Manual of the Stockholm Umeå Corpus version 2.0

Description of the content of the SUC 2.0 distribution, including the unfinished documentation by Gunnel Källgren

Sofia Gustafson-Capková and Britt Hartmann

December 2006

SUC 2.0 is a balanced corpus of 1 million words in written Swedish from the 1990's. The corpus is in an sgml format and annotated with parts of speech, morphological analysis and lemma (base form) as well as a range of structural tags and functionally interpreted tags. For a detailed description of the tags, please consult the Documentation of Stockholm Umeå Corpus.

The present distribution of SUC 2.0 in most aspects parallels the SUC 1.0 distribution. The text materials are the same, as well as the markup in terms of part of speech and morphological analysis. In addition, SUC 2.0 contains extended functionally interpreted markup, such as a detailed tagging of name expressions, and has also undergone more proof-reading than SUC 1.0.

1. CONTENT OF SUC 2.0

The present distribution of SUC 2.0 contains the following folders:

BIBLIOGRAPHY
BONUS
COPYRIGHT
CORPUS
DTD
HEADER
lib
LICENCE
MANUAL
WORDLISTS
The content of the folders is described below.

BIBLIOGRAPHY
The SUC 2.0 bibliography.

BONUS
This folder contains bonus materials included in the distribution:

TIGERSUC – SUC 2.0 converted to TIGER-xml by Martin Volk.
STORSUC – Additional SUC materials of 4 million words.

COPYRIGHT
Information about the copyright of SUC2.0.

CORPUS
This folder contains SUC 2.0 in three formats:

SUC 2.0c – The SUC corpus with SUC tags.
SUC 2.0d – The SUC corpus with PAROLE tags.
SUC 2.0x – The SUC corpus with SUC tags, without file headers.

DTD
The SUC 2.0 DTD:s.

HEADER
The SUC 2.0 corpus headers.

lib
The SGML libraries needed for SUC 2.0.

LICENSE
The license agreements for SUC 2.0.

MANUAL
The documentation of SUC 2.0.

WORDLISTS
Wordlists with frequency information based on SUC 2.0.

2. MOST PROMINENT DIFFERENCES BETWEEN SUC 1.0 AND SUC 2.0

Although SUC 2.0 has a richer structural markup than SUC 1.0, and a sligthly different reference system, the words and their morphosyntactic tags in the corpus are essentially the same.
An important difference between SUC1.0 and SUC 2.0 is that SUC 2.0 has a large number of functionally interpreted TEI-tags, with attribute values selected by human annotators. These TEI-tags follow the TEI guidelines P3.

The functionally interpreted tags specific for SUC 2.0 are the following:

<abbr>
Abbreviation. This tag spans over one or more words constituting an abbreviation.

<byline>
Byline, i.e. information about author, source etc. mainly in newspaper articles.

<distinct>
Word or sequence of words in non-standard Swedish.

<foreign lang=en>
Encloses one or more foreign, non-Swedish words. The tag is present also in SUC 1.0, but without the value specifying which language the tag contains. This value is specific to SUC 2.0.

To be able to distinguish lyrical language from prose, the tags <lg> and <l> have been introduced in SUC 2.0
<lg> is on the paragraph level and indicates a line group of poetic or verse material.
<l> is on the s-unit level and indicates a line of poem or verse supplied for s-unit in line groups.

List structures are marked in SUC 2.0:
<list> stands for a whole list.
<item> stands for an item within a list.
<label> shows the list item's label.

Name expressions are marked with the element name, and are also sub-classified with a value for the attribute type, e.g.:
<name type=person>
For a more elaborate attribute list, please consult the manual.

<mentioned>
This tag indicates a word or sequence of words referred to rather than used.

<ref>
Free format bibliographic citation.

The two tags below are present in SUC 2.0, but to a very small extent. They should be regarded as examples of this type of tagging, and not as mirroring the actual inventory of the phenomena.
<q> Direct speech or writing.
<quote> Quotation. Materials not attributed to the author of the suc-text.

Differences by structural tags:
The object language elements at the level of tokens are word tokens and punctuation tokens. These elements are formally distinguished by the tags "w" and "c" in the PAROLE format or "w" and "d" in the SUC-format (d for delimiter) respectively. The tag DL used in SUC 1.0 (suc1a) is not used in SUC 2.0.

Also the file headers contain a smaller difference being that the "title" element in SUC 2.0 has an added attribute level with values a, j or m (article, journal, monograph). This entails no modification of the TEI-header.

Differences in tokenisation:
All periods that occur in abbreviations have been included in the abbreviation token, rather than tokenised separately. Remaining free periods should be sentence final delimiters.

3. BONUS MATERIALS

SUC 2.0 contains two bonus materials: TIGERSUC and STORSUC. Since they are bonus materials, they are not formally included in the SUC2.0 corpus regarding the corpus format, and a description of them is not included in the corpus header. However, both materials are included in the SUC2.0 distribution and subject to the same copyright and license agreement as SUC 2.0.

TIGERSUC is SUC 2.0 converted to TIGERxml by Martin Volk. A big advantage with this material is that TIGERSUC is possible to load into the TIGERSearch Corpus tool <http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERSearch/> developed at IMS University Stuttgart. This eliminates the problem that no tailor-made search tool is present for SUC2.0. TIGERSUC is identical to SUC 2.0 regarding text materials and linguistic markup.

STORSUC contains some text materials originally included in the SUC texts, but only partially included in SUC 1.0 and SUC 2.0. in excerpt form. STORSUC is not annotated with elaborate markup as is SUC 2.0, but only structured into paragraph like segments. STORSUC is not balanced. The folder STORSUC contains a record of texts.

Documentation of the Stockholm – Umeå Corpus

Gunnel Källgren

*** Work in progress – unfinished***

This is very much in progress. Some parts, notably the description of the text categories and the SGML tagging manual are close to their final format, other parts are early sketches and others I have not even started on. My English has not been checked. Here and there various commands and questions directed to myself can be seen. In spite of its ragged outlook, I still think this overview with its appendices could give a picture of the corpus and the work behind it.
This version contains additions and changes relative to the version of 980202, which was submitted as a specimen to Stockholm University.

Gunnel Källgren

After Gunnel Källgren passed away in 1999, Britt Hartmann and I soon understood that we would never be able to finish this documentation in the way that Gunnel would have liked it to be done. Therefore, after many doubts, we decided to fix only the most important passages, where we were sure about the content, and leave the rest in the unfinished and ragged state Gunnel describes above. Both Britt and I are convinced that information in this ragged state is still better than no information at all. This especially since some of the information in this unfinished documentation, such as the information about the sampling of SUC, is impossible to get from other sources. We thus humbly ask the reader to excuse the ragged state and the missing parts, and make the best possible use of the present content.

Sofia Gustafson-Capková and Britt Hartmann
Stockholm December 2006

1 Introduction

1.1 Never more?

A colleague of mine, Matti Rahkonen at the University of Jyväskylä, once claimed that he could subsume his experiences of large-scale corpus work in two words: 'Never more!'

"Att koda en stor korpus är i många avseenden en otacksam uppgift. För det första förhåller det sig så att taggning inte är forskning i ordets egentliga betydelse. Det färdiga materialet är endast en förutsättning för empirisk forskning, hur stor arbetsinsats dess färdigställande än har krävt. Att vara tvungen att arbeta i år för att bara åstadkomma ett forskningsmaterial kan vara mycket frustrerande för vilken lingvist som helst. Mina egna erfarenheter av storskaligt korpusarbete kan sammanfattas med två ord: aldrig mer. För det andra är en korpus aldrig riktigt färdig. ..." (Rahkonen 1992, p. 3.)

I can wholeheartedly agree with Rahkonens description of the corpus-building process, but I am not quite so dejected. Rather, I could join in an old Louis Prima favourite that has given me consolation and encouragement: 'Next time ...'

This documentation of the SUC is deliberately written in the spirit of ‘Next time ...’ rather than just being a report of finished work. That means that I do not just describe what we have done but also discuss the decisions we have taken and what alternatives we had to choose between. I tell about some of the many mistakes we made and also describe solutions that worked out well and the reasons for both. Hopefully, the picture of the SUC will get fuller and the information to any future corpus builders will be more useful with a more complex and discursive presentation. Even if I am not exactly yearning to start building another corpus, I have some fairly good ideas about how I would do it ... next time.

In one sense the SUC work is really "never more". It is not probable that there will ever be resources for collection and manual annotation of another corpus of SUC's size and character. Especially the amount of manual processing of the SUC has been large and expensive (and definitely worth its price). The broad diversity of text types has also been a complicating and expensive factor. As I will argue throughout this report, both these efforts have given us knowledge and facts that we would not have had otherwise. All through the corpus-building project, we complained that our resources were too limited and that was true. It is also true that the SUC project got a lot more money than is usual in the humanities and we appreciate that. This, however, tells more about the conditions for research in the humanities than about this specific project and its funding. My hope is that the SUC with its unique features will come to much use in the field of humanities research - and outside it. This documentation is meant to facilitate that.

1.2 A short overview of the work process behind the SUC

Planning the Stockholm-Umeå Corpus began in 1987 and in autumn 1989 our work with it began. The background for it was a need felt by both its principals in spe, professor Eva Ejerhed at Umeå university and myself, for a generally accessible, annotated corpus of Swedish language. The focus in corpus linguistics has later come to be directed more towards very large corpora, containing many millions of words, or even monitor corpora, where large amounts of text pass by without being collected. Still, the one-million-word SUC is not outdated. Its semi-manual linguistic annotation and structural markup make it extraordinary. For much of the research that linguists want to carry out, access to a balanced and carefully annotated corpus is a necessary requirement. Every language needs its benchmark corpus and it is our hope that the SUC can fill that need for Swedish for years to come.

It is easier to understand the description of the corpus if one knows about how it came to be. The SUC project was carried out in co-operation between the departments of linguistics at the universities of Stockholm and Umeå and was jointly funded by the Swedish Council for Research in the Humanities and Social Sciences (HSFR) and the Swedish National Board for Industrial and Technical Development (NUTEK).

The process of building the SUC was divided into several steps with different distribution between the research groups in Stockholm and Umeå. This report is written from a Stockholm viewpoint. Of course I describe Umeå's parts as well but the steps for which Umeå had the main responsibility will be better documented elsewhere. Below is a list of steps in the construction of the SUC in approximate temporal order; many steps of course went on in parallel.

- 1. Basic definition of what the corpus should contain, preliminary contacts with presumptive text providers, handling the legal aspects. (Stockholm.)
- 2. Establishing and testing a system for the linguistic analysis and annotation of the material. This work resulted in a tagging manual (Ejerhed et al. 1992). (Stockholm and Umeå in co-operation with the main responsibility in Umeå).
- 3. Collecting texts in machine readable and printed form, conversion to a passably standardised format, manual data catch of some structural properties of the texts that would otherwise run a risk of being lost, back up of raw texts, choice and naming of the excerpts to be put into the corpus. (Stockholm.)
- 4. Lexicon lookup and pre- and postprocessing in connection with that. As we used a dictionary not specifically designed for SUC (the Helsinki SWETWOL, Karlsson et al. 1995) quite extensive processing was needed in this step. (Umeå.)
- 5. Manual disambiguation of the output from SWETWOL. This very time-consuming annotation was shared between Stockholm and Umeå.
- 6. Defining and testing a scheme for SGML markup of the corpus, creating a DTD, building headers and a bibliographic data base. (Stockholm.)
- 7. Postprocessing after the annotation, consistency check, bug handling. (Umeå.)
- 8. Automatic insertion of some of the SGML markup. (Umeå.)
- 9. Manual completion and check of the SGML markup. (Stockholm.)
- 10. Final formatting of the corpus. (Umeå for SUC 1.0, Stockholm for SUC 2.0, both in co-operation with Gothenburg.)

The description of the corpus will roughly follow the work flow, describing in turn the collecting of texts (Sect. 2) and the practical and legal intricacies in connection with that, the morphosyntactic tagging (Sect. 3) with its manual annotation phase, the SGML markup (Sect. 4) with a presentation of the DTD, the bibliography and the header. The Appendices contain much useful information, which can also be found on the CD-ROM. Last, but certainly not least: the URL of the home page is: http://www.ling.su.se/forskning/forskn.htm/SUC.

2 Collection of Texts

2.1 Governing principles for the composition of the SUC

To get the kind of broad coverage corpus that we wanted, we needed to decide some general principles for its composition. The Brown (Francis & Kucera 1964) and LOB (Johansson et al. 1978) corpora and the extensive use of them for various kinds of research have of course been an important source of inspiration for us, but so has also some early corpus work in Swedish (Allén 1970), (Teleman 1974). Work with the British National Corpus (BNC) began at about the same time as our work with SUC and we have also followed their progress. Their final documentation (BNC 1995) was not available to us until the summer 1995, but there are many clear parallels between their work and ours. Considering advantages and disadvantages of the works of our forerunners as well as our limited resources, we arrived at the following principles for the composition of the corpus:

- It should be a so-called balanced corpus, covering various text types and various stylistic levels.
- It should broadly mirror what a Swedish person might read in the early 1990's (but cf. below).
- A deliberate limitation to this is that all texts in the corpus are originally written in Swedish. We have tried to avoid translations, in spite of the fact that a large amount of what is read in Sweden has been translated from other languages.
- All texts in the corpus should be (made) freely accessible and distributable for non-profit research purposes, i.e., copyright questions had to be solved.
- The corpus only contains printed text, no spoken language and also no written, non-published texts, such as private letters.
- In so far as it does not conflict with other principles, the choice of texts for the corpus parallels that of the Brown and LOB corpora, so as to facilitate comparative studies.
- The texts had to already exist in computer readable form, as we had no resources for other kinds of data catch. This demand had to be weighed against the wishes for balance and parallelism expressed above. In a few instances we have had to scan in texts that we wanted.

The principles as stated above led to a fairly long and complicated process of gathering the texts. The corpus had to form a balanced whole, similar to its predecessors but with some well-motivated changes. Within these overall frames, we had to find texts that were computer readable, whose copyright holders were willing to let us have them, and, not least, that we could obtain physically without causing ourselves or the text providers too much trouble. To get the desired balance we had to find many short texts from many different sources. This all took an unforeseen amount of time and effort, no more about our hardships right now (the whimpering will start in Section 2.4), but blessed be in particular some editors at publishing companies who gave our odd problems so much of their precious time.

Seen in retrospect, it was worth the trouble. The texts give a broad coverage of modern written Swedish and the only restrictions connected to them is that the user must sign a form stating that the texts are not going to be used commercially. This legally secured accessibility is a characteristic of the SUC that opens it to various kinds of research and teaching purposes.

2.2 Definition of Text Categories

In the choice of texts, and thereby in the design of a corpus, there is always an element of subjectivity. Consulting various sources of information about e.g. which books, newspapers, and periodicals are published during a certain period and with what circulation, can diminish that subjectivity. Library statistics can be consulted as well as results from studies of people's reading habits etc. Such matters are thoroughly discussed in the BNC reference manual. (BNC 1995) or http://info.ox.ac.uk/bnc/getting/bncman.html.

The subjectivity there is in the composition of the SUC is entirely my own (Gunnel Källgren's) and I take the responsibility for it. After some preliminary explorations in 1989 of libraries' and booksellers' catalogues I decided that the Brown/LOB principles were not significantly less reliable than anything I might invent myself, so I decided to start out from their categories and redefine them where necessary to provide for the fact that the context of the corpus was Sweden in 1990, not the U.S. in 1961. By starting from the categories and then finding appropriate texts, I also had more freedom to find sources that were practically and legally easy to access than I would have had if I had started by, e.g. randomly picking out texts from a catalogue. In my search for texts, on the other hand, I had much help from catalogues, in particular the Massmedia 89-90 and later (Norstedts 1990), a catalogue of daily papers, journals, magazines, publishing companies etc. with some information about specialisation, circulation and other useful facts - including address and phone number, which was much needed in the actual text hunt that will be described in Section 2.4.

To decide what the Brown/LOB categories actually stood for, I not only read the descriptions of them but also went to the bibliographies of the chosen texts, which gave clarifying and sometimes surprising information. On the basis of this I defined a set of categories for which the task was to find suitable texts. The categories and subcategories were given names that are meant to be as informative as possible. In the tables in Section 2.3 both the Swedish and the English name of the categories are given together with the letter combinations that are used in the names of the texts.

In the SUC we have followed the general layout of Brown and LOB, with 500 samples of text with a length of about 2,000 words each. The texts can be excerpts from longer texts (as from books), single whole texts, or composed of several short texts. The latter is often the case with articles from newspapers and magazines, which rarely are as long as 2,000 words in themselves.

Each text is given a unique name, consisting of two letters designating the text category, a running enumeration of the texts within each category, and, in the case of composite texts, a letter for the individual text. Sometimes, but not always, Brown and LOB have subcategories. In SUC this system is fully developed and not only the main category but also the subcategory can be seen from the name of each text. Thus, kk27 is a single text taken from a book of fiction while ac02c is the third short text in a sample composed of financial articles taken from newspapers.

The total number of text samples is thus the same - 500 - in all three corpora, but sometimes the number of samples in each category differs between the Swedish and the two English corpora. This occurs when it is felt to be socio-culturally motivated. There are also some differences between Brown and LOB, for basically the same reasons.

In three instances I have made more far-reaching changes in the categories. The category D, religion, has been split up between categories E, F and J, occurring there as separate subcategories - ED, FD and JD, respectively. The other instance is AB, Society news, which in the English corpora turned out to be gossip. I have renamed it Community and redefined it to mean news about society in the sense of articles about social welfare, childcare, emancipation, immigrants, environment, etc., a type of articles that had no good coverage in the 1961 corpora. (For those interested, some gossip can be found in the category Spot News, AF.) The third case is fiction where I have fewer subcategories. I would have liked to add a couple of categories, e.g. schoolbooks and children's books, but have not done so.

With these explanations to be kept in mind, the set of texts in the three corpora Brown, LOB and SUC can be seen in Table 1. The categories will be further commented below and motivations for the variation in the number of texts will be given. All the categories and subcategories are also listed in the corpus header under <taxonomy>, and in Appendix A of this documentation. Bibliographic data for all the texts are given under <sourceDesc> in the file headers (TEI document headers) of suc2c and suc2d. For examples, see Appendix D/E.

	Brown	LOB	SUC	Diff. categ.	Diff. Accum.
I. Informative prose	374	374	373	-1	-1
A. Press: Reportage	44	44	44	0	0
B. Press: Editorial	27	27	17	-10	-10
C. Press: Reviews	17	17	27	+10	0
D. Religion	17	17	0*	-17	-17
E. Skills, Trades and Hobbies	36	38	58	+20	+3
F. Popular Lore	48	44	48	+4	+7
G. Belles Lettres, Biography, Memoirs	75	77	26	-51	-44
H. Miscellaneous	30	30	70	+40	-4
J. Learned and Scientific Writing	80	80	83	+3	-1
II. Imaginative prose	126	126	127	+1	0
KK. General fiction	29	29	82	+53	+52
KL. Mysteries (L) and Science fiction (M)	30	30	19	-11	+41
KN. Light reading (N+P)	58	58	20	-38	+3
KR. Humour	9	9	6	-3	0
Total number of text files	500	500	500	0	0

Table 1. Distribution of texts over main categories in the Brown, LOB and SUC corpora (cf. Garside et al. 1987); differences between LOB and SUC per category and accumulated.
* Category D has not disappeared but is spread out to E (3 texts), F (2 texts) and J (6 texts).

2.3 Description by category

2.3.1 Category A: Press, Reportage

The first three categories mainly comprise newspaper articles. Category A is newspaper reportage with subcategories Political, Community, Financial, Cultural, Sports, and Spot News. Brown/LOB treat daily/Sunday/weekly as separate subcategories, which we do not, but we take care to have Sunday papers in the material and let "weekly" be represented by regional papers which do not come daily (Sw. "fådagarstidningar"). Under "political" it is important to cover both foreign and domestic news in about the same proportions as in an average Swedish newspaper. "Cultural reportage" has been cut down in number as many such articles have been classified as reviews instead, while the number of "community" articles has been considerably increased, as stated above. The "society news" in the sense of gossip is treated as "spot news".

Almost all text samples in these categories are composite of shorter texts, as very few newspaper articles reach a length of 2,000 words. This is particularly characteristic of the spot news, where, e.g. AF01 contains 21 different texts. This is, however, exceptional. More normal are, e.g. AA01 with 3 articles about foreign affairs taken from Sweden's largest morning paper or AA12 with 5 articles about local matters from a couple of regional newspapers. Whenever such composite text samples are built, we have tried to make them homogeneous, keeping variables such as national/regional, morning/evening, daily/weekly as constant as possible within each sample.

Category	National	Regional	Swedish term
A. Press, Reportage			Tidningstext, reportage
AA. Political	8	5	Allmänpolitiska reportage
AB. Community	4	3	Samhälle, familj
AC. Financial	3	1	Ekonomi
AD. Cultural	2	2	Kultur
AE. Sports	4	3	Sport
AF. Spot News	5	4	Notiser
Total	26	18	

Table 2. Number of texts in category A.

2.3.2 Category B: Press, Editorials

Brown/LOB have three subcategories: Institutional Editorials, Personal Editorials, and Letters to the Editor. In Swedish newspapers "institutional editorials" are as frequent as in Anglo-Saxon press, while "personal editorials" and "letters to the editor" are less so. Swedish "letters to the editor" ("insändare") are normally short and strongly edited before publication. To use them, with their more or less unidentifiable authors, would also cause tricky copyright problems. After comparison with the texts of Brown/LOB I have chosen to merge the last two categories into one mainly containing debate articles (Sw. "debattartiklar").

Category	National	Regional	Swedish term
B. Press, Editorials			Tidningstext, ledare
BA. Institutional	4	3	Ledare
BB. Debate articles	6	4	Debattartiklar
Total	10	7	

Table 3. Number of texts in category B.

2.3.3 Category C: Press, Reviews

The category of reviews is large in SUC as we have put all reviews here, not only those from daily papers. We have taken care to have reviews from different subject fields, as there seems to be considerable linguistic variation depending on what is being reviewed. Our subclassification mirrors that. We have also used both national and regional papers as sources. It is to be noted that reviews are mostly quite short, so that what is registered as one or two text samples in a category may include 5-10 separate review articles or more.

Category	National	Regional	Swedish term
C. Reviews			Recensioner
CA. Books	4	3	Böcker
CB. Films	3	2	Filmer
CC. Art	2	1	Konst
CD. Theatre	2	1	Teater
CE. Music	2	1	Musik
CF. Artists, shows	2	1	Artister, shower
CG. Radio, TV	2	1	Radio, TV
Total	17	10	

Table 4. Number of texts in category C.

2.3.4 Category D: Religion

The category D, religion, which occurs in both Brown and LOB, is a somewhat odd category. No other main category is defined on the basis of its subject matter. Religious matters can be treated in many different styles and contexts so we have dissolved category D and spread it over other categories, where we let the texts about religion form separate subcategories. Thus, we have the subcategories ED with 3 texts, FD with 2 and JD with 6, altogether 11 texts of which several are composite.

2.3.5 Category E: Skills, trades and hobbies

This category is quite amorphous and its texts can be found in a wide variety of sources. We have used various kinds of books and handbooks, periodicals and some articles from daily papers. The number of excerpts in this category has been increased as the proportion of such texts has clearly increased since the early sixties. The texts are classified into four subcategories and many of them are composite from several shorter texts. The texts typically cover areas such as interior decoration, pets, sports, food and wine, travel, motor, outdoor activities, computers, gardening, private economy, non-governmental organisations of various types, trades and trade unions, and articles on religious matters.

Category		Swedish term
E. Skills, trades and hobbies		Arbete och fritid
EA. Hobbies, amusements	25	Hobbies, nöjen
EB. Society press	10	Föreningar, dock ej fackföreningar
EC. Occupational and trade union press	20	Yrkes- och fackföreningspress
ED. Religion	3	Religion
Total	58	

Table 5. Number of texts in category E.

2.3.6 Category F: Popular lore

To the category of popular lore we have brought in texts from natural science and technology, which, somewhat surprisingly, are not represented in Brown or LOB. Four texts from the fields of religion and complementary lifestyles have also been added.

Category		Swedish term
F. Popular lore		Populärvetenskap
FA. Humanities	5	Humaniora
FB. Behavioural sciences	5	Beteendevetenskap
FC. Social sciences	5	Samhällsvetenskap
FD. Religion	2	Religion
FE. Complementary life styles	2	Övrig livsåskådning
FF. History	3	Historia
FG. Health and medicine	7	Medicin, hälsa
FH. Natural science, technology	14	Övrig naturvetenskap och teknik
FJ. Politics	1	Politik
FK. Culture	4	Kultur
Total	48	

Table 6. Number of texts in category F.

2.3.7 Category G: Biographies, essays

In SUC's English predecessors this category is called "Belles lettres, biography, essays" and is fairly large. It also contains some more highbrow literary and arts critique. Belles lettres, however, seems not to be a common genre in the more brutish Swedish language. I have removed that part of the category entirely and cut down the number of texts considerably for the rest. What remains are biographies, memoirs and essays. Only more general essays are included, those that can be regarded as reviews are put in category C. The articles in G come from both books and periodicals and are mostly quite long, with one or in a few cases two excerpts per text sample.

Category		Swedish term
G. Biographies, essays		Biografier, essäer
GA. Biographies, memoirs	7	Biografier, memoarer
GB. Essays	19	Essäer
Total	26	

Table 7. Number of texts in category G.

2.3.8 Category H: Miscellaneous

Miscellaneous is, as could be expected, a heterogeneous mixture of texts that do not fit anywhere else but are still felt to be part of what should be in the corpus. I have widened its content and increased the number of texts. This concerns in particular the administrative texts of various kinds. A Swedish citizen is likely to meet quite much information and regulations from local and national authorities and such texts also play an important role in many people's daily work. Company information is also included here.

Category		Swedish term
H. Miscellaneous		Diverse
HA. Federal publications	30	Statliga publikationer
HB. Municipal publications	20	Kommunala publikationer
HC. Financial reports, business	4	Verksamhetsberättelser, företag
HD. Financial reports, non-profit organisations	4	Verksamhetsberättelser, organisationer
HE. Internal publications, companies	10	Företags interna publikationer
HF. University publications	2	Universitetskatalog
Total	70	

Table 8. Number of texts in category H.

2.3.9 Category J: Learned and scientific writing

This category turned out to be difficult. At this level Swedish people often write directly in English in order to reach a larger audience (just like I do myself right now). This is particularly striking in the natural sciences. Therefore the category JH is empty, while the category FH is correspondingly extended. The kinds of texts we have are doctoral theses, reports, articles from high-level professional journals, textbooks for academic teaching and some other books. Mostly they are long enough to allow unitary text samples. The subcategories in J are parallel to those in F, popular lore, barring subcategories E and F.

Category		Swedish term
J. Learned and scientific writing		Lärda och vetenskapliga skrifter
JA. Humanities	27	Humaniora
JB. Behavioural sciences	14	Beteendevetenskap
JC. Social sciences	22	Samhällsvetenskap
JD. Religion	6	Religion
JE. Technology	6	Teknik
JF. Mathematics	3	Matematik
JG. Medicine	5	Medicin
JH. Natural science	0	Naturvetenskap
Total	83	

Table 9. Number of texts in category J.

2.3.10 Category K: Imaginative prose

Under the heading "Imaginative prose" we have nearly the same number of excerpts, 127 as compared to 126, but have changed the distribution over different text types in relation to Brown/LOB. Rather than having six separate categories, we have a superordinate category K with four subcategories that will ensure that we get enough breadth in our choice of texts. KK is "General fiction". L, "Mystery and detective fiction", and M, "Science fiction", have been merged into KL and N, "Adventure and western", and P, "Romance and love", into KN. KR is "Humour". The texts come mostly from novels. From each book of some length we have taken two, sometimes three, excerpts chosen in a manner described below. Except for KR the text samples are never composite of shorter texts. We have taken care to cover different literary genres and levels and, among the light reading, to have texts aimed mainly at a male audience as well as texts aimed at a female audience. Most texts are, of course, unisex. In the category KN we had to take extra care to be sure that the books were not translations, as those books are often published under fictitious names.

Category		Swedish term
K. Imaginative prose		Skönlitteratur
KK. General fiction	82	Allmän skönlitteratur
KL. Mysteries and science fiction	19	Deckare och science fiction
KN. Light reading	20	Triviallitteratur
KR. Humour	6	Humor
Total	127	

Table 10. Number of texts in category K

In sections 2.2 and 2.3 the noble intentions behind the composition of the SUC corpus and the – slightly adjusted – results have been presented in commented tables.

While this may seem neat and reasonable the next section will by contrast tell about the down-to-earth teeth-grinding practical task of collecting text material from a large number of different sources.

