Fred Karlsson
Department of General Linguistics
University of Helsinki

LEXICOGRAPHY AND CORPUS LINGUISTICS

Opening Address at 5th Congress of EURALEX, Tampere, August 4, 1992.

(Published in Czech translation: "Lexikografie a korpusová lingvistika". Frantisek Cermák, Jana Klimová and Vladimir Petkevic, eds., *Studie z korpusové lingvistiky*. Acta Unviersitatis Carolinae. Philologica 3-4. 1997. Univerzita Karlova v Praze, Nakladatelství Karolinum, Praha 2000, 427-454.)

1. **Introduction**

Corpus linguistics is no newcomer on the linguistic scene. Many venerable schools of linguistics have based their descriptions of individual languages, and their theories of language, on observation and collection of natural data. Nevertheless, the early 1960s was an important turning-point because of the initiation of several large English corpus projects, notably the (British) Survey of English Usage and the (American) Brown Corpus (cf. Leech 1991, Svartvik 1992b). Several features were indicative of the new approach: the systematic design and collection of the corpora, their large size, the idea of making them generally available to the research community, careful evaluation of the problems of representativeness and sampling, and, especially in the case of the Brown Corpus, full-scale computerization.

In the subfield of computational linguistics, there has since the 1960s existed a well-established approach called "linguistic computing" or "linguistic data processing". This approach has been specifically geared towards the problems of large-scale text processing for linguistic and lexicographic purposes. One of the earliest lexicographic results in the field of linguistic computing was the *Frequency Dictionary of Present-Day Swedish* compiled by Sture Allen and his associates at the University of Gothenburg. The first volume was published in 1970.

The machine-readable corpora belonging to this first generation contained roughly 1 million word form tokens. A second generation was launched by the COBUILD project in lexical computing, at the University of Birmingham, starting in 1980. The COBUILD Main Corpus came to contain more than seven million words, supplanted with a reserve totalling 20 million words. Based on this corpus, the *Collins COBUILD Dictionary of the English Language* was compiled and published in 1987 by John Sinclair and his colleagues (cf. the report Sinclair 1987a, ed.). This marked a significant development in lexicography. The corpus was used as the main repository for information. The basic method was concordancing of word-forms (cf. Krishnamurthy 1987). The design of individual entries was primarily based upon corpus evidence. This is in contrast to classical lexicographical practice which often uses other dictionaries as main data and supplants these with less systematic corpus excerption.

One reason for the current corpus boom is thus the success of the two traditions of corpus linguistics briefly sketched above. There is another contributing factor. Recently, a growing number of grammarians and computational linguists of various theoretical persuasions have recognized the significance and usefulness of corpora, and have actively started using corpora and promoting corpus-based research.

Henceforth I shall include in the notion "corpus linguistics" both (i) the basic-level concern with collecting texts and using them for linguistic description, and (ii) the more recent concern with developing linguistically suitable computational tools for annotation and processing of large text corpora.

What can corpus linguistics, in the broad sense just defined, contribute to lexicography, on top of what COBUILD and other completed projects have already demonstrated by way of using raw concordances derived from large text corpora?

2. **Statistical corpus processing**

Statistical methods have a lot to offer. Church and Hanks (1990) have demonstrated how the information-theoretic notion **mutual information** can be used for finding significant collocations. Mutual information compares the probability of observing e.g. two words together (called joint probability), with chance, i.e. the probabilities of observing the words independently. The following data is borrowed from Church and Hanks (1990). Based on a 44 million word corpus of Associated Press news texts, it shows which words display significantly high mutual information values when co-occurring (to the right) with the verb *save*.

I(x,y)	f(x,y)	f(x)	x	f(y)	y
9.5	6	724	save	170	forests
9.4	6	724	save	180	$1.2
8.8	37	724	save	1697	lives
8.7	6	724	save	301	enormous
8.3	7	724	save	447	annually
7.7	20	724	save	2001	jobs
7.6	64	724	save	6776	money
7.2	36	724	save	4875	life
6.6	8	724	save	1668	dollars
6.4	7	724	save	1719	costs
6.4	6	724	save	1481	thousands
6.2	9	724	save	2590	face
5.7	6	724	save	2311	son
5.7	6	724	save	2387	estimated
5.5	7	724	save	3141	your
5.5	24	724	save	10880	billion
5.3	39	724	save	20846	million
5.2	8	724	save	4398	us
5.1	6	724	save	3513	less
5.0	7	724	save	4590	own
4.6	7	724	save	5798	world
4.6	7	724	save	6028	my
4.6	15	724	save	13010	them
4.5	8	724	save	7434	country
4.4	15	724	save	14296	time
4.4	64	724	save	61262	from
4.3	23	724	save	23258	more
4.2	25	724	save	27367	their
4.1	8	724	save	9249	company
4.1	6	724	save	7114	month

Several lexicographically useful observations can be made on the basis of such correlations. First, they give a principled ground for selecting example sentences demonstrating current usage. A good dictionary should of course contain frequent and typical example sentences, rather than rare, old-fashioned, or made-up ones. Second, the correlations point out morphosyntactically significant collocations, here, e.g. the syntagm *save from*. Third, they might even be of use in the most difficult of lexicographic tasks - that of discriminating between word senses, cf. *save forests - lives - jobs* vs. *save money - dollars*. Fourth, the probabilities aid in determining what the central senses are that should be put closer to the beginning of the completed lexical entry.

Obviously such data would be even more useful if it could be systematically linked to grammatical analysis and lemmatization of the context words. That would give the lexicographer a more condensed view of the currently relevant grammatical configurations (Church and Hanks have made such experients as well). The fragmentation caused by operating on the level of word-forms is a more tangible problem in languages that are more heavily inflected than English. E.g., Finnish nouns have maximally 2000 forms, and on some level of presentation one is likely to want to have data presented that generalize over all the forms of a lemma.

3. **Grammatical annotation of large corpora**

Lexicography would benefit from easy access to sufficiently large, morphosyntactically annotated (i.e. tagged) corpora, and to concordances derived from them. Such facilities would e.g. make it easier for the lexicographer to spot the **recurrent syntactic distributional patterns** of the word (s)he is working with. One of the practical problems with raw word-form concordances is that the lexicographer easily becomes flooded by thousands of examples. Given appropriate retrieval software, a tagged corpus could be interrogated for any structures that the lexicographer is interested in.

The classical ways of tagging unrestricted English text have been developed in the course of the compilation of the Brown and LOB corpora. Example (1) is the first sentence of the tagged version of the Brown Corpus, example (2) the first sentence of the tagged version of the LOB Corpus:

(1) the_AT *fulton_NP *county_NP *grand_NP *jury_NP said_VBD *friday_NR an_AT investigation_NN of_IN *atlanta's_NP$ recent_JJ primary_NN election_NN produced_VBD no_AT evidence_NN that_CS any_DTI irregularities_NNS took_VBD place_NN .

(2) a_AT move_NN to_TO stop_VB Mr_NPT Gaitskell_NP from_IN nominating_VBG any_DTI more_AP labour_NN life_NN peers_NNS is_BEZ to_TO be_BE made_VBN at_IN a_AT meeting_NN of_IN labour_NN MPs_NPTS tomorrow_NR .

Each word-form occurs together with a tag that uniquely identifies its part of speech and minor morphological features (number, case, tense, mood etc.), e.g. AT = article, NN = singular noun, VBD = verb in the past tense. These corpora and this simple type of representation have proven immensely useful. But of course we always long for more. Specifically, in addition to plain morphological analysis of word-forms, two more types of basic morphosyntactic information would be useful. First, for each word-form it should be indicated what its **lemma** is. This is equal to performing base-form reduction. Second, one would like to have at least a **surface syntactic analysis** of the corpus with a functional labelling of the basic constituents. The representation I have in mind is flat and simple, cf. (3); the sentence is borrowed from Hindle

(1989). Each word-form occurs at the far left margin and is associated with a set of automatically assigned **readings** (that are indented). Each reading consists of a **base-form** (lemma, e.g. *she* for *her*), a set of **intrinsic** and **morphological features** and finally one or more **syntactic labels** (here prefixed by "&") indicating the surface syntactic function of the word.

```
(3)
(*her
   (she * NonMod PRON PERS FEM GEN SG3 &GN>>))
(hand
   (hand N NOM SG &SUBJ))
(had
   (have PCP1:e/ing INF: SVO SVOC/A V PAST VFIN &+FAUXV))
(come
   (come PCP1:e/ing INF: SVC/A SV P/for PCP2 &-FMAINV))
(to
   (to INFMARK>> &INFMARK>>))
(rest
   (rest SVC/A SV SVO PCP1:ing INF: V INF &-FMAINV))
(on
   (on PREP &ADVL))
(that
   (that DET CENTRAL DEM SG &DN>>))
(very
   (very Attr A ABS &AN>>))
(book
   (book N NOM SG &<<P))
(.)
```

The features and labels are fairly self-explanatory. The syntactic model here entertained is a kind of dependency syntax. All modifiers are labelled at least by the parts of speech of their heads. Furthermore, the modifiers have an arrow ("<<" or ">>") indicating the direction where their head is to be found. Heads lack arrows. Thus, the label &P means "complement of the next preposition to the left", &DN>> means "determiner of the next head noun to the right" etc. Verbs are syntactically analyzed in regard to finiteness (+F, -F) and auxiliarity (MAIN, AUX). This simple flat notation makes it possible to compress and print out syntactic information equivalent to that of ordinary constituent structure trees. A further benefit is that word forms, base forms, morphological features, and syntactic labels all are objects of the same type. Thus, lexicographically relevant queries can be made even at the simple level of character strings using any combination of these entities.

