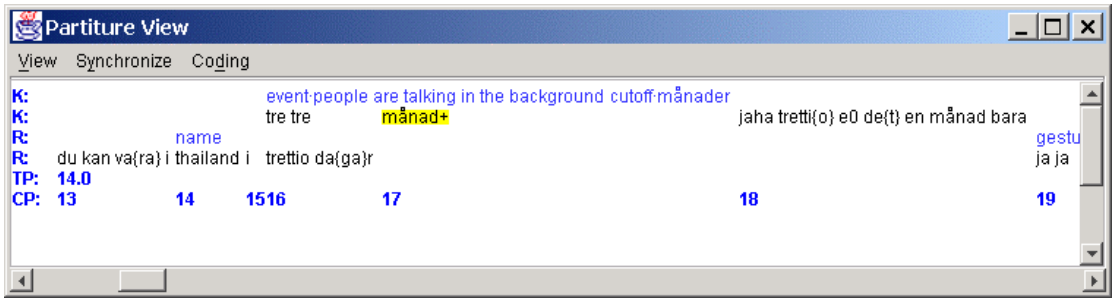
# NORDTALK

# Corpus-based Research on Spoken Language

## A NorFA Language Technology Network




## Interdisciplinary cooperation

## in the Nordic countries


## Edited by Jens Allwood and Elisabeth Ahlsén


## http://www.ling.gu.se/projekt/nordtalk/

# NORDTALK

# Corpus-based Research on Spoken Language

# Introduction

Below we will describe the network for corpus-based research on spoken language. First we will describe the background and goals of the network. Then we will describe what activities we have engaged in and are planning for future and in the third section the members of the network will present themselves.

# 1. Background and goals of the network

One of the problems facing us today is the widening gap between those who can make use of modern information technology and those who cannot. Those who can are people who have sufficient resources, skills and knowledge to do so (roughly speaking educated (often male) people in the western world). Those who cannot are people who lack sufficient resources, skills and knowledge (roughly speaking children, old people and most people in the non-western world). One of the reasons for the gap is the continuing difficulty of mastering the technology. Many people are blocked from access by keyboards, written language and flickering monitors. If the benefits of information technology in the future are going to be enjoyed by old people, people with poor writing skills etc, there has to be access to the technology through spoken language and multimodal face-to-face interaction.

For this to happen, we must have a better understanding of how naturalistic face-to-face spoken multimodal interaction works. One of the ways in which this can be done is through the collection and analysis of naturalistic spoken language corpora. Informed by this information, we can then gradually construct more accessible interfaces to technology.

Most people do not consider computers as the willing, diligent, frugal, and extremely competent assistants they are. Computers have to be addressed in artificial languages that do not at all resemble the plain vernacular we all master. Why, then, don't we talk to computers the way we do to a shop assistant, or a colleague? Because speech technology has not yet been able to produce a convincing general-purpose speech interface. This, in turn, has to do with the status of naturalistic speech in theoretical linguistics. Our understanding of the spoken language is still too metaphorical, too weak to crystalize into mathematical formulae and computer code. Above all there has been a lack of formally based studies into actual occuring speech, due to the absence of high quality speech corpora. The aim of the Nordtalk network is to produce, study, and provide corpora in all styles, from the prepared professional speech or sermon, to the syntactically relaxed informal conversation that we would ultimately like our computer to join.

The NORDTALK network has been formed to promote the following two goals:

(i) Cooperation between approaches from general linguistics, computational linguistics, language technology, speech technology and phonetics, and
(ii) The collection, annotation and analysis of spoken language corpora for the Nordic languages.

The Nordic countries are in a favourable position to become European leaders in language and speech technology. We have had considerable research in the field for almost forty years; we have a high general level of education, and we are well equipped with computers and telecommunicative hardware; we have many languages represented within a relatively small population. Moreover, we have a long tradition in keeping the democratic and educational institutions open for minorities such as the blind, the mentally disabled, and the functionally illiterate – challenged people to whom easy communication with computers would make a crucial difference.