2.4 Obtaining the Text Material

Once we had some idea about what kind of texts we wanted and where they might be found, the actual hunt began. We tried to send letters with a description of the corpus project and its need for texts to relevant places and ask people to get in touch. Of course they didn't. People have too much to do to do something by themselves even if they are in favour of the idea. The routine that emerged was to phone the presumptive text donors, ask at the switchboard for some suitable person (it is a bit tricky to decide and describe who is suitable in each case), and then try to coax this suitable person to help us. It is easier with someone you know, or someone who knows someone you know. I used all contacts and asked my friends and talked loudly at parties about the need for computer readable texts and before you knew it, there turned up somebody, or somebody's sister-in-law, and we could get some more texts of some crucial type. But it took time. One good strategy is to start as high up in a hierarchy as possible. It is much easier to phone someone and say: "I just talked to your boss and she gave me permission to use some of the texts you are working with" than to go in the opposite direction.

The large daily newspapers were different in the sense that once we got a good contact, we could have as much text as we wanted, but the hunt for the right person and the coaxing part was the same. We also needed the permission of the local section of the journalists' trade union. Sometimes this was easy, as with the small local paper where the editor-in-chief and the head of the trade union was the same person; a neat example of the Swedish model. This might have led to moral misgivings if it were not for the fact that he wrote most of the articles himself as well.

Books meant still another variety, as we had to have permissions from both the publisher and the author. (I will return to the legal intricacies in Section 2.6.) Here we first got in touch with central persons at some publishing houses to make sure they were positive, then tried to get the consent of some of their authors and then came back to the publishers to obtain the texts. One problem that we discovered was that even when publishers print books directly from a computer file, they often do not keep the files to make new editions from them but only keep the originals on paper. This meant that we could only have books that were actually in the process of being printed, not choose among all titles, and we had to get hold of the printer's diskettes before they were reformatted and used for new books.

For the 500 text files in the corpus, I would estimate that at least 1,500 phone calls were necessary. Of course, almost all sources gave us more than one text, but all sources also needed several phone calls and several letters and much bookkeeping to know who had promised what and who had kept their promises. Once we managed to get the first "yes" from someone, we sent an agreement form (see Appendix B) to be signed and returned to us in an enclosed stamped envelope with our address on it. If we did not get the form back, we had to phone again, send new forms (with new stamped envelopes that cost us lots of money) and keep track of where in this series of actions each text donor was, so as not to remind someone who had returned the first form and not to forget those guys who did not return anything and probably just stole our stamps.

This job was unscientific, unlinguistic and terribly boring but, alas, it turned out that it had to be done by someone with authority. To have shy students humbly asking for texts didn't work. Thus, I did a lot of the phone work myself, as the title "professor" has a surprisingly good effect as a door opener. Other members of the project group sent out forms and took care of the track keeping. They also did their share of the text collecting by tapping texts from friends and relatives. [1]

Actually, I think this was the most disgusting part of the corpus building business, as it so effectively destroyed other work. It was not possible to set aside one day a week or one hour each day for making phone calls. People have to be phoned when they can be reached. I made heaps of notes like "Monday 2 o'clock: Mr. Andersson is in a meeting, will be back by 4. 4 o'clock: Mr. Andersson is not back yet. 4.15: Mr. Andersson has left for the day, try again tomorrow. Tuesday 9 o'clock: Mr. Andersson is ill, try again by the end of the week." And this is the lucky situation when there is a switchboard operator who furthermore knows about Mr. Andersson's

[1] In fact some other members of the project group, particularly Gunnar Eriksson, did a great deal more of the text collecting work than tapping texts from friends and relatives.

whereabouts, quite as often this is not the case. The consequence is that the phoning business is always on the back of one's mind.

So, how would I do it the next time? Probably not try to cover so wide a variety of texts, although I now think it is good that we did. All the succeeding handicraft work with the collected texts strengthened my original conviction that there are large and important differences between text types, also in such a seemingly uniform language as modern written Swedish. The differences can be found on all linguistic levels and can have quite strong effects on some of the standard things one might want to do with a corpus. To claim things about language L after looking at a set of newspaper texts - or even a set of texts from the same newspaper - is a dangerous thing to do. At least one should have some conception of what features can vary strongly between text types and that information can only be found by looking at diversified corpora. Recent research has also shown genre to be an important parameter in, e.g. information retrieval and related areas. It is thus important to have access to corpora with various and well-defined genres for basic research. We who did the work of getting SUC fairly balanced would hardly be prepared to do the same thing again, but now it is done and we hope that our efforts will save other people from having to do it soon again. Admittedly, there is often not enough text in specific subgenres to allow strong generalisations, but there is enough text to formulate hypotheses to be tested on larger amounts of similar texts. See for instance the dissertation by Karlgren (2000) or the paper by Cutting and Karlgren (1994).

The practical business of finding the right person and getting all papers signed and archived is hard to avoid even when it is not done on such a grand scale as with SUC. The best advice would be not to underestimate the time and effort it takes, to start early, to assign a person with good self-confidence and immense patience to do it, and to meticulously keep track of every step. Most important of all is: Do it! Get all the legal stuff done correctly, in spite of my negative description above. If a corpus of reasonably modern texts is to be used by anyone else than its creator, the legal matters have to be cleared out. Copyright in the electronic era is a complicated matter and the awareness of it is growing among authors and publishers. To trespass in this respect may cause severe damage, not only to the one who does it but also to linguistic research in general. Copyright rules and related questions will be treated more in Section 2.6.

2.5 Data Catch and Pre-processing

When we started our text collecting in 1990 we had decided only to use texts that were available to us in computer-readable form. Of course that is the only sensible way of doing it, but still it did not save us as much trouble as we had thought. We were well aware of the problems of conversion between different formats and prepared to handle them, but by then we did not know how many wildly different formats there were and how many varying (non-)standards for representing characters outside the range of 7-bit ASCII.

This point might be a little easier today. There is a general movement towards more standardisation nowadays, which is highly desirable. Old and clumsy typesetting systems with idiosyncratic ways of representing characters are replaced by more modern systems adhering to one of a few standards. But also in a situation where everything is standardised – in principle – I would recommend that all files coming from external sources be run through a small check making sure that all characters in the file belong to a predefined set.

All incoming text files were converted to the desired format, which at this stage of the processing was a DOS-text format which could be fairly safely interchanged between PC and Mac. Excerpts with a length around 2,000 graphic word tokens were picked out in a way to be described below. They were put in separate files and given carefully composed SUC-names.

A SUC-file can be composed in three different ways, of which one is very rare, namely the case where one single text in its entirety is just about 2,000 words long. The other two cases are where an excerpt is cut out from a longer text, as with books, and where a file is composed of several shorter texts, as is almost always the case with, e.g. newspaper material. The TEI Guidelines have recommended ways of marking such properties. In SUC, however, we judged it unnecessary to introduce a special tag, as the difference can be seen anyway from the name of the text(s) in the file. The difference between excerpts from longer texts and the almost non-existent single texts of 2,000 words can not be told this way, but we did not think that motivated the introduction of another space-consuming SGML tag.

The SUC-names of texts consist of two letters, signifying main category and subcategory as presented in Sect. 2.3. This is followed by a two-digit enumeration of the texts in the concerned subcategory. In the case of entire texts or excerpts from longer texts, where all text in the file has the same source, this is all. In the very common situation where a file consists of several shorter texts, the file itself has a two-plus-two name but the single texts have an extra, lowercase letter added to their names. As the names of SUC files all look the same (like long text names) one has to look in the files themselves or in the bibliography to see if a file is composite or not.

Category		Number	Subtext	
Main	Sub-			
K	N	04		Long text, single source
A	F	06	a	One of several short texts with different sources

Table 11. The structure of the names of SUC texts.

As we foresee research on the discourse level, not only on levels such as lexicon and syntax, we have tried to choose excerpts that are in some sense whole even if taken from a larger context. Asking the first person we met in the hallway to mention a digit between 1 and 4 carried out this somewhat steered random choice from books. That digit decided which quarter of the book that would be used and we tried to pick out the 2,000-word excerpts in a way that was both intelligent and random.[2]

We have stored hard copies of all the corpus texts. And for most of the books there is a hard copy on the project bookshelf. This material should preferably be kept together some time after the project is formally finished.

The hard copies have proved useful all along for checking oddities while correcting the tagged corpus. Some peculiar words are due to clear printing errors and some are due to discrepancies between the electronic text files and the printed and published versions. All these should ideally have been found at the first proofreading of the incoming files, but of course were not always. The first proofreading was mainly aimed at comparing incoming files and printed text. If the electronic versions sent to us had been exported at an early stage in the production process the actually printed texts could be a lot different, due to late editing. This concerns particularly articles in newspapers and periodical journals. Since we did not have resources to compare everything letter by letter we would just browse through the texts to find cases with large discrepancies, and read these more thoroughly. In some cases the electronic text files arrived as Desktop program versions which would hopefully guarantee that files and printed versions are the same.

Thus our initial proofreading was not primarily aimed at finding all trivial spelling mistakes, since these could to some extent be corrected automatically or manually later in the SUC process. Some misspelled words appear as ‘non-words’ to the first part-of-speech tagger (cf. 3.2.1) and will not be tagged at all. Other misspelled words may be incorrectly tagged, and may be noticed by annotators checking the tags. And in a finally annotated corpus a parser may find oddities. So, proofreading really never ends – and is never perfect. (As anyone who ever reads books, newspapers and other printed matter may have noticed.)

2.6 Legal Aspects

In order to obtain permission not only to use but also to distribute texts for bona fide research purposes it is necessary to have the consent from both the publisher and the author. In the case of newspapers and all kinds of periodica, that means the editor-in-chief and the journalists' local trade union.

The agreement forms in Swedish and English are added as Appendix B. We consulted lawyers to make sure that the forms contain what they must, expressed in a legally relevant way.

Although I probably have the copyright of the agreement forms I am not going to claim it. Our forms have worked very well with the Swedish text donors. A couple of times, hesitant presumptive donors have become convinced that they could safely give their text away after reading the agreement they were to sign. If the formulations can help anyone, as they are or as models for something else, please use them. (Cf. Appendix B.)

[2] From here on the manuscript for this section became very sketchy. We have added a few paragraphs to meet at least some of the intentions of the author, inevitably in a different style.

3 The Morphosyntactic Tagging

3.1 Principles behind the Tag Set

The principles behind the morphosyntactic tagset have been extensively discussed in Ejerhed, Källgren, Wennstedt & Åström (1992).

3.2 Definition of the tags

3.2.1 The SUC Morphosyntactic Tagset

The morphosyntactic tagset used in SUC was designed at an early stage of the project. It was documented in Ejerhed, Källgren, Wennstedt & Åström (1992) and was used to tag the 300 000-word subpart of SUC published on the first ECI CD-ROM (1993). Since then the tagset has only undergone a few changes, most notably the introduction of a part-of-speech tag for verbal particles, PL.

The tagset is based upon that used by the SWETWOL (Karlsson 1992) with some modifications, e.g. in the subclassification of adverbs and in the relative order between elements. Throughout the project, the first part of the tagging of SUC texts was done in co-operation with the Department of General Linguistics at Helsinki, where the words were given SWETWOL tags. The tags were then transduced into SUC tags at the Department of Linguistics at Umeå.

There also exists a one-to-one mapping between the SUC tags and the more compact Parole tags (see ‘taggtabell’ at http://spraakbanken.gu.se/lb/parole for translation schemes in both directions). Mostly this mapping is quite straightforward. One exception is the treatment of participles, which SUC regards as a separate part-of-speech while Parole classifies them as adjectives.

Another difference is that the SUC format gives the text word at the beginning of the line and the base form at the end of it, while Parole has it the other way round. This makes SUC, with its verticalized running text, infinitely more readable to a human eye. Texts in Parole format are not supposed to be read by humans. Still, many scholars working with texts, e.g. doing analysis on discourse level, have to look at the texts with markup and read them as best they can. That is why we have kept this order between the elements in the SUC format of the corpus.

SUC 2.0 presently exists in three sgml-versions, two with SUC tags and one with Parole tags. Two of them are TEI-conformant, while the third has a local sgml-format not strictly TEI-conformant. The text words and their analyses are the same for all three versions. Below is a presentation of the current morphosyntactic SUC tags. For each part-of-speech tag the corresponding part of the Parole tag is also given. (Table 14)

3.2.2 The Structure of SUC Tags

Each SUC tag contains a part-of-speech label. (cf. Table 12). For many parts-of-speech, this is all there is of morphosyntactic information. For others, the part-of-speech tag is followed by one or more feature values for various properties of the tagged word. (cf. Table 13). Last in all tags, complex or simple, comes the base form of the word. In the SGML-format used, this will look like in the example below, the plural form arenor 'arenas' of the noun arena. (<w> is the SGML-tag used for words in SUC, <ana> stands for analysis, <ps> for part-of-speech, <m> for morphosyntactic information and <b> for base form.)

```
<w>arenor<ana><ps>NN<m>UTR PLU IND NOM<b>arena</w>
```

Code	**Swedish category**	**Example**	**English translation**
AB	Adverb	*inte*	Adverb
DT	Determinerare	*denna*	Determiner
HA	Frågande/relativt adverb	*när*	Interrogative/Relative Adverb
HD	Frågande/relativ determinerare	*vilken*	Interrogative/Relative Determiner
HP	Frågande/relativt pronomen	*som*	Interrogative/Relative Pronoun
HS	Frågande/relativt possessivt pronomen	*vars*	Interrogative/Relative Possessive
IE	Infinitivmärke	*att*	Infinitive Marker
IN	Interjektion	*ja*	Interjection
JJ	Adjektiv	*glad*	Adjective
KN	Konjunktion	*och*	Conjunction
NN	Substantiv	*pudding*	Noun
PC	Particip	*utsänd*	Participle
PL	Partikel	*ut*	Particle
PM	Egennamn	*Mats*	Proper Noun
PN	Pronomen	*hon*	Pronoun
PP	Preposition	*av*	Preposition
PS	Possessivt pronomen	*hennes*	Possessive
RG	Grundtal	*tre*	Cardinal number
RO	Ordningstal	*tredje*	Ordinal number
SN	Subjunktion	*att*	Subjunction
UO	Utländskt ord	*the*	Foreign Word
VB	Verb	*kasta*	Verb

Table 12. The 22 part-of-speech categories in SUC. The fairly mnemonic 2-letter-code is followed by the Swedish category name and a typical word where it can apply. English translations of category names in the last column.

In Table 13 below, all the morphosyntactic features used are given along with their possible values. The parts-of-speech to which each feature can be applied are also specified. Several parts-of-speech do not have any morphosyntactic features. Somewhat longer descriptions can be found in Ejerhed et al. (1992).

Feature	**Value**	**Legend**	**Parts-of-speech where feature applies**
Gender	UTR	Uter (common)	DT, HD, HP, JJ, NN, PC, PN, PS, (RG, RO)
	NEU	Neuter	
	MAS	Masculine	
Number	SIN	Singular	DT, HD, HP, JJ, NN, PC, PN, PS, (RG, RO)
	PLU	Plural	
Definiteness	IND	Indefinite	DT, (HD, HP, HS), JJ, NN, PC, PN, (PS, RG, RO)
	DEF	Definite	
Case	NOM	Nominative	JJ, NN, PC, PM, (RG, RO)
	GEN	Genitive	
Tense	PRS	Present	VB
	PRT	Preterite	
	SUP	Supinum	
	INF	Infinite	
Voice	AKT	Active	
	SFO	S-form (passive or deponential)	
Mood	KON	Subjunctive (Sw. konjunktiv)	
Participle form	PRS	Present	PC
	PRF	Perfect	
Degree	POS	Positive	(AB), JJ
	KOM	Comparative	
	SUV	Superlative	
Pronoun form	SUB	Subject form	PN
	OBJ	Object form	
	SMS	Compound (Sw. sammansättningsform)	All parts-of-speech

Table 13 . Morphosyntactic features (with 3-letter-values) and the parts-of-speech to which they apply. (Parentheses indicate that a feature only applies to some members of the p-o-s or that not all the values of a feature are applicable.)

3.2.3 Three tags that are not normal morphosyntactic tags: SMS, UO, AN

The tag SMS has a peculiar status. Originally (and as described in Ejerhed et al. 1992) it was meant for the particular forms that a handful of Swedish nouns have in compounds. Historically that is a case form and it can appear in isolation only in connection with conjoined compounds, as in *kvinno- och mansgrupper* 'women's and men's groups'. It is disturbing to call the form *kvinno-* either nominative or genitive and even more disturbing to evoke some obsolete early Swedish case. Instead we chose to indicate an omitted component through a hyphen in a contracted, conjoined compound and call it SMS (for Sw. *sammansättning* 'compound'). The human annotators, however, soon started to use it for all kinds of conjoined compounds, not only for nouns that do not have a special compound form (*café- och biovagn* 'café and movie train', SUC-text HE03) but also in constructions such as *över- och bottenvåning* 'upper and ground floor' (SUC-text KK34) with conjoined non-flexional adverbs. This actually turned out to be smart, as there have occurred, e.g. conjoined compounds involving verbs, where it is not possible to decide which form of the verb is being used (ex. *sov- och liggvagnar* 'wagon-lits and couchettes', lit. 'sleep- and liewagons', SUC-text HE03; *sov-* would, on morphological grounds, otherwise be classified as an imperative or a preterite). The tag SMS is thus allowed in all parts-of-speech on strings that end in a hyphen and are the first part of a compound whose second part comes later.

Examples of SMS-tagged items in the corpus follow.

aa04:

```
<w>skörde-<ana><ps>NN<m>UTR - - SMS<b>skörd</w>
<w>och<ana><ps>KN<b>och</w>
<w>upparbetningsmetoder<ana><ps>NN<m>UTR PLU IND NOM<b>upparbetningsmetod</w>
```

aa10:

```
<w>torsk-<ana><ps>NN<m>UTR - - SMS<b>torsk</w>
<w>och<ana><ps>KN<b>och</w>
<w>plattfiskyngel<ana><ps>NN<m>NEU PLU IND NOM<b>plattfiskyngel</w>
```

hb11:

```
<w>låg-<ana><ps>JJ<m>POS UTR - - SMS<b>låg</w>
<d>,<ana><ps>MID<b>,</d>
<w>mellan-<ana><ps>AB<m>SMS<b>mellan</w>
<w>och<ana><ps>KN<b>och</w>
<w>högstadiet<ana><ps>NN<m>NEU SIN DEF NOM<b>högstadium</w>
```

Foreign words and expressions are surrounded by the SGML-tag <foreign> (cf. 4.3.10) but they also have the part-of-speech tag UO (Sw. *utländskt ord* 'foreign word') without any further subclassification. The same holds for all foreign words, no matter if they can be expected to be wellknown by the reader or not. (Ex. cf03 and cc01 respectively.) Foreign names are not marked <foreign>, only <name>, and all words within such a <name>-tag are tagged PM. This may seem strange in languages where a normal reader easily can tell what is a proper noun and what is not (Ex. cb03) but in order to be consistent across languages and not have the analysis dependent on the annotator's knowledge of foreign languages, we have chosen the more simplistic solution. [3]

cf03:

```
<w n=1538>sant<ana><ps>AB<m>POS<b>sant</w>
<w n=1539>genuin<ana><ps>JJ<m>POS UTR SIN IND NOM<b>genuin</w>
<w n=1540>och<ana><ps>KN<b>och</w>
<d n=1541>"<ana><ps>PAD<b>"</d>
<foreign lang=en>
<w n=1542>basic<ana><ps>UO<b>basic</w>
</foreign>
<d n=1543>"<ana><ps>PAD<b>"</d>
```

cc01:

```
<foreign lang=el>
<w n=2152>kouroi<ana><ps>UO<b>kouroi</w>
</foreign>
```

cb03:

```
<name type=work>
<w n=2129>Howards<ana><ps>PM<m>GEN<b>Howard</w>
<w n=2130>end<ana><ps>PM<m>NOM<b>end</w>
</name>
```

An abbreviation can consist of one or more words. The part-of-speech of an abbreviation is decided from the syntactic function of the expression, not of the single words in it. Abbreviations have AN added as a morphological feature but are also surrounded by <abbr>-tags (cf. 4.3.12). The base form is either the abbreviation itself (cd03) or a spelling out of it (fa05). The choice of base form for each abbreviation is rather unsystematic in SUC, but for all occurrences of an abbreviation the base form is always the same. Initials in proper names are not treated as abbreviations.

cd03:

```
<w n=155>etc<ana><ps>AB<m>AN<b>etc</w>
```

fa05:

```
<w n=94>dvs<ana><ps>AB<m>AN<b>det_vill_säga</w>
```

[3] The attributes *lang* to foreign and *type* to name will be discussed in 4.3.10 and 4.3.6. The values used here are ‘en’ for an English word, ‘el’ for a Spanish word and ‘work’ for the name of a film.

3.2.4 A Comparison between the Part-of-Speech Tags used in SUC and Parole

The overview in Table 14 gives the part-of-speech tags in alphabetic order, the full grammatical term (only in Swedish), the corresponding tag fragment in the Parole system, the relevant SUC morphosyntactic features if any, and sometimes a short description of how the tag is used where this is not felt to be more or less self-evident. As for the values that the <m>-features can take, see Table 13.

Tables for converting entire suctags to corresponding Parole tags and vice versa can be found at http://spraakbanken.gu.se/lb/parole.

Here is just an example. The word *arenor*, which with suctags is:

```
<w n=1414>arenor<ana><ps>NN<m>UTR PLU IND NOM<b>arena</w>
```

(cf. 3.2.2) will with Parole tags be:

```
<w lem='arena' msd='NCUPN@IS' n=1414>arenor</w>.
```

SUC							**Parole**	**Example**
AB	adverb, icke-komparerbart						**RG**	*inte*
	Komparerbart	POS						*fullkomligt*
		KOM						
		SUV						
DT	Determinerare	NEU	SIN	DEF			**D**	*denna*
		UTR	PLU	IND				
		MAS		IND/DEF				
		UTR/NEU						
HA	frågande/relativt adverb						**RH**	*när*
HD	frågande/relativ determinerare	UTR	SIN	IND			**DH**	*vilken*
		NEU	PLU					
		UTR/NEU						
HP	frågande/relativt pronomen	NEU	SIN	IND			**PH**	*som, vad*
		UTR	PLU					
		UTR/NEU						
HS	frågande/relativt possessivt pronomen	DEF					**PE**	*vars*
IE	infinitivmärke						**C (CIS)**	*att*
IN	interjection						**I**	*ja*
JJ	adjektiv	POS	NEU	SIN	IND	NOM	**A**	*glad*
		KOM	UTR	PLU	DEF	GEN		
		SUV	UTR/NEU	SIN/PLU	IND/DEF			

			MAS					
KN	konjunktion						**C (CCS)**	*och*
NN	substantiv	NEU	SIN	IND	NOM		**N**	*hotell*
		UTR	PLU	DEF	GEN			
PC	particip							
	perfekt	PRF	NEU	SIN	IND	NOM	**A (AF)**	*utsänd*
			UTR	SIN/PLU	DEF	GEN		
			UTR/NEU	PLU	IND/DEF			
			MAS					
	presens	PRS	UTR/NEU	SIN/PLU	IND/DEF	NOM	**A (AP)**	*talande*
						GEN		
PL	partikel						**Q**	*under*
PM	egennamn	NOM					**N (NP)**	*Mats*
		GEN						
PN	pronomen	NEU	SIN	IND	SUB		**P**	*hon*
		UTR	PLU	DEF	OBJ			
		UTR/NEU	SIN/PLU	SUB/OBJ				
		MAS						
PP	preposition						**S (SP)**	*av*
PS	possessiva pronomen	NEU	SIN	DEF			**P (PS)**	*hennes*
		UTR	PLU					
		UTR/NEU	SIN/PLU					
RG	grundtal	NOM					**M (MC)**	*tre*
		GEN						
		NEU	SIN	IND	NOM			*ett, en*
		UTR			GEN			
RO	ordningstal	NOM					**M (MO)**	*tredje*
		GEN						
		MAS	SIN	IND/DEF	NOM			*förste, andre*
						GEN		
SN	subjunktion						**C (CS)**	*att*
UO	utländskt ord						**X (XF)**	*the*
VB	verb	KON	IMP	AKT			**V**	*ger*
			INF	SFO				
			PRS					
			PRT					
			SUP					

(Table 14. Caption on next page.)

Table 14. An overview of the SUC word tags. Swedish grammatical terms are given for the p-o-s tags, but not for their relevant morphological features, which were explained in table 13. The two rightmost columns contain the corresponding p-o-s-parts of Parole tags and examples of typical Swedish words where a tag is applicable. [4]

Delimiters also have "part-of-speech" tags that show their function. (Cf. 4.2.2)

SUC		**Parole**	**Example**
MAD	major delimiter	FE	.
MID	minor delimiter	FI	,
PAD	pairwise delimiter	FP	(
PAD	pairwise delimiter	FP	)

[4] For corresponding grammatical terms in English and a full discussion of the morphosyntactic features see Ejerhed et al. (1992).

4 The SGML Markup

4.1 Introduction

4.1.1 The Text Encoding Initiative and LE-PAROLE

In designing a markup system for the corpus, we have tried to follow the guidelines issued by the Text Encoding Initiative, TEI, in particular as presented in the so-called P2 (Burnard & Sperberg-McQueen 1993). Our manual was ready and we had started using it before the publication of ‘P3’ in April 1994, but we have taken into consideration some material from its earlier electronic version. In many cases, however, the guidelines were not (yet) elaborate enough or there existed no established TEI recommendations. This concerns especially the representation of the results of the linguistic analysis being carried out on the texts, where we invented a format and a set of tags of our own. In doing this, we have taken care to adhere to the general spirit of the TEI guidelines.

The SUC tags are wholly compatible with the LE-Parole tags and conversion between the formats is easy. Besides the compact LE-Parole tags, we want to keep our own, more substantial tags as an allowed format, as they are more convenient to people working closely with the texts. They are mnemonically based upon standard Swedish grammatical terminology. The format is described in 4.1.2.

4.1.2 The Application of SGML Markup to the SUC

Our linguistic and structural markup is entered directly into the texts in order to represent information that cannot be seen from the text surface itself. The type of markup used is presented in sections 4.2 and 4.3, and some conventions for its use in section 4.4.

A Document Type Declaration, DTD, for the SUC markup has been constructed. The DTD gives a kind of "formal syntactic" definition of the markup used. The SUC DTD is reproduced in Appendix C. The header of a text or a corpus is where all important bibliographical facts (in a wide sense) are given. The header is, so to speak, outside the text proper. The header of the SUC corpus (cf. Appendix F) is quite extensive and altogether follows the TEI recommendations. Besides the header of the entire corpus, each separate text in the corpus has its own document header with information that is unique to that text. The headers are presented in Appendix D, and the traditional bibliographical information about the texts included in the corpus is exemplified in Appendix E.

Something should be said about the terminological confusion in connection with the word *tag*. In SGML, *tag* is a term for any kind of markup, also typographic markup and markup in text and corpus headers. In linguistics, *tag* traditionally denotes information that is attached to a word as a result of some kind of linguistic analysis. In this section, *tag* is used in the more general (SGML) sense. Where this is not the case a modifying adjective has been added.

In sections 4.2-4.3 all markup used in the SUC 2.0 is described in a way that is meant to be so clear and exhaustive that the markup of the corpus can be easily understood and also so that the same system for markup can be applied to other texts by other people according to the same criteria. Examples are given, both of the general application of the tags and of the rules-of-thumb that we have settled on. Our decisions in tricky cases are discussed. The interaction between different tags, e.g. how they are nested, is also described. The examples are, as far as possible, authentic and drawn from the SUC. Where examples are constructed, this is pointed out. Sometimes we give examples with full markup; sometimes we use reduced markup to make the discussed issue more prominent. In some examples, we have added indentation to show the structure of, e.g. nested markup more clearly.

An important part of the markup of words is the representation of the results of the tagging and annotation process. Swedish is a language with a rich inflectional morphology. This means that, apart from the part-of-speech tag that all words are given, their uninflected base form is also given, and, depending on part-of-speech, up to five morphological properties may be specified. This could (perhaps should) be done by means of attribute-value pairs, but as all values are unique, we just state the values and regard them as a kind of shorthand for a full specification, to which they could easily be translated if need be. Thus, rather than writing *number=sin* we write *SIN*, knowing that *NUMBER* is the only attribute that can take *SIN* as a value.

We have tried to design a format that is as simple and general as possible, that can be applied to other languages as well - in particular to those with a rich morphology - and that can be easily extended to cover other needs and other kinds of information. It is our hope that the SUC format and tags for representing linguistic analysis can be used for other corpora as well.

To facilitate understanding for readers not acquainted with SGML, Section 4.1.3 can be read before plunging into the description of the markup proper in Sections 4.2-4.3.