Is it possible to have representations like (3) automatically assigned to unrestricted, running English text? The answer is positive. As for plain morphological tagging of English texts, there have been at least two stochastically working systems available for some years, CLAWS (Garside, Leech & Sampson 1987) and Parts (Church 1988). The somewhat richer representations (3), including lemmatization and syntactic coding, are possible to generate by **Constraint Grammar Parsing**.

The notion of Constraint Grammar Parsing has been developed by Karlsson (1990). The theory has been applied to English by Atro Voutilainen, Juha Heikkilä, and Arto Anttila (see the papers in Karlsson, Voutilainen, Heikkilä & Anttila 1991; id., in print). Constraint Grammar is a language-independent formalism for parsing unrestricted text. More precisely, CGP performs three basic types of **operations** on sentences consisting of morphologically analyzed word-forms having single or multiple readings:

- context-sensitive disambiguation, performed by **disambiguation constraints**,
- assignment of clause boundaries in complex sentences, performed by **clause boundary mappings**,
- assignment of surface-syntactic functions (grammatical labels such as finite main verb, subject, genitival premodifier to a noun), performed by **syntactic rules** of which there are two types: **morphosyntactic mappings** and **syntactic constraints** proper.

A full-fledged Constraint Grammar (CG) of a language contains rules of all these types. Such a grammar has been developed for English, and is being developed for Swedish and Finnish.

The roles of lexicon and morphology are much more prominent than in many recent rule-oriented approaches to parsing. But the basic innovations of CGP are (i) the explicit treatment of morphological ambiguities by constraints that discard improper alternatives, and (ii) the treatment of syntactic function assignment (parsing proper) by precisely the same mechanism. The constraints minimize impossible configurations by discarding either superfluous morphological readings or superfluous syntactic labels that occur in contexts where they are judged to be inappropriate. If the constraints succeed in eliminating all impossible alternatives, the unique correct analysis remains. If the constraints are lacking in some respect, or if the structure in question is genuinely ambiguous, several alternatives are left pending, leaving (spurious or true) ambiguity in the output. The overriding maxim of CGP is to discard as many impossible alternatives as possible as early as possible. The best way of discarding alternatives is to pick the correct one and discard the others. A formally weaker but valuable and frequently used method is to discard one alternative as improper, possibly leaving others waiting for final resolution by other constraints.

Full-scale CG parsing proceeds in five consecutive steps:

INPUT: raw ASCII text

1) PREPROCESSING which includes i.a. case conversion, sentence delimiter specification, and treatment of fixed syntagms,

2) LEXICON UPDATING, for spotting new words not included in the relevant **Master Lexicon** (which contains the core vocabulary of the language), and inserting them in one or more **domain-specific lexicons** to be used in conjunction with the Master Lexicon during morphological analysis,

3) MORPHOLOGICAL ANALYSIS which provides all word form tokens with their morphological readings, one or more,

4) LOCAL MORPHOLOGICAL DISAMBIGUATION: some morphological ambiguities arising due to compound formation may be discarded by mere local inspection of the configuration of readings derived for one word-form, without paying attention to contextual factors. Local disambiguation is relevant especially in languages such as Finnish, German, and Swedish, where productive processes of compound formation tend to overgenerate even if due care is taken in order to minimise such overgeneration.

5) CONSTRAINT GRAMMAR PARSING,

OUTPUT: morphologically and syntactically analyzed text.

The English morphological analyzer ENGTWOL, developed by Voutilainen and Heikkilä, works in conjunction with a Master Lexicon containing roughly 52,000 lexical entries which we take to be the English core vocabulary. ENGTWOL is capable of assigning proper morphological analyses to 97% of the word-form tokens in unrestricted English text. The residual normally consists of proper names not in the lexicon, or highly domain-specific words that are not part of the English core vocabulary.

For the purpose of properly analysing also the residual, two mechanisms are available. Either it is possible, by semiautomatic means, to construct a domain-specific additional lexicon and have it added to the ENGTWOL Master Lexicon before the texts are analyzed. The second option is to use heuristic morphological rules for predicting the morphological properties of unknown words. E.g., it is an almost safe guess that unknown words with an initial capital letter occurring non-initially in the sentence are names. The Constraint Grammar formalism contains mechanisms for expressing such heuristic decisions, working on top of the normal non-heuristic morphological analysis and CG disambiguation.

Now reconsider example (3). Above, this (laboratory!) sentence was presented in final analyzed form. But actually it is highly ambiguous on the morphological level. Here is the plain morphological analysis of this sentence, where ENGTWOL has generated all possible morphological analyses, but no disambiguation constraints have yet applied. It turns out that every word is at least two-ways ambiguous:

```
(4)
(*her
   (she * NonMod PRON PERS FEM GEN SG3)
   (she * NonMod PRON PERS FEM ACC SG3))
(hand
   (hand N NOM SG)
   (hand SVO SVOO PCP1:ing INF: V SUBJN VFIN)
   (hand SVO SVOO PCP1:ing INF: V IMP VFIN)
   (hand SVO SVOO PCP1:ing INF: V INF)
   (hand SVO SVOO PCP1:ing INF: V PRES -SG3 VFIN))
(had
   (have PCP1:e/ing INF: SVO SVOC/A V PAST VFIN)
   (have PCP1:e/ing INF: SVO SVOC/A PCP2))
(come
   (come PCP1:e/ing INF: SVC/A SV P/for PCP2)
   (come PCP1:e/ing INF: SVC/A SV P/for V SUBJN VFIN)
   (come PCP1:e/ing INF: SVC/A SV P/for V IMP VFIN)
   (come PCP1:e/ing INF: SVC/A SV P/for V INF)
   (come PCP1:e/ing INF: SVC/A SV P/for V PRES -SG3 VFIN))
(to
   (to PREP)
   (to INFMARK))
(rest
   (rest N NOM SG)
   (rest SVC/A SV SVO PCP1:ing INF: V SUBJN VFIN)
   (rest SVC/A SV SVO PCP1:ing INF: V IMP VFIN)
   (rest SVC/A SV SVO PCP1:ing INF: V INF)
   (rest SVC/A SV SVO PCP1:ing INF: V PRES -SG3 VFIN))
(on
   (on PREP)
   (on ADV))
(that
   (that **CLB CS)
   (that DET CENTRAL DEM SG)
   (that ADV AD-A)
   (that NonMod PRON DEM SG)
   (that NonMod **CLB Rel PRON SG/PL))
(very
   (very Attr A ABS)
   (very ADV AD-A))
(book
   (book N NOM SG)
   (book SVO PCP1:ing INF: V SUBJN VFIN)
   (book SVO PCP1:ing INF: V IMP VFIN)
   (book SVO PCP1:ing INF: V INF)
   (book SVO PCP1:ing INF: V PRES -SG3 VFIN))
(.)
```

In other words, the Constraint Grammar Parser is faced with a considerable task: (i) for each word-form, pick its correct morphological reading(s); (ii) assign to each reading all the possible syntactic functions it may have (e.g., a noun may occur as subject, object, indirect object, predicate complement of the subject, predicate complement of the object, premodifier of another noun, postcomplement of a preposition, etc.); and (iii), for each reading, pick the right syntactic function.

