GU-ISS-08-1 • 2008 ISSN-1401-5919
Research Reports from
the Department of Swedish
University of Gothenburg
<http://hum.gu.se/institutioner/svenska-spraket/publ/resrep/>

Habeas Corpus

A survey for SNK – a Swedish national corpus

Maia Andréasson
Lars Borin
Magnus Merkel

Språkdata, Institutionen för svenska språket
Göteborgs universitet, Box 200, SE-405 30 Göteborg

CONTENTS

1 INTRODUCTION

> ***habeas corpus***
> *1465, from L., lit. "(you should) have the person," in phrase* habeas corpus ad subjiciendum *"produce or have the person to be subjected to (examination)," opening words of writs in 14c. Anglo-Fr. documents to require a person to be brought before a court or judge, especially to determine if that person is being legally detained. From* habeas*, second pers. sing. pres. subjunctive of* habere *"to have, to hold" (see* ***habit****)* + corpus *"person," lit. "body" (see* ***corporeal****).*
>
> (Harper 2008)

In 2007, the the Research Infrastructure Committee (KFI) of the Swedish Research Council awarded a two-year planning grant to a national Swedish consortium in language technology, with 7 partner organizations:

- University of Gothenburg (coordinating partner)
- Chalmers University of Technology
- KTH (Royal Institute of Technology)
- Linköping University
- Lund University
- The Swedish Language Council
- Uppsala University

The planning grant was awarded for a proposal entitled *An infrastructure for Swedish language technology*, with the aim of preparing a project proposal or project proposals for creating an integrated basic Swedish language technology research infrastructure, consisting of:

1. a Swedish national corpus (*Svensk nationell korpus* – SNK);
2. a Basic LAnguage Resource Kit (BLARK) for Swedish.

The practical planning work has been carried out by two working groups, with researchers from Gothenburg (Maia Andréasson and Lars Borin) and Linköping (Magnus Merkel) responsible primarily for the work on SNK, and researchers from KTH (Rolf Carlson and Kjell Elenius) and Uppsala (Beáta Megyesi and Eva Forsbom) having worked mainly on the Swedish BLARK. The two groups have interacted constantly throughout the course of the work, both in physical meetings and by means of electronic communication, e.g. using a project wiki.

The main tasks of the working groups have been:

- to inventory and collect information about existing resources, their character, quality, and not least, availability for research and other purposes;
- to make a survey of the needs of the research community and industry;
- to collect information about similar initiatives – completed, ongoing and planned – in other countries, especially in Europe;
- on the basis of this information, to formulate a concrete description of the SNK and the Swedish BLARK, together with an outline work plan and budget for creating the resources.

Consequently, both working groups made questionnaire surveys, among potential users of the SNK and Swedish BLARK, respectively. This report is a summary of the SNK survey. The BLARK survey report has been published separately (Elenius, Forsbom and Megyesi 2008).

The SNK questionnaire was constructed on the basis of experience accumulated from a long tradition in corpus work at Gothenburg and Linköping. In addition, in mid-2007, the working group visitsed the British National Corpus group at the Univesity of Oxford and the Bank of English research unit at the University of Birmingham.

1.1 Acknowledgements

The work reported here was supported by a planning grant awarded by VR/KFI (VR Dnr 2006-6763) for the proposal entitled *An infrastructure for Swedish language technology*, coordinated by Lars Borin, University of Gothenburg.

The Faculty of Arts, University of Gothenburg has also contributed to this work in money and in kind, through its support of Språkbanken (the Swedish Language Bank).

We would like to thank Ylva Berglund, Lou Burnard and Martin Wynne at the University of Oxford (the British National Corpus group), and Pernilla

Danielsson, Susan Hunston and Oliver Mason at the University of Birmingham (the Bank of English research unit at the Centre for Corpus Research), for generously devoting their time to our questions when we visited Oxford and Birmingham in June 2007 in order to learn more about the two largest British corpus projects.

Thanks are also due to Robert Andersson of the Centre for Language Technology, Gothenburg, for setting up the web questionnaire form (see appendix A) and the database for automatically collecting and processing the respondents' answers so that they could be easily imported into a spreadsheet application.

2 THE QUESTIONNAIRE

The aim of this questionnaire has been to investigate how a range of researchers, teachers and other users of text corpora picture a Swedish national corpus (SNK). An invitation to fill in a web based questionnaire in Swedish was sent out by email in September 2007 to people with a presumable interest in using a future Swedish national corpus, primarily to researchers and teachers in Scandinavian languages/linguistics and secondarily to people working with language planning and language policy formulation.

The invitation was sent to one representative each in relevant departments in Swedish universities and university colleges, and also to some other institutions. Invitations were also sent to some researchers working on Scandinavian languages in institutions outside Sweden, and the respondents were also encouraged to forward the questionnaire to other people with a possible interest in using Swedish corpora.

The questionnaire has never been meant to be a statistical investigation, but rather was intended to give us a picture of the wishes and desires among some user groups of Swedish text corpora, primarily descriptive linguists (the computational linguistic/language technology community being catered for by the parallel BLARK survey; see section 1). Because of this, the selection of respondents has been a convenience sample, rather than a random sample. The results of this questionnaire as visualized in bar charts in the following sections, should consequently not be seen as statistical evidence but rather as an illustration of tendencies from the result of the investigation.

In the following section you find a short description of the respondents. The results concerning the composition of the corpus are summarized in section 4, the results about access to the corpus are found in section 5, the results about the corpus user interfaces in section 6, and the results about the respondents' views on the collection of data are summarized in section 7. The questionnaire in its entirety (in Swedish) is reproduced in appendix A.

3 THE RESPONDENTS

A total of 36 respondents answered the questionnaire. 25 of these work in Sweden: four researchers each from Uppsala University, Stockholm University and the University of Gothenburg. Two researchers from Växjö University have answered and one researcher from each of Chalmers University of Technology, Kristianstad University, Linköping University, Lund University and Umeå University. Two researchers work in Denmark (*Det danske sprog- og litteraturselskab* and University of Aarhus), one in Belgium (University of Gent) and one in Canada (Carleston University).