4.1.3 A few words about SGML format, tags and attributes

SGML stands for Standard Generalized Markup Language. Since 1986, it has been an international standard, ISO 8879. It is primarily intended to be used to obtain a representation of text format that is independent of the original mode of representation and stable enough to be truly portable between all various kinds of word processing programs, computers, operating systems, networks, etc. Given the right software, SGML markup can also be a good basis for information retrieval as well as printing and publishing. Transfer between SGML and HTML is easy, in principle. (There are many things in the world of computing that are said to be easy in principle.) A good introduction to SGML is van Herwijnen (1990) or Alschuler (1995).

SGML can be used for representing many kinds of information about texts. It can, e.g. be used to represent the structure of texts or parts of texts, as well as the results of some linguistic analysis. In the SUC project, we use SGML-type markup for all these purposes. This is also what the TEI, Text Encoding Initiative, recommends.

An important concept in SGML is that of a *tag*. A *tag* delineates a stretch of text and says something about what characterises it. A so-called *end tag* is used to show the range of the tag. Tags are written in angle brackets, < >, with slash, /, signalling end tags. In SUC, if something is a name the tag <name> will enclose it.

```
<name>Sweden</name>
<name>Ingrid Bergman</name>
```

As is seen from the example, more than one word can fall within the range of one tag. In fact, arbitrarily long sequences can be covered by one tag, as, e.g. when it is used to mark a long quotation or an entire chapter in a book. Under all circumstances, all material covered by one tag must of course be of the type indicated by the tag.

Sometimes a need to specify the tags further may arise. This can be done by means of what is called *attributes*. Attributes are given values, either from a predefined set or written freely. The names above can be classified according to the attribute *type* and be given values from a set of possible types of names. Attributes and their values are written inside the same angle brackets as the start tag:

```
<name type=place>Sweden</name>
<name type=person>Ingrid Bergman</name>
```

Tags are mostly like parentheses, in that they must be balanced, i.e. they must have both a start tag and an end tag. The actual insertion of end tags is, however, sometimes obligatory, as with names, and sometimes optional. The SUC tag for the basic unit <s> (mainly corresponding to graphic sentence) has an optional end tag, as the start of a new sentence normally signals the end of the foregoing one and thus can function as both a start tag and an end tag. Tags can also have hierarchical relations; a paragraph boundary <p>, e.g. serves to close both a foregoing paragraph and a foregoing sentence.

```
<p><s>This is the first paragraph and the first sentence.<s>This is the second
sentence.<p><s>This is the second paragraph and the third sentence.</p>
```

With full specification of both start tags and end tags, this will be written as below.

```
<p><s>This is the first paragraph and the first sentence.</s><s>This is the second
sentence.</s></p><p><s>This is the second paragraph and the third sentence.</s></p>
```

Some tags mark a single point in a text and thus need no end tags. The single tag in the example below shows the point in a text where something has been omitted. Its attributes tell us what was omitted (from a predefined set or in free text), the extent of the omitted material (written in free text) and the signature of the person responsible for the decision to omit the material.

```
<gap desc=table extent="15 lines" resp=gk>
```

For all tags that are described below, it is stated whether end tag is obligatory or optional, whether there are any attributes, where in the process the tags are entered, and whether it is done automatically or manually. All occurrences of tags are liable to changes during the manual annotation if they are found to be wrong in any way. In the present edition, SUC 2.0, not all of the tags have actually been used for lack of time and resources. What has been left out is stated in the description of the tags.

4.2 Markup for structural units in text

This section presents the SUC markup of the basic units in text, viz. words, punctuation, sentences (s-units), and paragraphs. All words, numeral expressions, punctuation and other (strings of) black characters in the texts are given separate tags. We have taken care to give the tags a uniform format regardless of what kind of element they tag.

4.2.1 Words

Tag: <w> (with embedded tags <ana>, <ps>, <m>, <b>)

The tag has no attributes (barring token number) but some embedded tags, see below. End tag, </w>, is not obligatory but strongly recommended for the sake of clarity. The <w> tags are automatically generated to surround each word in the text.

Every string consisting entirely of black characters that is not a punctuation is tagged <w>. To each word in a text, its analysis is attached by means of the tag <ana>, for analysis. The analysis has three sub-parts, one for the part-of-speech, <ps>, one for information (if there is any) about morphological properties, <m>, and one for the base form of the word, <b>. For the embedded tags, no end tag is necessary. None of the tags has any attributes.

Examples:

```
<w>för<ana><ps>PP<b>för</w>
```
preposition 'for'

```
<w>boken<ana><ps>NN<m>UTR SIN DEF NOM<b>bok</w>
```
noun 'book' in singular definite form

```
<w>arrogant<ana><ps>JJ<m>POS UTR SIN IND NOM<b>arrogant</w>
```
adjective 'arrogant'

```
<w>15<ana><ps>RG<m>NOM<b>15</w>
```
cardinal number '15'

```
<w>fördelar<ana><ps>VB<m>PRS AKT<b>fördela<ana><ps>NN<m>UTR PLU IND NOM<b>fördel</w>
```
the word token *fördelar* is ambiguous between the verb *fördela*, 'to share', and the noun *fördel*, 'advantage', and thus has two <ana> tags in non-disambiguated text

```
<w>boken<ana><ps>NN</ps><m>UTR SIN DEF NOM</m><b>bok</b></ana></w>
```
noun 'book' in singular definite form with optional end tags inserted to show the nesting more clearly

In the example above, e.g. *boken* is a text word, the word as it appears in the text. The first part of its analysis specifies its part-of-speech, in this case NN for noun. The morpho-syntactic information is given as a sequence of values of implicit attributes (cf. above). The example says that the gender of the word is uter, and that, in this occurrence, it is singular in number and definite and nominative in form. Finally, *bok* is the uninflected base form necessary for lemmatising. The full set of morpho-syntactic tags for single words, their exact meaning and the rules for their application are described in Section 3 and in Ejerhed et al. (1992).

The content of all the embedded tags is declared in the DTD (see Appendix C) as #PCDATA[5], i.e., in principle anything can be written in them. It is easy to specify which the allowed part-of-speech designations are, and it would also be possible to enumerate all allowed strings of morpho-syntactic data (they are about 260 and are listed in the annotation manual) and to permit those and only those as values of <m>. This solution would, however, have been more clumsy and inflexible, and is quite unnecessary as the correctness of the morpho-syntactic tags is checked elsewhere in the process. The base form must be #PCDATA, as there can be no restrictions on what is allowed there.

[5] #PCDATA means parsable character data. See e.g. Heerwijnen (1990).

All information about each word token could, as stated above, be given as values of attributes of the tag <w>, i.e. something like the following example.

```
<w ps="NN" m="UTR SIN DEF NOM" b="bok">boken</w>
```

This is more in the spirit of SGML, but makes texts even more repellent to the human user than our chosen format. We realise that although it is not meant to be so, it is unavoidable that people will actually read and use the corpus texts as they are, without access to sophisticated SGML software. When such is more generally available we may consider re-formatting our word tag.

As the format is now defined, it allows for flexible changes and additions. Entirely new kinds of information can also be entered in the form of new tags inside the superordinate tag <ana>. At present, the tag <ana> is in principle superfluous, but by having it as a welldefined delimiter between the text word and the inserted analysis, any insertions, deletions, or changed order among the analysis tags will not lead to any problems. As can be seen from the examples, it is also possible to have several tags for ambiguous words, with the readings listed and each new reading signalled by an <ana>. When tagging word groups and punctuation the tag <ana> also defends its position.

4.2.2 Punctuation

Tag: <d> (with embedded tags <ana>, <ps>, <b>)

No attributes (disregarding token number). End tag is obligatory. Markup can be inserted automatically according to the conventions given in the examples below, but, as delimiters are sometimes ambiguous, their markup has to be checked manually in the same way as for words.

Punctuation must be handled and classified according to function. We have chosen to identify a minimal set of three types of punctuation marks and to tag them in a way analogous to the markup of words.

The <d>, delimiter, tag contains the same kinds of information as a word tag, except <m>, morpho-syntactic information. The base form <b> is here always equal to the delimiter itself. Alternatively, the standard designation used in SGML for the delimiters could have been used as base form, or the entities discussed in chapter 6.2 in TEI P3 Part II.

The <ps> tag is here used for information about the function of the delimiter. It takes one of three values. The values are major (sentence) delimiter, called *mad*, minor delimiter, *mid*, and pairwise delimiter, *pad*. Of these, minor delimiter is the trashcan category, where things that are not clearly major sentence delimiters or pairwise delimiters are put. Otherwise, the most common punctuation marks are by default classified as in the examples. (Sometimes a carriage return may also function as a delimiter, but as it is not a 'black character' it will get no markup.) A classification like this one makes it possible to 'lemmatise' both occurrences of punctuation marks and their actual function, e.g. in order to derive statistics about how they are used. It also facilitates the delimitation of sentences. A sentence final delimiter is regarded as belonging to the sentence it closes.

Examples:

```
<d>.<ana><ps>MAD<b>.</d>                 (full stop) [6]
<d>?<ana><ps>MAD<b>?</d>
<d>!<ana><ps>MAD<b>!</d>
<d>...<ana><ps>MAD<b>...</d>             (treated as one single instance of mad)
<d>?!!<ana><ps>MAD<b>?!!</d>             (-"-)

<d>:<ana><ps>MAD<b>:</d>                 (':', ';' and ',' are ambiguous between
<d>;<ana><ps>MAD<b>;</d>                 mad and mid; these are, however,
<d>,<ana><ps>MID<b>,</d>                 their default values)

<d>'<ana><ps>PAD<b>'</d>                 (various kinds of quotation marks are tagged
<d>"<ana><ps>PAD<b>"</d>                 and can later be interpreted further)
```

[6] Full stops in abbreviations could be classified as *mid*, but in the SUC such stops are regarded as constituent parts of the word token.

```
<d>(<ana><ps>PAD<b>(</d>
<d>)<ana><ps>PAD<b>)</d>
```

The following delimiters are classified as mid, but can be changed in the rare cases when they are sentence delimiters:

```
<d>-<ana><ps>MID<b>-</d>                    (dash, not hyphen)
<d>*<ana><ps>MID<b>*</d>
<d>/<ana><ps>MID<b>/</d>
```

Hyphens do not receive any tag, as all soft hyphens are removed in the SUC and hard hyphens, as in proper names and compound words, are regarded as constituent parts of the word.

4.2.3 Sentences/S-units

Tag: <s>
Attribute: id
Possible values of id: a unique identification of each sentence

End tag is optional. Automatic markup in interaction with the classification of sentence delimiters. The TEI guidelines discuss this tag in P3 Part IV 15.1 Linguistic Segment Categories.

The *id* attribute takes as its value a unique identifier constructed from the SUC identifier of the text and a running enumeration of its sentences. By sentence is here meant graphic sentence, not clause. A graphic sentence is a string of text delimited by major delimiters. As noted above, it can sometimes also be delimited in other ways, e.g. by a carriage return as is often the case in headlines.

Example:

```
<s id=ab07c-003>
<w>Året<ana><ps>NN<m>NEU SIN DEF NOM<b>år</w>
<w>var<ana><ps>VB<m>PRT AKT<b>vara</w>
<date>
<w>1932<ana><ps>RG<m>NOM<b>1932</w>
</date>
<d>.<ana><ps>MAD<b>.</d>
</s>
```

In general, <s> should be interpreted as s-*unit* rather than as *sentence*, as it covers a wider range of phenomena. In the tagging of SUC, the <s> tag is a bit special, in that everything in an actual text must occur inside <s>. The only exception is in poetry, where lines instead occur inside <l>, cf. 4.3.4. This convention is discussed further in section 4.4.1.

4.2.4 Paragraphs

Tag: <p>

No attributes. End tag optional. Markup is done automatically but often presupposes some preprocessing of the original text. Discussed in TEI P3 PartII, chapter 6.1 Paragraphs.

The <p> tag is also a kind of basic unit in texts, but on the level above <s>, see 4.4.1. All running text in SUC must occur inside a tag<s>, which in turn occurs inside a tag <p> or an equivalent of <p>. Equivalents of <p> among the functionally interpreted units are <head>, <byline>, <list>, and <lg>, which are all described in sections 4.3.1-4.3.4. All these units close each other, in the sense described above, section 4.1.3. This means that all text in the corpus must be classified as belonging to one of the categories <p>, <head>, <byline>, <list>, or <lg> (*line group*, as in poetry, see 4.3.4). In a way - that is not totally arbitrary - <head>, <byline>, <list>, and <lg> can be seen as special cases of paragraphs. The automatic markup will tag all these categories as paragraphs, and it is later the task of an annotator to re-classify material and change the tags accordingly.[7]

[7] All files contain <p> and most files contain <head>. The tag <byline> is most frequent in categories A, B and C. The tags <list> and <lg> are rare: for <lg>see e.g. ca02 or ca03. The tag <list> can be found in aa08, ab02 or ac02, and also occurs in several files in categories HA and HB.

4.3 Markup for functionally interpreted units in text

This section deals with markup that is used for representing information about the function of various parts of the texts. Mostly, these tags can only be entered manually by a skilled annotator. There is always a choice of what phenomena should be tagged as well as a trade-off between the usefulness of a tag and the work needed to enter it. It is also fairly pointless to have tags that are so difficult to apply that annotators cannot add them with enough consistency. We have thus avoided markup that demands far-reaching semantic or pragmatic considerations.

What we have chosen to tag are mainly phenomena that may cause problems for any later analysis of the texts. Running text contains lots of material that is not foreseen by, e.g. syntactic parsers. Parsers will often run into trouble with names, lexicalized phrases, and expressions of date and time, as they have a structure that deviates from ordinary constituents. Headlines often have a peculiar syntax and many other phenomena may cause trouble because of their unpredictable appearances. Words and expressions in foreign languages should not be treated the same way as the surrounding text, etc. In many kinds of research it is also important to know whether something belongs to the ordinary text or is a rendering of direct speech or some other kind of quotation.

The section presents what we have empirically arrived at as a minimal set of markup for the texts of the SUC, given the just mentioned considerations. However, time and resources did not allow us to enter all of it in SUC 2.0. We have refrained from tagging dates and times, as it might be possible to derive and tag those expressions more or less automatically later on. Foreign words and phrases, meta-linguistic use, and various kinds of quotations are sometimes sparsely tagged, as they may be difficult to localise in a quick perusal of the texts. All other markup is (supposed to be) entered in the texts of SUC 2.0. The formats and the principles for application are given for all tags, whether they are used or not at present, and they are all in the DTD (Appendix C).

4.3.1 Headlines and leads

Tag: <head>

End tag is optional but recommended for the sake of clarity for human users. The tag takes no attributes. It must be inserted manually, either at the manual pre-processing or during annotation.
TEI P3 Part II discusses it in 7.2.1 Headings and Trailers.

All types of headlines are just tagged <head>, regardless of their relative level or their placement in the text. If the headline is marked typographically, e.g. by boldface or a larger font, this is not rendered in any way, as the information signalled typographically has already been caught when interpreting the concerned string as a headline. Headlines in upper case letters are, however, kept that way as it is stable to various kinds of format conversion (and to transform a string of upper case letters into lower case without losing motivated capitalisation is far from trivial).

Each headline is seen as equivalent to a paragraph in itself and as consisting of one or more sentences. This holds also for headlines consisting of single words. When two headlines appear in sequence, they are tagged as two separate units. The fact that they are two separate 'paragraphs', as it were, is stronger than the general wish for maximal range of tags, cf. 4.4.2.

Examples:

```
<head>
<s id=ce02a-001>
<w>En<ana><ps>DT<m>UTR SIN IND<b>en</w>
<w>måttstock<ana><ps>NN<m>UTR SIN IND NOM<b>måttstock</w>
<w>för<ana><ps>PP<b>för</w>
<w>andra<ana><ps>JJ<m>POS UTR/NEU PLU IND/DEF NOM<b>annan</w>
<w>tenorer<ana><ps>NN<m>UTR PLU IND NOM<b>tenor</w>
</s>
</head>

<head>
<s id=ab07c-021>
<w>STELT<ana><ps>JJ<m>POS NEU SIN IND NOM<b>stel</w>
</s>
</head>
```

In Swedish newspapers, the *leads*, i.e. the first paragraph(s) of an article, are often set in boldface to show their greater relative 'importance'. They are, however, treated as ordinary paragraphs and just tagged <p>.

4.3.2 Bylines

Tag: <byline>

End tag optional but recommended. No attributes. Bylines are marked up manually. TEI P3 Part II 7.2.2 Openers and Closers.

We use <byline> in a somewhat wider sense than is usual in journalism. We use it not only for the names of authors of, e.g. newspaper articles, but also for other information with similar function and position, such as the names of news agencies etc. Bylines normally occur immediately after a <head>, but if the same kind of material occurs at the very end of an article we tag that <byline> too. The linguistic content of a byline is furthermore classified as sentence(s), <s>, no matter how long or short it is. This is done to ensure a running identification of all linguistic material in the texts. <byline> can of course also contain other tags, in particular <name> (4.3.6).

Examples:

```
<byline>
<s id=ce02a-002>
<name type=place>
<w>LONDON<ana><ps>PM<m>NOM<b>London</w>
</name>
<d>(<ana><ps>PAD<b>(</d>
<name type=inst>
<abbr>
<w>SvD<ana><ps>PM<m>NOM<b>SvD</w>             8
</abbr>
</name>
<d>)<ana><ps>PAD<b>)</d>
</s>
</byline>

<byline>
<s id=ab07c-080>
<name type=person>
<w>HELENE<ana><ps>PM<m>NOM<b>Helene</w>
<w>REHN<ana><ps>PM<m>NOM<b>Rehn</w>
</name>
</s>
</byline>
```

4.3.3 Lists

Tags: <list>, <item>, <label>

No attributes. Optional but recommended end tag for <list>, optional for <item> and <label>. Lists are marked up manually. TEI P3 Part II 6.7 Lists.

Listed objects are tagged <item>. If the objects have some kind of identifier or key, this is tagged <label>. A <list> may thus consist only of instances of <item>, of instances of <label> and <item> (in pairs), but not only of instances of <label>. A consequence of the assignment of <s> as the basic unit in the SUC is that the content of not only <item> (which may sometimes be several sentences) but also <label> must be tagged <s>. Sometimes this may lead to rather awkward results, at other times it seems more motivated, but it is under all circumstances necessary in order to ensure a consistent identification of the linguistic material.

Examples:

```
<list>                                                          (constructed)
<label><s id=ex01-001>a)<item><s id=ex01-002>Mjölk /.../
</list>


<list>                                                          (constructed)
<label><s id=ex02-001>Punkt ett:
```

[8] SvD is the abbreviated name of a Swedish daily newspaper.

```
<item><s id=ex02-002>Köp mjölk och ställ den i kylskåpet. /.../
</list>


<list>
<item>
<s id=ab02b-020>
<w n=1823>Grundbidragsmodellen<ana><ps>NN<m>UTR SIN DEF NOM
<b>grundbidragsmodell</w>
</s>
</item>
<item>
<s id=ab02b-021>
<w n=1824>Barnbidragsmodellen<ana><ps>NN<m>UTR SIN DEF NOM
<b>barnbidragsmodell</w>
</s>
</item>
<item>
<s id=ab02b-022>
<w n=1825>Bostadsbidragsmodellen<ana><ps>NN<m>UTR SIN DEF NOM
<b>bostadsbidragsmodell</w>
</s>
</item>
</list>

<list>
<label>
<s id=ha18a-007>
<num>[9]
<w n=55>1.<ana><ps>RG<m>NOM<b>1.</w>
</num>
</s>
</label>
<item>
<s id=ha18a-008>
<w n=56>utgifter<ana><ps>NN<m>UTR PLU IND NOM<b>utgift</w>
<w n=57>för<ana><ps>PP<b>för</w>
<w n=58>rehabiliteringsåtgärder<ana><ps>NN<m>UTR PLU IND NOM
<b>rehabiliteringsåtgärd</w>
<w n=59>för<ana><ps>PP<b>för</w>
<w n=60>anställda<ana><ps>NN<m>UTR PLU IND NOM<b>anställd</w>
<w n=61>med<ana><ps>PP<b>med</w>
<w n=62>långvarigt<ana><ps>AB<m>POS<b>långvarigt</w>
<w n=63>nedsatt<ana><ps>PC<m>PRF UTR SIN IND NOM<b>nedsatt</w>
<w n=64>hälsa<ana><ps>NN<m>UTR SIN IND NOM<b>hälsa</w>
<d n=65>,<ana><ps>MID<b>,</d>
</s>
</item>
<label>
<s id=ha18a-009>
<num>
<w n=66>2.<ana><ps>RG<m>NOM<b>2.</w>
</num>
</s>
</label>
<item>
<s id=ha18a-010>
<w n=67>utgifter<ana><ps>NN<m>UTR PLU IND NOM<b>utgift</w>
<w n=68>för<ana><ps>PP<b>för</w>
<w n=69>åtgärder<ana><ps>NN<m>UTR PLU IND NOM<b>åtgärd</w>
<w n=70>för<ana><ps>PP<b>för</w>
<w n=71>att<ana><ps>IE<b>att</w>
<w n=72>nedbringa<ana><ps>VB<m>INF AKT<b>nedbringa</w>
<w n=73>anställdas<ana><ps>PC<m>PRF UTR/NEU PLU IND/DEF GEN<b>anställd</w>
<w n=74>sjukfrånvaro<ana><ps>NN<m>UTR SIN IND NOM<b>sjukfrånvaro</w>
<w n=75>samt<ana><ps>KN<b>samt</w>
</s>
</item>
/.../                                                      (a few items are left out)
</list>
```

[9] The structural tag *num*, for numeral, automatically applies to everything with p-o-s-tag RG or RO.

Lists are tagged only when they are typographically marked in some way, e.g. by indentation or by special characters signalling the items. This means that a simple enumeration in running text should not be tagged <list>. The sentence below reads 'Joan Sutherland, Itzhak Perlman and Zubin Mehta have performed at Rudas' arenas', and the Sutherland-Perlman-Mehta sequence is not regarded as a list.

Example of non-list:

```
<s id=ce02a-035>
<name type=person>
<w>Joan<ana><ps>PM<m>NOM<b>Joan</w>
<w>Sutherland<ana><ps>PM<m>NOM<b>Sutherland</w>
</name>
<d>,<ana><ps>MID<b>,</d>
<name type=person>
<w>Itzhak<ana><ps>PM<m>NOM<b>Itzhak</w>
<w>Perlman<ana><ps>PM<m>NOM<b>Perlman</w>
</name>
<w>och<ana><ps>KN<b>och</w>
<name type=person>
<w>Zubin<ana><ps>PM<m>NOM<b>Zubin</w>
<w>Mehta<ana><ps>PM<m>NOM<b>Mehta</w>
</name>
<w>har<ana><ps>VB<m>PRS AKT<b>ha</w>
<w>uppträtt<ana><ps>VB<m>SUP AKT<b>uppträda</w>
<w>på<ana><ps>PP<b>på</w>
<name type=person>
<w>Rudas<ana><ps>PM<m>GEN<b>Rudas</w>
</name>
<w>arenor<ana><ps>NN<m>UTR PLU IND NOM<b>arena</w>
<d>.<ana><ps>MAD<b>.</d>
</s>
```

4.3.4 Poetry

Tags: <lg>, <l>
Attribute of <lg>: type
Attribute of <l>: id

Possible values of id: a unique identification

End tag optional but recommended for <lg>, optional for <l>. Markup is entered manually. TEI P3 Part II 6.11 Passages of Verse or Drama.

Poems and other text types where line breaks carry special significance are tagged <lg>, *line group*, and each line in a line group is tagged <l>, *line*. If a line of poetry appears in running text without breaking the coherence of the text, it is, however, not tagged <l> but <quote>, as other quotations (cf. 4.3.14). The headline of a poem is tagged <head> and is regarded as belonging outside the <lg>.

<lg> is a unit on paragraph level, but its constituent parts are not tagged <s> but <l>. In poetry, sentence boundaries and significant line breaks often do not coincide. To avoid disallowed crossover (see 4.4.3) we have chosen to let <l> take priority and do not use <s> at all inside <lg>. To ensure the running identification, the enumeration in the *id* attribute in <l> follows the same series as in <s>. However, <lg> appears very rarely, if at all, in the material and should cause no problems or discrepancies.

Example:

```
<lg type="suc-line-group">[10]
<l id=ja03-004>
<w n=9>Ej<ana><ps>AB<b>ej</w>
<w n=10>fältherren<ana><ps>NN<m>UTR SIN DEF NOM<b>fältherre</w>
```

[10] The value “suc-line-group” is a default value for the required attribute *type* of lg.

```
<w n=11>blott<ana><ps>AB<m>POS<b>blott</w>
<w n=12>vinner<ana><ps>VB<m>PRS AKT<b>vinna</w>
<w n=13>slaget<ana><ps>NN<m>NEU SIN DEF NOM<b>slag</w>
<d n=14>,<ana><ps>MID<b>,</d>
</l>
<l id=ja03-005>
<w n=15>hans<ana><ps>PS<m>UTR/NEU SIN/PLU DEF<b>hans</w>
<w n=16>soldater<ana><ps>NN<m>UTR PLU IND NOM<b>soldat</w>
<w n=17>vinna<ana><ps>VB<m>PRS AKT<b>vinna</w>
<w n=18>det<ana><ps>PN<m>NEU SIN DEF SUB/OBJ<b>det</w>
<w n=19>ock<ana><ps>AB<b>ock</w>
<d n=20>,<ana><ps>MID<b>,</d>
</l>
<l id=ja03-006>
<w n=21>fast<ana><ps>SN<b>fast</w>
<w n=22>han<ana><ps>PN<m>UTR SIN DEF SUB<b>han</w>
<w n=23>är<ana><ps>VB<m>PRS AKT<b>vara</w>
<w n=24>den<ana><ps>PN<m>UTR SIN DEF SUB/OBJ<b>den</w>
<w n=25>ende<ana><ps>JJ<m>POS MAS SIN DEF NOM<b>ende</w>
<w n=26>i<ana><ps>PP<b>i</w>
<w n=27>laget<ana><ps>NN<m>NEU SIN DEF NOM<b>lag</w>
<d n=28>,<ana><ps>MID<b>,</d>
</l>
<l id=ja03-007>
<w n=29>som<ana><ps>HP<m>- - -<b>som</w>
<w n=30>får<ana><ps>VB<m>PRS AKT<b>få</w>
<d n=31>'<ana><ps>PAD<b>'</d>
<foreign lang=fr> [11]
<w n=32>Pour<ana><ps>UO<b>pour</w>
<w n=33>le<ana><ps>UO<b>le</w>
<w n=34>mérite<ana><ps>UO<b>mérite</w>
</foreign>
<d n=35>'<ana><ps>PAD<b>'</d>
<w n=36>på<ana><ps>PP<b>på</w>
<w n=37>sin<ana><ps>PS<m>UTR SIN DEF<b>sin</w>
<w n=38>rock<ana><ps>NN<m>UTR SIN IND NOM<b>rock</w>
<d n=39>.<ana><ps>MAD<b>.</d>
</l>
</lg>
```

Example of non-lg:

```
I och med samlingen "Fågelns öga" öppnas på allvar en annan möjlighet, ty jaget
känner "inte längre / hatets väg / inte heller begärets och sorgens".
```

Here the lines of poetry quoted to illustrate the review are clearly part of the running text, and should not be marked <lg> but possibly <quote>. However, <quote> is rarely used at all in SUC 2.0 - and the extent of this quotation is obvious anyway:

```
<s id=ca03a-026>
<w n=541>I<ana><ps>PP<b>i</w>
<w n=542>och<ana><ps>KN<b>och</w>
<w n=543>med<ana><ps>PP<b>med</w>
<w n=544>samlingen<ana><ps>NN<m>UTR SIN DEF NOM<b>samling</w>
<d n=545>"<ana><ps>PAD<b>"</d>
<name type=work>
<w n=546>Fågelns<ana><ps>NN<m>UTR SIN DEF GEN<b>fågel</w>
<w n=547>öga<ana><ps>NN<m>NEU SIN IND NOM<b>öga</w>
</name>
<d n=548>"<ana><ps>PAD<b>"</d>
<w n=549>öppnas<ana><ps>VB<m>PRS SFO<b>öppna</w>
<w n=550>på<ana><ps>PP<b>på</w>
<w n=551>allvar<ana><ps>NN<m>NEU SIN IND NOM<b>allvar</w>
<w n=552>en<ana><ps>DT<m>UTR SIN IND<b>en</w>
<w n=553>annan<ana><ps>JJ<m>POS UTR SIN IND NOM<b>annan</w>
<w n=554>möjlighet<ana><ps>NN<m>UTR SIN IND NOM<b>möjlighet</w>
<d n=555>,<ana><ps>MID<b>,</d>
<w n=556>ty<ana><ps>KN<b>ty</w>
<w n=557>jaget<ana><ps>NN<m>NEU SIN DEF NOM<b>jag</w>
<w n=558>känner<ana><ps>VB<m>PRS AKT<b>känna</w>
<d n=559>"<ana><ps>PAD<b>"</d>
<w n=560>inte<ana><ps>AB<b>inte</w>
```

[11] The value fr=French. Cf. 4.3.10 for allowed values of *lang*.

```
<w n=561>längre<ana><ps>AB<m>KOM<b>länge</w>
<d n=562>/</ana><ps>MID<b>/</d>
<w n=563>hatets<ana><ps>NN<m>NEU SIN DEF GEN<b>hat</w>
<w n=564>väg<ana><ps>NN<m>UTR SIN IND NOM<b>väg</w>
<d n=565>/</ana><ps>MID<b>/</d>
<w n=566>inte<ana><ps>AB<b>inte</w>
<w n=567>heller<ana><ps>AB<b>heller</w>
<w n=568>begärets<ana><ps>NN<m>NEU SIN DEF GEN<b>begär</w>
<w n=569>och<ana><ps>KN<b>och</w>
<w n=570>sorgens<ana><ps>NN<m>UTR SIN DEF GEN<b>sorg</w>
<d n=571>"<ana><ps>PAD<b>"</d>
<d n=572>.<ana><ps>MAD<b>.</d>
</s>
```

4.3.5 Word groups

Tag: <wg>

No attributes. End tag is obligatory. Not used in SUC 2.0. A defined set of word groups may be marked up automatically, but this has not been done in SUC 2.0. We have not found anything about word groups in TEI P3.

