UPPSALA UNIVERSITY

Project ref. no. *LE3-4239*

Project title *SCARRIE Scandinavian Proof-reading Tools*

Deliverable number *DEL 3.1.3*

Deliverable title ***A Swedish Text Corpus for Generating Dictionaries***

Number of pages *20*

WP/Task responsible *Anna Sågvall Hein, Department of Linguistics, Uppsala University, Box 513, S-751 20 Uppsala, Sweden*
Email: anna@ling.uu.se

Author(s) *Bengt Dahlqvist*

EC Project Officer *Pierre-Paul Sondag*

Keywords *Swedish, text corpus, dictionary generation*

Abstract *The main objectives of Work Package 3 of the Scarrie project are to produce extensive word form dictionaries to be used in word error recognition and correction. This paper describes the creation of a text corpus of published newspaper article texts. The corpus consists of all the articles published in 1995 and 1996 by the two prominent Swedish newspapers Svenska Dagbladet and Upsala Nya Tidning. The process of storing, segmenting and classifying word forms in this corpus for the purpose of obtaining word lists is described in detail as well as other descriptive information pertaining to the corpus.*

Executive summary

The primary users of the Swedish version of the Scarrie tool will be newspapers. Consequently, the Scarrie word form dictionary (DEL 3.2.3) will be based on a corpus of newspaper text, and such a corpus has been built. It comprises all the articles that were published in 1995 and 1996 by the two prominent Swedish newspapers *Svenska Dagbladet* (a contracted partner of the Scarrie consortium) and *Upsala Nya Tidning* (a sub-contractor of Scarrie). It is referred to as the SvD/UNT corpus.

The SvD/UNT corpus holds over 220 000 articles. The texts were dumped in ANSI-format from the database archives of the two newspapers. Each newspaper utilises a coding system of its own to mark up the article text and its different parts. Because of this, a substantial amount of the texts in the transferred files consists of format codes, section information codes and control characters. These codes are not considered to be part of the corpus proper and thus removed.

The SvD/UNT corpus consists of more than 70 million tokens and 1.5 million types. This huge corpus material was segmented into tokens and types, word lists were formed and sorted, and word forms were grouped into categories with respect to in-going types of characters. This sub-categorisation provided a starting-point for a rough process of approval or rejection for further dictionary processing. In this process, frequency was also taken into account.

Based on frequency and the character-based sub-categorisation, the word types were referred to one of the following basic groups:

a) Irrelevant for the dictionary
b) Dictionary candidates
c) Misspellings

Among the words marked with a b) proper noun candidates were sorted out and further classified (names of persons, companies, brands, geographical names, acronyms, and multiword units).

The most frequent types (350,000) of b) were extracted for further morphological processing. They constitute the basis for the first version of the Swedish word-form dictionary for Scarrie. Further testing will show the adequacy of the dictionary thus obtained. If needed, more words will be added from the store provided by b).

The character-based classification of the word types of the corpus turned out to be of good help in identifying well-formed and not so well-formed types for further morphological processing before entering of the words into the dictionary.

Uppsala University
Department of Linguistics
98-01-21

LE3-4239 SCARRIE PROJECT
Deliverable 3.1.3
Final 1.0

SCARRIE WP3:
The SvD/UNT newspaper corpus

Bengt Dahlqvist

Abstract

The main objectives of Work Package 3 of the Scarrie project are to produce extensive word form dictionaries to be used in word error recognition and correction. This paper describes the creation of a text corpus of published newspaper article texts. The corpus consists of all the articles published in 1995 and 1996 by the two prominent Swedish newspapers *Svenska Dagbladet* and *Upsala Nya Tidning*. The process of storing, segmenting and classifying word forms in this corpus for the purpose of obtaining word lists is described in detail as well as other descriptive information pertaining to the corpus.

Contents

Acknowledgements

Parts of the extracted word form material were manually inspected and classified. This work was carried out by students from the STP language engineering programme at the Dept. of Linguistics, Ulrika Hedström, Camilla Löfling, Per Sandhammar, Anna Staerner, Susanne Viestam and Elin Wålstedt. The classification of proper names was carried out partly under the supervision of senior lecturer Dr. Mats Dahllöf.

1 Introduction

The main objectives of Work Package 3 of the Scarrie project are to produce extensive word form dictionaries to be used in word error recognition and correction. Both approved (correctly spelled) words and non-approved (incorrectly spelled) words with recommendation for correction will be entered. A minimum of 250 000 approved word forms are to be included in each dictionary.

