Methods and Tools for Corpus Lexicography

Ole Norling-Christensen
København

Abstract

A survey is given of some technical aspects of the theory and practice of building and using text corpora for dictionary making. The survey builds on newer, especially Anglo-Saxon, literature, as well as on the experience of the editorial team of The Danish Dictionary.

1. Introduction

The work on the mainly corpus based, 6 volumes dictionary of contemporary Danish[1] was initiated in September 1991. Since then, a 40 mil. words (i.e. tokens) corpus of all kinds of general language has been collected, and each of the c. 40.000 text samples has been annotated according to a rather elaborated text typology.

In parallel, methods for reuse of existing lexical sources have been developed, and a database, The Word Bank (Duncker, forthcoming) of morphological, morphosyntactic, semantic and contextual information on more than 300.000 words (i.e. lemmas) has been extracted/constructed from the machine readable versions of some standard printed dictionaries, supplemented with corpus evidence. Work on word class tagging of the corpus is (September 1993) in its initial phase, as is the writing of dictionary entries.

2. Types of corpus - types of tool

As pointed out by the Text Encoding Initiative (TEI26 1993: 3), the term *language corpus* is used to mean a number of rather different things. However, for TEI, as well as for the purpose of this paper, the only distinguishing feature of a corpus that really matters is that its components have been selected or structured according to some conscious set of design criteria. A similar corpus definition is given by Atkins & al. (1992: 1) who, partly building on earlier work by Quémada, distinguish four types of machine readable text collection:

[1] Den Danske Ordbog, hereinafter called DDO.

Proceedings of NODALIDA 1993, pages 187-196

- *archive*: a repository of readable electronic texts not linked in any coordinated way;
- *electronic text library (ETL)*: a collection of electronic texts in standardized format with certain conventions relating to content etc, but without rigorous selectional constraints;
- *corpus*: a subset of an ETL, built according to explicit design criteria for a specific purpose; and
- *subcorpus*: a subset of a corpus, either a static component of a complex corpus or a dynamic selection from a corpus during on-line analysis.

This distinction mirrors an increasing grade of refining of the textual material; and as different tools are needed for the different levels of refinement, as well as for getting from one level to the next one, the distinction shows useful for classifying corpus tools as well. It is, thus, evident that computational tools are needed not only for corpus analysis, but also for the creation and preprocessing of corpora. And the more information in the form of linguistic as well as extralinguistic annotation according to the design criteria that is added to the text samples of the corpus during the preprocessing, the more sophisticated analyses can be made.[1]

Obviously, the tools used for preprocessing and the tools used for analysis must be compatible, by which is meant, among other things, that the analysis tools must conform to the format of the preprocessed texts and be able to make use of the complementary information of the annotations. Useful guidelines for defining such formats can be found in the SGML standard as it is utilized by the Text Encoding Initiative (TEI P2, 1992-).

2.1. From text archive to text library

The text archive may include all kinds of word processor and typesetter files. Changing them into the standardized format of a text library implies not only homogenizing the character set, but also making up one's mind about which of the features, represented by formatting codes in the files, should be preserved, and which not. It will, for instance, hardly be relevant to keep information on specific typefaces; on the other hand, it might be useful to keep some generic information on e.g. headlines and emphasis (leaving out information on whether the emphasis was represented by italics, underlining, boldface or small caps). For this job, text converting programs are needed. They may be supplied as functions

[1] One should, however, keep in mind that tagging a corpus according to e.g. a grammatical theory is likely to make the corpus less usable for testing other theories. Further, there is the risk of circularism: the evidence you get out of your corpus may be, more or less, only what you yourself did put into it.

of the source word processing program; but if this is not known or not available, rather much text specific programming may be needed.

At this stage, each text should also be annotated with at least the directly available bibliographical information, preferably in the form of an SGML header element. One may also chose even now to include text typological information like genre, subject and sender-reciever relationship, sociological information like sex, age and education of the author(s), linguistic information on e.g. language variants. Whether these annotations are made now or during the subsequent phase of corpus making, they have to be made in a standardized form and, whenever possible, with their values taken from a limited set of options. Only then, they will be useful for computational processing. Some kind of syntax and content checking device is, therefore, needed during the process of annotation.

2.2. From text library to corpus

Making one or more corpora out of a text library implies selecting in a balanced way samples of the library texts, in order to fit the specific purpose of the corpus. If information on text types etc. of the library is available in standardized, electronic form, such selection may be done more or less automatically. If this is not the case, it is now time for making these annotations. For checking the balance according to different criteria (i.e. annotations) and combinations of criteria, a statistical tool is needed for computing the relative sizes of texts belonging to different classes.