Cooperation is needed because the area is very clearly interdisciplinary. Phonetics and speech technology is needed for theory and data on how to analyze and synthesize speech. Language technology and linguistics is needed to create models of understanding, processing and generation which can cope with naturalistic spoken language. In order to gain ecological validity for the work, we believe it is essential at this stage to work with corpora of spoken language. To increase the realism and appreciation of the complexity of the task, it is important that significant parts of these corpora are collected in different naturalistic circumstances. Internationally, work on spoken language corpora is most advanced for English. Most research has so far been done for British English but several large scale initiatives are under way for American English as well. Although such corpora to some extent already exist for some of the Nordic languages, much more work is needed. This work is needed, since spoken interactive language in many significant ways is different from written language. Differences exist at all levels:

(i) In spoken language, there is often a radical reduction of morphemes and words in pronunciation.

(ii) In spoken language it is essential to gain a better understanding of the role of prosody for information structuring and expression of attitude and emotion.

(iii) The frequencies of words, collocations and grammatical constructions are highly different in spoken and written language.

(iv) The grammar and semantics of spoken language is also significantly different from that of written language; 30-40% of all utterances consist of short utterances of 1-2-3 words with no predicative verb.

(v) In spoken language, language is often used interactively, not monologically, which gives rise to a host of problems related to issues of dialog management in creating computer supported systems for interaction human-computer.

In the proposed network, we want to promote work on all Nordic languages in the exploration of issues such as the ones above. Since direct face to face human communication is not only spoken but also multimodal, the network will stay open to incorporation of work on multimodal communication, which is now rapidly growing as a field. If the Nordic languages are going to be used not only to localize systems originating in the English speaking countries, we think it is essential that we develop research for the spoken language of each Nordic language based on naturalistic corpora in the manner suggested above. The Scandinavian languages are similar enough to provide an initial common "critical mass". Comparative research, also involving Finnish, Icelandic, Estonian (and possibly also other languages) could focus also on typical differences that have consequences for different applications.

The aim is to build a Nordic network of researchers in general and computational linguistics, phonetics, language and speech technology. A special feature of this network is that large corpora of spoken interactive language are used as a basis for research which applies language and speech technology. Such corpora already exist for Swedish (1.3 million words), Danish (1.4 million words) and Norwegian Pilot projects are ongoing. Analysis methods already developed by the Swedish and Danish researchers provide one point of departure for the network. The aim is to achieve more integration of linguistics, language and speech technology in working on spoken language. This involves for example exchange and comparison of transcription and coding conventions for interchange of data.

The overall strategic importance of the network is that researchers from traditionally separate fields are brought together in a goal-oriented cooperation. Such cooperation can involve the investigation and comparison of the doctoral level education in these areas in the Nordic countries as well as efforts of establishing larger projects including researchers from different fields (like the German Verbmobil project, which brought together different fields of expertise working towards a common goal, as a locomotive for language and speech technology research). The results of this type of research are directly applicable in systems for human-computer interaction, where we could, for example, study the short utterances, e.g. questions, short answers and feedback phrases which play an important role in this kind of interaction. The network can, thus, coordinate comparative research and provide a basis for Ph.D. education by a) bringing together researchers from four Nordic countries representing languages which are typologically close (Swedish, Danish, Norwegian, Icelandic and Faeroese) and a few that are radically different (Finnish and Estonian) and b) bringing together researchers in general linguistics, computational linguistics/language technology and speech technology. Network activities are open to researchers in this area from the Nordic countries and Baltic-Northwest Russian area. The Nordic countries have, if they cooperate, a good chance of being innovative in language and speech technology in the coming century.

Spoken language corpora and work on spoken language corpora is potentially important for Nordic languages. Some of the reasons are the following:

Spoken language corpora are today becoming an important linguistic resource. They can be used as a basis for research on speech recognition/understanding and speech synthesis. They are a necessary prerequisite for realistic quantitative descriptions of words and collocations in spoken language. Existing quantitative descriptions are

usually based on written language which in many cases has different quantitative characteristics from spoken language (cf. Allwood 1996 and later editions and Allwood 1998). Grammatical descriptions of spoken language have a use in speech and language technology, but realistic quantitative descriptions of spoken language can also be used in language teaching to aid in selection of material. Furthermore, it can be used in creating material and standards for testing in work by psychologists or speech therapists.

