# SMALL CORPUS, GREAT INSTITUTION – AND AN ATTEMPT TO UNDERSTAND THEM

Vesa Heikkinen, Mikko Lounela, Research Institute for the Languages of Finland

The New Year's speech by the President of the Republic is one of the most important political speeches in Finland. We have gathered all the speeches from 1935 to 2007 into a corpus containing the speeches in writing. Our objective is to explore what the speeches are like in terms of linguistic choices and as a set or type of texts. We are also interested in the social dimensions of the speeches and the ideological meanings produced in them. This paper presents an analysis of our research questions and methods of analysis, rather than going into empirical results. We present the method and project we have decided to call "Teko" (from text to corpus), based on the compilation and structuring of small, mutually comparable corpora, as well as on detailed quantitative (corpus linguistic) and qualitative analysis (based on text analysis, applying, e.g., the process analysis of the Systemic Functional Grammar). We are considering the following research positions and questions of analysis related to them: the uniformity of the speeches as compared to another set of texts, i.e. that of news (e.g. based on their morphological features that have been analysed semi-automatically), the internal uniformity of the speeches judging by how the speakers refer to themselves (differences arising from the speakers on the one hand and the topics on the other hand) and the uniformity of the speeches on the basis of process analysis (distribution of processes by presidents and topics). Our fundamental question in this paper is how the quantitative analysis of a small corpus can be connected to a qualitative analysis of individual texts.

KEYWORDS: Finnish, New Year's speech, small corpus, Teko project and method, corpus and text analysis

## 1 INTRODUCTION[1]

In our Finnish presidential tradition, the most important annual public performance by the President of the Republic is the New Year's speech. It covers an overview of the year past and sketches an outline of the future. At the same time, the president attempts to shape conceptions of democracy and what it means to be a Finn. The president's executive power has been cut down drastically over the past few decades, and the New Year's speeches have gained more significance in the sense that they constitute a means by which the president can still address the citizens directly.

[1] Our warmest thanks to Marja Heikkinen for the translation into English.

As linguistic research material, the New Year's speeches form a consistent text set of a moderate size. The tradition of giving them begun in 1935 and has continued to the present day. The tradition has evolved during its history, yet it is one that represents stability in the Finnish national life.

Although the New Year's speeches represent the most important political speeches in Finland, they have only been studied to a limited degree so far. The presidents traditionally begin their speeches by addressing the audience as "Fellow Citizens", yet it has not been studied what kind of citizenship and leadership the speeches are used to construct. Thus, we decided to investigate at the Research Institute for the Languages of Finland what these nationally important speeches are really like. We have gathered the speeches into one corpus, and our objective in this paper is to analyse, above all, the research position and the methodical choices we made. We present the Teko project and method, centring on detailed quantitative and qualitative analysis of small corpora that have been encoded in a versatile manner and are mutually comparable. More empirical results on the research into the New Year's speeches and more detailed analysis of the material can be found in our other papers (e.g. Heikkinen 2006; Heikkinen & Lehtinen & Lounela 2005; Heikkinen & Lounela forthcoming). We have also considered questions related to the corpus and the methods in some of our other papers (e.g. Heikkinen & Lounela 2006; Heikkinen & Lounela forthcoming; Lehtinen & Lounela 2004; Lounela, forthcoming). The Teko method is part of a wider research project titled *Theory and Methods of Genres* (see KOTUS 2006).

We aim to continue both the research into the New Year's speeches – probably from a semantic/functional analysis – and the research into the other corpora obtained during the Teko project (*Teko* meaning 'from text to corpus', 'tekstistä korpukseksi' in Finnish), gradually also extending the variety of the corpora. Thus, the individual corpora of some tens of thousands of words will gradually turn into one large corpus of hundreds of thousands – even millions – of words, containing entire texts with their detailed metadata and morphological and other such analyses.

This paper covers New Year's speeches in particular. We have composed a comprehensive corpus of the New Year's speeches, and encoded it structurally and morphologically. On top of that, we have added special encoding covering the topics in the texts. As the corpus and its encoding and annotation follow a standard that has been used in previous research projects concerning, for example, newspaper texts (Heikkinen & Lehtinen & Lounela 2005), we have had the opportunity to make comparisons between certain structural and linguistic features in these small specialised corpora (see Hunston 2002, 14; see also de Beaugrande 2001; Sinclair 2001).