19 respondents teach at universities or university colleges. University of Gothenburg, with three respondents, Stockholm University, Uppsala University and Mid Sweden University, with two respondents each, dominate this category. Among the international respondents, University of Aarhus, Copenhagen Business School and University of Oslo are represented with teachers.

One respondent is also a secondary school teacher, one works with language policy and planning and one has checked the option “author of teaching materials” and “author of non-fiction”. Five respondents are lexicographers and one respondent represents the group “the general public with an interest in language” and nothing else.

4 CORPUS COMPOSITION

Under the heading *Composition of the corpus* the respondents answered questions about what kind of textual genres they wish to see in the SNK and what kinds of markup/annotation would be crucial to their needs.

4.1 Genres

The questions about what genres of written and spoken language the respondents wished to be included in the SNK were open questions.

> **Question:**
> *There will be many genres included in the SNK, in order for the corpus to cover as many varieties of Swedish as possible. We would like to know what genres are the most important for your research.*
>
> ***Spoken language***
> *Give some examples of genres that you would like to be included.*
>
> ***Written language***
> *Give some examples of genres that you would like to be included.*

In the answers, the request for spoken dialogue with two or more participants is predominant. As many as 25 of the 29 respondents who answered this question mentioned dialogue explicitly. Respondents ask for both informal and professional dialogues, as for instance in salesperson/customer, teacher/pupil, or doctor/patient contexts. Also more spontaneous speech in the form of monologues, as for example longer narratives, is mentioned.

Scripted spoken genres seem also to be of interest. Newscasts and reading of manuscripts, both fact and fiction, should be part of the corpus according to the respondents, even though some ranked this low compared to spontaneous speech. One comment on this genre stresses the need for readings of fiction to be “professional”. The respondents also advocated diversity of gen-

res, especially regarding age groups, speech situations and regional variants of Swedish. One respondent suggested that some L2 Swedish should be included.

Seven respondents chose not to answer this question and one of these expressed doubts about mixing spoken and written language in one corpus.

Concerning genres of written Swedish, the respondents express a desire to have more diversity among the genres. Their opinion seems to be that traditional corpora of written texts consist more or less exclusively of press text and fiction.

Many of the respondents mention the need for texts of more informal and personal genres, written by non-professional authors, as for example letters from day-care personnel to parents, diary entries, blogs, and chat-room texts. But there is also, according to the respondents, a need for more diversity among the formal genres. They suggest inclusion of, for example, advertising texts, business correspondences and negotiations. Subtitles are mentioned as a possible genre.

The diversity should also include texts written by different age groups, especially text written by children and adolescents. The respondents suggest that also children's literature should be included in the fiction part of the corpus. Parallel texts are mentioned, as well as translated texts and two respondents suggest that Finland-Swedish should be included as a variant of Swedish in the corpus.

Eight respondents chose not to answer the question about genres of written language. Below, the answers of one of the respondents are quoted (in English translation), to give an example of how the need for diversity was expressed:

> Press text, fiction (all variants including children's and adolescents' literature, in Swedish), non-fiction (all variants), official texts (community information, laws, from Internet etc.), advertising, text books on all levels of education, magazines (technical and popular), academic writing, annual reports from enterprises and institutions, weblogs (both private and public), email correpondence, students' essays, press releases from enterprises and institutions. More modern texts! Integrate all the historic
>
> corpora that are already out there, and also the Swedish Literature Bank. It would be wrong to go for older historical texts since this has been an important part in already existing corpora.

4.2 Metadata

Being asked which metadata about the author/speaker that would be most important to include in the corpus, the respondents ranked the alternatives *geo-*

graphic area, *sex*, *year of birth*, *Swedish as first or second language* and *other*. The respondents were asked to grade the relevance of the metadata from 1 to 5, where 1 represents the most important and 5 the least important metadata. See figure 1 for an overview of the answers.

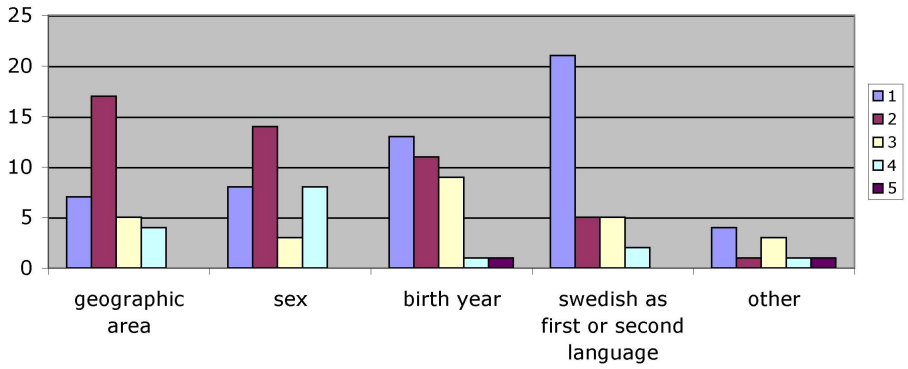

Figure 1: Relevance ranking of author metadata for inclusion in the SNK

Among the comments on the question about metadata, one concerned the metadata *Swedish as first or second language*. The respondent suggested that information about how many years the speaker/author had been in Sweden should be added. The specification of which "other" metadata the respondents want to see in the SNK are summarized in (1) below. The numbers in parentheses represent the grades.

(1) Other author/speaker metadata mentioned:

   (a) education, occupation, and/or social background (1, 3, 3)
   (b) sender/addressee relation (1)
   (c) some kind of unique identification (to give the possibility to search for other utterances of the same speaker) (1)
   (d) fluency in other languages etc. (3, 4)
   (e) country (Sweden or Finland) (1)

Two specifications to this question rather concerned metadata about the (authorship of the) text:

(2) Text authorship metadata:

   (a) one or several authors, named OR enterprise or public authority as responsible for the text (1),

(b) is the author speaking/writing on behalf of another, as a representative, as a customer (5)

In the answers to the question about metadata concerning the text, respondents ranked the alternatives in the order *time of publication*, *genre* and *other*, as indicated in figure 2 below. The respondents were asked to grade the importance from 1 (most important) to 3 (least important).