The English Constraint Grammar contains some 1.100 morphological disambiguation constraints developed by Atro Voutilainen, and some 400 syntactic constraints developed by Arto Anttila. 94-97% of the word-forms of unrestricted input text are fully disambiguated on the morphological level, with an error rate not exceeding 0.3%. This compares favourably to CLAWS and Parts both of which report an error rate around 3-4%. Furthermore, the Constraint Grammar Parser contains as an option, to be used on top of the ordinary safe constraints, the possibility of using heuristic constraints for disambiguating the remaining ambiguities. This option has a negligible influence on the error rate.

The syntactic constraints are presently less successful. Some 85% of the word-forms end up with a unique syntactic label, with an error rate of 3%. Work on the syntactic constraints is not yet completed.

Sentence (3) was rendered above in the form that the present version of the English Constraint Grammar yields, disregarding some technicalities. Arguably, there is a genuine ambiguity in (3) that ENG-CG has been "too effective" in obliterating, viz. the words *...the rest...* which also could be interpreted to form a noun phrase.

In Appendix 1, more complex and real examples, drawn from different text types, are provided of output produced by ENG-CG.

As of today, morphosyntactically annotated corpora are not available even for English that would be sufficiently large for full-scale lexicographic purposes. Recall that a 1 million word corpus provides reliable data only for the (central uses of the) 4000 or so most frequent words (Church & Hanks 1990, Sinclair 1991). Lexicography needs much more basic data.

The Cobuild group in Birmingham is creating a huge repository of English texts, called the **Bank of English**, which is to contain some 200 million words. This is a new magnitude of text corpora. Cobuild and the Department of General Linguistics at the University of Helsinki have recently agreed to collaborate on tagging the Bank of English, using the tools provided by the English Constraint Grammar Parser. The work starts in August, 1992. The first 20 million words should be tagged by the end of September, 1992, the first 100 million words by July, 1993, and the second 100 million words by July, 1994.

4. **Potential uses of a large annotated corpus**

A large tagged corpus where the word-forms have been disambiguated, lemmatized and identified in regard to morphological properties and surface syntactic function, has potential uses that range far beyond lexicography, e.g. in linguistic research and in several types of practical applications. Here we shall confine ourselves to some brief remarks on lexicographic use of large annotated corpora. What immediate benefits does such a corpus offer?

1) It makes possible the generation of frequency lists for lemmas. This is a good aid in selecting which words to include in the dictionary.

2) It makes possible comparisons concerning the lexical composition of text types on the lemma level. This sheds more light on the notions core vocabulary and specialized vocabulary. This too aids in the selection of words.

3) Lemmatization in conjunction with parts of speech and other morphological and syntactic features raise the level of abstraction somewhat and help the lexicographer in structuring the corpus data.

4) Collocation phenomena and syntactic frames are much easier to spot. As a case in point, consider the intransitive (objectless) use of the verb *sell*. By courtesy of Ken Church, AT&T Bell Labs, I have had recourse to the Wall Street Journal corpus containing some 43 million words. From this corpus I first picked (using standard Unix tools) all sentences containing the forms *sells* (N=332), *sold* (2409), *selling* (1607), and *sell* (2525). Then the sentences were analyzed by the English Constraint Grammar Parser, yielding output like that in Appendix 1. The parser offers various options for doing excerption, i.e. picking examples that conform to properties and restrictions specified by the user. Note that the full expressive power of CG output is available in the specification of the search key, i.e. word forms, base forms, morphological features, and syntactic functions. The hits of one simple search are presented in Appendix 2, all sentences with the finite verb form *sells* not followed by a direct object (N=49).

Note, by way of comparison, that the LOB Corpus contains 5 occurrences of the form *sells*. Because the Wall Street Journal Corpus is 43 times larger, we would by simple extrapolation expect some 215 instances of *sells*. The larger attested incidence of 332 is "natural" in view of the topics of WSJ.

Only two of the objectless instances in WSJ are such that *sells* is not accompanied by any adverbial. More than 30 instances include an indication of the price, strongly indicating that this is conventional usage and that the "price adverbial" is a salient or typical use of the objectless occurrences of *sell* that should be given some prominence in both dictionary and grammar.

Such corpus-based generalizations seem highly relevant. They have often been overlooked due to the lack of conveniently accessible data (however, cf. Atkins, Kegl & Levin 1988). In another connection, I have pointed out that the classical description of the verb *rain* as argumentless (zero-place) is slightly misleading. Some 40% of the corpus occurrences of *rain* are accompanied by a time adverbial (Karlsson 1983).

5) Typical, genuine example sentences are easier to spot (cf. Appendix 2).

6) There are lemma-internal preferences for particular word-forms that largely depend upon semantic factors (Sinclair 1991, Karlsson 1986). E.g., the average active/passive ratio for English verbs is around 9/1 (Svartvik 1966), but when I examined the WSJ instances of all forms of the lemma *sell*, the ratio turned out to be around 5/1. I.e., the passive is an especially typical form of

this lemma and this fact should arguably have some place in lexical and grammatical description.

As for the semantics of the selling transaction, the "reason" for the passive preference is obviously that the object being traded is a particularly salient participant, along with identification of the price, the buyer, and the time of the transaction. Such adverbials seem to roughly equally frequent in *sell*-passives. The seller, i.e. the surface subject, is a less important participant and is not reported at all (even as a *by*- or *through*-phrase) in more than half of the instances.

This paradigm-internal stratification of the inflected forms is even more clear in highly inflected languages. E.g., in Finnish, personal names favour the nominative case, mass nouns the partitive case, temporal nouns like *kesä* 'summer' the essive and adessive cases, etc. (Karlsson 1986). Detailed data on the incidence of the inflectional forms of the Finnish word *käsi* 'hand' are provided in Appendix 3. There is a "natural" correlation between the case form preferences, on the one hand, and the meaning of the lemma root and the cases, on the other. Presently, such facts concerning usage are not stated clearly either in dictionaries (which treat lemmas) nor in mainstream grammars (which strive for grammatical generalizations).

It seems reasonable that at least learners' dictionaries for inflected languages should contain some indications of what inflected forms are in active use, e.g. as entries of their own pointing to the main entry of the conventional citation form. Foreigners learning Finnish would get much help of a dictionary telling them that *kädessä* 'in the hand' is a form of *käsi*.

Presently, the need for including other than citation forms in dictionaries does not seem to have been recognized (cf. Mugdan 1989, Zgusta 1989, Bergenholtz & Mugdan 1990).

7) Annotated corpora might even aid in the difficult process of sense discrimination. Part of this task is facilitated if the lexicographer has access to collocation data and argument structures on a suitable level of abstraction. However, it is unlikely that any significant breakthroughs are to be expected in this domain. Meanings and boundaries between hypothesized senses of a word are so fuzzy and context-dependent that it is perhaps even unreasonable to be looking for a final solution (cf. Atkins 1991a, 1991b, Levin 1991, and Lenders 1991 for discussion).

References

Aijmer, K. & Altenberg, B. 1991 (eds.). *English Corpus Linguistics. Studies in Honour of Jan Svartvik*. London: Longman.

Atkins, B.T.S. 1991a. "Building a Lexicon: The Contribution of Lexicography". *International Journal of Lexicography* 4.3, pp. 167-204.

-- 1991b. "Corpus Lexicography: The Bilingual Dimension". CLAL, Vol. I, pp. 43-64.

Atkins, B.T., Kegl, J. & Levin, B. 1988. "Anatomy of a Verb Entry: from Linguistic Theory to Lexicographic Practice". *International Journal of Lexicography* 1:2, pp. 84-126.

Bergenholtz, H. 1989. "Probleme der Selektion im allgemeinen einsprachigen W|rterbuch". Hausmann & al. (1989, pp.772-779).

Bergenholtz, H. & Mugdan, J. 1990. "Formen und Probleme der Datenerhebung II: Gegenwartsbezogene synchrone W|rterbucher". Hausmann & al. (1990, pp. 1611-1625).

Boguraev, B. 1991. "Building a Lexicon: The Contribution of

Computers". *International Journal of Lexicography* 4.3, pp. 227-260.

Church, K.W. 1988. "A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text". *Proceedings of the Second Conference on Applied Natural Language Processing*, ACL, pp. 136-143.

Church, K.W. & Hanks, P. 1990. "Word Association Norms, Mutual Information, and Lexicography". *Computational Linguistics* 16:1, pp. 22-29.

CLAL 1991 = *Computational Lexicology and Lexicography*. I, II. Linguistica Computazionale VI, VII. Pisa: Giardini.