Tentatively, we have supplemented some of the New Year's speech material with analyses of the clause processes. In these analyses, we have applied process thinking that complies with the Systemic Functional Grammar (Halliday & Matthiessen 2004; for applications covering the Finnish Language, see Shore 1992; 1996; 2005), yet at a relatively rough level. There is still only very little research into Finnish texts carried out in the tradition of the SF Grammar; in this sense, every research project is grammatical

basic research, i.e. research yielding new information on the Finnish language from the perspective of the SF Grammar.

In this paper, we are focusing on the premises, data, and methods of analysis. In section 2 we shed light on the research questions we are seeking to answer in our project. Section 3 covers the Teko project and method, and section 4 our corpus of New Year's speeches. Section 5 shows what types of morphological (and other) features we are looking for in the quantitative analysis and includes a discussion of what our method of analysis yields. Section 6 demonstrates what information can be obtained from quantitative analysis comparing two text sets analysed in the same way: New Year's speeches and newspaper news. Section 7 includes a discussion of how our method can be used to analyse the uniformity of a text set: we aim to test the uniformity of the New Year's speeches as a text set when contemplated through certain variables – especially the linguistic choices referring to the speaker in this case. This involves testing the differences between the first-person-references by the presidents, as well as the differences detected in an analysis by topics. Finally, in section 8, we look at what happens when we supplement a partial corpus with systemic-functional process analysis. Are there differences between presidents and topics as regards these choices, and what does a method like this yield in the present research project and in a more general sense?

## 2 RESEARCH QUESTIONS

In our speech project we have both empirical and theoretical-methodical research questions. The empirical ones are as follows:

- What are the New Year's speeches like linguistically? Are they a linguistically uniform set of texts? What unites them? How do they differ from other sets of texts, if they do?
- Are there differences between the speeches by different presidents? What kinds of differences? What are they possibly related to?
- What other differences arising from external factors can be found in the texts?
- How do, e.g., topics affect the linguistic choices? What topics are dealt with? Are different topics spoken about in different ways?
- How have the speeches changed during their history? What could be the reason for these changes?
- Are we dealing with a genre, and if we are, on what grounds?
- In what way are the speeches part of influential use of language and power? What types of relations do they build between the speaker and the audience? Does the speaker represent more an individual than an institution?

Our methodical questions include the following:

- What answers to our empirical questions do the semi-automatic morphological analysis and other methods of quantitative corpus analysis provide? What answers does the qualitative differentiation of individual texts and linguistic choices made in compliance with, e.g., the SF Grammar

provide? How should different methods of analysis be combined in a sensible way?

- What presumptions and "theories" concerning language are we ready to accept whilst starting to use certain types of methods or, for example, certain automatic analysers? Are the presumptions the methods of analysis are based on mutually conflicting?
- What metadata should we supplement the texts with? How should we be prepared for possible future research needs (and possibly even for the future needs of other researchers) whilst compiling a corpus?
- What type of encoding (e.g. morphological, semantic, or having to do with the textual moves) should we add to the corpora? What accuracy, level, and quality of encoding and markup are sufficient?
- How can we combine simple morphological markup with structural encoding to obtain information meaningful for the purposes of text analysis?
- How can a linguistic genre be seen in this kind of quantitative-qualitative analysis? Does such an analysis yield data about the register, genre and ideology of the speeches – and can these data be differentiated from each other analytically?

To slightly oversimplify, we aim at combining two approaches: 1) a corpus linguistic (quantitative) one that concerns the text set and 2) a text analytical one (qualitative, partly based on the SF theory) that concerns individual texts, their structure and the individual linguistic choices in them. We presume that these approaches complete each other in the same way as we have seen the systemic theory and corpus linguistics to complete each other (see Halliday 2006, 294–295). In the end, we are faced with the question of what type of information is encoded in the corpus and what theoretical conception about language the encoded information is based on. Our basic assumption is that theory-neutral descriptions are impossible (see Matthiessen & Nesbin 1996).

The research we have carried out so far has already proved that combining the approaches is worthwhile (see Heikkinen & Lehtinen & Lounela 2005; Heikkinen & Lounela forthcoming), but there are still many points that need precision and improvement in our method. The present article is part of the process of that development.

## 3 TEKO PROJECT AND METHOD

In order to answer our research questions as validly as possible we have been working on a qualitative-quantitative research method. This method we are calling the Teko method. The name comes from the Finnish words "tekstistä korpukseksi", "from text to corpus" in English. The introduction of the Teko method has been partly based on reasons external to research: the entire Teko project is based on a practice where individual writers of academic papers and other researchers compile corpora of moderate size in accordance with given principles, which are then analysed according to

a joint model. Thus, both the corpora and the methods used to process and analyse them, together with the results, are mutually comparable.