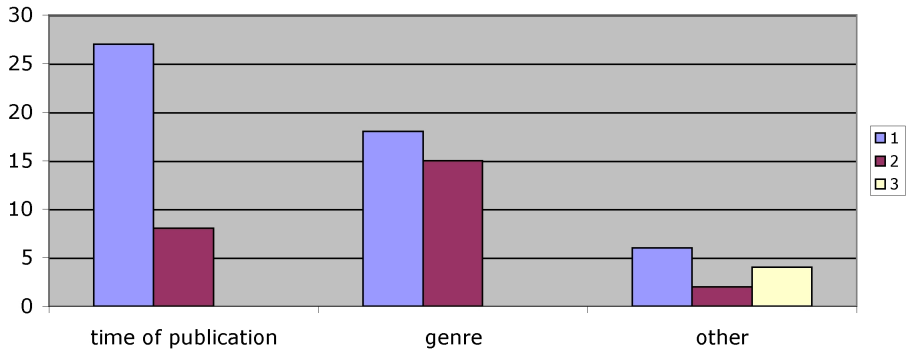

Figure 2: Which text metadata would be the most important to include in the corpus?

Other suggestions for text metadata to be included in the SNK are summarized in (3) below.

(3) Other text metadata mentioned by respondents:

(a) unique identification, for example ISBN number (1,1,1)
(b) translation or original (2)
(c) source (1)
(d) topic (gardening, computers, etc.) (1, 1, 3, 3)
(e) purpose (inform, advertise, negotiate, organize etc.) (1)
(f) country (Sweden/Finland) (1)

4.3 Annotation

Question:

How would you like the SNK to be annotated, that is, with which linguistic information should the linguistic units to be marked?

- parts of speech + inflection
- base forms
- name expressions
- syntactic information
- semantics (annotation for semantic relations between words in the text, semantic roles, annotation for meaning or semantic categories)
- anaphora (linking between noun phrases and anaphoric expressions)
- linking (between text and other media, as for example sound files)
- prosody
- other

Figure 3, below, shows that part of speech, base form and syntactic annotation are the three most interesting annotations for the respondents since these have the high scores of 1s and 2s. Also semantic and prosodic annotation have a high total score.

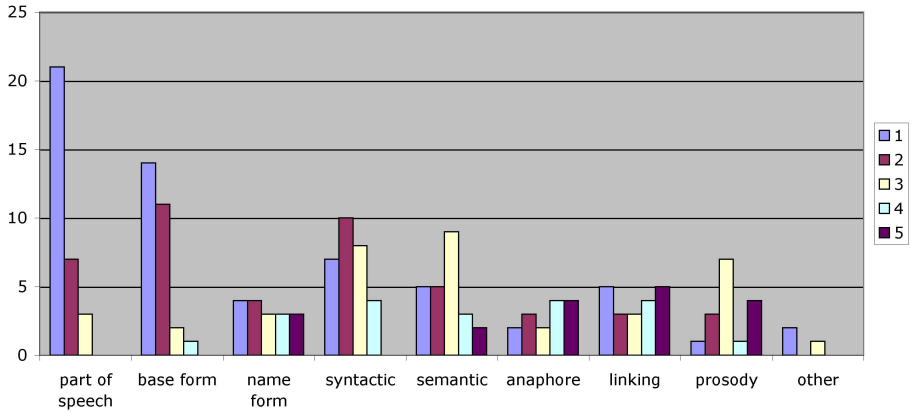

Figure 3: What linguistic annotation is most important?

Suggestions about additional types of annotation/markup for the texts include information about speech acts and text function, linking to other texts that are

related to the text (for instance that business protocols be linked to the previous and the next protocol), logical text structure (text, chapter, section, captions, etc), typographical characteristics (boldface, italics, etc). One Swedish respondent suggests the same kind of linking as that introduced in the Norwegian Speech Corpus (NoTa), namely linking to both sound and video files.

5 CORPUS ACCESS MODES

The respondents seem to agree that the SNK should be as widely available as possible (see figure 4 below). Some chose to check only the option *the general public*, which includes all the other options. Authorities and enterprises got somewhat lower scores than the other options.

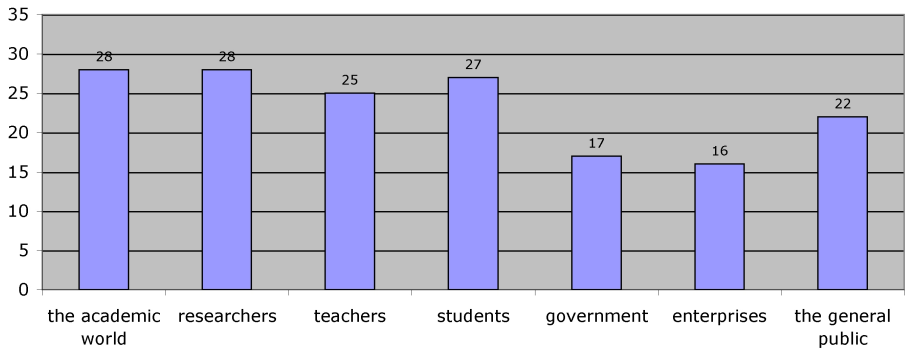

Figure 4: Who should have access to the corpus?

5.1 Would respondents be willing to pay to gain access to SNK?

The answer to the question about whether the respondents would be ready to pay to gain access to the SNK divide the respondents more or less into two camps: Some respondents find it reasonable to pay to get access to the corpus while others do not (see figure 5 below). Option A concerns paying for getting access to the text annotated with metadata, option B to the text, metadata and linguistic annotation and option C to the text, metadata, linguistic annotation and a search interface.