Garside, R., Leech, G. & Sampson, G. 1987 (eds.). *The Computational Analysis of English*. London and New York: Longman.

Hausmann. F.J. 1990. "Das Flexionsw|rterbuch". Hausmann & al. (1990, pp. 1311-1314).

Hausmann, F.J., Reichmann, O., Wiegand, H.E. & Zgusta, L. 1989. *Dictionaries. An International Encyclopedia of Lexicography*. First Volume. Berlin: Walter de Gruyter.

-- 1990. *Dictionaries. An International Encyclopedia of Lexicography*. Second Volume. Berlin: Walter de Gruyter.

Hindle, D. 1989. "Acquiring Disambiguation Rules from Text". *Proceedings of the 27th Annual Meeting of the ACL*, pp. 118-125.

Karlsson, F. 1983. "Prototypes as Models for Linguistic Structure". F. Karlsson (ed.), *Papers from the Seventh Scandinavian Conference of Linguistics*, Publications of the Department of General Linguistics, University of Helsinki, No. 11, pp. 583-604.

-- 1986. "Frequency Considerations in Morphology". *Zeitschrift fur Phonetik, Sprachwissenschaft und Kommunikationsforschung* 39, pp. 19-28.

-- 1990. "Constraint Grammar as a Framework for Parsing Running Text". H. Karlgren (ed.), *Papers Presented to the 13th International Conference on Computational Linguistics*, Helsinki 1990, Vol. 3, pp. 168-173.

Karlsson, F., Voutilainen, A., Heikkilä, J. & Anttila, A. 1991. "Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text, with an Application to English". *Natural Language Text Retrieval. Workshop Notes from the Ninth National Conference on Artificial Intelligence*. American Association for Artificial Intelligence, Anaheim, Cal.

-- (in print). Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text. 340 pp.

Krishnamurthy, R. 1987. "The Process of Compilation". Sinclair (1987a, pp. 62-85).

Leech, G. 1991. "The State of the Art in Corpus Linguistics". Aijmer & Altenberg (1991, pp. 8-29).

Lenders, W. 1991. "What's in a Lexical Entry". CLAL, Vol. II, pp. 45-63.

Levin, B. 1991. "Building a Lexicon: The Contribution of Linguistics". *International Journal of Lexicography* 4.3, pp. 205-226.

Mugdan, J. 1989. "Information on Inflectional Morphology in the General Monolingual Dictionary". Hausmann & al. (1989, pp.518-525).

Renouf, A. 1987. "Corpus Development". Sinclair (1987a, pp. 1-40).

Renouf, A. & Sinclair, J. McH. 1991. "Collocational Frameworks in English". Aijmer & Altenberg (1991, pp. 128-143).

Sinclair, J. 1987a (ed.). *Looking up. An Account of the COBUILD Project in Lexical Computing*. London: Collins ELT.

-- 1987b. "Grammar in the Dictionary". Sinclair (1987a, pp. 104-115).

-- 1987c. "The Nature of the Evidence". Sinclair (1987a, pp. 150-159).

-- 1991. Corpus, Concordance, Collocation. Oxford: Oxford

University Press.

Sinclair, J.M. & Kirby, D.M. 1991. "Progress in Computational Lexicography". CLAL, Vol. II, pp. 233-257.

Svartvik, J. 1966. *On Voice in the English Verb*. Mouton: The Hague.

-- 1992a (ed.). *Directions in Corpus Linguistics. Proceedings of the Nobel Symposium 82, Stockholm, 4-8 August 1991*. Mouton the Gruyter: The Hague.

-- 1992b. "Lexis in English Language Corpora". Plenary paper, 5th Congress of EURALEX, Tampere.

Zgusta, L. 1989. "The Influence of Scripts and Morphological Language Types on the Structure of Dictionaries". Hausmann & al. (1989, pp. 296-305).

APPENDIX 1. Application of ENGCG (Constraint Grammar for English) to short fragments from five types of text.

*Note: Understrokes "_" were inserted by the preprocessor. They mark compounds. Readings with the feature "?" have been heuristically predicted by the parser. Syntactic function labels are prefixed by "&". Syntactic errors are indicated by "***". The central syntactic labels are briefly explicated withing brackets, [...], the first time they occur.*

(Text 1: Grolier International Encyclopedia, "World War I")