A phenomenon that is notoriously tricky in all kinds of linguistic analysis is lexicalized expressions such as *in spite of*, with a meaning and function more like that of a single word. We call them *word groups* and want to identify them and mark them at an early stage of the analysis process, in order to avoid problems with them without losing any information. To decide when a certain string of words is to count as a word group is a difficult task in itself, but here we need not go into that. In general, what we call word groups are lexicalized expressions that consist of several words but function like a single word, mostly like a function word. It may just be worth noting that particle verbs, which are abundant in Swedish, are not regarded as word groups, as they are often discontinuous expressions.

The markup of word groups is structured so as to be parallel to that of single words. Inside <wg> are listed the component words with their full tags. This corresponds to the text word in single word tags. After that comes <ana> containing <ps> and possibly <m>, which here gives the part-of-speech and morpho-syntactic properties of the entire word group as used in the current context, followed by a 'base form' which is the lexicalized expression itself. The use of a base form also for a word group makes it possible to lemmatise lexicalized expressions. It is also in principle possible to lemmatise expressions containing variables, such as *pull someone's leg*. This variable facility is however not used in the SUC. More about word groups in SUC texts may be found in Lindberg (1999). But the word group tag has not been used in SUC 2.0. ‘Next time?…’

Example:

```
<wg>
<w>i<ana><ps>PP<b>i</w>
<w>stället<ana><ps>NN<m>NEU SIN DEF NOM<b>ställe</w>
<w>för<ana><ps>PP<b>för</w>
<ana><ps>PP<b>i stället för</wg>
('instead of', literally 'in the place of' and used as a preposition)
```

Both <w> and <wg> were invented for the needs of the SUC project, but they show similarities to the tags <l> and <lg>, line and line group respectively, recommended by the TEI and also used in SUC (cf. section 4.3.4). [12]

4.3.6 Names

Tag: <name>

[12] In a more bnc-like fashion (BNC reference manual 9.10) the example could be tagged:

```
<w>i_stället_för<ana><ps>PP<b>i_stället_för</w>
```

This requires a decision about the list of word groups already at the tokenization stage in the process, whereas a <wg> tag may be introduced at will by a corpus user around her/his favourite word groups.

(In BNC the s-units are numbered for reference but not the words, so it is probably not too inconvenient to make last minute changes in the tokenization.)

Attribute: type
Possible values of type: person|animal|myth|place|inst|product|work|event|other

End tag is obligatory. The tag is sometimes entered and always checked manually in the SUC, but it can be (and was) derived automatically (with a certain risk of errors) for names that are built up entirely by words with part-of-speech PM (proper name). TEI P3 Part II 6.4.1Referring strings.

This tag is used for words and phrases that in the text function as proper names in a wide sense. The attribute *type* can further specify them and each of its possible values is exemplified below. The value *inst* stands for 'institution', also in a wide sense, and *work* is used for all kinds of work of art. The other values should be self-explanatory. If no other value is applicable, *other* can be used. It is also possible to avoid specifying *type* in cases where it is not possible to decide.

The tag <name> plays an extra important role in all situations where an ordinary noun phrase is being used as a proper name. The words in the noun phrase can then keep their markup, giving base forms and morpho-syntactic information, at the same time as the <name> tag makes clear that the phrase functions as a name. Sometimes a <name>, in particular of *type=work*, can even be a full sentence. The SUC DTD does not allow recursive *s-units* (see 4.4.3), so they are not marked as such, but their constituent words keep their full markup.

Examples:

```
<name type=person>
<w>Joan<ana><ps>PM<m>NOM<b>Joan</w>
<w>Sutherland<ana><ps>PM<m>NOM<b>Sutherland</w>
</name>

<name type=place>
<w>Hyde<ana><ps>PM<m>NOM<b>Hyde</w>
<w>Park<ana><ps>PM<m>NOM<b>Park</w>
</name>

<name type=inst>
<w>Philharmonia<ana><ps>PM<m>NOM<b>Philharmonia</w>
<w>Orchestra<ana><ps>PM<m>NOM<b>Orchestra</w>
</name>

<name type=product>
<w>Steinway<ana><ps>PM<m>NOM<b>Steinway</w>
</name>

<name type=work>                             (the title of a book)
<w>Bilder<ana><ps>NN<m>UTR PLU IND NOM<b>bild</w>
<w>från<ana><ps>PP<b>från</w>
<w>Öreryd<ana><ps>PM<m>NOM<b>Öreryd</w>
</name>

<name type=event>                            (constructed; 'Second World War')
<w>Andra<ana><ps>RO<m>NOM<b>andra</w>
<w>Världskriget<ana><ps>NN<m>NEU SIN DEF NOM<b>världskrig</w>
</name>

<name type=person>                           (constructed; 'Charles the great')
<w>Karl<ana><ps>PM<m>NOM<b>Karl</w>
<w>den<ana><ps>DT<m>UTR SIN DEF<b>den</w>
<w>store<ana><ps>JJ<m>POS MAS SIN DEF NOM<b>stor</w>
</name>

<s id=ce03a-011>                             (‘Gustav the third’, a Swedish king)
<name type=person>
<w n=146>GUSTAV<ana><ps>PM<m>NOM<b>Gustav</w>
<num>
<w n=147>III<ana><ps>RO<m>NOM<b>3:e</w>
</num>
</name>
```

The names of the SUC have been studied separately in Wennstedt (1995), where the *type* subcategorization is presented and discussed. Some rules-of-thumb for their application are given below.

Epithets, numbers of kings, etc. belong to the name. We have also decided to regard (job) titles as parts of proper names. This applies only to titles in connection with names, not when used as common nouns, thus: *<name>doktor More</name>*, while *doktorn* '(the) doctor' is treated as a common noun also when used as a vocative. *<name type=person>prinsessan Sofia</name>* is regarded as a name in its entirety, but not hertigen av *<name type=place>Urbino</name>* 'the count of U.' Note that this only concerns job titles; titles of books, songs, paintings, etc. are tagged *<name type=work>*.

Examples:

```
<name type=person>
<w n=471>prinsessan<ana><ps>NN<m>UTR SIN DEF NOM<b>prinsessa</w>
<w n=472>Elena<ana><ps>PM<m>NOM<b>Elena</w>
</name>

<name type=person>
<w n=494>drottning<ana><ps>NN<m>UTR SIN IND NOM<b>drottning</w>
<w n=495>Sofia<ana><ps>PM<m>NOM<b>Sofia</w>
</name>

<s id=aa06a-042>
<w n=451>Kungen<ana><ps>NN<m>UTR SIN DEF NOM<b>kung</w>
<w n=452>av<ana><ps>PP<b>av</w>
<name type=place>
<w n=453>Spanien<ana><ps>PM<m>NOM<b>Spanien</w>
</name>
```

Names of groups of people, as sports teams and rock groups, are subcategorised as *type=inst*. For names of products, there may sometimes be a scale from a company (*Coca Cola Inc., type=inst*) via a product (*a Coca Cola, type=product*) to something that is more like an ordinary noun (*I got some Coca Cola on my jeans*). We have decided to tag cases like the latter as *<name type=product>*.

It is not a prerequisite that names are capitalised. It is not necessarily so that everything that is capitalised is a name, either. In doubtful cases, we have however used capitalisation as a last resort. When it is impossible to decide if something is a name or not, we let capitalisation decide. In particular with products and institutions, the decision can be difficult. (E.g. a word like *Utrikesdepartementet* 'the Foreign Ministry' can appear both with and without capitalisation in Swedish). We then let the typography decide, in the hope that it reflects the way the writer regards the concerned word. This may lead to inconsistency in the markup of words in the corpus, but it is definitely so that the same referent can sometimes be treated as something carrying a name, sometimes as a common noun, and the markup then reflects the various points of view.

Examples:

```
<name type=inst>
<w n=1932>Försvarets<ana><ps>NN<m>NEU SIN DEF GEN<b>försvar</w>
<w n=1933>forskningsanstalt<ana><ps>NN<m>UTR SIN IND NOM<b>forskningsanstalt</w>
</name>
<w n=1934>i<ana><ps>PP<b>i</w>
<name type=place>
<w n=1935>Umeå<ana><ps>PM<m>NOM<b>Umeå</w>
</name>

<name type=myth>
<w n=452>Atlantis<ana><ps>PM<m>NOM<b>Atlantis</w>
</name>
<d n=453>,<ana><ps>MID<b>,</d>
<w n=454>den<ana><ps>DT<m>UTR SIN DEF<b>den</w>
<w n=455>sjunkna<ana><ps>PC<m>PRF UTR/NEU SIN DEF NOM<b>sjunken</w>
<w n=456>kontinenten<ana><ps>NN<m>UTR SIN DEF NOM<b>kontinent</w>


<name type=myth>
<w n=1457>Oden<ana><ps>PM<m>NOM<b>Oden</w>
</name>

<w n=1491>den<ana><ps>DT<m>UTR SIN DEF<b>den</w>
<w n=1492>nya<ana><ps>JJ<m>POS UTR/NEU SIN DEF NOM<b>ny</w>
<w n=1493>svenska<ana><ps>JJ<m>POS UTR/NEU SIN DEF NOM<b>svensk</w>
<w n=1494>helaftonsoperan<ana><ps>NN<m>UTR SIN DEF NOM<b>helaftonsopera</w>
<name type=work>
<w n=1495>Balder<ana><ps>PM<m>NOM<b>Balder</w>
</name>
```

```
<name type=person>
<w n=1612>Sääf<ana><ps>PM<m>NOM<b>Sääf</w>
</name>
<w n=1613>låter<ana><ps>VB<m>PRS AKT<b>låta</w>
<w n=1614>skurken<ana><ps>NN<m>UTR SIN DEF NOM<b>skurk</w>
<name type=myth>
<w n=1615>Loke<ana><ps>PM<m>NOM<b>Loke</w>
</name>
<w n=1616>vara<ana><ps>VB<m>INF AKT<b>vara</w>
<w n=1617>förälskad<ana><ps>PC<m>PRF UTR SIN IND NOM<b>förälskad</w>
<w n=1618>i<ana><ps>PP<b>i</w>
<name type=myth>
<w n=1619>Balders<ana><ps>PM<m>GEN<b>Balder</w>
</name>
<w n=1620>flickvän<ana><ps>NN<m>UTR SIN IND NOM<b>flickvän</w>
<name type=myth>
<w n=1621>Nanna<ana><ps>PM<m>NOM<b>Nanna</w>
</name>
<d n=1622>.<ana><ps>MAD<b>.</d>
```

kk01:

```
<w n=2213>ätit<ana><ps>VB<m>SUP AKT<b>äta</w>
<w n=2214>av<ana><ps>PP<b>av</w>
<name type=myth>
<w n=2215>Kunskapens<ana><ps>NN<m>UTR SIN DEF GEN<b>kunskap</w>
<w n=2216>träd<ana><ps>NN<m>NEU SIN IND NOM<b>träd</w>
</name>

<name type=animal>                            (the name of a goose)
<w n=563>Akka<ana><ps>PM<m>NOM<b>Akka</w>
</name>

<name type=myth>
<w n=1154>Adam<ana><ps>PM<m>NOM<b>Adam</w>
</name>
<name type=myth>
<w n=1156>Eva<ana><ps>PM<m>NOM<b>Eva</w>
</name>

<name type=animal>                            (the name of a horse)
<w n=1922>J.<ana><ps>PM<m>NOM<b>J</w>
<w n=1923>R.<ana><ps>PM<m>NOM<b>R</w>
<w n=1924>Broline<ana><ps>PM<m>NOM<b>Broline</w>
</name>
```

4.3.7 Dates

Tag: <date>

End tag obligatory, no attributes. Not used in SUC 2.0, but can probably be entered manually with the aid of some search algorithm. TEI P3 Part II 6.4.4 Dates and Times.

To be marked as <*date*>, an expression must contain the name of a day or a month, or a numeral (year or date) expressed by digits or words, or a combination of these. If so, the entire expression is tagged. The name of a season in combination with a year is also tagged <*date*>, but not the season alone. In from-to expressions, *1990-1993* is tagged as one instance of <*date*>, while *1990 till 1993* is tagged as two instances separated by a preposition. We use the same tag for date ranges (like the last two examples) as for single dates.

Examples:

```
<date>                                        (constructed)
<w>den<ana><ps>DT<m>UTR SIN DEF<b>den</w>
<w>16<ana><ps>RG<m>NOM<b>16</w>
<w>augusti<ana><ps>NN<m>UTR SIN IND NOM<b>augusti</w>
</date>

<date>                                        (constructed)
<w>1676-1749<ana><ps>RG<m>NOM<b>1676-1749</w>
</date>
```

```
<date>                                        (constructed)
<w>mars<ana><ps>NN<m>UTR SIN IND NOM<b>mars</w>
</date>

<date>                                        (constructed)
<w>måndag<ana><ps>NN<m>UTR PLU IND NOM<b>måndag</w>
</date>
```

Words like *trettiotalet*, 'the thirties' *1800-talet*, 'the 1800s' are regarded as ordinary nouns, not as date expressions. Words that are normally classified as date expressions can sometimes have a purely nominal use and should not be tagged <*date*> in those instances. ('December was cold that year.' 'It was a rainy Monday.') Such use is not uncommon, and a reason why <*date*> cannot be generated automatically by means of regular expressions.

4.3.8 Time

Tag: <time>

End tag obligatory, no attributes. Not used in SUC 2.0, but to be entered in the same way as <*date*>. TEI P3 Part II 6.4.4 Dates and Times.

Numerals expressed in digits or words are tagged. For range expressions, the same conventions as for <*date*> are used.

Examples:

```
<time>                                        (constructed)
<w>klockan<ana><ps>NN<m>UTR SIN DEF NOM<b>klocka</w>
<w>21<ana><ps>RG<m>NOM<b>21</w>
</time>

<time>                                        (constructed)
<w>klockan<ana><ps>NN<m>UTR SIN DEF NOM<b>klocka</w>
<w>nio<ana><ps>RG<m>NOM<b>nio</w>
</time>
```

4.3.9 Bibliographic references

Tag: <bibl>, <ref>

End tag obligatory, no attributes. Entered manually. TEI P3 Part II 6.10.1 Elements of Bibliographic References

The tag <*bibl*> is meant for references referring to other works, 'Chomsky 1957', and <*ref*> for references referring to other parts of the work itself, e.g. 'see section 4.3.10'. The tag covers the entire reference. Further sub-tags for <*name*> and <*date*> are not necessary and have generally not been used in SUC 2.0. [13] Only proper references should be tagged. When a work is talked about rather than referred to, <*name type=work*> (sect. 4.3.6) should be used instead, and also whenever there is hesitation between the two.

[13] The tag <*ref*> is presently used for both types of reference in SUC 2.0 and now serves only to indicate a peculiar part of the text. In one of the examples (ja08) there is also a <*name*> tag inside <*ref*>. This is quite frequent because the annotators where told not to remove such <*name*> tags if already present.

Examples:
ja01:

```
        <ref>
        <w n=50>Anward<ana><ps>PM<m>NOM<b>Anward</w>
        <num>
        <w n=51>1986<ana><ps>RG<m>NOM<b>1986</w>
        </num>
        </ref>
```

ja08:

```
        <ref>
        <name type=person>
        <w n=284>Rietz<ana><ps>PM<m>NOM<b>Rietz</w>
        </name>
        <num>
        <w n=285>88a<ana><ps>RG<m>NOM<b>88a</w>
        </num>
        </ref>
```

ja09:

```
        <ref>
        <abbr>
        <w n=751>SAOB<ana><ps>PM<m>NOM<b>SAOB</w>
        </abbr>
        <w n=752>G<ana><ps>NN<m>NEU SIN IND NOM<b>g</w>
        <num>
        <w n=753>1239<ana><ps>RG<m>NOM<b>1239</w>
        </num>
        </ref>
```

fk04:

```
        <ref>
        <w n=506>P.<ana><ps>PM<m>NOM<b>P</w>
        <w n=507>Bourdieu<ana><ps>PM<m>NOM<b>Bourdieu</w>
        <d n=508>,<ana><ps>MID<b>,</d>
        <d n=509>"<ana><ps>PAD<b>"</d>
        <w n=510>La<ana><ps>PM<m>NOM<b>La</w>
        <w n=511>reproduction<ana><ps>UO<b>reproduction</w>
        <d n=512>"<ana><ps>PAD<b>"</d>
        <d n=513>,<ana><ps>MID<b>,</d>
        <abbr>
        <w n=514>s.<ana><ps>NN<m>AN<b>s</w>
        </abbr>
        <num>
        <w n=515>40<ana><ps>RG<m>NOM<b>40</w>
        </num>
        </ref>
```

ha01:

```
        <ref>
        <abbr>
        <w n=11>Dir.<ana><ps>NN<m>AN<b>dir</w>
        </abbr>
        <num>
        <w n=12>1990:4<ana><ps>RG<m>NOM<b>1990:4</w>
        </num>
        </ref>
```

4.3.10 Foreign words and phrases

Tag: <foreign>
Attribute: lang
Possible values of lang: an ISO conformant two-letter abbreviation, or 'other'

End tag is obligatory. The tag is entered manually. TEI P3 Part II 6.3.2.1 Foreign Words or Expressions.

Foreign material in the running text is tagged. Loan words that have been integrated and have Swedish inflection, spelling, etc. are not tagged. Foreign names are not tagged <foreign>, just <name>. Material that is within the scope of a <foreign> tag will get no further analysis.

As a list of values for <lang>, the list given in ISO 639 is used. So far, we have used the following abbreviations from this list in the SUC:

en=English
fr=French
de=German
es=Spanish
it=Italian

no=Norwegian
da=Danish
fi=Finnish
is=Icelandic
ru=Russian

la=Latin
el=Greek
cs=Czech

The list is expanded when needed. For singular instances of rare languages or in cases of a foreign word where the language is not known, *lang* will get the value *other*.

Examples:

aa06:

```
        <foreign lang=es>
        <w n=478>corridas<ana><ps>UO<b>corridas</w>
        </foreign>
```

aa11:

```
        <foreign lang=en>
        <w n=1974>red<ana><ps>UO<b>red</w>
        <w n=1975>nose<ana><ps>UO<b>nose</w>
        <w n=1976>disease<ana><ps>UO<b>disease</w>
        </foreign>
```

ae04:

```
        <foreign lang=other>
        <w n=1112>ippon<ana><ps>UO<b>ippon</w>
        </foreign>
```

ea01:

```
        <foreign lang=other>
        <w n=608>khanjar<ana><ps>UO<b>khanjar</w>
        </foreign>
```

eb01:

```
        <foreign lang=fr>
        <w n=1777>morpho-syntaxique<ana><ps>UO<b>morpho-syntaxique</w>
        </foreign>
```

ja14:

```
        <foreign lang=cs>
        <w n=1387>kejvat<ana><ps>UO<b>kejvat</w>
        </foreign>
```

ec07:

```
        <foreign lang=ru>
        <w n=973>lisjnij<ana><ps>UO<b>lisjnij</w>
        <w n=974>tjelovek<ana><ps>UO<b>tjelovek</w>
        </foreign>
```

4.3.11 Meta-linguistic use

Tag: <mentioned>

End tag obligatory, no attributes. The tag is entered manually. TEI P3 Part II 6.3.4 Terms, Glosses and Cited Words.

When a word is mentioned rather than used it is tagged <mentioned>. A word that is <mentioned> can also be <foreign>, with the <foreign> tag outside <mentioned>.

Examples:

ja17:

```
        <w n=1727>tidsuttryck<ana><ps>NN<m>NEU PLU IND NOM<b>tidsuttryck</w>
```

```
<w n=1728>som<ana><ps>HP<m>- - -<b>som</w>
<mentioned>
<w n=1729>just<ana><ps>AB<b>just</w>
<w n=1730>nu<ana><ps>AB<b>nu</w>
</mentioned>
<d n=1731>,<ana><ps>MID<b>,</d>
<mentioned>
<w n=1732>hela<ana><ps>JJ<m>POS UTR/NEU SIN DEF NOM<b>hel</w>
<w n=1733>tiden<ana><ps>NN<m>UTR SIN DEF NOM<b>tid</w>
</mentioned>
```

ja17:

```
<s id=ja17-084>
<w n=2006>En<ana><ps>DT<m>UTR SIN IND<b>en</w>
<w n=2007>konstmusikbeskrivning<ana><ps>NN<m>UTR SIN IND NOM
<b>konstmusikbeskrivning</w>
<w n=2008>med<ana><ps>PP<b>med</w>
<w n=2009>konventionella<ana><ps>JJ<m>POS UTR/NEU PLU IND/DEF NOM
<b>konventionell</w>
<w n=2010>skönhetssymboler<ana><ps>NN<m>UTR PLU IND NOM<b>skönhetssymbol</w>
<w n=2011>som<ana><ps>HP<m>- - -<b>som</w>
<mentioned>
<w n=2012>Gitarrens<ana><ps>NN<m>UTR SIN DEF GEN<b>gitarr</w>
<w n=2013>tonrankor<ana><ps>NN<m>UTR PLU IND NOM<b>tonranka</w>
<d n=2014>,<ana><ps>MID<b>,</d>
<w n=2015>lekfulla<ana><ps>JJ<m>POS UTR/NEU PLU IND/DEF NOM<b>lekfull</w>
<w n=2016>turneringar<ana><ps>NN<m>UTR PLU IND NOM<b>turnering</w>
<w n=2017>och<ana><ps>KN<b>och</w>
<w n=2018>meditativa<ana><ps>JJ<m>POS UTR/NEU PLU IND/DEF NOM<b>meditativ</w>
<w n=2019>reciterande<ana><ps>NN<m>NEU SIN IND NOM<b>reciterande</w>
<w n=2020>skimrar<ana><ps>VB<m>PRS AKT<b>skimra</w>
<w n=2021>som<ana><ps>KN<b>som</w>
<w n=2022>diskreta<ana><ps>JJ<m>POS UTR/NEU PLU IND/DEF NOM<b>diskret</w>
<w n=2023>diamanter<ana><ps>NN<m>UTR PLU IND NOM<b>diamant</w>
<w n=2024>mot<ana><ps>PP<b>mot</w>
<w n=2025>stråkarnas<ana><ps>NN<m>UTR PLU DEF GEN<b>stråke</w>
<w n=2026>mörka<ana><ps>JJ<m>POS UTR/NEU SIN DEF NOM<b>mörk</w>
<w n=2027>sammet<ana><ps>NN<m>UTR SIN IND NOM<b>sammet</w>
</mentioned>
<w n=2028>kan<ana><ps>VB<m>PRS AKT<b>kunna</w>
<w n=2029>jämföras<ana><ps>VB<m>INF SFO<b>jämföra</w>
<w n=2030>med<ana><ps>PP<b>med</w>
<w n=2031>det<ana><ps>DT<m>NEU SIN DEF<b>den</w>
<w n=2032>enklare<ana><ps>JJ<m>KOM UTR/NEU SIN/PLU IND/DEF NOM<b>enkel</w>
<w n=2033>metaforiska<ana><ps>JJ<m>POS UTR/NEU SIN DEF NOM<b>metaforisk</w>
<w n=2034>jazzuttrycket<ana><ps>NN<m>NEU SIN DEF NOM<b>jazzuttryck</w>
<w n=2035>för<ana><ps>PP<b>för</w>
<w n=2036>musikstycken<ana><ps>NN<m>NEU PLU IND NOM<b>musikstycke</w>
<d n=2037>,<ana><ps>MID<b>,</d>
<mentioned>
<w n=2038>pärlor<ana><ps>NN<m>UTR PLU IND NOM<b>pärla</w>
</mentioned>
```

ja20:

```
<s id=ja20-007>
<w n=105>Begreppet<ana><ps>NN<m>NEU SIN DEF NOM<b>begrepp</w>
<d n=106>'<ana><ps>PAD<b>'</d>
<foreign lang=en>
<mentioned>
<w n=107>argument<ana><ps>UO<b>argument</w>
</mentioned>
</foreign>
<d n=108>'<ana><ps>PAD<b>'</d>
<d n=109>(<ana><ps>PAD<b>(</d>
<d n=110>=<ana><ps>MID<b>=</d>
<d n=111>'<ana><ps>PAD<b>'</d>
<w n=112>argument<ana><ps>NN<m>NEU SIN IND NOM<b>argument</w>
<d n=113>'<ana><ps>PAD<b>'</d>
<d n=114>,<ana><ps>MID<b>,</d>
<d n=115>'<ana><ps>PAD<b>'</d>
<w n=116>diskussion<ana><ps>NN<m>UTR SIN IND NOM<b>diskussion</w>
<d n=117>'<ana><ps>PAD<b>'</d>
<d n=118>,<ana><ps>MID<b>,</d>
<d n=119>'<ana><ps>PAD<b>'</d>
<w n=120>gräl<ana><ps>NN<m>NEU SIN IND NOM<b>gräl</w>
<d n=121>'<ana><ps>PAD<b>'</d>
<d n=122>)<ana><ps>PAD<b>)</d>
<abbr>
<w n=123>t_ex<ana><ps>AB<m>AN<b>till_exempel</w>
</abbr>
<w n=124>struktureras<ana><ps>VB<m>PRS SFO<b>strukturera</w>
<w n=125>enligt<ana><ps>PP<b>enligt</w>
<w n=126>författarna<ana><ps>NN<m>UTR PLU DEF NOM<b>författare</w>
<w n=127>på<ana><ps>PP<b>på</w>
<w n=128>ett<ana><ps>DT<m>NEU SIN IND<b>en</w>
<w n=129>systematiskt<ana><ps>JJ<m>POS NEU SIN IND NOM<b>systematisk</w>
<w n=130>sätt<ana><ps>NN<m>NEU SIN IND NOM<b>sätt</w>
<w n=131>genom<ana><ps>PP<b>genom</w>
<w n=132>metaforen<ana><ps>NN<m>UTR SIN DEF NOM<b>metafor</w>
<d n=133>'<ana><ps>PAD<b>'</d>
<foreign lang=en>
<mentioned>
<w n=134>argument<ana><ps>UO<b>argument</w>
<w n=135>is<ana><ps>UO<b>is</w>
<w n=136>war<ana><ps>UO<b>war</w>
</mentioned>
</foreign>
<d n=137>'<ana><ps>PAD<b>'</d>
<d n=138>.<ana><ps>MAD<b>.</d>
</s>
```

4.3.12 Abbreviations

Tag: <abbr>

End tag is obligatory. No attributes. For standard abbreviations tags are entered automatically at the dictionary look-up, for other abbreviations tags have to be entered manually. TEI P3 Part II 6.4.5 Abbreviations and Their Expansions.

In the TEI guidelines, the tag <abbr> has a possible attribute *expan* where the value is an expansion of the abbreviation. The attribute *expan* is not used in the SUC, as all words and other strings of black characters have their tags, which include a base form. If the need to expand an abbreviation is felt, this can be done in the base form field of the tag, but normally it should be avoided. The processing in connection with dictionary look-up gives standard abbreviations a standardised base form, which is sometimes an expansion of the abbreviation, sometimes not. This base form can be used for 'lemmatising' abbreviations. Abbreviated names of various kinds of institutions etc. do not have to be expanded. Initials in people's names are not regarded as abbreviations at all.