The primary users of the Swedish version of the Scarrie tool will be newspapers. Consequently, the Scarrie dictionary will be based on a corpus of newspaper text. Text material for the corpus has been collected from the two prominent newspapers *Svenska Dagbladet* (a contracted partner of the Scarrie consortium) and *Upsala Nya Tidning* (a sub-contractor).The resulting joint SvD/UNT newspaper corpus contains the texts from all published articles in the years 1995 and 1996 from the two papers. This huge material has been analysed and categorised with respect to the ingoing words and their frequency. Resulting word lists have, according to a certain strategy, been manually inspected and labeled for further morphological analysis preceeding inclusion into the final dictionary.

This report describes how the corpus material was collected, segmented into word tokens and types, how word lists were formed and sorted, how word forms were grouped into categories and manually inspected and how a rough process of approval or rejection for dictionary inclusion was carried out. In short, all the work up to the morphological analysis, which will be the subject of another report (Del 3.2.3), is described here.

For the building of the corpus and processing of the text materials, more than 50 specialised programs in C and Perl were developed.

2 Text material

The SvD/UNT newspaper corpus contains all the articles that were published in the years 1995 and 1996 by the two Swedish newspapers *Svenska Dagbladet* and *Upsala Nya Tidning*. In total the corpus holds over 220 000 articles. The texts were dumped in ANSI-format from the database archive of each newspaper and transferred to a UNIX machine at the Dept. of Linguistics for further processing.

Table 1 presents a directory listing of the single original text files obtained from the newspapers, mostly comprising half a year of material each. The SvD material occupied about 345 MB in disk storage and the UNT about 162 MB. In total, the original text files then took more than 507 MB in storage.

```
SvD:
-rw-r-----   1     users    87646961 Nov 29 14:05 jan95junSvD.tfo
-rw-r-----   1     users    91171800 Nov 29 13:25 jul95decSvD.tfo
-rw-r-----   1     staff    93107014 Nov 29 12:05 jan96junSvD.tfo
-rw-r-----   1     staff    26388034 Nov 29 15:16 jul96augSvD.tfo
-rw-r-----   1     staff    48333436 Dec 06 14:28 sep96novSvD.tfo
-rw-r-----   1     staff    15310590 Jan 15 12:54 svd_dec_96.tfo

UNT:
-rw-r-----   1     staff    44235484 Dec 03 18:51 DL951.rdf
-rw-r-----   1     staff    42449510 Jan 10 13:39 9502.rdf
-rw-r-----   1     staff    83695770 Jan 14 21:37 1996.rdf
```

Table 1. Original text files transferred from the newspaper article archives of SvD and UNT.

Each newspaper utilises a coding system of its own to mark up the article text and its different parts. Because of this, a substantial amount of the texts in the transferred files consists of format codes, section information codes and control characters. These codes are not considered to be part of the corpus proper and were identified for elimination from the corpus.

In the SvD texts, delivered as tfo-files, all the codes start in column one and terminate with a ^-sign. A brief description of the different codes follows below and a short extract from a tfo-file is shown in figure 1.

Code R
signals the start of a new article.

Code F
provides information on text field and text. The numerical code for the field is given before the F, as in 1F or 12F. A total of 42 different field codes are used, for instance for headlines 1, for plain text 4 and so on. Field code 18, section, can take 28 values denoting content of text, e.g., sport, travels etc. See the appendix for a full list.

Codes P, S and A
are wrapped around the text parts (headline, ingress, plain text) and give certain information about change of paragraph, end of sentence and continuation.

```
R^
1F^
PFantasin flödar in i framtiden ^
4F^
PNio år, mer är det inte till år 2005.^
S Då kanske lurarna på bilden  här  till  vänster  finns  att köpa.^
S
^
PDe ingår nämligen i en fantasifull serie produkter som Philips designavdelning
tagit fram.
^
PVision of the Future är samlingsnamnet och tanken har varit att, med utgångspun
kt från vad  som tycks  vara  inom  möjligheternas ramar, visa upp framtiden.
^
PSE SID ^
S12^
5F^
PLåt inte storleken förvirra.^
S I verkligheten passar luren i örat och är en sladdlös telefon.^
9F^
P112^
10F^
P236^
11F^
PSVD^
12F^
P41^
13F^
P1996-08-31^
14F^
PJREICHER^
15F^
P1996-08-30^
16F^
PSvD^
17F^
PFoto^
18F^
PMAGASINET^
27F^
PASTROM^
28F^
P1996-09-02^
29F^
P364^
31F^
PSlutarkiv^
32F^
PPhilips designavdelning på lekhumör^
36F^
PFormgivning^
Pteknik^
40F^
Pphilips^
```

Figure 1. A text sample from the SvD corpus in tfo-format with original codes.