Only 3–5 respondents suggested prices for getting access to the corpus and among those who did the prices spanned from 200 to 3000 SEK for text and

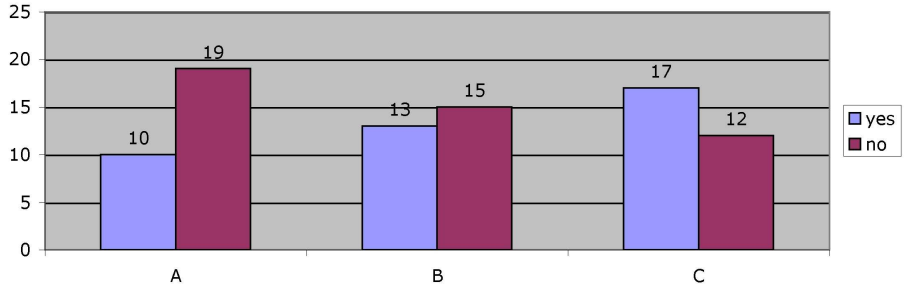

Figure 5: Would you consider it worth paying to get access to the SNK?

metadata, from 250 to 5000 SEK for text, metadata, and linguistic annotation, and 300–8000 SEK for text, metadata, linguistic annotation and user interface.

The comments on the other hand where many. One international respondent expresses his doubts about his institute being prepared to pay for him getting access to the corpus:

> I would say that you would serve the scientific community best if there were free access, possibly after some kind of application. The problem is that our institutes are not very keen on paying for getting access to international corpora, so we'd probably get to pay ourselves, if the access should cost money.

A Swedish respondent thinks that universities would be prepared to pay, but discusses how to get authorities and enterprises to do the same:

> It is difficult to estimate a cost for this. I think that universities generally are more inclined to pay, but at the same time they have the smallest economical resources. It will be necessary to stress the benefits for authorities and enterprises just as Presstext/Mediearkivet and Nationalencyklopedien. Which means marketing the SNK at the Gothenburg Book Fair, at seminars etc.

6 CORPUS USER INTERFACES

6.1 How do you want to get access to the SNK?

The results of the questionnaire show that it is vital to most respondents to get access to the corpus through a web interface, either in combination with the possibility to download the corpus or as the only way to access the corpus. Figure 6 shows that only five of the respondents view access to the corpus merely by downloading as a possible solution. Of these five only one chose this as the only option.

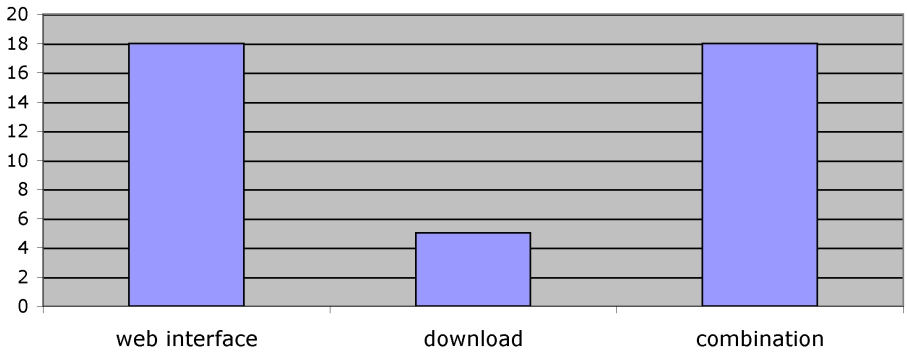

Figure 6: How would you like to get access to the corpus?

Three respondents chose not to answer this question. 15 respondents chose a combination as the only option. One chose downloading as the only option and 12 respondents do not find any need for downloading, but chose access merely through a web interface as the only option.

Comments to this question include for example a request for a “simple” web interface, where the user would get suggestions for alternative ways of formulating the search query if it is not correctly formulated, an alternative where the user may include the preceding and the following sentence in the

output, suggestions to look at other corpora: Oslo-korpuset (Norwegian), Korpus 2000 (Danish; a new web interface has recently been introduced for this corpus), the Sketch Engine (general corpus search tool), CQP (general corpus search tool), and a suggestion not to go for the same solution as the PAROLE corpus in Språkbanken.

6.2 Combinations of metadata

It seems to be of vital interest to the respondents to be able to combine metadata in searches in the corpus. Figure 7 below show that 30 out of 36 said that it is very important or rather important to be able to perform searches with such combinations.

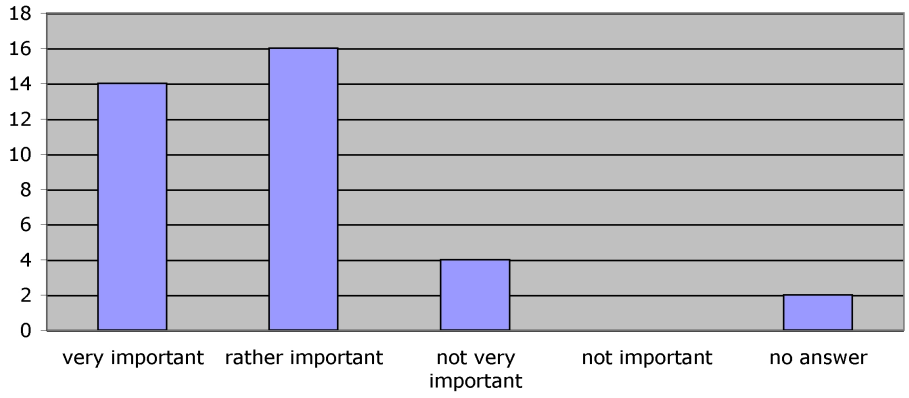

Figure 7: Is it important for you to be able to combine metadata in searches?

Most respondents (28) would like to have the possibility to combine all kinds of metadata in a search query. Figures 8 and 9, below, show what combinations of metadata are the most important to the respondents.

6.3 Display of results

Traditional ways of displaying results, such as concordances, frequency lists and hits in context, are the most popular in the questionnaire, see figure 10 below. Most important seems to be to get to see a hit in its context. 33 of the 36 respondents checked this option.