The assassination of the Austrian archduke Franz Ferdinand in Sarajevo in 1914 proved to be the spark that ignited World War I (1914-18). Called "the Great War," it quickly came to involve all the great powers of Europe and eventually most countries of the world, and cost the lives of more than 8 million soldiers. Among the causes of the war were rising nationalist sentiment (manifested both in the chauvinism of the great European powers and in the unrest among the subject peoples of the multinational European empires), colonial and economic rivalries, the formation of hostile alliance systems, and arms races, all of which contributed to the growing sense of international tension during the prewar years.

```
(*the
   (the * DEF DET CENTRAL ART SG/PL (&DN>>))) [modifier of next noun]
(assassination
   (assassination N:ASSASSINATION PCP1:ON/NG INF:ON/E N NOM SG
(&SUBJ))) [subject]
(of
   (of PREP (&NOM-OF))) [modifier of previous nominal]
(the
   (the DEF DET CENTRAL ART SG/PL (&DN>>)))
(*austrian
   (austrian ** NOMINAL A ABS (&AN>>))) [modifier of next noun]
(archduke
   (archduke TITLE N NOM SG (&N))) [appositive relation]
(*franz
   (franz ** PROPER N NOM SG (&N)))
(*ferdinand
   (ferdinand ** PROPER N NOM SG (&<<P))) [complement of previous
preposition]
(in
   (in PREP (&ADVL))) [adverbial]
(*sarajevo
   (*sarajevo ? N NOM SG (&<<P)))
(in
   (in PREP (&<<NOM &ADVL)))
(1914
   (1914" NUM CARD (&<<P)))
(proved
   (prove N:PROOF SVOC/N SVOC/A SVC/N SVC/A SVO SV PCP1:E/ING INF:/
V PAST VFIN (&+FMAINV))) [finite main verb]
(to
   (to INFMARK (&INFMARK))) [infinitive marker]
(be
   (be PCP1:/ING INF:/ SV SVC/N SVC/A V INF (&-FAUXV &-FMAINV)))
[nonfinite auxiliary, nonfinite main verb]
(the
   (the DEF DET CENTRAL ART SG/PL (&DN>>)))
(spark
   (spark N NOM SG (&PCOMPL-S))) [predicate complement of the
subject]
(that
   (that NONMOD **CLB REL PRON SG/PL (&SUBJ)))
(ignited
   (ignite N:IGNITION SVO SV PCP1:E/ING INF:/ V PAST VFIN
(&+FMAINV)))
(*world_*war_*i
   (world_*war_*i ** PROPER N NOM SG (&OBJ))) [direct object]
($()
(1914-18
   (1914-18" NUM CARD (&NPHR))) [nominal fragment]
($))
($.)
(*called
   (call * N:/ SVOC/N SVOC/A SVO SVOO SV P/FOR P/ON PCP1:/ING INF:/
PCP2 (&AN>>))) ***
($")
(the
   (the DEF DET CENTRAL ART SG/PL (&DN>>)))
(*great
   (great * A ABS (&AN>>)))
(*war
   (war * N NOM SG (&OBJ)))
($,)
($")
(it
   (it NONMOD PRON NOM SG3 SUBJ (&SUBJ)))
(quickly
   (quick DER:LY ADV (&ADVL)))
(came
   (come PCP1:E/ING INF:/ SVC/A SV P/FOR V PAST VFIN (&+FMAINV)))
(to
   (to INFMARK (&INFMARK)))
(involve
   (involve N:/MENT SVO P/IN P/WITH PCP1:E/ING INF:/ V INF
(&-FMAINV)))
(all
   (all QUANT DET PRE SG/PL (&QN>>))) [modifier of next noun]
(the
   (the DEF DET CENTRAL ART SG/PL (&DN>>)))
(great
   (great A ABS (&AN>>)))
(powers
   (power N NOM PL (&OBJ)))
(of
   (of PREP (&<<NOM-OF)))
(*europe
   (europe ** PROPER N NOM SG (&<<P)))
(and
   (and CC (&CC)))
(eventually
   (eventual DER:LY ADV (&ADVL &AD-A>>))) [adverbial, modifier of
next adjective]
(most
   (many QUANT DET POST SUP PL (&QN>>)))
(countries
   (country N NOM PL (&SUBJ &OBJ &<<P)))
(of
   (of PREP (&<<NOM-OF)))
(the
   (the DEF DET CENTRAL ART SG/PL (&DN>>)))
```

```
(world
   (world N NOM SG (&<<P)))
($,)
(and
   (and CC (&CC))) [conjunction]
(cost
   (cost PCP1:/ING INF:/ SV SVO SVOO V PAST VFIN (&+FMAINV))
   (cost PCP1:/ING INF:/ SV SVO SVOO V INF (&-FMAINV)))
(the
   (the DEF DET CENTRAL ART SG/PL (&DN>>)))
(lives
   (life N NOM PL (&OBJ)))
(of
   (of PREP (&<<NOM-OF)))
(more=than
   (more=than ADV (&ADVL &AD-A>>)))
(8
   (8 NUM CARD (&<<P &QN>>)))
(million
   (million NUM CARD (&QN>>)))
(soldiers
   (soldier N NOM PL (&<<P)))
($.)
(*among
   (among * PREP (&ADVL)))
(the
   (the DEF DET CENTRAL ART SG/PL (&DN>>)))
(causes
   (cause N NOM PL (&<<P)))
(of
   (of PREP (&<<NOM-OF)))
(the
   (the DEF DET CENTRAL ART SG/PL (&DN>>)))
(war
   (war N NOM SG (&<<P)))
(were
   (be PCP1:/ING INF:/ SV SVC/N SVC/A V PAST VFIN (&+FAUXV)))
(rising
   (rise PCP1:E/ING INF:/ SV PCP1 (&-FMAINV))) ***
(nationalist
   (nationalist N NOM SG (&NN>>))) [modifier of next noun]
(sentiment
   (sentiment N NOM SG (&SUBJ)))
($()
(manifested
   (manifest N:/ATION SVO P/IN PCP1:/ING INF:/ PCP2 (&-FMAINV
&AN>>)))
(both
   (both CC (&CC))
   (both NONMOD QUANT PRON NOM PL (&OBJ))) ***
(in
   (in PREP (&ADVL)))
(the
   (the DEF DET CENTRAL ART SG/PL (&DN>>)))
(chauvinism
   (chauvinism -INDEF N NOM SG (&<<P)))
(of
   (of PREP (&<<NOM-OF)))
(the
   (the DEF DET CENTRAL ART SG/PL (&DN>>)))
(great
   (great A ABS (&AN>>)))
(*european
   (european ** NOMINAL A ABS (&AN>>)))
(powers
   (power N NOM PL (&<<P)))
(and
   (and CC (&CC)))
(in
   (in PREP (&ADVL)))
(the
   (the DEF DET CENTRAL ART SG/PL (&DN>>)))
(unrest
   (unrest -INDEF N NOM SG (&<<P)))
(among
   (among PREP (&<<NOM &ADVL)))
(the
   (the DEF DET CENTRAL ART SG/PL (&DN>>)))
(subject
   (subject N NOM SG (&NN>>)))
(peoples
   (people N NOM PL (&<<P)))
(of
   (of PREP (&<<NOM-OF)))
(the
   (the DEF DET CENTRAL ART SG/PL (&DN>>)))
(multinational
   (multinational A ABS (&AN>>)))
(*european
   (european ** NOMINAL A ABS (&AN>>)))
(empires
   (empire N NOM PL (&<<P)))
($))
($,)
(colonial
   (colonial A ABS (&AN>>)))
(and
   (and CC (&CC)))
(economic
   (economic DER:IC A ABS (&AN>>)))
(rivalries
   (rivalry PCP1:RIVALLING INF:RIVAL N NOM PL (&SUBJ &OBJ &APP
&<<P))) [&app = apposition]
($,)
(the
   (the DEF DET CENTRAL ART SG/PL (&DN>>)))
(formation
   (formation PCP1:FORMING INF:FORM N NOM SG (&SUBJ &OBJ &<<P)))
(of
   (of PREP (&<<NOM-OF)))
(hostile
   (hostile A ABS (&AN>>)))
(alliance
   (alliance N NOM SG (&NN>>)))
(systems
   (system N NOM PL (&<<P)))
($,)
(and
   (and CC (&CC)))
(arms_races
   (arms_race N NOM PL (&SUBJ &OBJ &APP &<<P)))
($,)
```

```
(all
   (all NONMOD QUANT PRON SG/PL (&SUBJ)))
(of
   (of PREP (&<<NOM-OF)))
(which
   (which NONMOD REL PRON WH NOM SG/PL (&<<P)))
(contributed
   (contribute N:CONTRIBUTION SVO SV PCP1:E/ING INF:/ V PAST VFIN
(&+FMAINV)))
(to
   (to PREP (&ADVL)))
(the
   (the DEF DET CENTRAL ART SG/PL (&DN>>)))
(growing
   (grow PCP1:/ING INF:/ SVC/A SV SVO P/ON PCP1 (&NN>> &AN>>)))
(sense
   (sense N NOM SG (&<<P)))
(of
   (of PREP (&<<NOM-OF)))
(international
   (international A ABS (&AN>>)))
(tension
   (tension N NOM SG (&<<P)))
(during
   (during PREP (&<<NOM &ADVL)))
(the
   (the DEF DET CENTRAL ART SG/PL (&DN>>)))
(prewar
   (prewar A ABS (&AN>>)))
(years
   (year N NOM PL (&<<P)))
($.)
```

(Text 2: Lewis Carroll, Alice's Adventures in Wonderland)

Alice was beginning to get very tired of sitting by her sister on the bank, and of having nothing to do: once or twice she had peeped into the book her sister was reading, but it had no pictures or conversations in it, `and what is the use of a book,' thought Alice `without pictures or conversation?'

```
(*alice
   (alice ** PROPER N NOM SG (&SUBJ)))
(was
   (be PCP1:/ING INF:/ SV SVC/N SVC/A V PAST SG1,3 VFIN (&+FAUXV)))
[finite auxiliary]
(beginning
   (begin PCP1:/NING INF:/ SV SVO PCP1 (&-FMAINV))) [nonfinite main
verb]
(to
   (to INFMARK (&INFMARK)))
(get
   (get PCP1:/TING INF:/ SVOC/A SVC/A SVO SVOO SV V INF (&-FMAINV)))
(very
   (very ADV AD-A (&AD-A>>)))
(tired
   (tire SV SVO P/OF PCP1:E/ING INF:/ PCP2 (&AN>>))) ***
(of
   (of PREP (&ADVL)))
(sitting
   (sit PCP1:/TING INF:/ SV SVO P/ON PCP1 (&<<P-FMAINV))) [modifier
of previous preposition]
(by
   (by PREP (&<<NOM &ADVL)))
(her
   (she NONMOD PRON PERS FEM GEN SG3 (&GN>>)))
(sister
   (sister TITLE N NOM SG (&<<P)))
(on
   (on PREP (&<<NOM &ADVL)))
(the
   (the DEF DET CENTRAL ART SG/PL (&DN>>)))
(bank
   (bank N NOM SG (&<<P)))
($,)
(and
   (and CC (&CC)))
(of
   (of PREP (&<<NOM-OF)))
(having
   (have PCP1:E/ING INF:/ SVO SVOC/A PCP1 (&<<P-FMAINV)))
(nothing
   (nothing NONMOD COMP-PRON PRON NOM SG (&SUBJ &OBJ)))
(to
   (to INFMARK (&INFMARK)))
(do
   (do PCP1:/ING INF:/ SVO SVOO SV V INF (&-FMAINV)))
($:)
(once=or=twice
   (once=or=twice ADV (&ADVL)))
(she
   (she NONMOD PRON PERS FEM NOM SG3 SUBJ (&SUBJ)))
(had
   (have PCP1:E/ING INF:/ SVO SVOC/A V PAST VFIN (&+FAUXV)))
(peeped
   (peep N:/ SV SVO PCP1:/ING INF:/ PCP2 (&-FMAINV)))
(into
   (into PREP (&ADVL)))
(the
   (the DEF DET CENTRAL ART SG/PL (&DN>>)))
(book
   (book N NOM SG (&<<P)))
(her
   (she NONMOD PRON PERS FEM GEN SG3 (&GN>>))) [modifier of next
noun]
(sister
   (sister TITLE N NOM SG (&O-ADVL))) ***
(was
   (be PCP1:/ING INF:/ SV SVC/N SVC/A V PAST SG1,3 VFIN (&+FAUXV)))
(reading
   (read PCP1:/ING INF:/ SVO SV PCP1 (&-FMAINV)))
($,)
(but
   (but **CLB CC (&CC)))
(it
   (it NONMOD PRON NOM SG3 SUBJ (&SUBJ)))
(had
   (have PCP1:E/ING INF:/ SVO SVOC/A V PAST VFIN (&+FMAINV)))
(no
   (no QUANT DET CENTRAL SG/PL NEG (&QN>>)))
(pictures
   (picture N NOM PL (&OBJ)))
(or
   (or CC (&CC)))
(conversations
   (conversation N NOM PL (&SUBJ &OBJ)))
(in
   (in PREP (&ADVL)))
(it
   (it NONMOD PRON ACC SG3 (&<<P)))
($,)
($`)
(and
   (and CC (&CC)))
(what
   (what NONMOD **CLB PRON WH SG/PL (&SUBJ)))
(is
   (be PCP1:/ING INF:/ SV SVC/N SVC/A V PRES SG3 VFIN (&+FMAINV)))
(the
   (the DEF DET CENTRAL ART SG/PL (&DN>>)))
(use
   (use PCP1:USING INF:USE N NOM SG (&PCOMPL-S)))
(of
   (of PREP (&<<NOM-OF)))
(a
   (a INDEF DET CENTRAL ART SG (&DN>>)))
(book
   (book N NOM SG (&<<P)))
($,)
(')
(thought
   (think PCP1:/ING INF:/ VCOG SVOC/N SVOC/A SVO SV P/OF PCP2
(&<<NOM-FMAINV &-FMAINV &AN>>)) [&<<NOM-FMAINV = modifier of
previous nominal]
   (think PCP1:/ING INF:/ VCOG SVOC/N SVOC/A SVO SV P/OF V PAST VFIN
(&+FMAINV))
```

```
   (thought PCP1:THINKING INF:THINK N NOM SG (&NN>>)))
(*alice
   (alice ** PROPER N NOM SG (&SUBJ &OBJ &PCOMPL-O &<<P))) [&PCOMPL-O
= predicate complement of the object]
($`)
(without
   (without PREP (&ADVL)))
(pictures
   (picture N NOM PL (&<<P)))
(or
   (or CC (&CC)))
(conversation
   (conversation N NOM SG (&OBJ &<<P)))
($?)
(')
```