The texts from UNT were delivered as files in rdf-format. The codes in this file format are more explicit than their equivalents in the SvD files. Information fields and their values are given in plain text preceding ordinary article text. Typical fields are publishing date, section and page. Further, the articles are separated by line separators and each 31:th line has a standardised text (possible used for screen display). A short extract from an rdf-file is given in figure 2 below, which shows how article text and code fields are mixed together.

```
**************************** Doknr.: 17 ********************************

Publiceringsdatum: 950616          Avdelning: UNT'T'IN           Sida: 6

Rubrik: Höjt bensinpris och försämringar för tjänstebilar

Ingress: Höj bensinpriset successivt till drygt 9 kronor litern år 2000,
koppla förmånsvärdet på tjänstebilar till den privata körningen och öka
inte reseavdragen när bensinpriset höjs.  Det är huvudförslagen i trafik-
och klimatkommitténs slutbetänkande som efter två års utredande nu
överlämnats till regeringen.

Text: Det var en splittrad kommitté som på torsdagen presenterade sitt
slutliga förslag.  En av ledamöterna, direktör Nils-Erik Åhmansson,
reserverade sig mot de flesta av förslagen.
Grunden i förslaget är att koldioxidskatten höjs för alla fossila
drivmedel.
Bensinspriset föreslås höjas med 1:60 kr litern.  Det ska ske successivt
med 40 öre litern årligen från 1 januari 1997 till och med år 2000.
På det sättet tror kommittén att Sverige ska klara riksdagens mål att
koldioxidutsläppen från personbilstrafiken år 2000 inte ska vara högre än
1990 års nivåer.
                        Upsala Nya Tidning - Textarkivet

Enligt utredningens ordförande, professor Lars Nordström ska intäkterna
från den höjda koldioxidskatten inte användas för att förstärka den
ansträngda statskassan.
En annan tanke som förs fram är att sänka momsen.  Nu överlämnas frågan om
hur skatten ska återbetalas till skatteväxlingsutredningen.
Utredarna vill också komma åt de faktum att den som i dag kör tjänstebil
inte känner av ett höjt bensinpris eftersom det ingår fri bensin i
bilförmånen.  Ju mer man utnyttjar tjänstebilen privat, desto högre ska
förmånsvärdet bli är tanken.
Arbetsgivaren ska tala om för skattemyndigheterna hur mycket den anställde
utnyttjar bilen privat.  Några körjournaler blir det inte, eftersom dessa
skulle bli mycket svåra att kontrollera.  Bilägarna ska inte heller kunna
kompensera ett höjt bensinpris genom ett ökat reseavdrag.

**********************************************************************
```

Figure 2. A text sample from the UNT corpus in rdf-format with original codes.

The first task in building the corpus was to clean up the files and remove irrelevant characters from the texts. A number of filtering and segmenting programs were written and applied to the files. Table 2 presents the size reduction that was obtained after the filtering and cleaning up process.

File name	No of characters
svd9596.tfo	361 957 835
svd9596.txt	311 618 803
unt9596.rdf	170 380 764
unt9596.txt	149 857 738

Table 2. Reduction in file size after filtering away irrelevant characters.

3 Token extraction

The next task in the process of building the corpus was to count the number of ingoing articles, words and characters in the material. The results and the basic characteristics of the SvD/UNT text material are given in table 3 below.

	Articles	Tokens	Types	Characters
SvD	159 691	47 433 729	1 282 264	311 618 803
UNT	60 395	22 810 171	770 660	149 857 738
	220 086	70 243 900	1 672 993	461 476 541

Table 3. Statistics for the newspaper corpus created from the SvD and UNT text files.

From table 3 it is seen that the corpus consists of more than 70 million text words or tokens and about 1.6 million unique word types. The method for extracting token from a text is termed tokenising, and the principles underlying this process here may be stated as a BNF-grammar (see figure 3 below).

```
<token> ::= <alpha> | <alpha> : <alpha>
<alpha> ::= <alpha> <anums> | <anums>
<anums> ::= <letter> | <numeral> | <symb>
<letter> ::= a | A | ...
<numeral> ::= <numeral> <digit> | <digit>
<digit> ::=  0 | 1| 2 | 3 | 4 | 5 | 6| 7 | 8 | 9
<delim> ::=   | . | , | ? | ! | ” | ( | ) | / | _ | = | \t | \n
<symb> ::= & | % | ...
```

Figure 3. The tokenising process applied to the newspaper texts.