Combining qualitative and quantitative research methods is, of course, by no means a new invention as such; nor is our method unique. To some degree, even the most "qualitative" – e.g. systemic – linguistics has to take a stand towards how frequent the phenomena it analyses are and what is the relationship of an individual corpus with "language". Likewise, corpus researchers are always faced with qualitative and theoretical questions. As M. A. K. Halliday states, systemic linguists have always tried to base their descriptions on observable data, and – on the other hand – collecting, managing and interpreting corpus findings are a highly theoretical activity as such (2006, 295). From the perspective of a traditional division we can see that Systemic Functional Linguistics is a theory of language and Corpus Linguistics is a method for investigating language, yet this dichotomy is too simplistic in reality. Geoff Thompson and Susan Hunston put it this way: "SFL is increasingly concerned with methods of quantifying linguistic features, and CL is becoming more intent on developing theories to account for its findings." (2006, 1.)

A prominent part of the research into the Finnish language in the late 20th century was also qualitative-quantitative: e.g., linguistic style analysis of the "Oulu corpus" (the results are summarized in Saukkonen 1984; 2001) and syntactic research into textual clauses in the Finnish language (Hakulinen & Karlsson & Vilkuna 1980). In their extensive quantitative research project on the features of Finnish textual clauses, Auli Hakulinen, Fred Karlsson and Maria Vilkuna state that the validity of a quantitative analysis essentially has to do with the theoretical analysis on which the statistics are based (ibid: 2). By theoretical analysis, they are referring to qualitative syntactic research.

In Fennistics in general, and in the Research Institute for the Languages of Finland in particular, corpus work has been a traditional branch of activity. Early Finnish and dialects have been researched with large materials since the foundation of the institute. (See KOTUS 2008.) Critical text analysis is slightly more recent, and its introduction coincided with the first instances where the new opportunities of corpus work were used at our institute (using large storage methods; XML; morphological analysers). Thus, even the research projects carried out at the Research Institute for the Languages of Finland have been affected by technological development which makes linguistic description subject to many new requirements on the one hand, and enables descriptions much more comprehensive on the other (see Matthiessen & Nesbit 1996, 40).

As critical text analysis has used small sets of whole texts in the qualitative work, the thought of using carefully encoded text sets of moderate size to apply the qualitative method to enrich the analysis emerged naturally. Since then, text sets have been collected in connection with the Teko project (Lounela forthcoming) and used for text analysis (Heikkinen 1999; Heikkinen & Lehtinen & Lounela 2005; Kankaanpää 2006; Tiillilä 2007) while the method has been developed by, e.g., drafting precise instructions

for corpus construction and a detailed disambiguation guide, necessary for semi-automatic morphological analysis.

Our method involves composing a mini-corpus, encoding it in XML for basic text structure (special markup can be added for the needs of the research project), annotating semi-automatically to the morphological level (Lehtinen & Lounela 2004), and compiling morphological standard reports automatically from the text sets (Lounela 2005). In many cases, the texts in the corpus have already been made subject to preliminary qualitative text analysis, and on the basis of the quantitative reports, we can take the qualitative text analysis even further. Combining qualitative and quantitative analysis like this often makes it necessary to supplement the corpus with new analyses and conduct new calculations and cross-tabulations on their basis.

By increasing the number of the text sets we attain a large range of sets of readily compiled morphological and structural information which can be compared with each other. The research process is carried out by investigating the morphological reports and exploring the text in a qualitative manner at the same time. Many times the quantitative data directs the qualitative researcher to ask specific questions about the texts and this, in turn, may lead to a need to add some new encoding to the texts and perform new calculations to support the analysis.

After the collection work and basic digitising, the speeches are normalised and structured semi-automatically in the TEI P4 XML format (TEI 2004). After that, the text is run through a morphological analyser (Fintwol, see TWOL 2008), and disambiguated by hand. The method is somewhat time-consuming, but we hope that the result is more accurate than one based on using a readily disambiguating parser. Also, while disambiguating their material (or supervising the work), the researchers can take a close look at the texts from a new point of view, which is beneficial when forming and evaluating the research questions and the need for further markup and calculations. In the analysis of "small corpora" like these, it seems especially natural to combine qualitative and quantitative analysis. Following the calculation, we are faced with a comprehensive sample of numbers and distribution lists. These numbers and lists need to be interpreted in a qualitative manner, examining the texts closely, one by one.

