lancaster/IBM Group]
The examples of *grammatical tagging* are extracted from the Computer Manuals treebank of skeleton-parsed sentences:

(3) [N Files_NN2 N] [V[V& come_vvo [P into_II [N the_AT print_NN1 queue_NN1 N]P]V&] and_CC [V+either_LE[V& [V& match_VVO [N[G a_AT1 printer_NN1 's_SG] setup_NN1 N]V&] (_([V+ get_VVO [Tn printed_VVN Tn] V+])_) V&] or_CC [V+[V& do_VDO not_XX match_VVI V&] (_([V+ wait_VVO V+])_) V+]V+]V]._.

Additional symbols: in Constituent Labels for the UCREL Parsing Scheme
N – Noun phrase
V – Verb phrase & – Coordination – initial conjunct
+ – Coordination – non-initial conjunct

On top of grammatical parts-of-speech and syntactic tags, attempts are being made to mark the text with semantic and discourse labels (G. Leech). These techniques can bring about the refinement of the crude 'physical' tools for language analysis and introduce a more subtle methodology which can constrain the analysis to the level required for a number of computational applications such as e. g. automatic sense extraction, automatic abstracting, etc.

6. CONCLUSIONS

1. Computerized techniques of linguistic access and retrieval make it possible for the linguist to obtain a large spectrum of linguistic data in a relatively short time.
Lexical Knowledge Bases and their subdomains kept on-line and constantly updated, may be reused for different linguistic tasks (also bi- and multi--lingual).
Large linguistic corpora and MRDs provide data for automatic lexical data and knowledge acquisition.

2. Computerized language corpora, efficiently managed, and assisted by the automatic alignment software can be used for a number of tasks. In lexicography, CALL, translation, they provide: full lexical knowledge including frequencies and contextual modifications; collocations, associations, exploitations of conceptual and syntactic patterns (**microframe**); full infor-

mation on pragmatically-sensitive use (**macroframe**); information on similarities and contrasts in meaning.

3. Lexical semantic tagging supports the parts-of-speech and grammatized analysis and leads to automatic analysis of senses and its numerous applications such as automatic abstracting.

4. Statistically based technique in automatic annotation uncover the non-discrete nature of lexical senses and their inseparability from their knowledge frames.

REFERENCES

Anderson, J. (1977) "Induction augmented of transition networks". *Cognitive Science* 1: 125–157.

Black, E., Garsider, Leech, G. (eds) (1993), *Statistically-driven computer grammars of English: the IBM/Lancaster approach* [Language and Computers: Studies in Practical Linguistics, No. 8 ed. J. Aarts and W. Meijs] Amsterdam: Rodopi.

Fillmore, C. (1977), "Topics in lexical semmantics". In R. , W. Cole (ed.), *Current issues in Linguistic theory*, Bloomington: Indiana University Press, 76–138.

Gale, W. A., Church, K. W. (1993) "A Program for Aligning Sentences in Bilingual Corpora", *Computational Linguistics* 19, 1.

Holmqvist, K. (1993) *Implementing Cognitive Semantics.* Lund: Department of Cognitive Science, Lund University.

Jackendoff, R. (1983), *Semantics and cognition.* Cambridge, Mass.: MIT Press.

Jackendoff, R. (1992) "What is a concept?" In Lehrer, A. and E. F. Kitlay (eds). *Frames, fields and contrasts: New essays in semiotic and lexical organization.* Hillsdale, N. J.; Lawrence Erlbaum, 191–208.

Kučera, H. and W. N. Fracis (1967), *Computational analysis of present-day American English.* Providence: R. J. Brown University Press.

Lakoff, G. and M. Johnson (1980) *Metaphors we live by.* Chicago: University of Chicago Press.

Langacker, R. (1987) *Foundations of Cognitive Grammar*, vol. 1, Stanford: Stanford University Press.

Langacker, R. (1991) *Foundations of Cognitive Grammar*, Vol. 2. Stanford: Stanford University Press.

Leech, G. (1991) "The state of the art in corpus linguistics". In Aijmer, K. and B. Altenberg (eds). *English Corpus Linguistics* (Studies in honour of Jan Svartvik), Harlow: Longman, 8–29.

Makkai, A. (1989), "Theoretical and practical aspects of an associative lexicon for 20th century English:, In: L. Zgusta (ed.), *Theory and method in lexicography: Western and Non-Western Perspectives*, Columbia, S. Carolina: Horbeam Press, 125–46

McEnery, A. M. (1992) Computational Linguistics. Sigma Press.

Meijs, W. (ed.) (1987) *Corpus Linguistics and Beyond* (Proceedings of the Seventh International Conference on English Language Research on Computerized Corpora). Amsterdam: Rodopi.

Sinclair, J. (1991) *Corpus, condarce, collocation.* Oxford: Oxford University Press.

Vetulani, Z. (1991) "Polish activity in the domain of Language Technology (LT)". Paper presented at the meeting *Language and Technology*: Awarness Days in Luxembourg for Central and Eastern Europe, 13–14 January, Luxembourg.

Barbara Lewandowska-Tomaszczyk

JĘZYKOZNAWSTWO KORPUSOWE A LEKSYKA

Autorka analizuje miejsce i funkcje korpusów językowych w analizie leksykograficznej języka oraz w jej zastosowaniach leksykograficznych. Badana problematyka dotyczy akwizycji wiedzy leksykalnej z lingwistycznych danych korpusowych, wielokrotnego używania tej wiedzy w zadaniach leksykografii jedno- i wielojęzycznej oraz możliwych implikacji takich metodologii w analizie słownictwa języka naturalnego. W pracy poruszono zagadnienia automatycznej analizy językowych danych korpusowych i zaprezentowano ich przykłady na materiale języka angielskiego.