Examples:

ja14:

```
        <abbr>
        <w n=1122>osv.<ana><ps>AB<m>AN<b>och_så_vidare</w>         ('and so on')
        </abbr>
```

gb07:

```
        <w n=1694>ordbehandlare<ana><ps>NN<m>UTR PLU IND NOM<b>ordbehandlare</w>
        <w n=1695>och<ana><ps>KN<b>och</w>
        <w n=1696>bankomater<ana><ps>NN<m>UTR PLU IND NOM<b>bankomat</w>
        <d n=1697>,<ana><ps>MID<b>,</d>
        <abbr>
        <w n=1698>osv.<ana><ps>AB<m>AN<b>och_så_vidare</w>
        </abbr>
        <d n=1699>,<ana><ps>MID<b>,</d>
        <abbr>
        <w n=1700>osv<ana><ps>AB<m>AN<b>och_så_vidare</w>
        </abbr>
        <d n=1701>.<ana><ps>MAD<b>.</d>
        </s>
        </p>
```

aa12:

```
        <abbr>
        <w n=799>Tel<ana><ps>NN<m>AN<b>tel</w>                     ('telephone')
        </abbr>
```

aa03:

```
        <name type=inst>
        <abbr>
        <w n=2055>SvD<ana><ps>PM<m>NOM<b>SvD</w>
        </abbr>
        </name>
```

4.3.13 Direct speech/writing

Tag: <q> (and <in-q>)
Attribute: lang
Possible values of lang: see 4.3.10, Foreign words and phrases

End tag obligatory. Generally not used in SUC 2.0. The tag is entered manually, preferably already at pre-processing. TEI P3 Part II 6.3.3 Quotation.

<q> is used to tag what is said or written by persons in the text, i.e. for dialogues etc. Otherwise, <quote> is used. In a sentence like 'Hamlet's famous "to be or not to be" is still...' it is <quote> rather than <q> that should be used. It is, however, not necessary that the speech or writing be typographically signalled, e.g. by dash or quotation marks. Quotation marks are not removed, but the tag is entered outside the mark. When <q> covers a whole paragraph, <p> is entered outside <q>. When it covers more than one paragraph, each new paragraph gets a new instance of <p><q> at its beginning and </q></p> at its end.

Example:

```
        <s id=ja08-008>
        <name type=person>
        <w n=98>Linné<ana><ps>PM<m>NOM<b>Linné</w>
        </name>
        <w n=99>skriver<ana><ps>VB<m>PRS AKT<b>skriva</w>
        <w n=100>i<ana><ps>PP<b>i</w>
        <name type=work>
        <w n=101>Flora<ana><ps>PM<m>NOM<b>Flora</w>
        <w n=102>Svecica<ana><ps>PM<m>NOM<b>Svecica</w>
        </name>
        <d n=103>:<ana><ps>MAD<b>:</d>
        </s>
        <q>
        <s id=ja08-009>
        <d n=104>"<ana><ps>PAD<b>"</d>
        <w n=105>Till<ana><ps>PP<b>till</w>
        <w n=106>följd<ana><ps>NN<m>UTR SIN IND NOM<b>följd</w>
```

```
<w n=107>av<ana><ps>PP<b>av</w>
<w n=108>sina<ana><ps>PS<m>UTR/NEU PLU DEF<b>sin</w>
<w n=109>ständigt<ana><ps>AB<m>POS<b>ständigt</w>
<w n=110>dallrande<ana><ps>PC<m>PRS UTR/NEU SIN/PLU IND/DEF NOM<b>dallrande</w>
<w n=111>vippor<ana><ps>NN<m>UTR PLU IND NOM<b>vippa</w>
<w n=112>är<ana><ps>VB<m>PRS AKT<b>vara</w>
<w n=113>detta<ana><ps>DT<m>NEU SIN DEF<b>denna</w>
<w n=114>gräs<ana><ps>NN<m>NEU SIN IND NOM<b>gräs</w>
<w n=115>det<ana><ps>PN<m>NEU SIN DEF SUB/OBJ<b>det</w>
<w n=116>av<ana><ps>PP<b>av</w>
<w n=117>lantmännen<ana><ps>NN<m>UTR PLU DEF NOM<b>lantman</w>
<w n=118>mest<ana><ps>AB<m>SUV<b>mest</w>
<w n=119>kända<ana><ps>PC<m>PRF UTR/NEU PLU IND/DEF NOM<b>känd</w>
<d n=120>"<ana><ps>PAD<b>"</d>
</s>
</q>
```

When a quotation is in a foreign language, the attribute *lang* is used with the language concerned as its value, cf. the conventions described in 4.3.10.

For sequences telling who said what, we have invented the rather unorthodox tag <in-q> to be used in all cases when it belongs to the same graphic sentence as (part of) the direct speech/writing. [14] We have chosen to do like this because we want to avoid discontinuous sentences and also to avoid having several parallel systems of markup. Our predominant interest is in linguistic structure and, from a syntactic point of view, both the utterance and the information about the speaker can be said to form part of the same sentence. Quotations have a corresponding tag <in-quote>.

Examples: (constructed)

```
<q><s id=ex03-001>Hej, <in-q>sa hon.</in-q></q>            ('Hello, she said.')

<q><s id=ex03-002>Så trevligt, <in-q>sa han,</in-q> att du kunde komma. </q>
'How nice, he said, that you could come.'
```

But:

```
<s id=ex03-003>Han reste sig och sa:                        (constructed)
<q><s id=ex03-004>Är du hungrig? <s id=ex03-005>Vill du ha middag?</q>
('He stood up and said: Are you hungry? Would you like to have dinner?')
```

In the last example, the speaker information is terminated by colon followed by capital letter, and such a colon should be classified as a major delimiter.

4.3.14 Quotations

Tag: <quote> (<in-quote>)
Attribute: lang
Possible values of lang: see 4.3.10, Foreign words and phrases

End tag obligatory. Generally not used in SUC 2.0. As with <q>, the tag is entered manually and preferably already at preprocessing. TEI P3 Part II 6.3.3 Quotation.

All quotations, long or short, in the running text or separated from it, are tagged. For <quote> to be applicable it must be clear that the material concerned has a source other than the writer but is not rendered in the text as direct speech/writing (4.3.13). The presence of quotation marks is neither a necessary nor a suffficient condition. If there are quotation marks, the <quote> tag is entered outside them. For the combination with <p>, the same principles as for <q> (4.3.13) hold. When quotations are in a foreign language, the attribute *lang* is used according to the principles described in 4.3.13 and 4.3.10.

When information about the source is embedded in the quotation, the <quote>-sequence can be interrupted by an <in-quote> in the same way as for <q> - <in-q> above.

[14] Not used in SUC 2.0.

There sometimes occur a kind of pseudo-quotations, where quotation marks rather signal distance or irony and they should not be tagged at all, or possibly as <mentioned> (4.3.11) or <distinct> (4.3.17) if there are separate reasons for that. (Cf. also the tag <soCalled> in TEI P3 Part II 6.3.3.) If the quotation is in a foreign language, the tag <foreign> is entered inside the <quote> tag.

Examples:

jf02:

```
<quote>
<s id=jf02-084>
<w n=1449>Säg<ana><ps>VB<m>IMP AKT<b>säga</w>
<w n=1450>mig<ana><ps>PN<m>UTR SIN DEF OBJ<b>jag</w>
<w n=1451>hur<ana><ps>HA<b>hur</w>
<w n=1452>mycket<ana><ps>JJ<m>POS NEU SIN IND NOM<b>mycken</w>
<w n=1453>guld<ana><ps>NN<m>NEU SIN IND NOM<b>guld</w>
<d n=1454>,<ana><ps>MID<b>,</d>
<w n=1455>mässing<ana><ps>NN<m>UTR SIN IND NOM<b>mässing</w>
<d n=1456>,<ana><ps>MID<b>,</d>
<w n=1457>tenn<ana><ps>NN<m>NEU SIN IND NOM<b>tenn</w>
<w n=1458>och<ana><ps>KN<b>och</w>
<w n=1459>järn<ana><ps>NN<m>NEU SIN IND NOM<b>järn</w>
<w n=1460>du<ana><ps>PN<m>UTR SIN DEF SUB<b>du</w>
<w n=1461>måste<ana><ps>VB<m>PRS AKT<b>måste</w>
<w n=1462>använda<ana><ps>VB<m>INF AKT<b>använda</w>
<w n=1463>för<ana><ps>PP<b>för</w>
<w n=1464>att<ana><ps>SN<b>att</w>
<w n=1465>hela<ana><ps>JJ<m>POS UTR/NEU SIN DEF NOM<b>hel</w>
<w n=1466>kronan<ana><ps>NN<m>UTR SIN DEF NOM<b>krona</w>
<w n=1467>skall<ana><ps>VB<m>PRS AKT<b>ska</w>
<w n=1468>väga<ana><ps>VB<m>INF AKT<b>väga</w>
<num>
<w n=1469>sextio<ana><ps>RG<m>NOM<b>sextio</w>
</num>
<foreign lang=la>
<w n=1470>minae<ana><ps>UO<b>minae</w>
</foreign>
<d n=1471>.<ana><ps>MAD<b>.</d>
</s>
</quote>
```

jf03:

```
<quote>
<s id=jf03-023>
<w n=319>Primtalen<ana><ps>NN<m>NEU PLU DEF NOM<b>primtal</w>
<w n=320>är<ana><ps>VB<m>PRS AKT<b>vara</w>
<w n=321>flera<ana><ps>JJ<m>POS UTR/NEU PLU IND NOM<b>flera</w>
<w n=322>än<ana><ps>KN<b>än</w>
<w n=323>varje<ana><ps>DT<m>UTR/NEU SIN IND<b>varje</w>
<w n=324>uppgiven<ana><ps>PC<m>PRF UTR SIN IND NOM<b>uppgiven</w>
<w n=325>mångfald<ana><ps>NN<m>UTR SIN IND NOM<b>mångfald</w>
<w n=326>av<ana><ps>PP<b>av</w>
<w n=327>primtal<ana><ps>NN<m>NEU PLU IND NOM<b>primtal</w>
<d n=328>.<ana><ps>MAD<b>.</d>
</s>
</quote>
```

ja23:

```
<quote>
<s id=ja23-060>
<d n=1413>"<ana><ps>PAD<b>"</d>
<w n=1414>För<ana><ps>PP<b>för</w>
<distinct type=obsolete>
<w n=1415>hvart<ana><ps>DT<m>NEU SIN IND<b>var</w>
</distinct>
<w n=1416>honom<ana><ps>PN<m>UTR SIN DEF OBJ<b>han</w>
<w n=1417>tillsände<ana><ps>PC<m>PRF MAS SIN DEF NOM<b>tillsänd</w>
<distinct type=obsolete>
<w n=1418>bref<ana><ps>NN<m>NEU SIN IND NOM<b>brev</w>
</distinct>
<d n=1419>,<ana><ps>MID<b>,</d>
<d n=1420>...<ana><ps>MID<b>...</d>
<w n=1421>honom<ana><ps>PN<m>UTR SIN DEF OBJ<b>han</w>
<distinct type=obsolete>
<w n=1422>gifvas<ana><ps>VB<m>INF SFO<b>giva</w>
</distinct>
```

```
<w n=1423>och<ana><ps>KN<b>och</w>
<w n=1424>betalas<ana><ps>VB<m>INF SFO<b>betala</w>
<num>
<w n=1425>två<ana><ps>RG<m>NOM<b>två</w>
</num>
<distinct type=obsolete>
<w n=1426>öra<ana><ps>NN<m>NEU PLU IND NOM<b>öre</w>
</distinct>
<d n=1427>,<ana><ps>MID<b>,</d>
<w n=1428>dem<ana><ps>PN<m>UTR/NEU PLU DEF OBJ<b>de</w>
<w n=1429>han<ana><ps>PN<m>UTR SIN DEF SUB<b>han</w>
<w n=1430>för<ana><ps>PP<b>för</w>
<w n=1431>sin<ana><ps>PS<m>UTR SIN DEF<b>sin</w>
<w n=1432>flitige<ana><ps>JJ<m>POS MAS SIN DEF NOM<b>flitig</w>
<w n=1433>beställning<ana><ps>NN<m>UTR SIN IND NOM<b>beställning</w>
<w n=1434>njuta<ana><ps>VB<m>INF AKT<b>njuta</w>
<w n=1435>och<ana><ps>KN<b>och</w>
<w n=1436>behålla<ana><ps>VB<m>INF AKT<b>behålla</w>
<w n=1437>skall<ana><ps>VB<m>PRS AKT<b>ska</w>
<d n=1438>.<ana><ps>MAD<b>.</d>
<d n=1439>"<ana><ps>PAD<b>"</d>
</s>
</quote>
```

4.3.15 Footnotes and other notes

Tag: <note>
Attributes: type, resp
Possible values of type: footnote|suc
Possible values of resp: the signature of a person

End tag is obligatory. The tag is entered manually. TEI P3 Part II 6.8.1 Notes and Simple Annotation .

Notes can be of (at least) two different kinds: notes that appear in the original text and notes that are entered by the staff while working with the text. In accordance with TEI guidelines we use the same tag for both kinds. The difference between different kinds of notes is shown in the value of the attribute *type*. For notes in the original text, *footnote* is used (for both footnotes and endnotes). For notes entered to point out or comment something found in the processing of the text, the type value *suc* is used. For the latter kind of notes, but not for the footnotes/endnotes, the attribute *resp* should also be specified, telling who is responsible for the note and its contents.

The texts of the SUC core corpus are as far as possible chosen so as to avoid any footnotes. We have also refrained from adding our own comments to the texts, so the <note> tag is sparsely used in the core corpus. In a larger corpus containing entire books, it can, however, be put to use. Footnotes are then given at the end of texts with the place of the reference in the text to the note signalled by a pointer tag, <ptr>. For details concerning the use of <ptr>, the reader is referred to TEI P3 Part II 6.6, Simple Links and Cross References.

Example (constructed):

```
<note type=suc resp=gk>This is a note.</note>
```

4.3.16 Omitted material

Tag: <gap>
Attributes: desc, extent, resp
Possible values of desc: table|diagram|picture|formula|other
Possible values of extent: the size of the omission
Possible values of resp: the signature of the person responsible for the omission

No end tag. The tag should be entered manually, mostly at preprocessing. TEI P3 Part II 6.5.3 Additions, Deletions and Omissions. The tag has not been used in SUC 2.0.

It is often necessary or desirable to leave out material in the corpus texts. As the corpus is primarily intended for linguistic research, it should mainly consist of running text. Pictures, tables, diagrams, etc. are thus removed at the stage of preprocessing. If it can be decided where in the text the diagram etc. comes in (this is not always the case, in particular not with pictures) <omit> is entered at the place of the omission. <omit> has to be used only when it is necessary for the comprehension of the text, e.g. when reference is being made in the text to some diagram or other omitted material. This means that most omissions of such material as, e.g. pictures in newspapers and illustrative boxes in (popular) scientific texts are not recorded at all in the SUC.

4.3.17 Remaining phenomena of a linguistic nature

Tag: <distinct>
Attribute: type
Possible values of type: obsolete|nonstandard|emph|term|formula|other

End tag is obligatory. The tag is entered manually. TEI P3 Part II 6.3.2.3 Other Linguistically Distinct Material.

The tag <distinct> can be seen as superordinate to all the more fine-grained classification above. For material which is felt to be in some way deviant from ordinary running text but which is not covered by any other tag, <distinct> can be used. It is also the only way of marking language that is old-fashioned (*type=obsolete*), dialectal or slang (*type=nonstandard*), emphasized (*type=emph*), specialized terms (*type=term*), and formulas etc. (*type=formula*) appearing in running text.

To decide what is an instance of <distinct type=term> is not easy and clearly depends on the text type - a word that would be an advanced technical term in a newspaper article can belong to the standard vocabulary of a scientific article. The recommendation is to use <distinct type=term> only when there are clear signals that the author of the text treats a word like that, either by the use of typographic means or by, e.g. defining the term. Otherwise, the tag <mentioned> should be used.

The attribute *emph* is also used only when the emphasis is typographically signalled.

If it is not possible to decide between the given alternatives among possible tags and type values, <distinct> can be used without a specification of the attribute *type*. If it seems clear that the deviation is of a type other than those suggested, the value *other* is used. In general, <distinct> is the default category for deviant material that cannot be classified.

Examples:

ja23:

```
<w n=817>då<ana><ps>HA<b>då</w>
<w n=818>jag<ana><ps>PN<m>UTR SIN DEF SUB<b>jag</w>
<distinct type=obsolete>
<w n=819>imedlertid<ana><ps>AB<b>emellertid</w>          (oldfashioned spelling)
</distinct>
<w n=820>näst<ana><ps>AB<b>näst</w>
<name type=myth>
<w n=821>Guds<ana><ps>PM<m>GEN<b>Gud</w>
</name>
<w n=822>tillhjälp<ana><ps>NN<m>UTR SIN IND NOM<b>tillhjälp</w>
<w n=823>och<ana><ps>KN<b>och</w>
<w n=824>mina<ana><ps>PS<m>UTR/NEU PLU DEF<b>min</w>
<w n=825>mågars<ana><ps>NN<m>UTR PLU IND GEN<b>måg</w>
<distinct type=obsolete>
<w n=826>assistence<ana><ps>NN<m>UTR SIN IND NOM<b>assistans</w>
</distinct>
<d n=827>,<ana><ps>MID<b>,</d>
<w n=828>verket<ana><ps>NN<m>NEU SIN DEF NOM<b>verk</w>
<w n=829>så<ana><ps>AB<b>så</w>
<distinct type=obsolete>
<w n=830>drifva<ana><ps>VB<m>INF AKT<b>driva</w>         (oldfashioned spelling)
</distinct>
<w n=831>skall<ana><ps>VB<m>PRS AKT<b>ska</w>
```

ca02:

```
<l id=ca02b-018>                                            (dialectal)
<w n=2292>di<ana><ps>DT<m>UTR/NEU PLU DEF<b>den</w>
<w n=2293>små<ana><ps>JJ<m>POS UTR/NEU PLU IND/DEF NOM<b>liten</w>
<distinct type=nonstandard>
<w n=2294>fulana<ana><ps>NN<m>UTR PLU DEF NOM<b>fågel</w>
<w n=2295>sjong<ana><ps>VB<m>PRT AKT<b>sjunga</w>
```

```
</distinct>
<w n=2296>så<ana><ps>SN<b>så</w>
<distinct type=nonstandard>
<w n=2297>de<ana><ps>PN<m>NEU SIN DEF SUB/OBJ<b>det</w>
<w n=2298>ronga<ana><ps>VB<m>PRT AKT<b>runga</w>
</distinct>
<d n=2299>.<ana><ps>MAD<b>.</d>
</l>
```

fh11:

```
<s id=fh11-070>
<w n=1304>De<ana><ps>PN<m>UTR/NEU PLU DEF SUB<b>de</w>
<w n=1305>kallade<ana><ps>VB<m>PRT AKT<b>kalla</w>
<w n=1306>den<ana><ps>DT<m>UTR SIN DEF<b>den</w>
<w n=1307>nya<ana><ps>JJ<m>POS UTR/NEU SIN DEF NOM<b>ny</w>
<w n=1308>partikeln<ana><ps>NN<m>UTR SIN DEF NOM<b>partikel</w>
<distinct type=term>
<w n=1309>&PSgr;<ana><ps>PM<m>NOM<b>&PSgr;</w>             (‘psi, a particle symbol’)
</distinct>
<d n=1310>.<ana><ps>MAD<b>.</d>
</s>
```

ec01:

```
<w n=282>nitrat<ana><ps>NN<m>NEU SIN IND NOM<b>nitrat</w>
<d n=283>(<ana><ps>PAD<b>(</d>
<distinct type=formula>                                  (‘chemical’)
<w n=284>NO3-<ana><ps>PM<m>NOM<b>NO3-</w>
</distinct>
<d n=285>)<ana><ps>PAD<b>)</d>
<w n=286>och<ana><ps>KN<b>och</w>
<w n=287>därefter<ana><ps>AB<b>därefter</w>
<w n=288>till<ana><ps>PP<b>till</w>
<w n=289>kvävgas<ana><ps>NN<m>UTR SIN IND NOM<b>kvävgas</w>
<d n=290>(<ana><ps>PAD<b>(</d>
<distinct type=formula>
<w n=291>N2<ana><ps>PM<m>NOM<b>N2</w>
</distinct>
```

jb01:

```
<w n=1722>uttal<ana><ps>NN<m>NEU SIN IND NOM<b>uttal</w>
<d n=1723>:<ana><ps>MID<b>:</d>
<d n=1724>"<ana><ps>PAD<b>"</d>
<distinct type=other>
<w n=1725>prajmat<ana><ps>PC<m>PRF NEU SIN IND NOM<b>prajmad</w>
</distinct>
<d n=1726>"<ana><ps>PAD<b>"</d>
```

4.3.18 Cases of remaining typographic marking

Tag: <hi>

End tag is obligatory, no attributes. Entered automatically or manually. TEI P3 Part II 6.3.2.2 Emphatic Words and Phrases.

<hi> stands for 'highlighted'. This tag is mainly of a temporary nature and is used at an early stage of the processing to capture things that might otherwise be lost. It covers all kinds of typographic variation; differences in font or size, italics, boldface, capitals, underlining, etc. Normally all this is used to signal something and as soon as that has been interpreted and expressed by any of the tags described in 4.3.1-4.3.16 the need to keep the tag <hi> disappears.

The temporary nature of <hi> also explains why we do not use the attribute rend for rendition and with, e.g. *italics* and *bold* among its possible values although this attribute of <hi> is suggested in the TEI guidelines. To us as linguists, it is of no importance whether a headline is typeset in 14 points or boldface or both, the only thing that matters is that it is a headline. Thus, if something is tagged <head> or <foreign> or <distinct> or whatever, it is not at the same time tagged <hi>. <hi> is only used for phenomena that are not yet interpreted or that for some reason cannot be interpreted. Furthermore, <hi> is only used for phenomena that cannot be rendered in pure text (ASCII) format, i.e., capitalisation is never tagged <hi> but is kept even after it has been interpreted.

All this means that <hi> is never seen in the final version of the SUC core corpus, but in intermediate versions it could be an important cue for the annotators.

4.4 Some conventions for the markup

4.4.1 Basic units in the markup of texts

To ensure the unique identification of all strings of linguistic material, we have chosen to prescribe that all text should belong to exactly one <s> unit. This means that <s> should not be interpreted as (graphic) sentence, although it mostly coincides with that, but as a more general and more abstract unit that forms the basic building block in the corpus.

Each <s>, as described above (4.2.3), has a unique *id*-value composed of the SUC identification of the text and a running three-digit enumeration starting from 001 in each text. All excerpts from the corpus will thereby have an identification that makes it easy to retrieve the context it was taken from.

After some thought and discussion we have also assigned another tag this basic status, namely the <l> tag for lines of poetry. In a poem, it is important that the line breaks are saved. This is done by means of the <l> tag (4.3.4). It is, however, often so that line breaks and sentence boundaries do not coincide in poetry. A sentence can start in the middle of one line and end in the middle of another. As it is not allowed to have tags cross (see 4.4.3), we have chosen to let <l>, instead of <s>, be the basic structural unit in sequences of poetry. To get the identification that was the reason for having a basic unit in the first place, we let the *id* of <l> tags have the same running enumeration as the surrounding <s> tags. Thus:

```
<p>
   <s id=ex04-007> ...
   <s id=ex04-008> ...
<lg>
   <l id=ex04-009> ...
   <l id=ex04-010> ...
   <l id=ex04-011> ...
   <l id=ex04-012> ...
</lg>
<p>
   <s id=ex04-013> ...
   <s id=ex04-014> ...
```

In most text types, however, it is rare that poems are reproduced in the middle of a running text, so this mechanism seldom has to be applied.

The convention about <s> as a basic unit sometimes may have odd consequences, as, e.g. in connection with <label> and <item> in lists (sect. 4.3.3). Especially what appears as <label> may be felt superfluous to also be tagged as <s>, but we want to stick to this convention in order to get all linguistic material segmented into uniquely identifiable strings.

The <p> tag is also a kind of basic unit in texts, along with its 'equivalents' <head>, <byline>, <list>, and <lg>, but on the level above <s>. This means that everything in a text first belongs to exactly one <s> (or, rarely, <l>). Every instance of <s> is then classified as belonging to exactly one of the categories <p>, <head>, <byline>, <list>, or <lg>. These categories are never nested into each other (cf. 4.4.3). A text thus consists of a coherent sequence of such <p> level units. This convention should not cause any problems. In most texts, only <p> and <head> occur. In newspaper texts <byline> can be frequent, while <list> and <lg> are always very infrequent.

Another way of doing this would have been to treat all <p> level units as functionally interpreted instances of <p> and then separate them by means of the attribute *type*, e.g. <p type=head>. That would have been an easier way of handling it in the SUC, as the automatic mark-up generates a <p> tag for all these cases. It would then be more natural to add the value of an attribute than to change a tag. This is, however, not the way it is done in the TEI guidelines, where they are all kept apart, and we have chosen to adhere to the guidelines.

4.4.2 The range of tags

Tags are normally applied to maximal strings of words, i.e. a row of words that all have the same tag and the same reference is given one pair of tags. Dates can be long and composite, as can names, but they are still given one single pair of tags. In enumeration of names, each name will of course get its own <name> tag, as the names there have different referents, cf. the example of a non-list in 4.3.3. Also, the name of an institution followed by the name of the place where it is situated is tagged as two names even if they can sometimes be felt to form a unity. As stated above (4.3.8), periods of time in from-to expressions are tagged as two separate instances, while expressions containing a dash count as one.

Headlines sometimes consist of two parts, e.g. on different lines or with different fonts. We then tag them as two consecutive occurrences of <head>, even if we otherwise do nothing to render various levels of headlines. Below are some examples of range.

Composite name:

```
<name>                                                     (constructed)
<w>Tekniska<ana><ps>JJ<m>UTR/NEU SIN/PLU IND/DEF NOM<b>teknisk</w>
<w>Högskolan<ana><ps>NN<m>UTR SIN DEF NOM<b>högskola</w>
</name>
<w>i<ana><ps>PP<b>i</w>
<name>
<w>Stockholm<ana><ps>PM<m>NOM<b>Stockholm</w>
</name>
```

Dates and times:

```
<date>                                                     (constructed)
<w>måndagen<ana><ps>NN<m>UTR SIN DEF NOM<b>måndag</w>
<w>den<ana><ps>DT<m>UTR SIN DEF<b>den</w>
<num>
<w>16<ana><ps>RG<m>NOM<b>16</w>
</num>
<w>augusti<ana><ps>NN<m>UTR SIN IND NOM<b>augusti</w>
<num>
<w>1993<ana><ps>RG<m>NOM<b>1993</w>
</num>
</date>

<w>från<ana><ps>PP<b>från</w>                              (constructed)
<time>
<w>klockan<ana><ps>NN<m>UTR SIN DEF NOM<b>klocka</w>
<num>
<w>21<ana><ps>RG<m>NOM<b>21</w>
</num>
</time>
<w>till<ana><ps>PP<b>till</w>
<time>
<w>klockan<ana><ps>NN<m>UTR SIN DEF NOM<b>klocka</w>
<num>
<w>22<ana><ps>RG<m>NOM<b>22</w>
</num>
</time>

<date>
<num>
<w>1676-1749<ana><ps>RG<m>NOM<b>1676-1749</w>
</num>
</date>
```

Two-part headline:

```
<head>
<s id=ab07c-001>
<w>Byskolan<ana><ps>NN<m>UTR SIN DEF NOM<b>byskola</w>
<w>som<ana><ps>HP<b>som</w>
<w>överlevt<ana><ps>VB<m>SUP AKT<b>överleva</w>
<w>samhällsreformer<ana><ps>NN<m>UTR PLU IND NOM<b>samhällsreform</w>
<w>firar<ana><ps>VB<m>PRS AKT<b>fira</w>
<num>
<w>50<ana><ps>RG<m>NOM<b>50</w>
</num>
<w>år<ana><ps>NN<m>NEU PLU IND NOM<b>år</w>
<d>:<ana><m>MAD<b>:</d>
</s>
</head>
```

```
<head>
<s id=ab07c-002>
<name type=person>
<w>Märta<ana><ps>PM<m>NOM<b>Märta</w>
</name>
<w>minns<ana><ps>VB<m>PRS SFO<b>minnas</w>
<num>
<w>första<ana><ps>RO<m>UTR/NEU SIN/PLU IND/DEF NOM<b>första</w>
</num>
<w>dagen<ana><ps>NN<m>UTR SIN DEF NOM<b>dag</w>
<w>i<ana><ps>PP<b>i</w>
<name type=place>
<w>Örserum<ana><ps>PM<m>NOM<b>Örserum</w>
</name>
</s>
</head>
```

4.4.3 Combinability and embedding of tags

As seen from the above presentation, tags very often combine and are embedded into each other. In order to reach consistency in the SUC, we have decided on some conventions for the embedding of tags. When implementing an annotation scheme using the basic mechanisms of the TEI guidelines, one of the few restrictions on the nesting of tags is that they are not allowed to cross, i.e., given that <tag1> and <tag2> have the same range, the sequences <tag1><tag2>... </tag2></tag1> and <tag2><tag1>...</tag1></tag2> are both accepted while <tag1><tag2>...</tag1></tag2> is not. In the SUC, however, we want to regulate the combinations of tags a little more and also to regulate their relative order. [15]

Instances where one tagged sequence is entirely embedded into another are unproblematic, as the tag with the longest range will then by necessity also be the outermost one. Quotation marks are not removed, but remain inside the tags that express an interpretation of them. [16]

Examples (constructed):

```
<name type=work>                                              (the title of a book)
         <w>Nu<ana><ps>AB<b>nu</w>
         <w>var<ana><ps>VB<m>PRT AKT<b>vara</w>
         <w>det<ana><ps>PN<m>NEU SIN DEF NOM<b>det</w>
<date>
      <num>
         <w>1914<ana><ps>RG<m>NOM<b>1914</w>
      </num>
   </date>
</name>

<name type=work>                                              (the title of a book)
   <d>"<ana><m>PAD<b>"</d>
   <w>En<ana><ps>DT<m>UTR SIN IND<b>en</w>
   <w>kakelsättares<ana><ps>NN<m>UTR SIN IND GEN<b>kakelsättare</w>
   <w>eftermiddag<ana><ps>NN<m>UTR SIN IND NOM<b>eftermiddag</w>
   <d>"<ana><m>PAD<b>"</d>
</name>
```

<w> with its contents <ana>, <ps>, <m> and <b> is of course the innermost tag. No other tags are allowed inside a <w>.