(Text 3: OS/2 Computer Manual)

In addition to running programs from the DOS Command Prompt that is provided in the OS/2 operating system, you can set up your machine to run a version of the DOS perating system on the same system as OS/2. Some DOS programs may not run under OS/2 Standard Edition Version 1.3. If you need to run DOS programs that are time-dependent (such as communication and real-time programs) or hardware-dependent (such as networking and driver programs), see A.2, "Using Dual Boot" on page 195 before installing the OS/2 operating system.

```
(*in=addition=to
   (in=addition=to * COMPPP PREP (&<<NOM &ADVL)))
(running
   (run PCP1:/NING INF:/ SVC/A SV SVO PCP1 (&NN>> &<<P-FMAINV
&AN>>)))
(programs
   (program N NOM PL (&SUBJ &OBJ &<<P)))
(from
   (from PREP (&<<NOM &ADVL)))
(the
   (the DEF DET CENTRAL ART SG/PL (&DN>>)))
(*d*o*s
   (*d*o*s ? ** NOBASEFORMNORMALISATION N NOM SG/PL (&NN>>)))
(*command
   (command * N NOM SG (&NN>>)))
(*prompt
   (prompt * N NOM SG (&<<P)))
(that
   (that NONMOD **CLB REL PRON SG/PL (&SUBJ)))
(is
   (be PCP1:/ING INF:/ SV SVC/N SVC/A V PRES SG3 VFIN (&+FAUXV)))
(provided
   (provide N:PROVISION SVO SV P/FOR P/WITH PCP1:E/ING INF:/ PCP2
(&-FMAINV)))
(in
   (in PREP (&ADVL)))
(the
   (the DEF DET CENTRAL ART SG/PL (&DN>>)))
(*o*s/2
   (*o*s/2" ? ** N NOM SG (&<<P)))
(operating
   (operate N:OPERATION SV SVO DER:ATE PCP1:E/ING INF:/ PCP1
(&<<NOM-FMAINV))) ***
(system
   (system N NOM SG (&OBJ))) ***
($,)
(you
   (you NONMOD PRON PERS NOM SG2/PL2 (&SUBJ)))
(can
   (can V AUXMOD VFIN (&+FAUXV)))
(set
   (set PCP1:/TING INF:/ N:/TING SVO SVOO SV P/ON V INF (&-FMAINV)))
(up
   (up ADV (&ADVL)))
(your
   (you NONMOD PRON PERS GEN SG2/PL2 (&GN>>)))
(machine
   (machine N NOM SG (&OBJ &I-OBJ)))
(to
   (to INFMARK (&INFMARK)))
(run
   (run PCP1:/NING INF:/ SVC/A SV SVO V INF (&-FMAINV
&<<NOM-FMAINV)))
(a
   (a INDEF DET CENTRAL ART SG (&DN>>)))
(version
   (version N NOM SG (&OBJ)))
(of
   (of PREP (&<<NOM-OF)))
(the
   (the DEF DET CENTRAL ART SG/PL (&DN>>)))
(*d*o*s
   (*d*o*s ? ** NOBASEFORMNORMALISATION N NOM SG/PL (&<<P))) ***
(perating
   (perating ? PCP1 (&NN>> &AN>>)))  [misspelling in text, NB
correct morphological prediction but incorrect prediction of base
form!]
(system
   (system N NOM SG (&<<P)))
(on
   (on PREP (&<<NOM &ADVL)))
(the
   (the DEF DET CENTRAL ART SG/PL (&DN>>)))
(same
   (same A ABS (&AN>>)))
(system
   (system N NOM SG (&<<P)))
(as
   (as PREP (&ADVL)))
(*o*s/2
   (*o*s/2" ? N NOM SG (&<<P)))
($.)
(*some
   (some * QUANT DET CENTRAL SG/PL (&QN>>)))
(*d*o*s
   (*d*o*s ? ** NOBASEFORMNORMALISATION N NOM SG/PL (&NN>>)))
(programs
   (program N NOM PL (&SUBJ)))
(may
   (may V AUXMOD VFIN (&+FAUXV)))
(not
   (not NEG-PART (&NEG)))
(run
   (run PCP1:/NING INF:/ SVC/A SV SVO V INF (&-FMAINV)))
(under
   (under PREP (&ADVL)))
(*o*s/2
   (*o*s/2" ? ** N NOM SG (&NN>> &<<P)))
(*standard
   (standard * N NOM SG (&NN>> &<<P)))
(*edition
   (edition * N NOM SG (&NN>>)))
(*version
   (version * N NOM SG (&<<P)))
(1.3
   (1.3" NUM CARD (&<<P))) ***
($.)
(*if
   (if * **CLB CS (&CS)))
(you
   (you NONMOD PRON PERS NOM SG2/PL2 (&SUBJ)))
(need
   (need N:/ SVO SV PCP1:/ING INF:/ V PRES -SG3 VFIN (&+FMAINV)))
(to
   (to INFMARK (&INFMARK)))
(run
   (run PCP1:/NING INF:/ SVC/A SV SVO V INF (&-FMAINV)))
(*d*o*s
   (*d*o*s ? ** NOBASEFORMNORMALISATION N NOM SG/PL (&NN>>)))
(programs
   (program N NOM PL (&OBJ)))
(that
   (that NONMOD **CLB REL PRON SG/PL (&SUBJ)))
(are
   (be PCP1:/ING INF:/ SV SVC/N SVC/A V PRES -SG1,3 VFIN (&+FMAINV)))
(time-dependent
   (time-dependent A ABS (&PCOMPL-S)))
($()
(such=as
   (such=as COMPPP PREP (&ADVL)))
(communication
   (communication N:COMMUNICATION PCP1:ON/NG INF:ON/E N NOM SG
(&<<P))) ***
(and
   (and CC (&CC)))
(real-time
   (real-time N NOM SG (&NN>>)))
(programs
   (program N NOM PL (&<<P)))
($))
(or
   (or CC (&CC)))
(hardware-dependent
   (hardware-dependent ? N NOM SG (&SUBJ &<<P))) ***
($()
(such=as
   (such=as COMPPP PREP (&ADVL)))
(networking
   (networking -INDEF N NOM SG (&<<P)) ***
   (network SVO RARE PCP1:/ING INF:/ PCP1 (&<<P-FMAINV))) ***
(and
   (and CC (&CC)))
(driver
   (driver PCP1:E/ING INF:/ DER:ER N NOM SG (&NN>> &<<P)))
(programs
   (program N NOM PL (&SUBJ &OBJ &<<P)))
($))
($,)
(see == morphological error ==
   (see PCP1:/ING INF:/ AS/SVOC/A SVO SV INFCOMP V INF (&-FMAINV
&P-FMAINV))
   (see PCP1:/ING INF:/ AS/SVOC/A SVO SV INFCOMP V PRES -SG3 VFIN
(&+FMAINV)))
(*a.2
   (*a.2" ? ** N NOM SG (&OBJ)))
($,)
($")
(*using
   (use * N:USAGE AS/SVOC/A SVO SV PCP1:E/ING INF:/ PCP1 (&OBJ
&PCOMPL-O &<<P &-FMAINV &AN>>)))
(*dual
   (dual * A ABS (&AN>>)))
(*boot
   (boot * N NOM SG (&OBJ &<<P)))
($")
(on
   (on PREP (&ADVL)))
(page
   (page N NOM SG (&<<P)))
(195
   (195" NUM CARD (&SUBJ))) ***
(before
   (before PREP (&ADVL)))
(installing
   (instal N:/MENT SVO PCP1:/ING INF:/ PCP1 (&<<P-FMAINV)))
(the
   (the DEF DET CENTRAL ART SG/PL (&DN>>)))
(*o*s/2
   (*o*s/2" ? ** N NOM SG (&OBJ))) ***
(operating
   (operate N:OPERATION SV SVO DER:ATE PCP1:E/ING INF:/ PCP1 (&NN>>
&<<NOM-FMAINV &-FMAINV &AN>>)))
(system
   (system N NOM SG (&OBJ)))
($.)
```