When it comes to saving the results, plain text files are the most popular. The possibility of getting a text file that easily opens in for example Excel (the

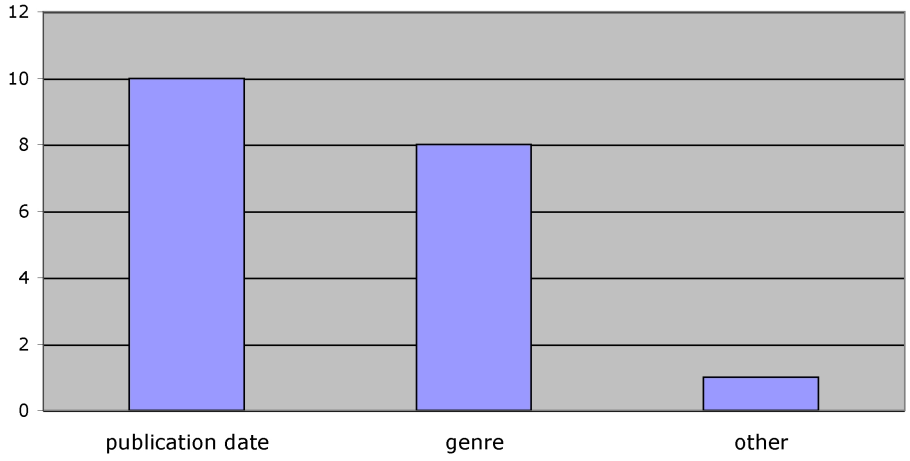

Figure 8: What types of text metadata would you like to be able to combine in a web-based search interface?

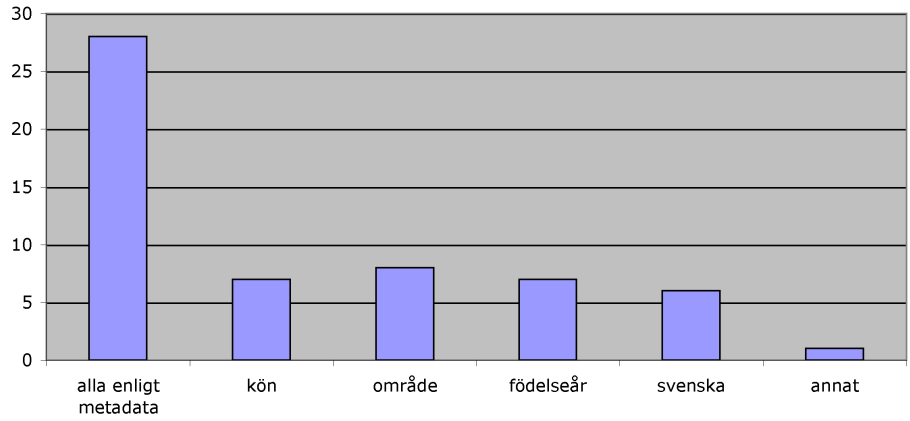

Figure 9: What types of author/speaker metadata would you like to be able to combine in a web-based search interface?

option *textfil++* in figure 11) is also very popular.

One respondent stresses the need for users to be able to download the results as a "raw" text file for further processing with his own software.

Suggestions about alternative ways of saving the text include saving comma or tab-separated text files and saving as xml. One respondent comments that SPSS files are possible to open in Excel.

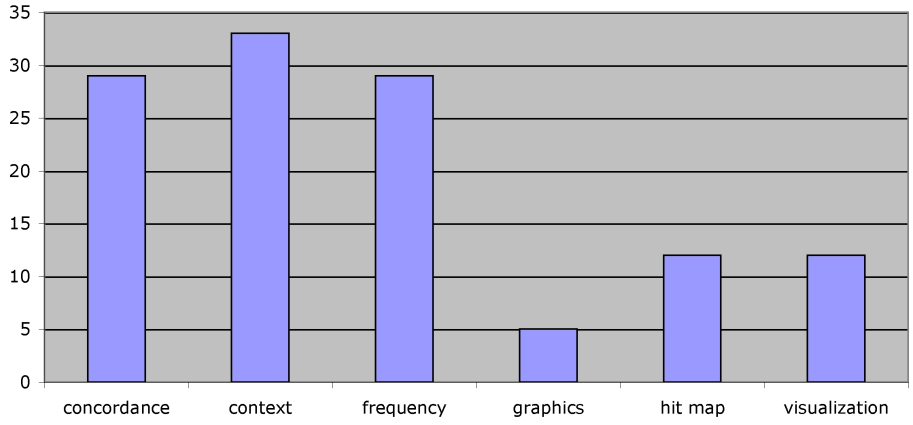

Figure 10: How would you prefer the results of your searches to be displayed on the web?

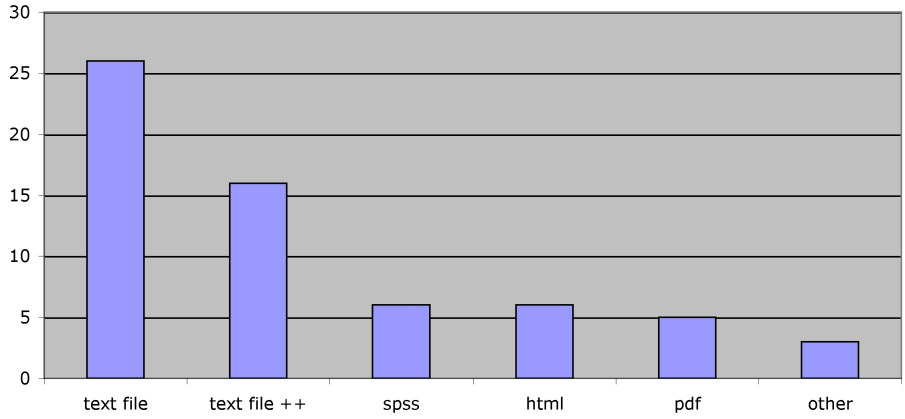

Figure 11: How would you prefer to save the results of your searches?

7 DATA COLLECTION METHODOLOGY

It is more important to the respondents that the material in the spoken language part is transcribed according to the same principles, than that all the spoken material is from the same period in time, see figures 12 and 13 below.