<gap> is punctual, i.e. it has no range and never combines with other tags.

The problematic cases occur when two tags have the same range. Below are some rules for individual tags. The rules only apply in cases where two or more tags have the same range.

<s> is inside <p>, <head>, <byline>, <list>, <label>, <item>, and <note>.

<l> is inside <lg>.

[15] Cf. also the DTD in appendix C.

[16] This is not always the case in SUC 2.0, where quotation marks frequently occur outside the name tags.

<s> is outside <wg>, <name>, <date>, <time>, <ref>, <foreign>, <mentioned>, <abbr>, <distinct>, and <hi>.

<foreign> is outside <mentioned>

<abbr> is always innermost, containing only <w>.

Example (constructed):

```
    <name type=inst><abbr>UNESCO</abbr></name>
```

<num> is always innermost, containing only <w>.

<p>, <head>, <byline>, <list>, and <lg> are, by definition, mutually exclusive, as are <s> and <l>.

A paragraph, <p> (and its equivalents), can entirely consist of material that is quoted, <q> or <quote>, or that is <foreign>. The <p> is then outermost. Also when several consecutive paragraphs in their entirety are quoted or written in a foreign language, the <p> tag should be outside the more linguistic/semantic tag, which thus has to be repeated for each new paragraph.

4.5. The Markup Process

The morphosyntactic annotation took place before the texts were converted into SGML format.[17] In that conversion process, some markup was automatically inserted (cf. chapter 4.2). In a later step, it was the task of the annotators to check automatically inserted SGML markup and to add such tags as require human interpretation of the functions of various parts of the texts. In doing this, they had access to hard copies of the texts.

Annotation was thus done in two steps, one for disambiguation and one for SGML markup. This was due to the fact that disambiguation started long before we had settled on an apparatus for TEI conformant markup. The optimal way of doing it is probably to enter as much as possible of the structural and functional information about the texts already at data capture and then have annotators perform disambiguation and add SGML tags at the same time. ‘Next time? …’

[17] See corpus header, section ‘project description’ in appendix F.

References

Allén, S (1970): Nusvensk frekvensordbok baserad på tidningstext. Frequency Dictionary of Present Day Swedish. (Part 1), Almqvist & Wiksell Förlag AB, Stockholm 1970.

Alschuler, L (1995): ABCD … SGML: A User's Guide to Structured Information. International Computer Press, 1995.

BNC (1995) Users Reference Guide British National Corpus Version 1.0, ed. Lou Burnard, Published for the British National Corpus Consortium by Oxford University Computing Services, May 1995. (http://info.ox.ac.uk/bnc/getting/bncman.html)

ECI-CD = Edinburgh CD

Ejerhed, E., G. Källgren, O. Wennstedt & M. Åström (1992): The Linguistic Annotation System of the Stockholm-Umeå Corpus Project. Publications from the Department of General Linguistics, University of Umeå, no. 33. 1992.

Francis, W.N. & H. Kucera. 1979. Manual of information to accompany a Standard Sample of Present-day Edited American English, for use with digital computers. Original ed. 1964, revised 1971, revised and augmented 1979. Providence, R.I.: Department of Linguistics, Brown University.

van Herwijnen, E.(1990): Practical SGML. Kluwer Academic Publishers 1990.

Johansson, S., G. Leech & H. Goodluck. 1978. Manual of information to accompany the Lancaster-Oslo/Bergen Corpus of British English, for use with digital computers. Department of English, University of Oslo.

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

Karlgren, J. and Cutting, D.(1994): Recognizing textgenres with simple metrics using discriminant analysis. Coling 1994, Kyoto, July 1994.

Karlgren, J. (2000) Stylistic Experiments for Information Retrieval. Dissertation, Stockholm 2000.

Karlsson, F. "SWETWOL: A Comprehensive Morphological Analyzer for Swedish". Nordic Journal of Linguistics 15, 1992, 1-45.

Lindberg, J (1999): Automatic Detecting of Lexicalised Phrases in Swedish: Nodalida '99. Proceedings from the 12th "Nordiske datalingvistikkdager". Trondheim, 9-10 December 1999. Torbjørn Nordgård, editor, Trondheim Department of Linguistics, NTNU 2000, pp 103-114.

Massmedia 89-90 and later : Massmedia handbok för journalister, informatörer och andra som följer press radio och TV. Guide till svenska medier. Kunskapsförlaget/Norstedt Stockholm. (Utkommer årligen, från 1998 även på CD-ROM. Senaste årgång, Massmedia 2000.)

Rahkonen, M. (1992): Studentsvenska 79/80 - en korpus över finska abiturienters inlärarsvenska. Meddelanden från Institutionen för nordiska språk vid Jyväskylä universitet 8, Jyväskylä 1992, s 3.

Sperberg-McQueen, C. M. & Burnard, L. (eds.): Guidelines for Electronic Encoding and Interchange. Chicago, Oxford 1990-1993.

Sperberg-McQueen, C. M. & Burnard, L. (eds.): TEI Guidelines for Electronic Encoding and Interchange (P3) (http://etext.lib.virginia.edu/TEI.html)

Teleman, U.: Manual för grammatisk beskrivning av talad och skriven svenska. Studentlitteratur, Lund. Lundastudier, 1974.

Wennstedt, O (1995): Annotering av namn i SUC-korpusen. The Nordic Languages and Modern Linguistics. Proceedings of the Ninth International Conference of Nordic and General Linguistics. University of Oslo. January 11-12, 1995. Edited by Kjartan G. Ottósson, Ruth V. Fjeld and Arne Torp. Novis forlag. Oslo 1996, pp 315-324.

Appendices

A. SUC Text Categories

A. Press: Reportage

AA Political
AB Community
AC Financial
AD Cultural
AE Sports
AF Spot News

B. Press: Editorial

BA Institutional
BB Debate articles

C. Press: Reviews

CA Books
CB Films
CC Art
CD Theatre
CE Music
CF Artists, shows
CG Radio, TV

E. Skills and Hobbies

EA Hobbies, amusements
EB Society press
EC Occupational and trade union press
ED Religion

F. Popular Lore

FA Humanities
FB Behavioural sciences
FC Social sciences
FD Religion
FE Complementary life styles
FF History
FG Health and medicine
FH Natural science, technology
FJ Politics
FK Culture

G. Biographies, essays

GA Biographies, memoirs
GB Essays

H. Miscellaneous

HA Government publications
HB Municipal publications
HC Financial reports, business
HD Financial reports, non-profit organisations
HE Internal publications, companies
HF University publications

J. Learned and scientific writing

JA Humanities
JB Behavioural sciences
JC Social sciences
JD Religion
JE Technology
JF Mathematics
JG Medicine
JH Natural science, technology

K. Imaginative prose

KK General fiction
KL Science fiction and mystery
KN Light reading
KR Humour

B. SUC agreement forms

Avtal om nyttjande av maskinellt läsbara texter för forskningsändamål

Mellan Institutionen för Lingvistik, Stockholms universitet, Institutionen för Lingvistik, Umeå universitet (gemensamt benämnda IL)

och .... (Rättighetsinnehavaren)

har följande avtal träffats avseende rätt att nyttja maskinläsbara texter för forskningsändamål.

Avtalet innefattar följande bilagor:

Bilaga 1: Specifikation av de maskinläsbara texter som Rättighetsinnehavaren tillhandahåller.

Bilaga 2: Beskrivning av nyttjandet för forskningsändamål.

Bilaga 3: Utfästelse från enskilda forskare angående nyttjande av maskinläsbara texter.

1 §

Rättighetsinnehavaren upplåter till IL nyttjanderätt till maskinläsbara texter uteslutande för forskningsändamål. En allmän beskrivning av medgivet nyttjande framgår av bilaga 2 till detta avtal.

IL tillåts att kopiera erhållna texter samt att vidareupplåta exemplar av texterna till enskilda forskare, vilka dock skall underteckna en utfästelse, vilken framgår av bilaga 3 till detta avtal.

IL förbinder sig att inte nyttja erhållna maskinläsbara texter för något som helst kommersiellt ändamål.

Rättighetsinnehavaren äger inga som helst rättigheter till forskningresultatet.

2 §

Rättighetsinnehavaren garanterar att nyttjanderätten, reglerad i detta avtal, får upplåtas och sålunda inte utgör intrång i annans immateriella rättigheter.

3 §

IL åtar sig att endast lämna ut exemplar av erhållna maskinläsbara texter till enskilda forskare sedan dessa undertecknat utfästelse enligt bilaga 3 till detta avtal.

IL skall på begäran tillställa Rättighetsinnehavaren en kopia av varje sådan utfästelse. Utfästelsen i original skall förvaras hos IL.

4 §

Detta avtal gäller från dess undertecknande under 5 år.

Om avtalet inte sägs upp senast sex månader före avtalstidens slut löper avtalet vidare 3 år i taget med sex månaders uppsägningstid varje gång.

Uppsägning skall för att gälla ske skriftligen.

Vid avtalets upphörande skall, om Rättighetsinnehavaren så bestämmer i den skriftliga uppsägningen av avtalet, samtliga exemplar av upplåtna maskinläsbara texter raderas. IL och enskilda forskare med nyttjanderätt äger dock rätt att för enbart arkivändamål behålla erforderlig del av de maskinläsbara texterna samt fritt förfoga över analys- och forskningsresultat.

5 §

Ekonomisk ersättning till Rättighetsinnehavaren skall endast utgå i form av skäligt engångsbelopp motsvarande självkostnader för att till IL överlämna en kopia av de avtalade maskinläsbara texterna.

6 §

Detta avtal har undertecknats i två exemplar, av vilka parterna erhållit var sitt.

Ort datum Ort datum

för Institutionen för Lingvistik

______________________________ ______________________________

Avtalsbilaga 1

Specifikation av de maskinläsbara texter som Rättighetsinnehavaren tillhandahåller

Avtalsbilaga 2

Beskrivning av nyttjandet för forskningsändamål

Texterna i den samlade korpusen kommer att lagras och bearbetas på dator för att sedan analyseras både manuellt och med dator. Analysen kommer att omfatta ordklasser, satsdelar och större enheter, liksom samtliga enheters egenskaper, betydelse och funktion i texten. Jämförelser mellan olika texter och olika texttyper kommer att göras.

Resultaten kommer huvudsakligen att presenteras i form av ordlistor och statistiska och andra uppställningar där de ingående texterna inte går att identifiera. En annan typ av resultat är de datorprogram som byggs upp för analysen.

I samband med muntlig och skriftlig presentation och diskussion av resultaten kan kortare avsnitt ur texterna citeras. Källan ska då alltid anges.

Avtalsbilaga 3

Utfästelse avseende nyttjande av anförtrott textmaterial i maskinläsbar form för forskningsändamål

Undertecknad forskare har tagit del av gällande avtal mellan
Institutionen för Lingvistik, Stockholms universitet,
Institutionen för Lingvistik, Umeå universitet och

textgivarna till Stockholm-Umeå Corpus (SUC)

jämte bilagor och förbinder mig härmed att hantera mig anförtrodda maskinläsbara texter helt i enlighet med vad som där reglerats beträffande Institutionen för Lingvistik, Stockholms universitet och Institutionen för Lingvistik, Umeå universitet.

Vidare förbinder jag mig att inte distribuera vidare, eller lämna annan tillgång till, något exemplar av de maskinläsbara texterna. Jag får dock fritt förfoga över mina analys- och forskningsresultat.

Vid all publicering av forskningsresultat baserade på denna textkorpus skall anges att korpusen använts.

Ort och datum

....

Namn

....

Namnförtydligande

....

Adress

....

....

Agreement concerning the Use of Machine-readable Texts for Research Purposes

The following Agreement concerning the right to use machine-readable texts for research purposes has been entered into between the Department of Linguistics, Stockholm University, the Department of Linguistics, Umeå University (hereinafter both referred to as "IL")

and ....(Holder of the Right)

The Agreement is accompanied by the following Appendices:

Appendix 1: Specification of the machine-readable texts supplied by the Holder of the Right

Appendix 2: Description concerning the right of use for research purposes

Appendix 3: Declaration submitted by each researcher concerning the use of machine-readable texts

§ 1

The Holder of the Right grants IL the legal right of using and enjoying machine-readable texts solely for research purposes. A general description of the granted usufruct is supplied in Appendix 2 accompanying this Agreement.

IL is permitted to copy the acquired texts and put them at the disposal of individual researchers. In this case each researcher must sign a declaration specified in Appendix 3 to this Agreement.

IL undertakes not to use the acquired texts for any commercial purposes whatsoever.

The Holder of the Right has no right whatsoever to any research results.

§ 2

The Holder of the Right guarantees that the usufruct regulated in this Agreement may be granted to others, and that it shall thus not constitute an infringement of another person's intellectual property rights.

§ 3

IL accepts that copies of acquired machine-readable texts will be handed over to individual researchers only after they have signed a Declaration in accordance with Appendix 3 to this Agreement.

At the request of the Holder of the Right, IL shall submit a copy of such a Declaration to the Holder. The original of the Declaration shall be stored by IL.

§ 4

This Agreement is valid from the date of its signature for the period of 5 years.

If notice of termination is not given at least six months before the expiration of the Agreement, the Agreement continues to be valid for a term of 3 years, with a six-months' period of notice each time.

To be valid, notice of termination must be given in writing.

At the termination of the Agreement, if the Holder of the Right so decides in a written notice of termination of the Agreement, all the acquired copies of machine-readable texts shall be erased. IL and individual researchers with the right of usufruct are entitled, however, to keep, solely for archival purposes, the necessary portion of the machine-readable texts, as well as having free disposal of results from statistical analysis and research.

§ 5

Financial compensation shall be paid to the Holder of the Right only in the form of a one-time payment of a reasonable sum equivalent to the prime cost of a copy of the machine-readable texts agreed upon to be handed over to IL.

§ 6

This Agreement has been signed in duplicate originals, each party having received one of them.

Place	Date	Place	Date

For the Department of Linguistics

…………………………………

Appendix 1 to Agreement .

Appendix 1

Specification of the Machine-readable Texts supplied by the Holder of the Right

(Under this heading the texts supplied by one Holder of Right are listed.)

Appendix 2 to Agreement.

Appendix 2

Description concerning the Right of Use for Research Purposes

The texts contained in the assembled corpus are going to be stored on and processed by the computer in order to be analyzed, both manually and with the help of the computer. The analysis of the texts will embrace word classes, clause elements and larger units, as well as the characteristics of all the specified units, their meaning and function in the text. Comparisons between different texts and different text types are going to be made.

The results will be presented primarily in the form of word lists, statistical tabulations and other kinds of classifications, where it will be impossible to identify the texts making them up. Another type of results will consist in computer programs written for analytical purposes.

In connection with oral and written presentations and discussions of the results, short sections of the texts can be quoted. In such cases the source shall always be stated.

Appendix 3 to Agreement

Appendix 3

Declaration concerning the Use of the Entrusted Text Material in Machine-readable Form for Research Purposes

I, the undersigned, have acquainted myself with the contents of the existing Agreement, including the accompanying Appendices, between the Department of Linguistics, Stockholm University, the Department of Linguistics, Umeå University

and the Holders of Right of the texts in the Stockholm - Umeå Corpus (SUC)

and I hereby undertake to handle the machine-readable texts entrusted to me in accordance with the terms of the Agreement concerning the Department of Linguistics, Stockholm University and the Department of Linguistics, Umeå University.

I further undertake not to distribute or provide access to any copy of the machine-readable texts. My analytical and other research results are, however, totally at my disposal.

In the event of publication of the results based on this text corpus the source shall be stated.

Place and date

....

Name

....

Name in Typescript

....

Address

....

....

C. SUC DTD

```
<!ENTITY % ISOlat1 SYSTEM "ISOlat1.ent">
 %ISOlat1;

<!ENTITY % ISOlat2 SYSTEM "ISOlat2.ent">
 %ISOlat2;

<!ENTITY % ISOgrk1 SYSTEM "ISOgrk1.ent">
 %ISOgrk1;

<!ENTITY %  doctype "suc-file">

<!-- *********************************************************** -->
<!-- 1. Content model as entity declarations       -->
<!-- *********************************************************** -->

<!-- Word level:                                                      -->

<!ENTITY %  words           "w|d"         >

<!-- Level above words, but below s-units:                            -->

<!ENTITY %  wgs             "abbr|num|(%words;)+"      >

<!ENTITY %  ndts            "name|date|time"            >
<!ENTITY %  ndt-tyg         "%wgs;"       >

<!ENTITY %  mds                           "mentioned|distinct"        >
<!ENTITY %  md-tyg          "%ndts;|%ndt-tyg;"          >

<!ENTITY %  for-tyg         "%mds;|%md-tyg;"            >

<!ENTITY %  refs            "ref|bibl"    >
<!ENTITY %  ref-tyg         "foreign|%for-tyg;"         >

<!ENTITY %  s-tyg           "%refs;|%ref-tyg;"          >

<!--Level above s-units but below paragraphs:                         -->

<!ENTITY %  qs                            "quote|q"     >

<!-- Paragraph equivalent level:                                       -->

<!ENTITY %  t-tyg           "p|head|byline|list|lg"     >

<!-- ************************************************ -->
<!-- 2. Tag definitions                          -->
<!-- ************************************************ -->

<!-- 2.1 Top nodes                         -->
<!ELEMENT suc-file          - o           (suctext)+  >

<!ELEMENT suctext                         - -           (%t-tyg;)+                  >
<!ATTLIST suctext
              id            ID            #REQUIRED                    >

<!--2.2 Paragraphs and their equivalents                              -->

<!ELEMENT p                 - -           (%qs;|s)+ >
```

```
<!ELEMENT head                                  - -             (s)+ >

<!ELEMENT byline                - -             (s)+ >

<!ELEMENT list                                  - -             (label?,item)+ >
<!ELEMENT lg                                    - -             (l)+ >
<!ATTLIST lg
                type            (suc-line-group) #CURRENT>

<!--2.3 Optional elements in paragraphs, outside s-units -->
<!ELEMENT quote                                 - -             (s)+>
<!ATTLIST quote
                lang            (en|fr|no|da|la|el|de|es|it|fi|is|ru|cs|other) #IMPLIED >
<!ELEMENT q                     - -             (s)+ >

<!--2.3.b List elements. Label is optional, item is mandatory in lists -->

<!ELEMENT label                                 - -             (s)+ >
<!ELEMENT item                                  - -             (s)+ >

<!--2.4 s-units and equivalents - all words must belong to one of these-->

<!ELEMENT l                     - -             (%s-tyg;)+ >
<!ATTLIST l
                id              ID              #REQUIRED >

<!ELEMENT s                     - -             (%s-tyg;)+ >
<!ATTLIST s
                id              ID              #REQUIRED >

<!--2.5 Optional elements inside s-units-->

<!ELEMENT ref                                   - -             (%ref-tyg;)+ >
<!ELEMENT bibl                                  - -             (%ref-tyg;)+ >
<!ELEMENT foreign               - -             (%for-tyg;)+ >
<!ATTLIST foreign
                lang            (en|fr|no|da|la|el|de|es|it|fi|is|ru|cs|other)          #REQUIRED >


<!ELEMENT mentioned             - -             (%md-tyg;)+ >

<!ELEMENT distinct              - -             (%md-tyg;)+ >
<!ATTLIST distinct
                type            (obsolete|nonstandard|emph|irony|term|formula|other) #REQUIRED >

<!ELEMENT name                                  - -             (%ndt-tyg;)+ >
<!ATTLIST name
                type            (person|place|animal|myth|inst|product|work|event|other) #REQUIRED>

<!ELEMENT date                                  - -             (%ndt-tyg;)+ >
<!ELEMENT time                                  - -             (%ndt-tyg;)+ >
<!ELEMENT num                                   - -             (%words;)+ >
<!ELEMENT abbr                                  - -             (%words;)+ >


<!--2.6 Word level elements - mandatory-->

<!ELEMENT w                                     - -             (#PCDATA, ana) >
<!ATTLIST w
```

```
                n               NUTOKEN     #REQUIRED >

<!ELEMENT d                     - -         (#PCDATA, ana) >
<!ATTLIST d
                n               NUTOKEN     #REQUIRED >

<!--2.7 Elements inside w or d - ana, ps and b are mandatory-->

<!ELEMENT ana                               - o             (ps, m?, b) >

<!ELEMENT ps                                - o             (#PCDATA) >
<!ELEMENT m                                 - o             (#PCDATA) >
<!ELEMENT b                                 - o             (#PCDATA) >

<!-- end of  DTD                                            -->
```

D. SUC headers

The SUC headers are the corpus header (cf. Appendix F) and the 500 document headers.

The SUC corpus header is a rather long formal document structured as recommended in the guidelines in TEI P3. It will accompany the 2 TEI compatible versions of the SUC 2.0 as 2 separate text files. Since they are TEI corpus headers they are essentially described by the teihdr2.dtd with a few extensions added to the SGML library that can be downloaded from http://etext.lib.virginia.edu.

The document headers found at the beginning of SUC 2.0 files look like this:

```
<tei.2>                                                    (The document is a tei2-document)
<teiheader id=h.fc01>                       (The teiheader has an id referring to the text file name)
<fileDesc>
<titleStmt>
<title>suc-fc01</title>
</titleStmt>
<extent words=2017>2017 word tokens</extent>
<publicationStmt>
<distributor>SUC</distributor>
</publicationStmt>
<sourceDesc>                                          (Bibliographic information about the text)
<listBibl id="file-fc01">
<biblFull id="bib-fc01">
<titleStmt>
        <title level=a>Kan Afrika försörja dubbelt så många om 25 år?</title>
        <title level=j>Forskning och Framsteg 3/92</title>
        <author>Björn Fjaestad</author>
</titleStmt>
<extent words=2017>pp 26-30</extent>
<publicationStmt>
        <publisher>Forskning och Framsteg</publisher>
        <pubPlace>Stockholm</pubPlace>
        <idno type="issn">1102-6537</idno>
        <date>1992</date>
</publicationStmt>
</biblFull>
</listBibl>
</sourceDesc>                                                    (End of bibliographic part)
</fileDesc>
<profileDesc>
<textClass>
<catRef target=SUC.FC>                   (Refers to taxonomy part of the corpus header)
</textClass>                             (cf. also Appendix A)
</profileDesc>
</teiheader>                             (This is the end of the document header)
<text id=fc01>                           (The text fc01 is about to begin)
<body>
- - - - - - - - -                        (The body of the text is contained here)
</body>
```

Since the document is tei-conformant, the tei2.dtd, with extensions referenced in the corpus header is appropriate.

E. SUC Bibliography

The rather long SUC bibliography file (for the 500 documents) will not be included here, but what kind of information it contains should be apparent from some examples taken from the 'sourceDesc' in document headers.

1. Bibliographic information for a composite text:

```
<sourceDesc>
<listBibl id="file-ec11">                    (cf. Appendix A for an interpretation of EC)
<biblFull id="bib-ec11a">                    (concerns the first text in the document ec11)
<titleStmt>
      <title level=a>Fackklubben på ISS i Luleå har fullt upp</title>        (the title of an article)
      <title level=j>Fastighetsfolket 3/4 93</title>                         (the title of a journal)
      <author>Minna Pyykölä</author>
</titleStmt>
<extent words=1052>pp 22-23</extent>         (1052 words on pages 22 and 23 in the journal)

<publicationStmt>
      <publisher>Fastighetsanställdas förbund</publisher>
      <pubPlace>Stockholm</pubPlace>
      <idno type="issn">0345-326X</idno>
      <date>1993</date>
</publicationStmt>
</biblFull>
<biblFull id="bib-ec11b">                    (concerns the second text in the document ec11)
<titleStmt>
      <title level=a>Växande kritik mot utredning: Nytt pensionssystem gynnar männen</title>
      <title level=j>Fastighetsfolket 6/7 93</title>
      <author>Minna Pyykölä</author>
</titleStmt>
<extent words=1092>pp 18-19</extent>
<publicationStmt>
      <publisher>Fastighetsanställdas förbund</publisher>
      <pubPlace>Stockholm</pubPlace>
      <idno type="issn">0345-326X</idno>
      <date>1993</date>
</publicationStmt>
</biblFull>
</listBibl>
</sourceDesc>
```

(In the example above the 2 parts of the composite text are from different issues of the journal, but the author is the same. That is not always the case.)