Before the tokenising process, all ingoing characters of the texts were classified as follows:

1. Control characters (ASCII code < 32)
2. Word delimiters:
   general delimiters (e.g. space, point, comma)
   positional delimiters (e.g. colon)
3. Letters:
   Upper case = "ABCDEFGHIJKLMNOPQRSTUVWXYZÅÄÖÀÁÂÃÆÇÈÉÊËÌÍÎÏÑÒÓÔÕØÙÚÛÜ"
   Lower case = "abcdefghijklmnopqrstuvwxyzåäöàáâãæçèéêëìíîïñòóôõøùúûüýÿß"
4. Numerals (0…9)
5. Symbols (all non-alphanumerical and non-control ANSI/ASCII characters).

The *control characters* were for the most part treated as word delimiters, e.g. line feed, carriage return and tabulator. The *word delimiters* proper include such characters as exclamation mark, point and space. Delimiters like ‘:’ or ‘=’ were eliminated only if found first or last in a token and are termed *positional delimiters* since their status as delimiters depend on their position in the token. All delimiters were removed from the identified tokens.The letter set in the alphabet used for the corpus consists of all Latin-1 characters that were ANSI/ASCII encoded. *Numerals* are the digits, and the *symbol set* includes all other characters that were found.

As a result of the tokenising process, a segmented file with all the text words was obtained. This huge list of tokens was sorted and reduced to a list of types together with frequency information. This word index was further sorted in two different ways to yield one file with

the types listed in descending frequency order and one file with the types listed in initial alphabetic order. The specific command used in UNIX for doing this are:

```
SORT 9596.seg -o 9596.sg
UNIQ -c 9596.sg 9596.nq
LANG=sv_SE SORT -T tmp -d +1 9596.nq -o 9596.ini
SORT +0 -r -n 9596.nq -o 9596.fre
```

The 20 most frequent tokens found in the corpus are listed below. From this list of raw tokens it is evident that a lot of processing remains to be done before true words are obtained. Especially, in the list we see that the word ‘det’ occurs both with the first letter as lower case and with upper case. Further, the token ‘-‘ reminds us that the tokeniser accepts tokens consisting of individual symbols only.

```
1984949 i
1851851 och
1613000 att
1215877 som
1073019 en
1066489 på
 993203 är
 875212 av
 851159 för
 837300 det
 791490 med
 753079 till
 644748 har
 609229 den
 528539 inte
 509518 de
 489230 om
 476291 ett
 357612 Det
 335976 –
```

Figure 4. Descending frequency sort list of tokens from the corpus.

Clearly, there is a need for a classification and grouping of tokens into categories with respect to their types of ingoing characters. This issue will be treated in the next section.

4 Type categorisation

In the stage of the text processing presented hitherto, it is unknown how many of the tokens that were found represent correct and approved words that should be included in a dictionary and how many of them should be disregarded or marked as unapproved. To make these decisions, we need more knowledge about the single words.

Now, this large amount of tokens does not easily lend itself to manual inspection one by one. Here, a strategy was adopted for grouping the tokens into categories that could be handled more easily. Basically, the grouping intended to decide what words could be:

a) discarded from future dictionary work
b) forwarded (to the morphological analysis) as candidates to the dictionary
c) regarded as examples of misspellings

To accomplish this, the tokens were classified depending on their ingoing characters. For this purpose, six distinct categories were defined as described and shown in table 4.

Name of category	Expression with	Examples
NUM1	only numerals	1997 747
NUM2	numerals and letters	E4 32-åringen
NUM3	numerals and symbols	+30% 100$
WRD1	only letters	ABB Thage kWh
WRD2	letters and symbols	<tt> EU:s
OTH1	only symbols	- @

Table 4. Token categories.

The result when applying this strategy to the token set of the corpus is shown in table 5 below. Not unexpectedly, the greater parts of the tokens are to be found in the numeral category and the letters-only category.

Tokens	SvD	UNT	Corpus
NUM1	1 315 436	344 485	1 659 921
NUM2	154 615	58 583	213 198
NUM3	174 635	6 245	180 880
WRD1	44 972 098	22 148 151	67 120 249
WRD2	447 637	222 599	670 236
OTH1	369 308	30 108	399 416
	47 433 729	22 810 171	70 243 900

Table 5. The result of the division of the text tokens in the SvD/UNT corpus.

The same procedure was also applied to the types of the corpus, which yielded the result shown in table 6 below. Compared with the result for the tokens, the numerals slip back in prominence, and the category of types consisting of a mixture of letters and symbols rises in importance.