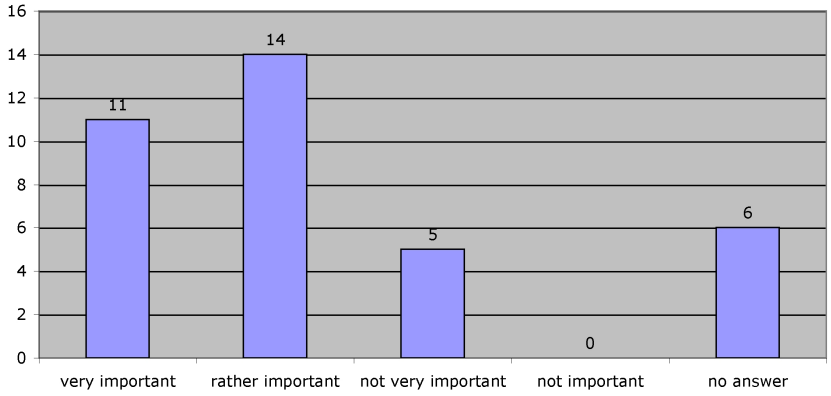

Figure 12: How important is it that the spoken language part is transcribed according to the same principles?

Figure 14 shows that the respondents do not consider it very important that the spoken part and the written part should be collected during the same period of time. One comment to this question stresses that this holds given that there is metadata about production/recording time in the corpus.

The respondents did not find it important that all written material should be collected during the a limited period of time, see figure 15.

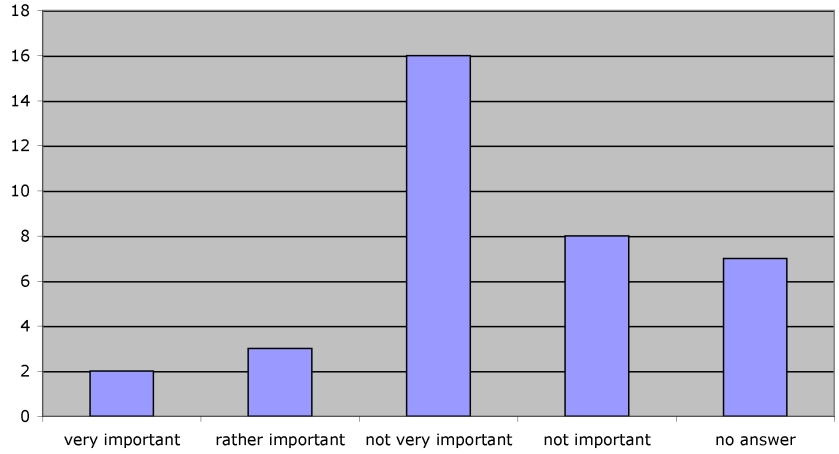

Figure 13: How important is it that all spoken material is collected during a limited period of time?

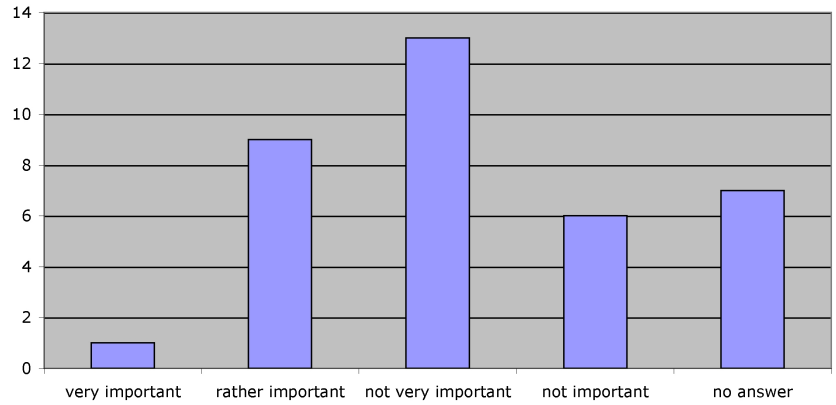

Figure 14: How important is it that the spoken part and the written part are collected during the same period of time?

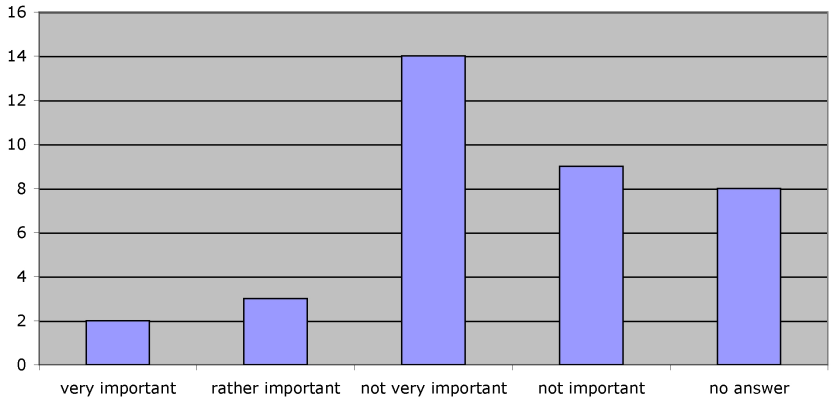

Figure 15: How important is it that all written material is collected during a limited period of time?

8 SUMMARY

In this report the answers from 36 respondents to a questionnaire on a future Swedish national corpus (SNK) have been summarized. Most of the respondents are university teachers and researchers in Scandinavian languages and linguistics.

In section 4, we saw that it is of great importance to the respondents that both the spoken and the written part of the Swedish national corpus include as many different genres as possible. The key word here is *variety*. For the spoken part, most respondents mention dialogue and suggest that both formal and informal dialogues be included. For the written part, the respondents stresses the need for other genres than press text and fiction. Both for spoken and written texts the respondents request some variation in age for speakers/authors and for addressees/readers.