(Text 4: Ford Consul Grenada, User's Manual)

The cast iron cylinder block is integral with the upper half of the crankcase, the lower half of which is formed by the pressed steel sump. The overhead valves are mounted at a slight angle across the cylinder head and are operated by an overhead camshaft driven by a reinforced toothed rubber belt from the crankshaft. A spring-loaded follower on the unloaded side of the belt serves to tension the belt. The crankshaft runs in five steel-backed shell bearings in the crankcase main journals.

```
(*the
   (the * DEF DET CENTRAL ART SG/PL (&DN>>)))
(cast_iron
   (cast_iron -INDEF N NOM SG (&NN>>)))
(cylinder
   (cylinder N NOM SG (&NN>>)))
(block
   (block N NOM SG (&SUBJ)))
(is
   (be PCP1:/ING INF:/ SV SVC/N SVC/A V PRES SG3 VFIN (&+FMAINV)))
(integral
   (integral A ABS (&PCOMPL-S)))
(with
   (with PREP (&<<NOM &ADVL)))
(the
   (the DEF DET CENTRAL ART SG/PL (&DN>>)))
(upper
   (upper ATTR A ABS (&AN>>)))
(half
   (half N NOM SG (&<<P)))
(of
   (of PREP (&<<NOM-OF)))
(the
   (the DEF DET CENTRAL ART SG/PL (&DN>>)))
(crankcase
   (crankcase N NOM SG (&<<P)))
($,)
(the
   (the DEF DET CENTRAL ART SG/PL (&DN>>)))
(lower
   (low A CMP (&AN>>)))
(half
   (half N NOM SG (&SUBJ)))
(of
   (of PREP (&<<NOM-OF)))
(which
   (which NONMOD REL PRON WH NOM SG/PL (&<<P)))
(is
   (be PCP1:/ING INF:/ SV SVC/N SVC/A V PRES SG3 VFIN (&+FAUXV)))
(formed
   (form N:FORMATION SVO SV PCP1:/ING INF:/ ER PCP2 (&-FMAINV)))
(by
   (by PREP (&ADVL)))
(the
   (the DEF DET CENTRAL ART SG/PL (&DN>>)))
(pressed
   (press N:PRESSURE SVO SV PCP1:/ING INF:/ PCP2 (&AN>>)))
(steel
   (steel -INDEF N NOM SG (&NN>>)))
(sump
   (sump N NOM SG (&<<P)))
```

```
($.)
(*the
   (the * DEF DET CENTRAL ART SG/PL (&DN>>)))
(overhead
   (overhead ATTR A ABS (&AN>>)))
(valves
   (valve N NOM PL (&SUBJ)))
(are
   (be PCP1:/ING INF:/ SV SVC/N SVC/A V PRES -SG1,3 VFIN (&+FAUXV)))
(mounted
   (mount N:/ SVO SV P/ON PCP1:/ING INF:/ PCP2 (&-FMAINV)))
(at
   (at PREP (&ADVL)))
(a
   (a INDEF DET CENTRAL ART SG (&DN>>)))
(slight
   (slight A ABS (&AN>>)))
(angle
   (angle N NOM SG (&<<P)))
(across
   (across PREP (&<<NOM &ADVL)))
(the
   (the DEF DET CENTRAL ART SG/PL (&DN>>)))
(cylinder
   (cylinder N NOM SG (&NN>> &<<P)))
(head
   (head N NOM SG/PL (&<<P))
   (head SVO SV P/FOR PCP1:/ING INF:/ V PRES -SG3 VFIN (&+FMAINV)))
(and
   (and **CLB CC (&CC)))
(are
   (be PCP1:/ING INF:/ SV SVC/N SVC/A V PRES -SG1,3 VFIN (&+FAUXV)))
(operated
   (operate N:OPERATION SV SVO DER:ATE PCP1:E/ING INF:/ PCP2
(&-FMAINV)))
(by
   (by PREP (&ADVL)))
(an
   (an INDEF DET CENTRAL ART SG (&DN>>)))
(overhead
   (overhead ATTR A ABS (&AN>>)))
(camshaft
   (camshaft N NOM SG (&<<P)))
(driven
   (drive PCP1:E/ING INF:/ SVOC/A SVO SV PCP2 (&<<NOM-FMAINV)))
(by
   (by PREP (&ADVL)))
(a
   (a INDEF DET CENTRAL ART SG (&DN>>)))
(reinforced
   (reinforce N:/MENT SVO PCP1:E/ING INF:/ PCP2 (&AN>>)))
(toothed
   (tooth RARE SVO SV PCP1:/ING INF:/ PCP2 (&-FMAINV &AN>>)))
(rubber
   (rubber N NOM SG (&OBJ &NN>> &<<P)))
(belt
   (belt N NOM SG (&SUBJ &OBJ &<<P)))
(from
   (from PREP (&<<NOM &ADVL)))
(the
   (the DEF DET CENTRAL ART SG/PL (&DN>>)))
```

```
(crankshaft
   (crankshaft N NOM SG (&<<P)))
($.)
(*a
   (a * INDEF DET CENTRAL ART SG (&DN>>)))
(spring-loaded
   (spring-loaded A ABS (&AN>>)))
(follower
   (follower PCP1:ER/ING INF:ER/ DER:ER N NOM SG (&SUBJ)))
(on
   (on PREP (&<<NOM &ADVL)))
(the
   (the DEF DET CENTRAL ART SG/PL (&DN>>)))
(unloaded
   (unload SVO SV P/ON PCP1:/ING INF:/ PCP2 (&AN>>)))
(side
   (side N NOM SG (&<<P)))
(of
   (of PREP (&<<NOM-OF)))
(the
   (the DEF DET CENTRAL ART SG/PL (&DN>>)))
(belt
   (belt N NOM SG (&<<P)))
(serves
   (serve N:SERVICE SVOO SVOC/A SVO SV P/IN P/FOR P/WITH P/ON
PCP1:E/ING INF:/ V PRES SG3 VFIN (&+FMAINV)))
(to
   (to INFMARK (&INFMARK)))
(tension
   (tension N:/ SVO PCP1:/ING INF:/ V INF (&-FMAINV)))
(the
   (the DEF DET CENTRAL ART SG/PL (&DN>>)))
(belt
   (belt N NOM SG (&OBJ)))
($.)
(*the
   (the * DEF DET CENTRAL ART SG/PL (&DN>>)))
(crankshaft
   (crankshaft N NOM SG (&SUBJ)))
(runs
   (run PCP1:/NING INF:/ SVC/A SV SVO V PRES SG3 VFIN (&+FMAINV)))
(in
   (in PREP (&ADVL)))
(five
   (five NUM CARD (&QN>>)))
(steel-backed
   (steel-backed A ABS (&AN>>)))
(bearings
   (bear PCP1:/ING INF:/ DER:ING N NOM PL (&<<P)))
(in
   (in PREP (&<<NOM &ADVL)))
(the
   (the DEF DET CENTRAL ART SG/PL (&DN>>)))
(crankcase
   (crankcase N NOM SG (&NN>>)))
(main
   (main A ABS (&AN>>)))
(journals
   (journal N NOM PL (&<<P)))
($.)
```

APPENDIX 2. Sentences with objectless uses of the verb form "sells" in the Wall Street Journal Corpus.