Types	SvD	UNT	Corpus
NUM1	37 641	7 697	41 602
NUM2	43 573	9 017	49 097
NUM3	44 684	2 479	45 999
WRD1	1 035 034	666 841	1 343 959
WRD2	121 196	84 586	192 180
OTH1	136	40	156
	1 282 264	770 660	1 672 993

Table 6. The result of the division of the text types in the SvD/UNT corpus.

After this division of the word forms in distinct categories, some of the sets are still very large and contain more words that one would like to have for manual inspection.

Because of this, a refined division of the largest category was called for. This category consists of words only containing letters and is denoted wrd1. After consideration, a break-up of the group into seven subcategories seemed natural.

Firstly, words which always occur with all their letters in either upper or lower case were determined. These groups were denoted v4 and g6, respectively. Then we have an important group consisting of words which only varies in the first letter, occurring in upper and lower case, while the rest of the word is in lower case. This group was denoted gvg. A small variation of this is the gvv group in which at least some letter after the first one is in upper case.

Another important group was denoted vin. Here, the first letter always occurred as an upper case letter, and the rest of the token in lower case. The variation here, where one allows at least one upper case letter after the initial one, was denoted vro.

Name of subcategory	Expression with	Examples
wrd1.v4	only upper case	ABF, OBS
wrd1.gvg	w/o uc 1st pos, rest lc	Per/per, Uppsala/uppsala
wrd1.gvv	w/o uc 1st pos, more uc	MHz/mHz, PostNet/postNet
wrd1.g6	only lower case	abchazisk, último
wrd1.vin	with uc 1st pos, rest lc	Abacus, Östhammmar
wrd1.vro	with uc 1st pos, more uc	AfterShave, ThageG
wrd1.unk2	with lc 1st pos, more uc	börsVECKAN, dB

Table 7. Token sub-categories.

The resulting statistics of this operation are summarised in table 8. It shows that the most prominent group is the gvg one. This is the group with case variation in the first letter only.

```
2169161 i   1984949 184212
1914992 och 1851851  63141
1650315 att 1613000  37315
1237005 som 1215877  21128
1194912 det  837300 357612
1183182 en  1073019 110163
1123412 på  1066489  56923
1002608 är   993203   9405
 907401 för  851159  56242
 886363 av   875212  11151
 818659 med  791490  27169
 773603 till 753079  20524
 736365 den  609229 127136
 651352 har  644748   6604
 614912 de   509518 105394
 545160 inte 528539  16621
 525049 om   489230  35819
 518980 ett  476291  42689
 372672 han  281142  91530
 341679 men  183816 157863
```

Figure 5. A listing of the 20 most frequent members of gvg.

A listing of the 20 most frequent words in the category gvg is found in figure 5. Only the word form variant with the initial letter in lower case is printed out. The total frequency together with the frequencies of the initial lower and the upper case occurrence is also presented.

File	Types	Percent	Tokens	Percent	Examples
num1	41 602	2.48	1 659 921	2.36	1814, 007
num2	49 097	2.93	213 198	0.30	TV4, 1700-talet
num3	45 999	2.75	180 880	0.26	4:50, +2, 2-1
wrd1.v4	42 509	2.54	690 105	0.98	SAS, LO, AB
wrd1.gvg	2 x 163 245	19.52	61 121 282	87.01	Via/via, SvD/svd
wrd1.gvv	2 x 41	0.00	31 584	0.05	euroShell, mHz
wrd1.g6	682 512	40.80	2 701 054	3.85	måndag, msk
wrd1.vin	279 819	16.73	2 547 473	3.63	Malmö, Nato
wrd1.vro	9 106	0.54	23 482	0.03	McDonalds, TWh
wrd1.unk2	3 441	0.21	5 269	0.01	pH, iDAG, kW
wrd2	192 180	11.49	670 236	0.95	S:t, SE-banken
oth1	156	0.01	399 416	0.57	©, -
	1 672 993	100.00	70 243 900	100.00	

Table 8. The result of the division of the corpus tokens into sub-categories.

The classification in Table 8 gives a good basis for selecting these entries that are to go into the Scarrie word form dictionary. It is evident that some are not appropriate for further consideration. Such a group is num1, the numerals. Here, the dictionary will only include the digits, 0 – 9. Larger expressions will be handled by means of rules.

The group num2, which holds tokens with at least one letter and at least one digit, contains mostly acronyms (e.g., E4) and compounds (e.g., 32-åringen). For the compounds, only the letter-part will be forwarded to the dictionary. Also, the group num3 will be handled by special procedures and not entered into the dictionary in full.