What we want to grasp is – neither more nor less than – the meaning potential activated in the speeches (see, e.g., Butt 1996, xv–xvi; Halliday 1996, 4–5; Thompson & Hunston 2006, 2). We trust that the countable features "are the manifestations of fundamental grammatical properties" (Halliday ibid: 25). We are interested in, e.g., what it means that certain linguistic features are general or rare. In fact, our research has indicated that even a very rare feature, e.g., an open first-person reference, may weigh heavily when the meanings of a textual entity are being interpreted. We have also managed to show that individual features, such as first-person references, may have ideational, textual, and interpersonal functions in the New Year's speeches. (See Heikkinen & Lounela forthcoming.) We accept as one of our premises the fact that morphological features are immediately connected with the activation of the meaning potential (see Halliday ibid. 25; see also Butt ibid.).

An important part of the Teko method is that relevant metadata are added to the corpora. Thus, e.g., every New Year's speech is marked with metadata concerning the speaker and the point of time at which the speech was given. Yet, information of the metadata type can also be added inside the individual texts. The special characteristics of the New Year's speech corpus led us to supplement the texts with information on the topics. The fact is – and this might be a generic feature – that the New Year's speeches display a relatively established structure based on the topics dealt with. We have divided the topics into three categories at the most general level of abstraction: home country; world; general. The topics are encoded in each text passage. A text passage is a typographic element in the original written text. By topic, we refer to speech topic, i.e. what the text passages are about.

Besides going from the quantitative to the qualitative and vice versa in the analysis of individual Teko corpora, we also move between different Teko corpora and compare partial corpora with each other. The Teko method provides a steady basis for such comparisons. Essentially, the Teko project has consisted and still consists in the gradual compilation of small comparable special corpora (see table 1). These corpora can be continuously supplemented with new encoding according to the relevant research questions. The different partial corpora need to be supplemented with relevant additional encoding; yet it is possible to make the partial corpora more uniform by supplementing as many of them as possible with the same additional encodings.

Table 1: Teko materials

| **Text set** | **Size (words)** | **Time span** | **Special encodings** |
|---|---|---|---|
| Texts from a weekly journal | 680 936 | 1917–1972 | - |
| Presidents' New Year's speeches | 63 110 | 1935–2007 | Text passage topic |
| Administrative press releases | 19 065 | 1979–1999 | Named entity references, addresses |
| News on plain language | 14 530 | 2001–2003 | - |
| Guidelines given by church administration | 7 639 | 2002 | Modal verb chains |
| Short news from local newspapers | 97 325 | 2002 | - |
| Handbooks by tax administration | 18 591 | 2002 | - |
| Laws and directives | 232 449 | 2002–2003 | - |
| Communal introductory www-pages | 23 256 | 2004 | - |
| Benefit decisions by the social security administration | 10 690 | 1992–2003 | Functional stages |

## 4 THE NEW YEAR'S SPEECH CORPUS

The corpus of the presidents' New Year's speeches is small by size, yet at the same time it is (almost) as big as it can be. It includes all the New Year's speeches held by presidents of Finland (or, in certain specific cases, their substitutes) from the beginning of the tradition to the year 2007 and, since the tradition continues, our corpus grows by one text every year. The New Year's speech corpus, just as any other Teko corpus can be used as an example of specialised corpus, i.e. a corpus of texts of a particular type, aiming to be representative of a given type of text (Hunston 2002, 14). Another good thing about the Teko corpora is that they consist of "real language", and can be used "to establish the probability profiles of major grammatical systems", as well as "to investigate register variation in grammatical terms" (Halliday 1996, 25–26).

The corpus includes 73 speeches, altogether 60 485 words. Twelve people have given the speeches; the most prominent of them being President Urho Kekkonen, who gave as many as 25 New Year's speeches. (Table 2.) The speeches in the corpus are in written form: we have collected them from newspapers, archives and the internet. The New Year's speech corpus interests even other people than researchers, which is why we have decided to make all the texts available at RILF's corpus service on the internet (see KAINO 2008; Lounela 2007).