In the same section we saw that the respondents agree that the most important metadata to be included concerning the *speaker/author* was if he or she is an L1 speaker of Swedish. Information about place/area of origin, sex and year of birth of the speaker/author were also of great importance to the respondents. The most important kind of metadata about the *texts* was time of publication/recording. The respondents found part of speech, base form and syntactic information to be the most important linguistic annotations that should be made to the texts.

It is crucial to the respondents that the SNK should be accessible to the general public (see section 5). Paying for getting access to the SNK was an option only for about half of the respondents.

Most of the respondents want to be able to get access to the SNK through a web interface (see section 6), preferably combined with the alternative to download it as a file. It is very important for the respondents to be able to perform searches on combinations of several kinds of metadata.

When it comes to the collection of data – see section 7 – it is important to the respondents that all of the spoken part is transcribed according to the same principles. Given that the different parts of the corpus are annotated with

information about publication/recording time, the respondents do not consider it important that written or spoken texts are from the same time period.

In short, the respondents want the SNK to be filled with a great variety of genres of spoken and written Swedish and they want to have a great range of options when it comes to performing searches in the corpus. Several of the respondents have made clear that they would be willing to assist in the coming work with the construction of the SNK by participation in a user panel.

REFERENCES

Elenius, Kjell, Eva Forsbom and Beáta Megyesi 2008. Survey on Swedish language resources. <http://www.speech.kth.se/prod/publications/files/3151.pdf>. Speech, Music and Hearing, School of Computer Science and Communication, KTH; Department of Linguistics and Philology, Uppsala University.

Harper, Douglas 2008. Dictionary.com. Online Etymology Dictionary, s.v. *habeas corpus*. <http://dictionary.reference.com/>. Accessed 2008-05-15.

A THE SNK QUESTIONNAIRE

On the following pages, the SNK survey web questionnnaire is reproduced. The language of the questionnaire is Swedish, and no version was prepared in English or any other language, the intended primary target group of the questionnaire being scholars in Swedish and Scandinavian linguistics.

Most of the questions – together with a synopsis of the answers – have been covered in the preceding sections of this report, but we reproduce the questionnaire here for completeness' sake.

Kontaktinformation

Namn

Telefon

E-post

Jag kan tänka mig att svara på fördjupade frågor i en telefonintervju som tar cirka 30 minuter.
○ ja
○ nej
◉ inget svar

Jag kan tänka mig att ingå i en referensgrupp för SNK under uppbyggnaden.
○ ja
○ nej
◉ inget svar

Jag är...

☐ forskare/forskarstuderande
- universitet/liknande:
- ämne:

☐ universitetslärare
- universitet:

☐ annan lärare
- nivå:

- ämne:

☐ språkvårdare

☐ journalist

☐ författare av...

läromedel
facklitteratur
skönlitteratur
inget svar

(Högerklicka (PC) eller håll nere äppleknappen (Mac) för att kunna markera mer än ett alternativ.)

☐ ordboksredaktör/-producent

☐ språkintresserad allmänhet

☐ annat:

Uppbyggnad

Vad tycker du ska finnas med i korpusen?

Genrer

I SNK kommer många olika genrer att finnas med, för att korpusen ska vara så täckande som möjligt. Vi vill gärna veta vilka genrer som är viktigast för din forskning.

Talat språk. Ge exempel på vad du vill ha med:

Skrivet språk. Ge exempel på vad du vill ha med:

Metadata

Vilka metadata är viktiga att ha med? Markera det viktigaste med 1, det näst viktigaste med 2 osv. Om några metadata är lika viktiga, markera dem då med samma nummer.

...om talaren/författaren

☐ område i landet

☐ man/kvinna

☐ födelseår

☐ svenska som modersmål/andraspråk

☐ annat: ☐

...om texten

☐ utgivningstid

☐ genre

☐ annat: ☐

Annotering

Hur skulle du vilja att SNK är annoterad, dvs. med vilken språklig information ska språkliga enheter vara uppmärkta med? Markera det viktigaste med 1, det näst viktigaste med 2 osv. Om några är lika viktiga, markera dem då med samma nummer.

☐ Ordklass + böjningsform

☐ Grundform, för ord

☐ namnuttryck

☐ Syntaktisk information (för fraser, satser och meningar)

☐ Semantik (markering av semantiskt släktskap mellan ord i texten, t.ex. semantiska roller, eller markering av ordbetydelse eller semantiska kategorier)

☐ Anaforer (länkning mellan nominalfraser och anaforiska uttryck)

☐ Länkning (mellan text och annan media, som t.ex. ljudfiler

☐ Prosodi

☐ Annat

Tillgänglighet

Vilka ska ha tillgång till SNK?

☐ den akademiska världen

☐ forskare

☐ lärare

☐ studerande

☐ myndigheter

☐ företag

☐ allmänheten

Kan du tänka dig att betala för att få tillgång till SNK?

Det är inte ovanligt att man får betala för att få tillgång till stora nationella korpusar. Nedan finns ett antal olika scenarier. Fyll i om du tycker det är värt att betala för tillgång till SNK, med eller utan annotering och sökgränssnitt.
(Utan annotering och sökgränssnitt är det en balanserad korpus som en textfil med uppmärkning av genre, författares kön, ålder, födelseort osv. som säljs. Sökgränssnitt är här ett färdigt webbgränssnitt där sökningar kan göras utan större tekniska kunskaper.)

Om du har en uppfattning om hur mycket det skulle vara rimligt att betala får du gärna fylla i det också.

A. Text + metataggning: Tillgång till en balanserad korpus om 100 miljoner ord skriven och (transkriberad) talad text. Uppmärkt med information om genre, författares kön, ålder, födelseort osv. Du får ordna annotering (uppmärkning av ord med grammatiska taggar) och ett eventuellt sökgränssnitt själv.