The group v4 holds upper case letter words. These typically constitute acronyms of proper names that will go into the dictionary. The group gvg holds lower case words that also occur with an initial upper case. Mostly, the upper case form indicates the beginning of a sentence. The less frequent variation is due to a misspelling of a proper name. A number of proper name candidates have been sorted out from this group, i.e. items that vary grossly in frequency between the two forms.

The gvv group differs from gvg only in respect to the following. It contains an upper case letter somewhere in the token position two further on in the string. The gvv group contains almost exclusively misspellings or wrongly concatenated words. The g6 group words, which have only lower case letters, are forwarded to morphological analysis without further notice.

The vin group with initial upper case letter consists basically of proper names. This group was inspected manually down to the frequency two. A set of proper names (for persons, companies, brands and geographical locations) from this group was manually classified. As mentioned above, added to this set were names collected from the gvg group, where the proportion of occurrences was very skew (1 to 20 or more). A total of more than 50 000 proper names were found and classified in this way.

The vro group, like vin but with at least one extra upper case, includes some proper names and many wrongly concatenated words.

Finally, the unk2 group contains mainly garbage, with some exceptions, not suitable for inclusion into the dictionary.

5 Test material

A separate collection was made of tokens without the tokens in group num1, num3 and oth1 and with the gvg group presented one entry for each upper/lower case with frequencies. The resulting set contains mostly well-formed tokens. Further, from this collection all tokens occurring only once were removed. These two collections were named fx.fre and fx.fre2, respectively. Statistics for these files are found in table 9.

Filename	Types	Tokens
fx.fre	1 421 950	68 003 683
fx.fre2	618 099	67 199 832

Table 9. The statistics for the sets fx.fre and fx.fre2.

For testing the coverage of the existing dictionary and its extended versions, a set of 350 000 tokens from the set fx.fre was extracted. This set consisted of the most frequent tokens, including items of frequency 3. These high-frequent words are being morphologically analysed.

During the work with the SvD/UNT corpus a sub-corpus of only text from headlines was also extracted. For this, from UNT was sorted out a total of 55 065 headlines and from SvD a total of 159 045 headlines.

6 Conclusion

The merged SvD/UNT corpus was found to comprise more than 70 000 000 tokens and 1 500 000 types. From this set of types a selection was to be made for further inclusion in the Scarrie word form dictionary. A procedure for a character-based classification of the types was implemented. The resulting classification turned out to be of good help in identifying well-formed and not so well-formed types for further morphological processing before the entering of the words into the dictionary.

Appendix

Fältnamn	Nr	Typ	Del	Obl	Indx	Orig	Pris	Kommentar
RUBRIK	1	TEXT	N	N	J	J	0	Rubrik
FÖRINGRESS	2	TEXT	N	N	J	J	0	
INGRESS	3	TEXT	N	N	J	J	0	
BRÖDTEXT	4	TEXT	N	N	J	J	0	
BILDTEXT	5	TEXT	N	N	J	J	0	
BYLINE	6	TEXT	N	N	J	J	0	
VINJETT	7	FRAS	N	N	J	J	0	
TIMELINE	8	TEXT	N	N	J	J	0	
ÅRGÅNG	9	HELTAL	N	N	J	N	0	
NUMMER	10	HELTAL	N	N	J	N	0	
UTGÅVA	11	FRAS	N	N	J	N	0	
SIDA	12	HELTAL	N	N	J	N	0	
PUBL_DATUM	13	DATUM	N	N	J	N	0	
SKAPAD_AV	14	FRAS	N	N	J	N	0	
SKAPAD_DATUM	15	DATUM	N	N	J	N	0	
FÖRFATTARE	16	FRAS	N	N	J	N	0	
FOTOGRAF	17	FRAS	N	N	J	N	0	
AVDELNING	18	FRAS	N	N	J	N	0	
SERIE	19	FRAS	N	N	J	N	0	
PERSONNAMN	20	FRAS	N	N	J	N	0	
EGENNAMN	21	FRAS	N	N	J	N	0	
ÄMNE_1	22	FRAS	N	N	J	N	0	
ÄMNE_2	23	FRAS	N	N	J	N	0	
ÄMNE_3	24	FRAS	N	N	J	N	0	
ANMÄRKNINGAR	25	TEXT	N	N	J	J	0	
RÄTTELSETEXT	26	TEXT	N	N	J	J	0	
RÄTTAD_AV	27	FRAS	N	N	J	N	0	
RÄTTELSEDATUM	28	DATUM	N	N	J	N	0	Senaste rättelsedatum.
ARTIKELLÄNGD	29	HELTAL	N	N	J	N	0	
GRÄNSDATUM	30	DATUM	N	N	J	N	0	
STATUS	31	FRAS	N	N	J	N	0	Arkivstatus, tex förbas,arkiverad...
UNDERRUBRIK	32	TEXT	N	N	J	J	0	Ny 930811
UPPLAGA	33	FRAS	N	N	J	N	0	Anpassning för GP
PUBLIKATION	34	FRAS	N	N	J	N	0	Anpassning för GP
ARTIKELTYP	35	FRAS	N	N	J	N	0	Anpassning för SvD & GP
ÄMNE	36	FRAS	N	N	J	N	0	Anpassning för SvD
GEOGRAFISKT_NAMN	37	FRAS	N	N	J	N	0	Anpassning för SvD
URSPRUNG	38	FRAS	N	N	J	N	0	Anpassning för SvD
ANDRA_TIDNINGAR	39	TEXT	N	N	J	J	0	Anpassning för SvD
FÖRETAG	40	FRAS	N	N	J	N	0	Anpassning för SvD
BEVAKNINGSDAG	41	DATUM	N	N	J	N	0	Anpassning för SvD
SPÄRR	42	FRAS	N	N	J	N	0	Anpassning för SvD