In addition to the annotations related to selectional criteria, the scope of which will typically vary from the entire corpus down to an entire text sample, the corpus design criteria may also define kinds of additional information that must be added at lower levels, the scope being individual sentences, phrases, words, or even parts of words. The object of these kinds of annotation will normally be some kind of computational and/or computer-aided analysis; consequently, the annotation system has to be thoroughly formalized.

3. A standard format for corpus samples[1]

The international standard SGML (ISO 8879, 1986) for generic description of textual structures and for marking up the texts accordingly, is used by The Danish Dictionary for describing and tagging not only the

[1] An earlier version of this section was part of Norling-Christensen 1992.

dictionary but also the corpus. Readers who are not familiar with SGML and with terms like DTD, element, attribute, entity reference, may consult the brilliant introduction in (TEI 1990: 9-32).

For the corpus an SGML document type *CorpusEntry* has been defined. It provides a suitable form for registration of the necessary (extralinguistic) information *about* the text as well as a means for unambiguous tagging of those (linguistic) features of the text proper that we have decided to represent in the corpus. Each sample (element *CorpusEntry*) consists of: A *Header*, that contains information on the kind and provenience of the text, and the *Text* proper. In the language of SGML:

```
<!DOCTYPE   CorpusEntry  [
<!ELEMENT   CorpusEntry     ( Header, Text ) > ] >
```

3.1. The Header

For designing the Header and deciding which information types should form part of it, we found much inspiration in Atkins (1992). The Header of each corpus entry is divided into two main parts, *viz.* information on the source (SourceInfo), and information on the text sample proper (TextDescription). *SourceInfo* consists of an unambiguous identification (TextGroup/TextNumber), notes on restrictions of use imposed by the supplier of the text (private or confidential texts), information on those who produced the text (LanguageUser = authors, speakers), and on title, publisher, date of origination, and location (*e.g.* page number). There is one element *LanguageUser* for each person involved in the production of a given text sample. Especially in spoken language there usually will be more than one. The element describes the person's name, role (*e.g.* interviewer or interviewee), sex, education, occupation, year and place of birth, and language variant (*i.e.* standard or regional Danish). The element *TextDescription* gives an account of the language type (general or special), whether it is written or spoken, and public (reception) or private (production), the age relation between sender and receiver of the text (adult-adult, adult-juvenile, adult-child, juvenile-adult, etc.), medium (book, newspaper, television etc.), genre, subject field, size of the sample.

The full structure and contents of the header can be explained in the following way. An interrogation mark (?) after an element name means that the element is facultative, *i.e.* it shall only be there if it is relevant, and if the information in question is known. The plus (+) after LanguageUser means that there may be one or more of these elements in a single header.

Header
  SourceInfo
    TextGroup *Unambiguous identifier of a group of (related) corpus entries*
    TextNumber *Serial number inside the text group*
    Restriction?
      RestrictA *Proper names in text must be altered: "Y[es]"/"N[o]"*
      RestrictB *Text must only be used for the dictionary: "Y[es]"/"N[o]"*
      Expiration *of Restriction B, e.g. "1998"*
    LanguageUser+ *(one element for each author/speaker)*
      Role? *Esp. when more language users are involved; e.g. "teacher", "pupil"*
      Identification? *A unique three character string, referred to by SpeakerTurns in the Text*
      LastName? *If known*
      FirstName? *If known*
      ***Sex** "m"/"f"/"u[nknown]"*
      Education? *if known*
      Occupation? *if known*
      ***YearOfBirth** a number between 1880 and 1990*
        Precise? *"?", if not known exactly*
      PlaceOfBirth? *if known*
      ***LanguageVariant** "standard"/"regional"*
    TextTitle? *if any*
    VolTitle? *Name of Anthology, Newspaper, Magazine, etc., if any*
    Publisher? *Book publisher or Radio or TV station, if any*
    Date
      Day? *if known*
      Month? *if known*
      Year *number between 83 and 92*
        Precise? *"?", if not known exactly*
    Location? *e.g. Section, page, column of Newspaper; (Vol.,) page of book*
  TextDescription
    ***LanguageType** "general"/"special purpose"*
    ***Written_Spoken** "written"/"spoken" or one of two intermediate types*
    ***Aspect** "reception"/"production"*
    ***AgeRelation** "child-child"/"child-juvenile"/"child-adult"/../"adult-adult"/"unknown"*
    ***Medium** taken from a list of 12 different media, e.g. book, journal, radio, film*
    ***Genre?** taken from a list of 124 partly medium-dependent genres, like novel, letter, comic*
    ***Subject?** taken from a list of 64 different subject areas, like biology, literature, physics*
    Size *Number of words (tokens) in this text sample*

The elements marked by an asterisk (*) above are standardized descriptors that play a special role in corpus search and analysis. For each of the descriptors a restricted list of legal values is defined. Different text types, and corresponding subcorpora, can be defined in terms of one or more of these descriptors, *e.g.* "Women born before 1940 speaking to

children" or "Newspaper texts on politics". Besides, the decriptors are used for studies of the distribution of all kinds of linguistic features over the different text types.