Besides the practical task of resource establishment, the computational part of the project also has an innovative side, viz. developing new computational methods for grammatical analysis of spoken language. Traditional grammatical frameworks are often rooted in the presupposition that utterances can and should be described as well-formed sentences, or at least as constituents in concordance with a normative set of syntax rules. Spontaneous speech does not typically consist of sentences or such (less so, the more 'spontaneous' the speech). This does not, however, imply that speech is basically disorderly, but as it seems the ordering principles of unplanned speech have more to do with regularities in information flow and in discourse coherence, than with conformance with strict rules of syntax. It hence seems natural to investigate frameworks that take these fundamental differences into consideration even in their theoretical outset. Such frameworks should strive for natural analyses of linguistic phenomena involving attitudes, feedback, anaphoric linking, information state updating, etc. The project will include comparisons of both kinds of computational grammatical frameworks: modified traditional ones and freshly developed new ones - with respect to descriptive coverage, explanatory adequacy, and computational properties.

**Goals for the project period**

The network aims to

1) Establish contacts and cooperation between researchers and Ph.D. students from the different countries and disciplines through workshops, visits and courses in cooperation with other initiatives in the Nordic countries. Researchers from speech technology and general linguistics, language technology/computational linguistics with focus on speech do not communicate and exchange information as much as would be desirable, given many overlapping areas. Cooperation is needed.

2) Make it possible for researchers and students to exchange short visits.

3) Establish bilateral and multilateral research cooperation (e.g. for developing spoken language corpora in Norwegian and Finnish and for working on spoken language grammar).

4) Initiate work to promote the availability and interchange of spoken language corpus data from the Nordic countries. A catalogue of "established" semi-orthographical forms of speech representation, comparisons and ways of transferring data between systems will be explored. Focus will be on language from spoken dialogs for dialog systems.

5) Interact with portals for written language corpora and try to make them take spoken language work and corpora are taken into consideration. Provide expertise on spoken language that is needed, in order to extend the usability of corpora

6) Establish contacts and exchange of information between academic research groups and industry.

# 2. Activities

**What we have done so far**

**2001**

• We have had an introductory planning meeting and seminar for coordinators and group leaders. "Spoken Language: Dialog and Corpus Establishment" in May 2001, in connection with the NODALIDA conference in Uppsala.

• We have had a second workshop on “Corpus formats, transciption and exchange of data” in September 2001 in connection with the Eurospeech conference in Aalborg.

Both meetings were lively and interesting and have led to several new members joining the network.

• The network has expanded so that we now have members also in Iceland, The Faeroe islands and Estonia.

• The NORDTALK network has presented at the OFTI (Spoken language interaction) conference in Växjö, in September 2001. The network met with interest also from researchers in sociology, psychology and business.

• Student and researcher exchange: 2 Scholarships for short visits to other Nordic research groups for Ph.D. students or teachers/researchers have been announced.

**Plans for the future**

• A Nordic Ph.D. course on "Utilizing Spoken Language Corpora" will be held in Göteborg in August 2002. The course will be organized around themes of corpus research.

• Two Workshops will be held in 2001, in Trondheim and Stockholm. The first workshop will focus on so called “Dysfluencies” or mechanisms through which speakers regulate their own talk, such as hesitations, repairs etc. The second conference will focus on “Dependencies between utterances”.

• Student and researcher exchange: 2 Scholarships for short visits to other Nordic research groups for Ph.D. students or teachers/researchers will be announced in 2002.

**Continued cooperation after the period**

- The network will probably apply for a third year, in order to broaden participation and establish long term contact with more researchers from more of the Nordic countries, Baltic and North West Russian researchers. Contacts with Finland are established.

- The network will devote time to planning interdisciplinary projects in Nordic subgroups, which can apply for projects to continue cooperation from national, Nordic and European funding sources.

- The network will investigate the possibilities of continuous contacts for Ph.D. guidance and education. It will also apply for future NorFA course funding.

- A plan for continued meetings in conjunction with conferences in the Nordic countries will be made.

- The project will plan a possible framework for constructing, maintaining and making available spoken language corpora for the Nordic languages and establish a reference group for future corpus work.