Table A1. Listing of the field values used in the SvD text material.

T=1	<327>	BREV TILL REDAKTÖREN
T=2	<241>	BRÄNNPUNKT
T=3	<602>	CITY
T=4	<1497>	ETTAN
T=5	<277>	IDAG
T=6	<3830>	INRIKES
T=7	<35>	KARRIÄR
T=8	<2439>	KULTUR
T=9	<1085>	LEDARE
T=10	<408>	MAGASINET
T=11	<1118>	MARGINALEN
T=12	<141>	MAT
T=13	<1296>	NAMN FAMILJ
T=14	<6258>	NÄRINGSLIV
T=15	<1639>	POLITIK
T=16	<266>	RESOR
T=17	<105>	SAMTIDER
T=18	<21>	SCENHÖST 96
T=19	<80>	SIDAN FEM
T=20	<29>	SJUKVÅRD
T=21	<1>	SOMMAR
T=22	<3469>	SPORT
T=23	<2119>	STOCKHOLM
T=24	<271>	SÖNDAG
T=25	<295>	TV RADIO
T=26	<147>	TÄVLINGAR
T=27	<2658>	UTRIKES
T=28	<172>	VETENSKAP

Table A2. Listing of the values to the field ’Avdelning’ (text section) in the SvD material.

- Doknr:
- Publiceringsdatum:
- Avdelning:
- Sida:
- Rubrik:
- Ingress:
- Text:
- Bildtext:
- Anm:
- Korr:

Table A3. Listing of the fields used in the UNT material.

UNT'T	Text (dvs diverse)
UNT'T'	Text (dvs diverse)
UNT'T'AFF	affärer
UNT'T'E0	Edition noll (Uppsalaeditionen)
UNT'T'E1	Edition ett (Sydeditionen)
UNT'T'E2	Edition två (Nordeditionen)
UNT'T'EK	ekonomi
UNT'T'FA	familj
UNT'T'FD	För Dagen
UNT'T'FOR	Forskning
UNT'T'FS	Första sidan
UNT'T'GOL	Go Lördag
UNT'T'HEM	hem
UNT'T'IN	Inrikes
UNT'T'KU	kultur
UNT'T'LA	Lager (artiklar och telegram "vid behov")
UNT'T'LF	Läsarnas Forum
UNT'T'LO	Likt och Olikt
UNT'T'MAR	I markerna
UNT'T'NJ	Nöje
UNT'T'PO	Politik (ledare och debatt)
UNT'T'RES	resor
UNT'T'TEM	Tema
UNT'T'TRA	Trafik
UNT'T'UA	Uppsala
UNT'T'UNG	ung
UNT'T'UT	utland
UNT'T'tem	Tema

Table A4. Listing of the values to the field ’Avdelning’ (text section) in the UNT material.