3.2. The Text

The structure of the *Text* element depends on whether it consists of written language or of (transcribed) spoken language. Written language is split up into paragraphs (the element *p*) that are subdivided into sentences (the element *s*). Sentences are mostly non-tagged strings of characters[1] (the SGML category #PCDATA); these may, however, be interspersed with elements of special types of text, viz. the elements *Highlighted* and *Note* . The tag *Highlighted* covers all kinds of accentuation in the original text: underlining, boldface, italics, spacing, bigger or deviant fonts; *Note* are foot- or endnotes.

Spoken language is normally not cut into paragraphs; instead, they may be divided up into speaker turns. Most of the spoken texts are conversations or interviews with more persons involved. Consequently, the header contains two or more instances of the element *LanguageUser*. Each of them contains in the subelement *Identification* a different three letter string. Each element *SpeakerTurn* contains an attribute *id* that refers to the *Identification*. The *SpeakerTurn* element consists of #PCDATA interspersed with entity references like {hesitation} representing non-verbal sounds like 'eh', 'mmm'; {pause}; {uf} that represents a passage that was incomprehensible to the transcriber; {laughter}; and with the elements *Comment* (the transcriber's "stage directions" that are not part of the speech), and *Doubtful*: a word or passage that the transcriber was not sure about.

4. Use of the corpus

4.1 Two problems: abundance and scarcity

The lexicographer working with corpora runs into two basic problems: the theoretical problem of the significance of sparse or none instances of some linguistic phenomena, and the practical problem of being flooded with too many instances of others. The former problem can only be solved by making the corpus even larger, or by relying on sources external to the corpus. To cope with the latter, however, computational

[1]The ongoing word class tagging splits the sentences up into single words, each with a word class attribute, leaving only the interpunction untagged.

tools are needed in order to structure the flood; without such tools, large corpora will not be of much use.

4.2. Interactive analysis

"Structuring the flood" can be seen as the repetitive process of asking ever more specific questions to the material. The basic, first question is "Give me all the instances of the lemma X". The following questions include contextual restrictions which can be made the more precise the more annotated the corpus is. The questioning is repeated until some characteristic behaviour of the word (lemma) crystallizes. Once such behaviour (e.g. one meaning; one valency frame) has been recognized and described by the lexicographer, the instances of it are thrown away, and the procedure is repeated for the rest of the instances.

4.3. Statistical analysis

There is, however, one class of important questions that cannot be meaningfully asked just to the immediate context of the instances of a lemma. Exploration of the collocational behaviour of a word is not possible without some knowledge of the corpus as a whole. The mere observation that one word seems to be frequently occuring in the neighbourhood of another word does not in itself indicate a collocational connection between the two, neither does a seemingly infrequent occurrence indicate the absence of such connection. Only a statistical calculation that takes into account the total numbers of occurrence of the words in question can give a reliable indication.

In Church (1991) three statistical methods for collocational studies are discussed. They all ought to be part of toolboxes for corpus analysis, even though rather big corpora are needed in order to make reliable statistics. "Mutual Information" reveals positional interdependence between two words by comparing the observed frequency of a co-occurrence to the calculated frequency for co-occurrence by chance. "Scale Statistics" calculates the mean and the standard deviation of the distance between such pairs. The more sophisticated "T-score" test looks for significant differences between the immediate neighbourhood of two different words, typically pairs of near synonyms like "strong"/"powerful" or "his"/"her". The observed neighbouring words, e.g. in the position immediately to the right of the two, are ranged on a scale spanning from those having greatest affinity to one of the synonyms, through those which are neutral, to those with greatest affinity to the other synonym.