☐ ja

(rimligt pris för en enanvänderlicens för 5 år ☐)

(rimligt pris för en 5-användarlicens för 5 år ☐)

☐ nej

B. Text + metataggning + grammatisk taggning: Tillgång till en balanserad korpus om 100 miljoner ord skriven och (transkriberad) talad text. Uppmärkt med information om genre, författares kön, ålder, födelseort osv. Dessutom är korpusen

annoterad dvs. uppmärkt med grammatiska taggar. Du får ordna sökgränssnitt själv. <

☐ ja

(rimligt pris för en enanvänderlicens för 5 år ☐)

(rimligt pris för en 5-användarlicens för 5 år ☐)

☐ nej

C. Text + metataggning + grammatisk taggning + sökgränssnitt: Tillgång till en balanserad korpus om 100 miljoner ord skriven och (transkriberad) talad text. Uppmärkt med information om genre, författares kön, ålder, födelseort osv. Dessutom är korpusen annoterad dvs. uppmärkt med grammatiska taggar och inkluderar ett sökgränssnitt .

☐ ja

(rimligt pris för en enanvänderlicens för 5 år ☐)

(rimligt pris för en 5-användarlicens för 5 år ☐)

☐ nej

Kommentar:

Gränssnitt

Den här sektionen handlar om hur du vill att SNK ska göras tillgänglig för sina användare. Det handlar alltså om ett användarvänligt gränssnitt.

(Om du vill se hur ett gränssnitt mellan användare och korpus kan fungera kan du gå in på t.ex. Språkbankens hemsida och se på de olika sökgränssnitt som används där. Klicka på namnen prova att göra sökningar i PAROLE/SUC-korpusen och i Språkbankens konkordanser.)

Hur vill du få tillgång till SNK?

☐ webbgränssnitt
(korpusen ligger på internet, alla sökningar görs on-line och resultat kan hämtas hem från gränssnittet)

☐ hämtning till egen dator
(korpusen finns i den egna datorn, alla sökningar görs i ett eget eller ett medföljande gränssnitt)

☐ en kombination av ovanstående
(sökningar med begränsat antal träffar kan göras via gränssnittet medan mer avancerade sökningar kräver att hela korpusen finns i användarens dator)

Sökningen i gränssnittet

Har du några särskilda önskemål om hur själva sökningen i SNK ska gå till?

Vilka typer av **metadata** vill du kunna söka på?

☐ alla, som markerats ovan under metadata, eller kryssa i nedan

...om talaren/författaren
☐ man/kvinna
☐ område i landet
☐ födelseår
☐ svenska som modersmål/andraspråk
☐ annat:

...om texten
☐ utgivningstid
☐ genre
☐ annat:

Kombinationer av metadata
Det är ofta intressant att kunna söka på kombinationer av metadata, så att man t.ex. kan göra sökningar i tidningstext skriven av män födda på 40-talet. Markera hur värdefullt du anser att det skulle vara för dig att kunna göra sådana kombinationssökningar.

○ mycket värdefullt
○ ganska värdefullt
○ inte särskilt viktigt
○ helt ointressant
◉ ingen åsikt

Vilka olika metadata vill du i så fall helst kunna kombinera med varandra. Ge förslag i rutan:

```
[Skriv t.ex. kön + område i landet + födelseår]
```

Visning av träffar
Hur vill du att träffarna ska kunna visas för dig på webbsidan?

☐ konkordans
☐ möjlighet att se träffar i kontext
☐ frekvenslistor
☐ grafiska illustrationer (som t.ex. diagram)
☐ träffkartor (som visar var i korpusarna träffarna finns)
☐ visualiseringar (som t.ex. att visa olika typer av träffar med olika färg)

Spara träffar
I vilken form vill du kunna spara träffarna?

☐ textfil
☐ textfil, som är färdig att öppnas som ett kalkylblad, t.ex. i Excel
☐ SPSS-fil (statistikprogram)
☐ html-fil
☐ pdf-fil, där man kan se och läsa träffarna men inte vidarebearbeta dem
☐ annat

☐ oviktigt att spara träffar

Talspråksdelen

Talat språk är, som bekant, särskilt kostsamt att samla in eftersom det kräver extra mycket bearbetning, inspelning, transkription och eventuell anonymisering, innan det kan märkas upp på samma sätt som den skrivna delen av SNK. Här följer frågor som gäller insamling och uppmärkning av talat språk.

Transkription

Hur viktigt är det att allt talspråksmaterial är transkriberat enligt samma principer?

○ mycket viktigt
○ ganska viktigt

○ inte särskilt viktigt
○ helt ointressant
◉ ingen åsikt

Insamlingstid

Det finns en hel del mindre talspråkskorpusar i landet. En del av dem är flera tiotal år gamla, andra är ganska nya. En möjlighet att samla talspråksmaterial till SNK är att få tillgång till redan insamlat och transkriberat material.

Hur viktigt är det att allt talspråksmaterial i SNK är insamlat under en begränsad tid?

○ mycket viktigt
○ ganska viktigt
○ inte särskilt viktigt
○ helt ointressant
◉ ingen åsikt

Hur viktigt är det att talspråksmaterialet och skriftspråksmaterialet i SNK är insamlat under en samma tid?

○ mycket viktigt
○ ganska viktigt
○ inte särskilt viktigt
○ helt ointressant
◉ ingen åsikt

Skriftspråksdelen

Hur viktigt är det att allt skriftspråksmaterial är skrivet under samma tid?

○ mycket viktigt
○ ganska viktigt
○ inte särskilt viktigt
○ helt ointressant
◉ ingen åsikt

Om du tycker att det är viktigt att skriftspråksmaterialet är insamlat under samma tid: Hur gammalt skriftspråksmaterial tycker du kan få finnas med i SNK?

○ 2 år gammalt
○ 5 år gammalt
○ 10 år gammalt
○ 15 år gammalt
○ äldre
◉ ingen åsikt

Bidrag till SNK

Har du en talspråkskorpus som du kan ställa till SNK:s förfogande?
○ ja, se kommentar
○ nej
○ kanske, se kommentar
◉ inget svar

Har du en skriftspråkskorpus som du kan ställa till SNK:s förfogande?
○ ja, se kommentar
○ nej
○ kanske, se kommentar
◉ inget svar

Skicka dina svar

Ta bort all inmatad information

Skicka mina svar till SNK