Code		Frequency
10		7385538
32		67284667
33	!	43648
34	"	382622
35	#	921
36	$	540
37	%	1990
38	&	18627
39	'	178761
40	(	264040
41	)	294811
42	*	5875
43	+	7240
44	,	2749080
45	-	1136239
46	.	4554969
47	/	67953
48	0	1062587
49	1	1036179
50	2	587964
51	3	452088
52	4	402365
53	5	488902
54	6	339864
55	7	310461
56	8	319545
57	9	587046
58	:	377164
59	;	29971
60	<	33656
61	=	1469
62	>	33566
63	?	109693
64	@	440
65	A	803263
66	B	643457
67	C	308361
68	D	1137869
69	E	828078
70	F	593327
71	G	410854
72	H	632567
73	I	573176
74	J	377267
75	K	504873
76	L	586987
77	M	852690
78	N	670368
79	O	436351
80	P	483919
81	Q	11221
82	R	527804
83	S	1466097
84	T	758829
85	U	422823
86	V	492454
87	W	142218
88	X	10072
89	Y	52118
90	Z	28906
91	[	16813
92	\	9
93	]	16815
94	^	2567781
95	_	173343
96	`	118
97	a	33442597
98	b	4694754
99	c	4395124
100	d	14652771
101	e	35210610
102	f	6848555
103	g	11219906
104	h	6643305
105	i	20259526
106	j	2253617
107	k	11414555
108	l	18611457
109	m	11700287
110	n	31019064
111	o	15244721
112	p	6384833
113	q	36175
114	r	31326855
115	s	22759849
116	t	30687551
117	u	6176939
118	v	7817928
119	w	153122
120	x	394891
121	y	2243967
122	z	115651
123	{	19
124	|	49
125	}	13
126	~	206
127	•	18
128	□	41
129	□	5
131	ƒ	1
138	Š	14
139	‹	39
140	Œ	13
146	’	4
149	•	1
157	□	7
158	□	8
161	¡	8
162	¢	4
163	£	35
164	¤	11233
165	¥	1
167	§	185
168	¨	191
169	©	4686
170	ª	150
171	«	5
172	¬	57
173	-	19516
174	®	4419
175	¯	2
176	°	71
177	±	194
178	²	104
180	´	9050
181	µ	2
182	¶	19
183	·	12973
184	¸	65
185	¹	983
187	»	41
188	¼	1
191	¿	2
192	À	46
193	Á	57
194	Â	1
195	Ã	19
196	Ä	76515
197	Å	74308
198	Æ	72
199	Ç	345
200	È	197
201	É	2018
202	Ê	29
203	Ë	6
204	Ì	1
205	Í	38
206	Î	7
207	Ï	1
208	Ð	1
209	Ñ	2
210	Ò	4
211	Ó	52
212	Ô	5
213	Õ	15
214	Ö	123552
215	×	555
216	Ø	153
217	Ù	25
218	Ú	16
219	Û	18
220	Ü	471
223	ß	30
224	à	1575
225	á	2961
226	â	258
227	ã	62
228	ä	7083414
229	å	5795668
230	æ	1073
231	ç	845
232	è	3785
233	é	106154
234	ê	337
235	ë	857
236	ì	29
237	í	630
238	î	152
239	ï	114
241	ñ	187
242	ò	66
243	ó	1308
244	ô	470
245	õ	46
246	ö	5331229
247	÷	41
248	ø	1861
249	ù	10
250	ú	203
251	û	37
252	ü	10170
253	ý	59
255	ÿ	27

Total no of chars = 461476541

Table A5. Listing of the character frequency distribution in the SvD/UNT material.

Length	Frequency				
1	3225100	45	13	102	4
2	8370116	46	13	108	2
3	16555548	47	13	109	1
4	7720999	48	56	111	1
5	7668507	49	9	113	1
6	6546302	50	4	114	2
7	4742100	51	7	115	1
8	3919260	52	2	117	1
9	3187981	53	3	119	1
10	2432196	54	6	120	4
11	1672300	55	3	126	4
12	1145687	56	2	127	1
13	794604	57	4	129	1
14	614136	58	3	131	1
15	480165	59	1	132	3
16	336460	60	10	135	1
17	248670	61	1	138	2
18	196146	62	2	139	1
19	126847	63	2	144	2
20	85824	64	7	150	1
21	63066	65	1	153	1
22	37018	66	3	155	1
23	23924	67	1	156	1
24	27479	68	1	159	1
25	8358	69	2	162	3
26	4422	70	3	168	2
27	5082	71	2	169	2
28	1816	72	2	174	1
29	1233	76	2	183	1
30	641	77	2	186	3
31	349	78	6	188	18
32	283	79	1	189	2
33	177	81	2	192	3
34	111	82	1	199	1
35	113	84	5	204	1
36	86	87	2	205	1
37	45	89	1	209	46
38	41	90	3	210	1
39	45	93	2	214	117
40	43	95	1	215	77
41	22	96	7		
42	30	97	1	Total = 70243900	
43	23	99	1		
44	13	101	1		

Table A6. Listing of the frequency distribution of token length in the SvD/UNT material.