4.4. Subcorpora

In so far as the individual text samples that make up the corpus have been annotated with text typological information etc. (cf. section 3.1.2.1) it shall be possible to use (boolean combinations of) the annotations for the selection of subsets like e.g. "texts on science or medicine written by women born before 1950". The result may be a new corpus of its own; but a flexible corpus management system will also allow for creating temporary virtual subcorpora by inserting filters between the corpus and the user. The full range of analytical methods specified for a corpus must also be appliable to a virtual subcorpus as well as to the collections of corpus examples (e.g. sentences or KWIC lines) that result from searches and subsequent annotating and sorting.

5. Computational tools

As much as possible of the Header-information, as well as the identification and tagging of the entity references and subelements of the Text proper, is made (semi)automatically. This means, that for each group of texts of a given provenience or type, a customized conversion program is written. A toolbox of Borland Pascal units, called DICONV, programmed by the author, makes such programming fast and efficient; it was originally made for conversion and adjustment of dictionary texts for a publishing house. Not only does the program convert a given wordprocessor format into our standard format; in some cases it also makes use of the authors' idiosyncratic ways of marking those features we are interested in. In other cases these features are marked up manually, using word processor macros. The rest of the header information is keyed in using a customized database application.

5.1. Grammatical tagging

The only "syntactical" tagging that is done for the moment is the delimiting of paragraphs and sentences. For this, Jann Scheuer has written a program that analyses surface information like Newline, interpunction, and the use of uppercase letters. The biggest problems in delimiting sentences are the well known ambiguities of the full stop character (abbreviation mark, ordinal number mark, sentence delimiter, or both sentence delimiter and one of the other functions at the same time), and of uppercase initial being either a proper name marker or conventionally put after a full stop, or both at the same time. As a by-product, the program produces lists of identified proper names and abbreviations. It further makes use of such lists during the analysis. The best result are, therefore, obtained by running the program twice.

After the delimiting of sentences, each sentence is run through a word class disambiguating system made by Jørg Asmussen. The system was originally made for German as part of the authors thesis; but he now has accomodated it for the needs of The Danish Dictionary. The analyses is based on the likelihoods of different sequences of word classes; these are established during training sessions, where the program asks the human trainer for advice when in doubt. The system's dictionary of "homograph classes" is derived from the Word Bank.

5.2. Corpus analysis

For corpus search and interactive analysis, a tool called Corpus°Bench has been developed by the Danish software house TEXTware A/S according to specifications made jointly by Longman Publishers (UK) and The Danish Dictionary. Concordances can be built in real time according to complex search criteria in the form of Boolean combinations of a keyword (lemma) with neighbouring words and/or text type specifications. The concordance lines can be tagged by the user according to up to eight different, user defined criteria. The lines can be sorted according to any combination of key word, left context, right context, user defined tags, and text type information. Besides the statistically based methods, mentioned above, for collocational analysis are available. Further, frequency information, including frequency distribution over e.g. text types, can be obtained.

References

Atkins, Sue, Jeremy Clear and Nicholas Ostler. 1992. *Corpus Design Criteria.* pp.1–16, LITERARY AND LINGUISTIC COMPUTING, Volume 7, Number 1, Oxford University Press.

Church, Kenneth, William Gale, Patrick Hanks and Donald Hindle. 1991. *Using Statistics in Lexical Analysis.* In Zernik (ed.).

Zernik, Uri (ed.). 1991. *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon.* Erlbaum, New Jersey.

Duncker, Dorthe and O. Norling Christensen. Forthcoming. *Genbrug af ordbogsdata. Den Danske Ordbank.* To appear in the *Proceedings of the 2nd Scandinavian conference on Lexicography, Copenhagen 1993.*

Kiefer, Ferenc et al. (ed). 1992. *Papers in Computational Lexicography. COMPLEX '92.* Linguistics Institute, Hungarian Academy of Sciences, Budapest.

Norling-Christensen, Ole. 1992. *Preparing a Text Corpus - Computational Tools and Methods for Standardizing, Tagging and Structuring Text Data.* In Kiefer (ed.).

TEI P1. 1990: C.M. Sperberg-McQueen and Lou Burnard (ed.s): *ACH - ACL - ALLC. Guidelines for the Encoding and Interchange of Machine-Readable Texts. TEI P1. Draft Version 1.1.* Chicago and Oxford, 1 November 1990.

TEI P2. 1992-: C.M. Sperberg-McQueen and Lou Burnard (ed.s): *ACH - ACL - ALLC. Guidelines for Electronic Text Encoding and Interchange. TEI P2. Part I–VIII.* Chicago, Oxford. In preparation; partly released 1992–93.

TEI26. 1993: *Additional Tag Set for Language Corpora.* In *TEI P2 1992-, part IV,* chapter 26, released 11 March 1993.