2. Bibliographic information for a unitary text:

```
<sourceDesc>
<listBibl id="file-kk01">
<biblFull id="bib-kk01">
<titleStmt>
        <title level=a>18-20</title>            (chapter heading, in this case just numbers)
        <title level=m>Livets ax</title>        (the title of the book)
        <author>Sven Delblanc</author>
</titleStmt>
<extent words=2173>pp 50-59</extent>
<publicationStmt>
        <publisher>Albert Bonniers förlag AB</publisher>
        <pubPlace>Stockholm</pubPlace>
        <idno type="isbn">91-0-055201-1</idno>
        <date>1991</date>
</publicationStmt>
</biblFull>
</listBibl>
</sourceDesc>
```

F. A SUC2.0 d corpus header [18]

```
<!DOCTYPE teiCorpus.2 SYSTEM "tei2.dtd" [
<!ENTITY % TEI.extensions.ent SYSTEM "suc2d.ent">
<!ENTITY % TEI.extensions.dtd SYSTEM "suc2d.dtd">
]>
<teiCorpus.2>
<teiHeader type=corpus>
    <fileDesc>
        <titleStmt>
            <title>Stockholm Umeå Corpus Version 2, SUC 2.0</title>
            <principal>Eva Ejerhed, Umeå University (UmU)</principal>
            <principal>Gunnel Källgren, Stockholm University (SU)</principal>
            <funder>HSFR (Swedish Council for Research in the Humanities and Social Sciences)</funder>
            <funder>STU/NUTEK (The Swedish National Board for Industrial and Technical
            Development)</funder>
            <funder>The Faculty of Humanities, Umeå University</funder>
            <funder>The Department of Linguistics, Stockholm University </funder>
            <respStmt>
                <resp>Project management at SU and UmU respectively</resp>
                <name>Gunnel Källgren, Eva Ejerhed</name>
                <resp>Compilation of corpus and text type taxonomy</resp>
                <name>Gunnel Källgren</name>
                <resp>Creation of the SUC tagset for morphosyntactic descriptions</resp>
                <name>Eva Ejerhed</name>
                <resp>Data acquisition and legal agreements</resp>
                <name>Gunnel Källgren</name>
                <resp>English translation of legal agreements</resp>
                <name>Teresa Bjelkhagen</name>
                <resp>Bibliography for SUC texts</resp>
                <name>Britt Hartmann</name>
                <resp>Programming (SU)</resp>
                <name>Gunnar Eriksson, Sune Magnberg</name>
                <resp>Selection of text samples and creation of raw text</resp>
                <name>Gunnel Källgren, Britt Hartmann</name>
                <resp>Preprocessing texts for manual annotation, assigning lexical analyses to word
                tokens</resp>
                <name>Eva Ejerhed and Magnus Åström in collaboration with Fred Karlsson, University
                of Helsinki</name>
                <resp>Programming library for SUC format in SUC 1.0, corpus production tools</resp>
                <name>Magnus Åström</name>
                <resp>Manual annotation, morphosyntactic descriptions</resp>
                <name>SU: Janne Lindberg, Cecilia Lyckow, Ulrika Kvist, Carin Svensson-
                Lindberg</name>
                <name>UmU: Joana Arnesson, Eva Ejerhed, Ola Wennstedt, Anna-Lena Wiklund</name>
                <resp>Post-processing manually annotated text</resp>
                <name>Eva Ejerhed, Joana Arnesson, Anna-Lena Wiklund, Magnus Åström, Fredrick
                Backman, Rolf Sandberg</name>
                <resp>Preparing SUC 1.0 for distribution</resp>
                <name>Eva Ejerhed, Magnus Åström, Fredrick Backman, Anders Arnholm, in
                collaboration with Daniel Ridings and Pernilla Danielsson, Gothenburg University</name>
                <resp>Construction of the SGML (TEI) tag set for SUC 2.0</resp>
                <name>Gunnar Eriksson, Gunnel Källgren</name>
                <resp>DTD for manual SGML-markup of SUC 2.0</resp>
                <name>Gunnar Eriksson</name>
                <resp>manual SGML markup</resp>
```

[18] This version of the corpus header has been slightly edited to make it easier to read. If it is to be used as a proper TEI corpusheader the file must be converted to text format, and some lines must be marked as comments.

```
            <name>Maria Arnstad, Harald Berthelsen, Christina Ericsson, Malin Ericson, Tove
            Gerholm, Sofia Gustafson-Capkova, Sara Rydin</name>
            <resp>Management of hard copies</resp>
            <name>Britt Hartmann</name>
            <resp>Project management in Stockholm 1999-2006</resp>
            <name>Benny Brodda, Sofia Gustafson-Capkova</name>
        </respStmt>
    </titleStmt>
    <editionStmt>
        <p>The present edition of the Stockholm-Umeå Corpus, Version 2.0, SUC 2.0. The SGML
        annotated corpus exists in 3 formats. The 2.0c version of the corpus has morphosyntactic
        descriptions in SUC format, while the 2.0d version is in PAROLE format. The third format has
        SUC morphosyntactic descriptions, and elaborate structural markup but the files lack
        bibliographic headers. They are SGML-conformant but not TEI-conformant documents, and
        bibliographic information must be provided in a separate file.</p>
        <p>The first edition of the complete Stockholm-Umeå Corpus (SUC 1.0) was distributed in
        1997. A subset of the annotated SUC corpus of approximately 300 000 words (swe01), created
        october 31, 1992, was distributed in 1994 as a part of the ACL European Corpus Initiative.</p>
    </editionStmt>
    <extent bytes=54617223 punctuations=134099 words=1032494> The number of bytes of the SUC
    format of the corpus (2.0c) is 59691192. A punctuation is defined as a token that is assigned part of
    speech F in PAROLE format, respectively MAD, MID or PAD in SUC format. A word is defined as a
    token that is not a punctuation token.</extent>
    <publicationStmt>
        <distributor>Department of Linguistics</distributor>
        <address>
            <addrline>Stockholm University</addrline>
            <addrline>SE-106 91, Stockholm, Sweden</addrline>
        </address>
        <availability status=restricted>
            <p>Copyright (c) 2006 Dept of Lingustics, Stockholm University, and Dept of Linguistics,
            Umeå University </p>
            <p>Available only for non-commercial purposes and only in accordance with the terms
            stated in the SUC 2.0 (and SUC 1.0) user agreement. Copies of the user agreement in
            Swedish and English are provided in the file LICENCE. (Cf.also AppendixB).</p>
        </availability>
        <Date>2006-12-24</Date>
    </publicationStmt>
    <sourceDesc>
        <biblStruct>
            <monogr>
            <author>Eva Ejerhed, Gunnel Källgren and Benny Brodda</author>
            <title>Stockholm Umeå Corpus Version 2.0, SUC 2.0</title>
            <imprint>
            <pubPlace>Stockholm</pubPlace>
            </imprint>
            </monogr>
            <idno type=isbn>91-631-5876-0</idno>
        </biblStruct>
        <p>The full bibliographic references of the SUC 2.0 texts can be found in the file
        bibliography.suc2.</p>
    </sourceDesc>
</fileDesc>
<encodingDesc>
    <projectDesc>
        <p>Project description 1. The period 1989-1997.</p>
        <p>The SUC corpus was created as part of the joint research project "Corpus based research on
        models for processing unrestricted Swedish Text" ("Korpusbaserad utveckling av modeller för
        datoranalys av löpande svensk text") between the Departments of Linguistics at Stockholm
        University and Umeå University respectively. The principal investigators were Gunnel Källgren
```

in Stockholm and Eva Ejerhed in Umeå. The project was financed by grants from HSFR and STU/NUTEK during 1989/90-1993/94, and 1995/96, and by funds allocated to General Linguistics in Umeå by the Umeå Faculty of Humanities during the same period. The project was part of the Swedish Language Technology Program 1990-1996. </p>

<p>The SUC corpus was created for the purpose of serving as the basis for development, training and testing of various analyzers for unrestricted Swedish text. The objective with respect to the corpus was to create a collection of modern Swedish prose texts of at least 1 million word tokens, in which each word token is tagged by human annotators, i.e. annotated with information about its part of speech, inflectional form, and lemma. The corpus was to contain texts from different genres, i.e. both informative prose and imaginative prose, following the principles used in the composition of the Brown and LOB corpora. </p>

<p>The production of the SUC 1.0 corpus proceeded in three stages which are briefly described below. </p>

<p>The initial stage, covering roughly the period 1989/90-1990/91, consisted of the following: 1) The two project groups agreed on using the Brown LOB principles for corpus compilation, adapted to Swedish circumstances. 2) The Stockholm group worked out a text type taxonomy for the corpus, set goals for how much text to include of each type, and had legal agreements drawn up for use in data aquisition. 3) The Umeå group created and tested the SUC tagset for morphosyntactic description, the principles for its use, and an initial lexicon called wordlist.</p>

<p>During the period 1991/92-1993/94, project work consisted of the following iterated processes:

1) Data acquisition and negotiation of legal agreements with authors and publishers.

2) Construction of a bibliographic database for acquired texts.

3) Selection of the text samples to include in the corpus, and the creation of normalized 7bit raw text. Raw texts were manually marked up for "head" (by multiple @'s), and "p" (by single @).

4) Preprocessing texts for manual annotation by assigning lexical analyses to word tokens. The SUC project gratefully acknowledges that the University of Helsinki morphological analyzer SWETWOL + transduction provided the lexical analyses of 63% of the SUC vocabulary (67376/107101), and 25,5% of the corpus (297648/1166902). 20% of the SUC vocabulary (21751/107101) was unanalyzed by SWETWOL. An expanded version of the wordlist lexicon created by Eva Ejerhed in combination with special lexical lists (namelist, abbreviations) and lookup tools by Magnus Åström provided lexical analyses of 37% of the SUC vocabulary (39725/107101), and 74,5% of the corpus (869254/1166902). The SUC project also acknowledges that the format here called SUC format is an adaptation and extension by Åström of the format of SWETWOL output. The SUC format differs in using the SUC tagset, in placing lemmas last and making them case sensitive, and in the possible inclusion of frequencies of analyses. It has proved to be a flexible format, useful for representing corpora as well as lexicons.

5) Manual annotation in which morphosyntactic analyses were assigned to word tokens. Annotators were given text in which each word had been assigned one or more analyses, and they were instructed to select the appropriate analysis by marking it with their inital. If no analysis that was offered was appropriate, they were instructed to mark no analysis, leaving unmarked words for postprocessing. When annotators noted problems with the lexical analyses, or experienced uncertainty, they were instructed to note observations and questions in a problem file associated with each file that was annotated, and problem files were checked, and questions in them were answered by persons supervising the annotation process at SU and UmU. Annotators were asked to do no editing of the data files other than that of marking appropriate analyses. Annotation was carried out on a variety of computers, and using a variety of editors. Files that had been manually annotated were returned to a designated directory at UmU.</p>

<p> The following statistics shows how many words each annotator annotated, and the division of labour between SU and UmU with respect to annotation:
292327 O = Ola Wennstedt, UmU
276305 J = Joana Arnesson, UmU
166332 C = Cecilia Lyckow, SU
121528 F = Fredrick Backman, UmU
115410 U = Ulrika Kvist, SU
66833 E = Eva Ejerhed, UmU
64013 S = Carin Svensson-Lindberg, SU
38544 L = Janne Lindberg, SU
19884 A = Anna-Lena Wiklund, UmU

```
        2395 B = Britt Hartmann, SU
        2291 R = Rolf Sandberg, UmU
        801 T = Magnus Åström, UmU
        166 G = Gunnel Källgren, SU
        </p>
        <p>The final stage of the project covered the period 1995/96-1996/97 and it consisted of the
        following two parts: 1) Post-processing of manually annotated texts included analyzing
        remaining unanalyzed words, adding required but missing analyses to analyzed words, marking
        unmarked words, editing lemma forms, correcting errors in tokenization, in tagging, and in
        lemma forms, doing consistency checks using automatic support, and finally processing
        replacement texts of approximately 35 000 tokens for texts that had to be withdrawn from the
        corpus for copyright reasons.
        2) Preparing SUC 1.0 for distribution in two formats. This phase did not only consist of fully
        automatic conversion of the corpus from SUC format to SGML format (by the program
        suc2sgml by Fredrick Backman), it also involved implementing the distinction between "head"
        and "p" in all files, correcting bugs in the annotated corpus that had been discovered when using
        it for training purposes, introducing automatically derived extra markup of document structural
        elements and the elements "s", "num", and "foreign", and implementing file headers. The SUC
        project gratefully acknowledges the assistance of Daniel Ridings in this phase of the work. He
        wrote the ENT and DTD for the SGML version of the corpus, and he and Pernilla Danielsson
        assisted in defining a suitable text structure for the corpus. </p>
        <p>All of the 500 files of the corpus have passed through the following eight versions:
        *.text.raw.7bit, *.UT, *.IN, *.ep0, *.ep1, *ep2, *.suc1a, and *.suc1b. The formal correctness of
        the *.suc1a files in the distributed version have been validated by the script suc-check-format,
        and the formal correctness of the *.suc1b files have been validated by the nsgmls parser. </p>
</projectDesc>
<projectDesc>
        <p>Project description 2. The period 1995-2006.</p>
        <p>The early stages of the SUC-project are amply accounted for in the SUC1.0 project
        description above. The present project description therefore mainly concerns the final stages
        particular to SUC 2.0. </p>
        <p>One important difference between SUC1.0 and SUC 2.0 is that SUC 2.0 has a large number
        of functionally interpreted TEI-tags, with attribute values selected by human annotators. </p>
        <p> An extensive report on the preliminary work on TEI-tags can be found in the documentation
        of the Stockholm-Umeå corpus written by Gunnel Källgren. Her work on the construction of an
        SGML tag set for SUC began during the early stages of the TEI-project, i.e. before the extensive
        TEI Guidelines (P2 and P3) were published in 1993 and 1994. </p>
        <p>A tentative set of instructions for the use of TEI-tags in SUC was issued in manuscript form
        (TEI-manual for SUC) before the publication of TEI P3. They were tried out on a few SUC-files
        by Gunnar Eriksson, who also took part in the construction of the tag set, and wrote the SUC-
        dtd. The application of some of the tags proved to be intricate and time consuming, and it was
        decided that they should not be included in the first edition of the TEI-marked corpus. This
        pertains particularly to ”q” and ”quote”, which were eventually used only in a few example files
        in SUC 2.0.</p>
        <p>A TEI-conformant document (P3) not only needs SGML-tagged text, but also a document
        header containing some bibliographical information about the text(s) contained in the document.
        Relevant information about the SUC text samples had been gathered along with the work to
        collect them. (cf. GK-documentation). This information had been stored in a bibliographical
        database constructed by Bo Nathorst-Westfelt in 1990. The bibliographic information was
        subsequently squeezed into one of the alternate structures described in the TEI guidelines. A
        DTD for this bibliographic header was written by Gunnel Källgren. </p>
        <p>However, during the final preparation of SUC 1.0, a somewhat different structure for the
        document header was chosen, and we now find it wise to use more or less the same in SUC 2.0,
        one small difference being that the "title" element in SUC 2.0 has an added attribute with values
        a, j or m (article, journal, monograph). This entails no modification of the TEI-header. </p>
        <p>A TEI-conformant corpus (P3) requires a corpus header. A corpus header was sketched by
        Gunnel Källgren, and was subsequently completed in Umeå for SUC 1.0 (cf project description
        for SUC 1.0 above). </p>
        <p>The final application of the elaborate SGML-tags to the SUC files proceeded as follows.
        </p>
```

<p>All of the 500 files of the corpus were automatically converted to suc1c-format by Fredrik Backman in Umeå in 1996. The suc1c-format is an sgml format with morphosyntactic SUC-tags and tei-tags such as name, abbr, foreign inserted without attributes. This format provided a suitable base for manual tei-annotation using the SUC-dtd written by Gunnar Eriksson. The annotators checked and completed tags with Author/Editor 3.0 for Windows, a SoftQuad program that supports SGML-tagging. The annotated files were validated in A/E as far as concerns structural markup. The annotation process also included checking the files against hard copies of the printed texts. Some discrepancies were found and comments were added where changes had to be made. </p>

<p>In March 1997 Daniel Ridings generously provided sgml-libraries with entity declarations and dtd:s for the parole-version of suc-files, as well as perl programs for converting the tei-marked suc1c-files to TEI2 document files, with bibliographic headers and parole-tags. Slightly modified such programs have later been used in Stockholm to create documents with bibliographic headers and suc-tags. </p>

<p>During the summer of 1997 Sofia Gustafsson-Capkova, who had also done a lot of tei-annotating in 1996 began the work of checking the appropriateness of the teimarking, which requires good judgment and a discriminating eye. She now had to read and handle all the comment lines, and remove them. Christina Ericson and Maria Arnstad also did some of this work, which continued during the autumn term of 1997, and spring 1998. In august 1997 SUC 1.0 was released. From then on we were able to compare the Parole version of SUC 2.0 files with the parole-version SUC 1.0b. Although SUC 2.0 has a richer structural markup than SUC 1.0, and a sligthly different reference system, the words and their morphosyntactic tags in the corpus are of course essentially the same. </p>

<p>The formal correctness of the SUC 2.0 files has been validated by the nsgmls parser. </p>

<p>In January 1999 the SUC 2.0 work received a terrible blow, when the project leader (GK) became seriously ill. She died in February, and SUC 2.0 seemed further than ever from release.</p>

<p>However, after a while professor Benny Brodda volunteered to act as project leader. Thanks to this and to a number of devoted graduate students and a generous attitude at the department of linguistics, SUC 2.0 is still in progress. </p>

<p>In the period between 1997 and 2002, some clear errors have been corrected, not only in the TEI-tag markup, but also in the morphosyntactic tags and base forms in the tagged SUC-files. Most of these are accidental, and are presumably caused by an annotator just pressing the wrong button. Many errors have been pointed out by people using SUC 1.0 for some purpose, which is gratefully acknowledged. Just a few names will be mentioned here. Janne Lindberg who is studying word groups has commented on many doubtful cases. Nikolaj Lindberg and Gunnar Eriksson have from time to time sent quiet (or more boisterous) comments on peculiar tags or base forms, and Johan Carlberger at NADA has pointed out several errors and inconsistencies.</p>

<p>After a rather long period of silent progress with Britt Hartmann working mainly with preparing the text materials not included in SUC1.0 and SUC2.0 but subject to the same license agreement, Sofia Gustafson-Capková and Britt Hartmann 2006 decided to get the corpus published, even though all parts of the originally planned content were not finished. The decision was based on a need from users to be able to correctly refer to the SUC 2.0 version.</p>

</projectDesc>

<samplingDecl>

<p>The SUC 2.0 corpus consists of text samples grouped in 500 files of an average size of 2065 word tokens per file (1032499/500). A file is either one entire article, or one excerpt from a longer text, in which case the file constitutes a single text with a single bibliographic identity, or a file is a composite of several shorter texts, each having a distinct bibliographic identity. There are 1040 bibliographically distinct texts in SUC 2.0 (the same texts as in SUC 1.0).[19] </p>

<p>The language of the texts in the corpus is modern Swedish prose. The texts must have been first published in 1990 or later. Acquisition of texts for the SUC project took place between 1990 and 1994. In order to get a balanced corpus of informative prose as well as fiction, the texts have been selected and classified according to criteria corresponding to those of the Brown and LOB corpora. The taxonomy of text types used for SUC is given in "classDecl" below. The selection

[19] The classification of one of the text samples has for formal reasons been changed from kl to kk. If you use both SUC1.0 and SUC 2.0, you must remember that the text in'kl20' of SUC1.0 is found in 'kk82' of SUC 2.0.

```
        of text samples has been done at random, but as far as possible, coherent and delimited stretches
        of text have been chosen. </p>
        <p>The corpus is based mostly on texts that were made available in machine readable form, and
        relies only to a lesser extent on scanning or typing in of material. </p>
</samplingDecl>
<editorialDecl>
        <correction>
                <p>Clear typographical errors in the raw texts have been corrected in the annotated
                versions of the corpus texts, without recording the corrections.</p>
        </correction>
        <quotation>
                <p>All quotation marks in the raw texts are retained in the annotated versions.</p>
        </quotation>
        <hyphenation>
                <p>All 'soft' hyphens at the ends of lines in the raw text have been removed. Remaining
                hyphens in words should be 'hard' hyphens. </p>
        </hyphenation>
        <segmentation>
                <p>The object language elements at the level of tokens are word tokens and punctuation
                tokens. These elements are formally distinguished by the tags "w" and "c" in the PAROLE
                format or "w" and "d" in the SUC-format (d for delimiter) respectively. The DL used in
                SUC 1.0 (suc1a) is not used in SUC 2.0. </p>
                <p>All periods that occur in abbreviations have been included in the abbreviation token,
                rather than tokenized separately. Remaining free periods should be sentence final
                delimiters.</p>
                <p>Text structural elements are itemized in "tagsDecl", where frequencies are also
                provided. In general the tags used are 'canonical' TEI tags. A more elaborate description of
                the tags and the rules for their application can be found in the written documentation of the
                corpus.</p>
        </segmentation>
        <interpretation>
                <p>In a "tagUsage" declaration in the corpus header section "tagsDecl"the frequency of
                each morphosyntactic description in SUC 2.0 is provided, and one linguistic example.</p>
                <p>Information about values of the msd attribute of w elements in the PAROLE format is
                provided in a separate file in the directory MANUAL.</p>
                <p>Information about tags for functionally interpreted units in SUC 2.0 documents is
                provided in the written documentation of the corpus, section 4.3.</p>
        </interpretation>
</editorialDecl>
<tagsDecl>
        <!--structural tags common to SUC1.0 and SUC 2.0. -->
        <tagUsage gi="group" occurs=170>Comprises a composite text file</tagUsage>
        <tagUsage gi="text" occurs=1210>Occurs 170 times for composite suc-texts, implied by "group", and 1040 times for
        unitary suc-texts, implied by "body".</tagUsage>
        <tagUsage gi="body" occurs=1040>Occurrence implied by unitary suc-text</tagUsage>
        <tagUsage gi="div" occurs=3581>Occurs 3550 times implied by "head" and 3 times where a paragraph follows a
        "byline" at the end of a text and 28 times where a paragraph precedes a "byline" at the beginning of a text</tagUsage>
        <tagUsage gi="head" occurs=3550>Heading preceding a paragraph</tagUsage>
        <tagUsage gi="p" occurs=23939>Paragraph</tagUsage>
        <tagUsage gi="s" occurs=74161>S-unit, occurs in paragraphs, headings, bylines, list labels and list items.</tagUsage>
        <tagUsage gi="w" occurs=1032494>Word token</tagUsage>
        <tagUsage gi="c" occurs=134099>Punctuation token</tagUsage>
        <tagUsage gi="num" occurs=18125>Encloses one or more cardinal or ordinal numbers</tagUsage>
        <tagUsage gi="foreign" occurs=1180>Encloses one or more foreign, non-Swedish words. Attribute lang specified in
        SUC 2.0</tagUsage>
        <!--SUC 2.0 tags for functionally interpreted structures, in alphabetical order-->
        <tagUsage gi="abbr" occurs=8400>Abbreviation</tagUsage>
        <tagUsage gi="byline" occurs=654>Byline, i.e. information about author, source etc. mainly in newspaper
        articles</tagUsage>
        <tagUsage gi="distinct" occurs=383>Word or sequence of words in non-standard Swedish</tagUsage>
        <tagUsage gi="item" occurs=971>List item</tagUsage>
        <tagUsage gi="l" occurs=82>Line of poem or verse supplied for s-unit in line groups.</tagUsage>
        <tagUsage gi="label" occurs=508>List item's label</tagUsage>
        <tagUsage gi="lg" occurs=17>Line group of poetic or verse material</tagUsage>
        <tagUsage gi="list" occurs=202>List</tagUsage>
```

```
<tagUsage gi="mentioned" occurs=157>Word or sequence of words referred to rather than used</tagUsage>
<tagUsage gi="name" occurs=34189>Names in SUC 2.0 are subclassified</tagUsage>
<tagUsage gi="ref" occurs=735>Free format bibliographic citation</tagUsage>
<!--some examples of the use of quote and q can be found in ja07, ja08, ja16, ja23, ja26, jf02 and jf03-->
<!-- The occurrence numbers are not significant. They would be very much higher if we had used q and quote in the whole corpus.-->
<tagUsage gi="q" occurs=1>Direct speech</tagUsage>
<tagUsage gi="quote" occurs=17>Quotation. Not attributed to the author of the suc-text</tagUsage>
<tagUsage gi ="w" occurs=59100 msd="RG0S">example: inte</tagUsage>
<rendition msd="RG0S">AB</rendition>
<tagUsage gi ="w" occurs=2320 msd="RG0A">example: kl</tagUsage>
<rendition msd="RG0A">AB AN</rendition>
<tagUsage gi ="w" occurs=3788 msd="RGCS">example: oftare</tagUsage>
<rendition msd="RGCS">AB KOM</rendition>
<tagUsage gi ="w" occurs=14883 msd="RGPS">example: ofta</tagUsage>
<rendition msd="RGPS">AB POS</rendition>
<tagUsage gi ="w" occurs=35 msd="RG0C">example: upp-</tagUsage>
<rendition msd="RG0C">AB SMS</rendition>
<tagUsage gi ="w" occurs=2037 msd="RGSS">example: oftast</tagUsage>
<rendition msd="RGSS">AB SUV</rendition>
<tagUsage gi ="w" occurs=1 msd="D0@00@A">example: d</tagUsage>
<rendition msd="D0@00@A">DT AN</rendition>
<tagUsage gi ="w" occurs=49 msd="DF@MS@S">example: denne</tagUsage>
<rendition msd="DF@MS@S">DT MAS SIN DEF</rendition>
<tagUsage gi ="w" occurs=4 msd="DI@MS@S">example: samme</tagUsage>
<rendition msd="DI@MS@S">DT MAS SIN IND</rendition>
<tagUsage gi ="w" occurs=5198 msd="DF@NS@S">example: det</tagUsage>
<rendition msd="DF@NS@S">DT NEU SIN DEF</rendition>
<tagUsage gi ="w" occurs=8861 msd="DI@NS@S">example: ett</tagUsage>
<rendition msd="DI@NS@S">DT NEU SIN IND</rendition>
<tagUsage gi ="w" occurs=197 msd="D0@NS@S">example: allt</tagUsage>
<rendition msd="D0@NS@S">DT NEU SIN IND/DEF</rendition>
<tagUsage gi ="w" occurs=6899 msd="DF@0P@S">example: de</tagUsage>
<rendition msd="DF@0P@S">DT UTR/NEU PLU DEF</rendition>
<tagUsage gi ="w" occurs=1281 msd="DI@0P@S">example: några</tagUsage>
<rendition msd="DI@0P@S">DT UTR/NEU PLU IND</rendition>
<tagUsage gi ="w" occurs=1175 msd="D0@0P@S">example: alla</tagUsage>
<rendition msd="D0@0P@S">DT UTR/NEU PLU IND/DEF</rendition>
<tagUsage gi ="w" occurs=9 msd="DF@0S@S">example: vardera</tagUsage>
<rendition msd="DF@0S@S">DT UTR/NEU SIN DEF</rendition>
<tagUsage gi ="w" occurs=724 msd="DI@0S@S">example: varje</tagUsage>
<rendition msd="DI@0S@S">DT UTR/NEU SIN IND</rendition>
<tagUsage gi ="w" occurs=864 msd="DI@00@S">example: samma</tagUsage>
<rendition msd="DI@00@S">DT UTR/NEU SIN/PLU IND</rendition>
<tagUsage gi ="w" occurs=10943 msd="DF@US@S">example: den</tagUsage>
<rendition msd="DF@US@S">DT UTR SIN DEF</rendition>
<tagUsage gi ="w" occurs=19627 msd="DI@US@S">example: en</tagUsage>
<rendition msd="DI@US@S">DT UTR SIN IND</rendition>
<tagUsage gi ="w" occurs=191 msd="D0@US@S">example: all</tagUsage>
<rendition msd="D0@US@S">DT UTR SIN IND/DEF</rendition>
<tagUsage gi ="w" occurs=9323 msd="RH0S">example: när</tagUsage>
<rendition msd="RH0S">HA</rendition>
<tagUsage gi ="w" occurs=100 msd="DH@NS@S">example: vilket</tagUsage>
<rendition msd="DH@NS@S">HD NEU SIN IND</rendition>
<tagUsage gi ="w" occurs=216 msd="DH@0P@S">example: vilka</tagUsage>
<rendition msd="DH@0P@S">HD UTR/NEU PLU IND</rendition>
<tagUsage gi ="w" occurs=213 msd="DH@US@S">example: vilken</tagUsage>
<rendition msd="DH@US@S">HD UTR SIN IND</rendition>
<tagUsage gi ="w" occurs=12752 msd="PH@000@S">example: som</tagUsage>
<rendition msd="PH@000@S">HP - - -</rendition>
<tagUsage gi ="w" occurs=2541 msd="PH@NS0@S">example: vad</tagUsage>
<rendition msd="PH@NS0@S">HP NEU SIN IND</rendition>
<tagUsage gi ="w" occurs=1 msd="PH@NS0@C">example: vad-</tagUsage>
<rendition msd="PH@NS0@C">HP NEU SIN IND SMS</rendition>
<tagUsage gi ="w" occurs=223 msd="PH@0P0@S">example: vilka</tagUsage>
<rendition msd="PH@0P0@S">HP UTR/NEU PLU IND</rendition>
<tagUsage gi ="w" occurs=426 msd="PH@US0@S">example: vem</tagUsage>
<rendition msd="PH@US0@S">HP UTR SIN IND</rendition>
<tagUsage gi ="w" occurs=160 msd="PE@000@S">example: vars</tagUsage>
<rendition msd="PE@000@S">HS DEF</rendition>
<tagUsage gi ="w" occurs=12907 msd="CIS">example: att</tagUsage>
<rendition msd="CIS">IE</rendition>
<tagUsage gi ="w" occurs=1593 msd="I">example: javisst</tagUsage>
<rendition msd="I">IN</rendition>
<tagUsage gi ="w" occurs=74 msd="AQ00000A">example: St</tagUsage>
```

```
<rendition msd="AQ00000A">JJ AN</rendition>
<tagUsage gi ="w" occurs=3 msd="AQC00G0S">example: svagares</tagUsage>
<rendition msd="AQC00G0S">JJ KOM UTR/NEU SIN/PLU IND/DEF GEN</rendition>
<tagUsage gi ="w" occurs=4277 msd="AQC00N0S">example: större</tagUsage>
<rendition msd="AQC00N0S">JJ KOM UTR/NEU SIN/PLU IND/DEF NOM</rendition>
<tagUsage gi ="w" occurs=2 msd="AQC0000C">example: äldre-</tagUsage>
<rendition msd="AQC0000C">JJ KOM UTR/NEU SIN/PLU IND/DEF SMS</rendition>
<tagUsage gi ="w" occurs=69 msd="AQPMSGDS">example: unges</tagUsage>
<rendition msd="AQPMSGDS">JJ POS MAS SIN DEF GEN</rendition>
<tagUsage gi ="w" occurs=906 msd="AQPMSNDS">example: unge</tagUsage>
<rendition msd="AQPMSNDS">JJ POS MAS SIN DEF NOM</rendition>
<tagUsage gi ="w" occurs=181 msd="AQPNSN0S">example: eget</tagUsage>
<rendition msd="AQPNSN0S">JJ POS NEU SIN IND/DEF NOM</rendition>
<tagUsage gi ="w" occurs=1 msd="AQPNSGIS">example: angelägets</tagUsage>
<rendition msd="AQPNSGIS">JJ POS NEU SIN IND GEN</rendition>
<tagUsage gi ="w" occurs=9101 msd="AQPNSNIS">example: stort</tagUsage>
<rendition msd="AQPNSNIS">JJ POS NEU SIN IND NOM</rendition>
<tagUsage gi ="w" occurs=58 msd="AQP0PG0S">example: vuxnas</tagUsage>
<rendition msd="AQP0PG0S">JJ POS UTR/NEU PLU IND/DEF GEN</rendition>
<tagUsage gi ="w" occurs=19513 msd="AQP0PN0S">example: stora</tagUsage>
<rendition msd="AQP0PN0S">JJ POS UTR/NEU PLU IND/DEF NOM</rendition>
<tagUsage gi ="w" occurs=1354 msd="AQP0PNIS">example: flera</tagUsage>
<rendition msd="AQP0PNIS">JJ POS UTR/NEU PLU IND NOM</rendition>
<tagUsage gi ="w" occurs=19 msd="AQP0SGDS">example: allmännas</tagUsage>
<rendition msd="AQP0SGDS">JJ POS UTR/NEU SIN DEF GEN</rendition>
<tagUsage gi ="w" occurs=11822 msd="AQP0SNDS">example: nya</tagUsage>
<rendition msd="AQP0SNDS">JJ POS UTR/NEU SIN DEF NOM</rendition>
<tagUsage gi ="w" occurs=3214 msd="AQP00N0S">example: framtida</tagUsage>
<rendition msd="AQP00N0S">JJ POS UTR/NEU SIN/PLU IND/DEF NOM</rendition>
<tagUsage gi ="w" occurs=19 msd="AQP00NIS">example: rätt</tagUsage>
<rendition msd="AQP00NIS">JJ POS UTR/NEU SIN/PLU IND NOM</rendition>
<tagUsage gi ="w" occurs=6 msd="AQP0000C">example: utrikes-</tagUsage>
<rendition msd="AQP0000C">JJ POS UTR/NEU - - SMS</rendition>
<tagUsage gi ="w" occurs=425 msd="AQPUSN0S">example: egen</tagUsage>
<rendition msd="AQPUSN0S">JJ POS UTR SIN IND/DEF NOM</rendition>
<tagUsage gi ="w" occurs=2 msd="AQPUSGIS">example: enskilds</tagUsage>
<rendition msd="AQPUSGIS">JJ POS UTR SIN IND GEN</rendition>
<tagUsage gi ="w" occurs=19533 msd="AQPUSNIS">example: stor</tagUsage>
<rendition msd="AQPUSNIS">JJ POS UTR SIN IND NOM</rendition>
<tagUsage gi ="w" occurs=35 msd="AQPU000C">example: låg-</tagUsage>
<rendition msd="AQPU000C">JJ POS UTR - - SMS</rendition>
<tagUsage gi ="w" occurs=1 msd="AQSMSGDS">example: äldstes</tagUsage>
<rendition msd="AQSMSGDS">JJ SUV MAS SIN DEF GEN</rendition>
<tagUsage gi ="w" occurs=60 msd="AQSMSNDS">example: yngste</tagUsage>
<rendition msd="AQSMSNDS">JJ SUV MAS SIN DEF NOM</rendition>
<tagUsage gi ="w" occurs=279 msd="AQS0PNDS">example: flesta</tagUsage>
<rendition msd="AQS0PNDS">JJ SUV UTR/NEU PLU DEF NOM</rendition>
<tagUsage gi ="w" occurs=13 msd="AQS0PNIS">example: flest</tagUsage>
<rendition msd="AQS0PNIS">JJ SUV UTR/NEU PLU IND NOM</rendition>
<tagUsage gi ="w" occurs=2094 msd="AQS00NDS">example: största</tagUsage>
<rendition msd="AQS00NDS">JJ SUV UTR/NEU SIN/PLU DEF NOM</rendition>
<tagUsage gi ="w" occurs=447 msd="AQS00NIS">example: störst</tagUsage>
<rendition msd="AQS00NIS">JJ SUV UTR/NEU SIN/PLU IND NOM</rendition>
<tagUsage gi ="w" occurs=53137 msd="CCS">example: och</tagUsage>
<rendition msd="CCS">KN</rendition>
<tagUsage gi ="w" occurs=125 msd="CCA">example: &</tagUsage>
<rendition msd="CCA">KN AN</rendition>
<tagUsage gi ="c" occurs=68038 msd="FE">example: .</tagUsage>
<rendition msd="FE">MAD</rendition>
<tagUsage gi ="c" occurs=49898 msd="FI">example: -</tagUsage>
<rendition msd="FI">MID</rendition>
<tagUsage gi ="w" occurs=208 msd="NC000@0S">example: fjol</tagUsage>
<rendition msd="NC000@0S">NN - - - -</rendition>
<tagUsage gi ="w" occurs=2905 msd="NC000@0A">example: kr</tagUsage>
<rendition msd="NC000@0A">NN AN</rendition>
<tagUsage gi ="w" occurs=3 msd="NCN00@0S">example: orda</tagUsage>
<rendition msd="NC000@0C">NN - - - SMS</rendition>
<tagUsage gi ="w" occurs=354 msd="NCNPG@DS">example: barnens</tagUsage>
<rendition msd="NCNPG@DS">NN NEU PLU DEF GEN</rendition>
<tagUsage gi ="w" occurs=4311 msd="NCNPN@DS">example: problemen</tagUsage>
<rendition msd="NCNPN@DS">NN NEU PLU DEF NOM</rendition>
<tagUsage gi ="w" occurs=222 msd="NCNPG@IS">example: seklers</tagUsage>
<rendition msd="NCNPG@IS">NN NEU PLU IND GEN</rendition>
<tagUsage gi ="w" occurs=14752 msd="NCNPN@IS">example: intressen</tagUsage>
<rendition msd="NCNPN@IS">NN NEU PLU IND NOM</rendition>
```

```
<tagUsage gi ="w" occurs=1648 msd="NCNSG@DS">example: landets</tagUsage>
<rendition msd="NCNSG@DS">NN NEU SIN DEF GEN</rendition>
<tagUsage gi ="w" occurs=18086 msd="NCNSN@DS">example: arbetet</tagUsage>
<rendition msd="NCNSN@DS">NN NEU SIN DEF NOM</rendition>
<tagUsage gi ="w" occurs=432 msd="NCNSG@IS">example: lands</tagUsage>
<rendition msd="NCNSG@IS">NN NEU SIN IND GEN</rendition>
<tagUsage gi ="w" occurs=28135 msd="NCNSN@IS">example: intresse</tagUsage>
<rendition msd="NCNSN@IS">NN NEU SIN IND NOM</rendition>
<tagUsage gi ="w" occurs=202 msd="NCN00@0C">example: pris-</tagUsage>
<rendition msd="NCN00@0S">NN NEU - - -</rendition>
<tagUsage gi ="w" occurs=62 msd="NC000@0C">example: trav-</tagUsage>
<rendition msd="NCN00@0C">NN NEU - - SMS</rendition>
<tagUsage gi ="w" occurs=85 msd="NCU00@0S">example: dags</tagUsage>
<rendition msd="NCU00@0S">NN UTR - - -</rendition>
<tagUsage gi ="w" occurs=1130 msd="NCUPG@DS">example: myndigheternas</tagUsage>
<rendition msd="NCUPG@DS">NN UTR PLU DEF GEN</rendition>
<tagUsage gi ="w" occurs=10746 msd="NCUPN@DS">example: dagarna</tagUsage>
<rendition msd="NCUPN@DS">NN UTR PLU DEF NOM</rendition>
<tagUsage gi ="w" occurs=444 msd="NCUPG@IS">example: timmars</tagUsage>
<rendition msd="NCUPG@IS">NN UTR PLU IND GEN</rendition>
<tagUsage gi ="w" occurs=33504 msd="NCUPN@IS">example: kronor</tagUsage>
<rendition msd="NCUPN@IS">NN UTR PLU IND NOM</rendition>
<tagUsage gi ="w" occurs=4381 msd="NCUSG@DS">example: kommunens</tagUsage>
<rendition msd="NCUSG@DS">NN UTR SIN DEF GEN</rendition>
<tagUsage gi ="w" occurs=43344 msd="NCUSN@DS">example: tiden</tagUsage>
<rendition msd="NCUSN@DS">NN UTR SIN DEF NOM</rendition>
<tagUsage gi ="w" occurs=707 msd="NCUSG@IS">example: persons</tagUsage>
<rendition msd="NCUSG@IS">NN UTR SIN IND GEN</rendition>
<tagUsage gi ="w" occurs=69134 msd="NCUSN@IS">example: plats</tagUsage>
<rendition msd="NCUSN@IS">NN UTR SIN IND NOM</rendition>
<tagUsage gi ="w" occurs=514 msd="NCU00@0C">example: bostads-</tagUsage>
<rendition msd="NCU00@0C">NN UTR - - SMS</rendition>
<tagUsage gi ="c" occurs=16163 msd="FP">example: (</tagUsage>
<rendition msd="FP">PAD</rendition>
<tagUsage gi ="w" occurs=2 msd="AF00000A">example: adj</tagUsage>
<rendition msd="AF00000A">PC AN</rendition>
<tagUsage gi ="w" occurs=6 msd="AF0MSGDS">example: knivhuggnes</tagUsage>
<rendition msd="AF0MSGDS">PC PRF MAS SIN DEF GEN</rendition>
<tagUsage gi ="w" occurs=96 msd="AF0MSNDS">example: avlidne</tagUsage>
<rendition msd="AF0MSNDS">PC PRF MAS SIN DEF NOM</rendition>
<tagUsage gi ="w" occurs=2294 msd="AF0NSNIS">example: taget</tagUsage>
<rendition msd="AF0NSNIS">PC PRF NEU SIN IND NOM</rendition>
<tagUsage gi ="w" occurs=22 msd="AF00PG0S">example: deporterades</tagUsage>
<rendition msd="AF00PG0S">PC PRF UTR/NEU PLU IND/DEF GEN</rendition>
<tagUsage gi ="w" occurs=4153 msd="AF00PN0S">example: intresserade</tagUsage>
<rendition msd="AF00PN0S">PC PRF UTR/NEU PLU IND/DEF NOM</rendition>
<tagUsage gi ="w" occurs=11 msd="AF00SGDS">example: förflutnas</tagUsage>
<rendition msd="AF00SGDS">PC PRF UTR/NEU SIN DEF GEN</rendition>
<tagUsage gi ="w" occurs=1470 msd="AF00SNDS">example: kallade</tagUsage>
<rendition msd="AF00SNDS">PC PRF UTR/NEU SIN DEF NOM</rendition>
<tagUsage gi ="w" occurs=2 msd="AF0USGIS">example: underordnads</tagUsage>
<rendition msd="AF0USGIS">PC PRF UTR SIN IND GEN</rendition>
<tagUsage gi ="w" occurs=4815 msd="AF0USNIS">example: skriven</tagUsage>
<rendition msd="AF0USNIS">PC PRF UTR SIN IND NOM</rendition>
<tagUsage gi ="w" occurs=8 msd="AP000G0S">example: sovandes</tagUsage>
<rendition msd="AP000G0S">PC PRS UTR/NEU SIN/PLU IND/DEF GEN</rendition>
<tagUsage gi ="w" occurs=5590 msd="AP000N0S">example: liknande</tagUsage>
<rendition msd="AP000N0S">PC PRS UTR/NEU SIN/PLU IND/DEF NOM</rendition>
<tagUsage gi ="w" occurs=11722 msd="QS">example: ut</tagUsage>
<rendition msd="QS">PL</rendition>
<tagUsage gi ="w" occurs=3948 msd="NP00G@0S">example: Stockholms</tagUsage>
<rendition msd="NP00G@0S">PM GEN</rendition>
<tagUsage gi ="w" occurs=38018 msd="NP00N@0S">example: Europa</tagUsage>
<rendition msd="NP00N@0S">PM NOM</rendition>
<tagUsage gi ="w" occurs=39 msd="NP000@0C">example: Göteborgs-</tagUsage>
<rendition msd="NP000@0C">PM SMS</rendition>
<tagUsage gi ="w" occurs=67 msd="PF@MS0@S">example: denne</tagUsage>
<rendition msd="PF@MS0@S">PN MAS SIN DEF SUB/OBJ</rendition>
<tagUsage gi ="w" occurs=16311 msd="PF@NS0@S">example: detta</tagUsage>
<rendition msd="PF@NS0@S">PN NEU SIN DEF SUB/OBJ</rendition>
<tagUsage gi ="w" occurs=2885 msd="PI@NS0@S">example: ingenting</tagUsage>
<rendition msd="PI@NS0@S">PN NEU SIN IND SUB/OBJ</rendition>
<tagUsage gi ="w" occurs=1661 msd="PF@0PO@S">example: dem</tagUsage>
<rendition msd="PF@0PO@S">PN UTR/NEU PLU DEF OBJ</rendition>
<tagUsage gi ="w" occurs=3881 msd="PF@0PS@S">example: de</tagUsage>
```