Cristal at $ 90 a bottle __ sells out around the country and Taittinger's Comtes de Champagne Blanc
That package now __ sells for about $ 2,099
and a 100-megabyte hard_disk now __ sells for $ 5,699, down from $ 6,799
a specific dress style that __ sells briskly
The 186-seat 757 normally __ sells for $ 46 million to $ 52 million

is $ 2,700 and which actually __ sells for $ 1,600 to $ 1,700
, Ford's biggest car, __ sells for about $ 30,000
to leverage a company that __ sells to Uncle_Sam, as defense contractor Tracor's financial problems show
One of its two systems __ sells for between $ 40,000 and $ 65,000 ; the other sells for
drug, 5-FU, already __ sells at a low price, Dr. Moertel said
not cheap " as it __ sells for more=than 30 times the 60 cents a share
buried within it " and __ sells for a more reasonable multiple
it can_ _n't reuse it __ sells directly to the public, opening its " showroom "
Because an investor who __ sells short is _n't entitled to dividends on those shares,
information, which Mr. Hatch __ sells to the public for a modest fee, is free
about 160,000 last year that Chrysler __ sells under its Dodge, Plymouth and Eagle marques, and
of staccato prose, and __ sells for $ 18.95 at those places that still get full price
a dollar to make and __ sells for more=than $ 3, he says
a 256-kilobit DRAM that now __ sells for something over $ 3
of several patented biochemicals Biopharm __ sells to pharmaceutical labs
the financial instruments that ITG __ sells are legitimate
Mazda uses in vehicles it __ sells in Japan and exports to other countries, as=well=as in
) The stock __ sells at more=than 20 times estimated 1989 earnings, even after
The average European luxury car __ sells for $ 15,000 more=than the most lavish Lincolns and Cadillacs
GXE with the air-conditioning option __ sells for $ 15,074, or $ 299 more=than a 1990 GXE that
" If the coffee __ sells well, I_ _'ll finish off the coca, "
North Hills __ sells to Boeing Co., McDonnell Douglas Corp., Grumman Corp.
Co. and fine crystal that __ sells in Macy's and Neiman Marcus
; Time Warner's Book Digest __ sells for $ 11.95 a volume
Americans as a large Safeway __ sells on a summer Saturday
for the AS/400, which __ sells for an average price of $ 150,000 to $ 175,000
" Generally, the stuff __ sells on hype up there
to $ 100 if the option __ sells for less=than $ 500
maker of the goods it __ sells either owns or has other substantial interests in South_African companies
A speculator buys cheap and __ sells dear ; here, that_ _'s illegal

Moreover, A&W now __ sells for a lofty 24 times projected 1989 earnings, while
It __ sells for 38 times operating cash_flow, he says, while

The 1989 Allante __ sells for $ 57,183
idea of a low-end flute __ sells for $ 5,700
(Prodigy's software __ sells for less=than $ 50 and a modem needed to connect a
pack of 10 Cartier Vendomes __ sells for the same price as a standard 20-cigarette pack
According=to Mr. Moynihan, it __ sells at a favorable price-earnings ratio, pays a $ 1 dividend
the camp fresh bread normally __ sells for between three and five lire
store, a crocodile belt __ sells for $ 815
Group used the formula that __ sells well in the land of the insomniacs

If UAL __ sells for $ 250 a share, as many analysts predict it The average Cuban canvas __ sells at auction for only $ 3,000 to $ 5,000

one American food-processing joint_venture that __ sells within China says he will be struggling this year to
of Gallo's jug wines usually __ sells for $ 4.99
The product __ sells at a stiff premium to alternative drugs whose performance have
He says the group __ sells at cost plus overhead
goods from China as it __ sells to that country
" Obviously, pride __ sells in Texas, " says a spokeswoman for Bozell Inc.
something that grows faster and __ sells at a comparable ä price-earnings } multiple
M4 __ sells to the original equipment manufacturer market world-wide and has about
The factory __ sells to workers at cost, which means prices are about

If the new Cheer __ sells well, the trend toward smaller packaging is likely to
, the average stock now __ sells for about 12.5 times companies' earnings
If the toy __ sells well, it could offset sluggish sales at its Milton

APPENDIX 3. Attested forms of the Finnish noun "käsi" 'hand' in the Suomen Kuvalehti 1975 -corpus (600,000 word-forms).

	N	%
LOCATIVE	308	42.3
EXT.LOC.DIR.	16	2.2
INSTRUMENTAL	130	17.9
NOM., GEN.	196	26.9
PARTITIVE	59	8.1
TRANSLATIVE	19	2.6
SUM	728	100.0

POSSESSIVE	187	25.7
NON-POSS.	541	74.3

INSTRUMENTAL 130

85	käs+i+n	with your hands (instr. pl.)
23	käde+llä	with hand (adess. sg.)
17	käs+i+llä	with hands (adess. pl.)
3	käs+i+llä+än	with his, her, their hands (adess. pl. poss.)
2	käde+llä+än	with his, her, their hand (adess. sg. poss.)

LOCATIVE: 157+106+45=308

LOCATIVE STATIC 157

68	käde+ssä	in hand (iness. sg.)
39	käs+i+ssä	in hands (iness. pl.)
21	käde+ssä+än	in his, her, their hand (iness. sg. poss.)
18	käs+i+ssä+än	in his, her, their hands (iness. pl. poss)
4	käs+i+ssä+mme	in our hands (iness. pl . poss.)
4	käde+ssä+ni	in my hand (iness. pl. poss.)
2	käs+i+ssä+nne	in your hands (iness. pl. poss.)
1	käs+i+ssä+mme+kin	in our hands (iness. pl. poss. encl.)

LOCATIVE DIRECTIONAL TOWARDS 106

29	käs+i+in	in hands (illat. pl.)
28	käte+en	in hand (illat. sg.)
25	käs+i+i+nsä	in his, her, their hands (illat. pl. poss.)
9	käte+e+nsä	in his, her, their hand (illat. sg. poss.)
6	käs+i+i+ni	in my hands (illat. pl. poss.)
5	käte+e+ni	in my hand (illat. sg. poss.)
2	käte+e+nne	in your hand (illat. sg. poss.)
2	käs+i+i+nne	in your hands (illat. pl. poss.)

LOCATIVE DIRECTIONAL FROM 45

17	käde+stä	from hand (elat. sg.)
14	käs+i+stä	from hands (elat. pl.)
4	käs+i+stä+än	from his, her, their hands (elat. pl. poss.)
4	käde+stä+än	from his, her, their hand (elat. sg. poss.)
2	käde+stä+ni	from my hand (elat. sg. poss.)
2	käs+i+stä+si	from your hands (elat. pl . poss.)
1	käs+i+stä+nne	from your hands (elat. pl. poss.)
1	käs+i+stä+mme	from my hands (elat. pl. poss)

NOMINATIVE, GENITIVE 196

58	käde+t	hands (nom. pl.)
46	käsi	hand (nom. sg.)
40	käte+nsä	his, her, their hand(s) (nom. sg. or pl. poss.)
30	käde+n	hand, hand's (nom. sg. as obj., gen. sg.)
6	käte+ni poss.	
4	käte+mme poss.	
4	käs+i+en pl.	

3 kät+ten pl.
2 käte+si poss.
1 kät+te+nsä pl.
1 käte+nne poss.
1 käde+n+kin

PARTITIVE 59

23 kät+tä	hand (part. sg.)
13 käs+i+ä+än	hands (part. pl. poss.)
10 kät+tä+än	hand (part. sg. poss.)
9 käs+i+ä	hands (part. pl.)
3 käs+i+ä+ni	hands (part. pl. poss.)
1 käs+i+ä+nsä	hands (part. pl. poss.)

TRANSLATIVE 19
19 käs+i+ksi pl.

EXTENAL LOCATIVES 16
3 käde+ltä
3 käde+lle+en poss.
3 käde+lle
2 käs+i+lle+en pl. poss.
2 käs+i+lle pl.
1 käs+i+lle+ns pl. poss.
1 käde+lle+ni poss.
1 käde+lle+mme poss.