<rendition msd="PF@0PS@S">PN UTR/NEU PLU DEF SUB</rendition>
<tagUsage gi ="w" occurs=587 msd="PF@0P0@S">example: dessa</tagUsage>
<rendition msd="PF@0P0@S">PN UTR/NEU PLU DEF SUB/OBJ</rendition>
<tagUsage gi ="w" occurs=1536 msd="PI@0P0@S">example: alla</tagUsage>
<rendition msd="PI@0P0@S">PN UTR/NEU PLU IND SUB/OBJ</rendition>
<tagUsage gi ="w" occurs=5810 msd="PF@00O@S">example: sig</tagUsage>
<rendition msd="PF@00O@S">PN UTR/NEU SIN/PLU DEF OBJ</rendition>
<tagUsage gi ="w" occurs=873 msd="PF@UPO@S">example: oss</tagUsage>
<rendition msd="PF@UPO@S">PN UTR PLU DEF OBJ</rendition>
<tagUsage gi ="w" occurs=4251 msd="PF@UPS@S">example: vi</tagUsage>
<rendition msd="PF@UPS@S">PN UTR PLU DEF SUB</rendition>
<tagUsage gi ="w" occurs=4176 msd="PF@USO@S">example: henne</tagUsage>
<rendition msd="PF@USO@S">PN UTR SIN DEF OBJ</rendition>
<tagUsage gi ="w" occurs=21118 msd="PF@USS@S">example: hon</tagUsage>
<rendition msd="PF@USS@S">PN UTR SIN DEF SUB</rendition>
<tagUsage gi ="w" occurs=3178 msd="PF@US0@S">example: denna</tagUsage>
<rendition msd="PF@US0@S">PN UTR SIN DEF SUB/OBJ</rendition>
<tagUsage gi ="w" occurs=4576 msd="PI@USS@S">example: man</tagUsage>
<rendition msd="PI@USS@S">PN UTR SIN IND SUB</rendition>
<tagUsage gi ="w" occurs=1327 msd="PI@US0@S">example: ingen</tagUsage>
<rendition msd="PI@US0@S">PN UTR SIN IND SUB/OBJ</rendition>
<tagUsage gi ="w" occurs=124433 msd="SPS">example: i</tagUsage>
<rendition msd="SPS">PP</rendition>
<tagUsage gi ="w" occurs=4 msd="SPA">example: inkl.</tagUsage>
<rendition msd="SPA">PP AN</rendition>
<tagUsage gi ="w" occurs=1 msd="PS@000@A">example: h:s</tagUsage>
<rendition msd="PS@000@A">PS AN</rendition>
<tagUsage gi ="w" occurs=1559 msd="PS@NS0@S">example: ditt</tagUsage>
<rendition msd="PS@NS0@S">PS NEU SIN DEF</rendition>
<tagUsage gi ="w" occurs=1789 msd="PS@0P0@S">example: sina</tagUsage>
<rendition msd="PS@0P0@S">PS UTR/NEU PLU DEF</rendition>
<tagUsage gi ="w" occurs=3082 msd="PS@000@S">example: hennes</tagUsage>
<rendition msd="PS@000@S">PS UTR/NEU SIN/PLU DEF</rendition>
<tagUsage gi ="w" occurs=3648 msd="PS@US0@S">example: sin</tagUsage>
<rendition msd="PS@US0@S">PS UTR SIN DEF</rendition>
<tagUsage gi ="w" occurs=8 msd="MC00G0S">example: fyras</tagUsage>
<rendition msd="MC00G0S">RG GEN</rendition>
<tagUsage gi ="w" occurs=276 msd="MCNSNIS">example: ett</tagUsage>
<rendition msd="MCNSNIS">RG NEU SIN IND NOM</rendition>
<tagUsage gi ="w" occurs=15361 msd="MC00N0S">example: fyra</tagUsage>
<rendition msd="MC00N0S">RG NOM</rendition>
<tagUsage gi ="w" occurs=83 msd="MC0000C">example: 1930-</tagUsage>
<rendition msd="MC0000C">RG SMS</rendition>
<tagUsage gi ="w" occurs=463 msd="MCUSNIS">example: en</tagUsage>
<rendition msd="MCUSNIS">RG UTR SIN IND NOM</rendition>
<tagUsage gi ="w" occurs=7 msd="MO00G0S">example: xii:s</tagUsage>
<rendition msd="MO00G0S">RO GEN</rendition>
<tagUsage gi ="w" occurs=1 msd="MOMSG0S">example: andres</tagUsage>
<rendition msd="MOMSG0S">RO MAS SIN IND/DEF GEN</rendition>
<tagUsage gi ="w" occurs=48 msd="MOMSN0S">example: förste</tagUsage>
<rendition msd="MOMSN0S">RO MAS SIN IND/DEF NOM</rendition>
<tagUsage gi ="w" occurs=2024 msd="MO00N0S">example: tredje</tagUsage>
<rendition msd="MO00N0S">RO NOM</rendition>
<tagUsage gi ="w" occurs=3 msd="MO0000C">example: andra-</tagUsage>
<rendition msd="MO0000C">RO SMS</rendition>
<tagUsage gi ="w" occurs=17106 msd="CSS">example: innan</tagUsage>
<rendition msd="CSS">SN</rendition>
<tagUsage gi ="w" occurs=2571 msd="XF">example: of</tagUsage>
<rendition msd="XF">UO</rendition>
<tagUsage gi ="w" occurs=82 msd="V@000A">example: jfr</tagUsage>
<rendition msd="V@000A">VB AN</rendition>
<tagUsage gi ="w" occurs=1337 msd="V@M0AS">example: tänk</tagUsage>
<rendition msd="V@M0AS">VB IMP AKT</rendition>
<tagUsage gi ="w" occurs=7 msd="V@M0SS">example: låtsas</tagUsage>
<rendition msd="V@M0SS">VB IMP SFO</rendition>
<tagUsage gi ="w" occurs=33514 msd="V@N0AS">example: kunna</tagUsage>
<rendition msd="V@N0AS">VB INF AKT</rendition>
<tagUsage gi ="w" occurs=4633 msd="V@N0SS">example: finnas</tagUsage>
<rendition msd="V@N0SS">VB INF SFO</rendition>
<tagUsage gi ="w" occurs=55 msd="V@SPAS">example: vare</tagUsage>
<rendition msd="V@SPAS">VB KON PRS AKT</rendition>
<tagUsage gi ="w" occurs=149 msd="V@SIAS">example: vore</tagUsage>
<rendition msd="V@SIAS">VB KON PRT AKT</rendition>
<tagUsage gi ="w" occurs=1 msd="V@SISS">example: funnes</tagUsage>
<rendition msd="V@SISS">VB KON PRT SFO</rendition>

```
    <tagUsage gi ="w" occurs=67968 msd="V@IPAS">example: kommer</tagUsage>
    <rendition msd="V@IPAS">VB PRS AKT</rendition>
    <tagUsage gi ="w" occurs=8540 msd="V@IPSS">example: används</tagUsage>
    <rendition msd="V@IPSS">VB PRS SFO</rendition>
    <tagUsage gi ="w" occurs=40825 msd="V@IIAS">example: kunde</tagUsage>
    <rendition msd="V@IIAS">VB PRT AKT</rendition>
    <tagUsage gi ="w" occurs=4108 msd="V@IISS">example: lyckades</tagUsage>
    <rendition msd="V@IISS">VB PRT SFO</rendition>
    <tagUsage gi ="w" occurs=20 msd="V@000C">example: läs-</tagUsage>
    <rendition msd="V@000C">VB SMS</rendition>
    <tagUsage gi ="w" occurs=11396 msd="V@IUAS">example: kommit</tagUsage>
    <rendition msd="V@IUAS">VB SUP AKT</rendition>
    <tagUsage gi ="w" occurs=2239 msd="V@IUSS">example: lyckats</tagUsage>
    <rendition msd="V@IUSS">VB SUP SFO</rendition>
    <tagUsage gi ="w" occurs=0 msd="QC">example:</tagUsage>
    <!--medtagen, ty den finns i SUC1-->
    <rendition msd="QC">PL SMS</rendition>
    <tagUsage gi ="w" occurs=0 msd="MCMSNDS">example:</tagUsage>
    <!--medtagen, ty den finns i SUC1-->
    <rendition msd="MCMSNDS">RG MAS SIN DEF NOM</rendition>
    <tagUsage gi ="w" occurs=0 msd="MC0SNDS">example:</tagUsage>
    <!--medtagen, ty den finns i SUC1-->
    <rendition msd="MC0SNDS">RG UTR/NEU SIN DEF NOM</rendition>
</tagsDecl>
<refsDecl>
    <p>The value of the id attribute of each of the 1040 text elements that make up the corpus is a
    unique identifying code for that text. The code is constructed from a two letter code for text type
    (see "classDecl") and a two digit consecutive enumeration within each text type. In the 170 cases
    where a text file comprises several unitary texts, these are ordered alphabetically by the addition
    of a lower case letter to the code. Thus, e.g. "text id=ca05a" uniquely identifies the first text in
    the fifth file of the text type CA, book reviews. </p>
    <p>In SUC 2.0, SUC format and PAROLE format, each s-unit (or l-unit) has a unique
    identification code made up from the value of the text id attribute and a three-digit running
    enumeration of the s-units (or l-units) within this text. All words and delimiters (word tokens and
    punctuation tokens) within each of the 500 files are consecutively numbered. Every word or
    delimiter thus has a unique identification code consisting of the id value of the s-unit and the
    value of the attribute n of the word or delimiter. The combined locators "s id=ca05a-004" and "w
    n=36" tell that the token numbered 36 in the file ca05 is a word in the fourth s-unit of the first
    text ca05a in the file. </p>
    <p>Other text structural elements in SUC 2.0, such as abbr, byline, distinct, foreign, head, list,
    mentioned, name, num or p are not numbered. Nor are any of the end tags. Since every item
    either belongs to an s-unit or l-unit or contains one or more such units, the s-unit identification
    code should give a sufficient reference base for linguistic purposes. [20]</p>
    <p>Bibliographic information is listed in the document header of each suc-file as the contents of
    a "listBibl" element which includes one or more "biblFull" elements, each concerning one text.
    (Cf. also Appendix E (or D) in the written documentation of the corpus).</p>
    <p>In SUC 2.0 (as distinct from SUC1.0) the title element in "biblFull" has been given a level
    attribute with value a, j or m. For a newspaper article, the first title element has level=a and
    contains the title of the article, while the second one has level=j and contains the title of the
    newspaper. If a text excerpted from a monograph has two title elements, the first with level=a
    contains the heading of a chapter, while the second with level=m contains the title of the
    book.</p>
</refsDecl>
<classDecl>
    <taxonomy id=SUC>
        <category id=SUC.Genre>
            <catDesc>Genre</catDesc>
            <category id=SUC.A>
                <catDesc>Press: Reportage</catDesc>
                <category id=SUC.AA>
                    <catDesc>Political</catDesc>
                    </category>
```

[20] As apparent from the BNC, it would in fact even be possible to do without word numbers.

```
        <category id=SUC.AB>
                <catDesc>Community</catDesc>
                </category>
        <category id=SUC.AC>
                <catDesc>Financial</catDesc>
                </category>
        <category id=SUC.AD>
                <catDesc>Cultural</catDesc>
                </category>
        <category id=SUC.AE>
                <catDesc>Sports</catDesc>
                </category>
        <category id=SUC.AF>
                <catDesc>Spot News</catDesc>
                </category>
</category>
<category id=SUC.B>
        <catDesc>Press: Editorial</catDesc>
        <category id=SUC.BA>
                <catDesc>Institutional</catDesc>
                </category>
        <category id=SUC.BB>
                <catDesc>Debate articles</catDesc>
                </category>
</category>
<category id=SUC.C>
        <catDesc>Press: Reviews</catDesc>
        <category id=SUC.CA>
                <catDesc>Books</catDesc>
                </category>
        <category id=SUC.CB>
                <catDesc>Films</catDesc>
                </category>
        <category id=SUC.CC>
                <catDesc>Art</catDesc>
                </category>
        <category id=SUC.CD>
                <catDesc>Theater</catDesc>
                </category>
        <category id=SUC.CE>
                <catDesc>Music</catDesc>
                </category>
        <category id=SUC.CF>
                <catDesc>Artists, shows</catDesc>
                </category>
        <category id=SUC.CG>
                <catDesc>Radio, TV</catDesc>
                </category>
</category>
<category id=SUC.E>
        <catDesc>Skills and Hobbies</catDesc>
        <category id=SUC.EA>
                <catDesc>Hobbies, amusements</catDesc>
                </category>
        <category id=SUC.EB>
                <catDesc>Society press</catDesc>
                </category>
        <category id=SUC.EC>
                <catDesc>Occupational and trade union press</catDesc>
        </category>
```

```
    <category id=SUC.ED>
        <catDesc>Religion</catDesc>
        </category>
</category>
<category id=SUC.F>
    <catDesc>Popular Lore</catDesc>
    <category id=SUC.FA>
        <catDesc>Humanities</catDesc>
        </category>
    <category id=SUC.FB>
        <catDesc>Behavioral sciences</catDesc>
        </category>
    <category id=SUC.FC>
        <catDesc>Social sciences</catDesc>
        </category>
    <category id=SUC.FD>
        <catDesc>Religion</catDesc>
        </category>
    <category id=SUC.FE>
        <catDesc>Complementary life styles</catDesc>
    </category>
    <category id=SUC.FF>
        <catDesc>History</catDesc>
        </category>
    <category id=SUC.FG>
        <catDesc>Health and medicine</catDesc>
        </category>
    <category id=SUC.FH>
        <catDesc>Natural science, technology</catDesc>
    </category>
    <category id=SUC.FJ>
        <catDesc>Politics</catDesc>
        </category>
    <category id=SUC.FK>
        <catDesc>Culture</catDesc>
        </category>
</category>
<category id=SUC.G>
    <catDesc>Biographies, essays</catDesc>
    <category id=SUC.GA>
        <catDesc>Biographies, memoirs</catDesc>
        </category>
    <category id=SUC.GB>
        <catDesc>Essays</catDesc>
        </category>
</category>
<category id=SUC.H>
    <catDesc>Miscellaneous</catDesc>
    <category id=SUC.HA>
        <catDesc>Government publications</catDesc>
    </category>
    <category id=SUC.HB>
        <catDesc>Municipal publications</catDesc>
    </category>
    <category id=SUC.HC>
        <catDesc>Financial reports, business</catDesc>
    </category>
    <category id=SUC.HD>
        <catDesc>Financial reports, non-profit organisations</catDesc>
    </category>
```

```
                    <category id=SUC.HE>
                        <catDesc>Internal publications, companies</catDesc>
                    </category>
                    <category id=SUC.HF>
                        <catDesc>University publications</catDesc>
                    </category>
                </category>
                <category id=SUC.J>
                    <catDesc>Learned and scientific writing</catDesc>
                    <category id=SUC.JA>
                        <catDesc>Humanities</catDesc>
                        </category>
                    <category id=SUC.JB>
                        <catDesc>Behavioral sciences</catDesc>
                    </category>
                    <category id=SUC.JC>
                        <catDesc>Social sciences</catDesc>
                        </category>
                    <category id=SUC.JD>
                        <catDesc>Religion</catDesc>
                        </category>
                    <category id=SUC.JE>
                        <catDesc>Technology</catDesc>
                        </category>
                    <category id=SUC.JF>
                        <catDesc>Mathematics</catDesc>
                        </category>
                    <category id=SUC.JG>
                        <catDesc>Medicine</catDesc>
                        </category>
                    <category id=SUC.JH>
                        <catDesc>Natural science, technology</catDesc>
                    </category>
                </category>
                <category id=SUC.K>
                    <catDesc>Imaginative prose</catDesc>
                    <category id=SUC.KK>
                        <catDesc>General fiction</catDesc>
                        </category>
                    <category id=SUC.KL>
                        <catDesc>Science fiction and mystery</catDesc>
                    </category>
                    <category id=SUC.KN>
                        <catDesc>Light reading</catDesc>
                        </category>
                    <category id=SUC.KR>
                        <catDesc>Humour</catDesc>
                        </category>
                </category>
            </category>
        </taxonomy>
    </classDecl>
</encodingDesc>
<profileDesc>
    <creation>2006-12-24</creation>
    <langUsage>
    <language id=se>Modern Swedish Prose</language>
    </langUsage>
    <langUsage>
        <language id=da>Danish</language>
```

```
         <language id=en>English</language>
         <language id=fr>French</language>
         <language id=de>German</language>
         <language id=es>Spanish</language>
         <language id=cs>Czech</language>
         <language id=la>Latin</language>
         <language id=el>Greek</language>
         <language id=fi>Finnish</language>
         <language id=no>Norwegian</language>
         <language id=is>Icelandic</language>
         <language id=it>Italian</language>
         <language id=ru>Russian</language>
         <language id=other>other</language>
      </langUsage>
   </profileDesc>
</teiHeader>
```