Table 2: New Year's speech corpus: years, speakers and sources of material

| **Year** | **Speaker** | **Source** |
|---|---|---|
| 2007–2001 | Halonen | Internet |
| 2000–1995 | Ahtisaari | Internet |
| 1994 | Koivisto | Press release |
| 1993 | Aho (PM) | Press release |
| 1992–1983 | Koivisto | Press release |
| 1982 | Koivisto (PM) | Press release |
| 1981–1968 | Kekkonen | Press release |
| 1967–1957 | Kekkonen | Extracted from the work Speeches and Writings 2 |
| 1956 | Paasikivi | Newspaper cutting and archived writing |
| 1955–1947 | Paasikivi | Newspaper cutting |
| 1946 | Paasikivi (PM) | Newspaper cutting |
| 1945 | Pekkala (Minister) | Newspaper cutting |
| 1944 | Linkomies (PM) | Newspaper cutting |
| 1943 | Ryti | Newspaper cutting |
| 1942 | Hakkila (Speaker of the Parliament) | Newspaper cutting |
| 1941 | Ryti | Newspaper cutting |
| 1940 | Kallio | Archived writing |

| | | |
|---|---|---|
| 1939–1938 | Kallio | Newspaper cutting and archived writing |
| 1937–1935 | Svinhufvud | Newspaper cutting and archived writing |

## 5 FEATURES IN THE QUANTITATIVE ANALYSIS

As mentioned above, the Teko method involves the chance to run standard morphological reports for each material (or a part of them), and to make comparisons between these reports. The calculations in the standard reports may raise further questions, in which case special reports can be created. This requires extra programming work, but as the procedure and the tools are available, it can be done with reasonable effort. The function of the reports in connection with the qualitative interpretational analysis is two-fold: on the one hand, it directs the process of forming the research questions, and on the other hand it may support or challenge the hypotheses formed during the qualitative analysis.

In this section we will be considering the morphological features available in the process. First we will concentrate on the standard features, e.g., the features that are readily available in the standard reports. After that we will focus on the special features, i.e. the features we have chosen to pay special attention to, along with the features especially calculated for this particular project. (See Appendix.)

### 5.1 STANDARD FEATURES AND THEIR INTERPRETATION

The standard morphological and structural features of the Teko system are organised in four different reports. The reports describe
1) the material in general
2) the properties of the nominals in the material
3) the properties of the verbs in the material
4) the distributions of the words according to the parts-of-speech.

The general standard features (table 3) include general information on the number and length of different structural units in the texts. These units include the texts themselves; sentences; clauses; punctuation; and words. The further morphological features include frequency lists of parts of speech; possessive suffixes; compound word lengths (word parts); and the most common words.

Table 3: General standard features

| **Feature** | **Comment** |
|---|---|
| Number of texts in the corpus | |
| Number of sentences in the corpus | |
| Number of clauses in the corpus | The number of clauses is same as the number of finite verbs |

| | |
|---|---|
| Number of unknown words in the corpus | Unknown words are words that are not understood by the Fintwol morphological analyser |
| Number of punctuation marks in the corpus | |
| Average text length in sentences | |
| Average text length in clauses | |
| Average text length in words | |
| Average clause length in words | |
| Average sentence length in words | |
| Average clause length in words | |
| Frequency list of punctuation marks | |
| Frequency list of parts of speech | |
| Frequency list of possessive suffixes | |
| Frequency list of compound word lengths | How many parts the compound has |
| Frequency list of most general words | Both lemmas and word-forms |

To illustrate a frequency list, table 4 presents a partial frequency list of parts of speech in President Ahtisaari's New Year's speeches. The list gives the parts of speech, the number of their occurrences, and their proportion of all the words, in order of their total number in the text set. The part-of-speech analysis materializes in the fact that no linguistic analysis or description is free from theory. When implementing a given automatic program of analysis, we also acquire a conception of language and, for example, a conception of the part-of-speech system. It is methodically interesting that the different programs of analysis offer different conceptions of a variable as general and basic as the part of speech. (See Heikkinen & Lounela 2006.)

Table 4: Example of a frequency list

| **Type** | **Number** | **Proportion** (%) |
|---|---|---|
| Noun | 2043 | 39.24 |
| Verb | 933 | 17.92 |
| Adjective | 650 | 12.49 |
| Pronoun | 373 | 7.16 |

The standard feature report for nominals (see table 5) includes some of the general features counted from the set of nominals along with some specific features that are meaningful only in connection with them. Nominals include all the nouns, numerals, adjectives, and pronouns in the text.

Table 5: Standard features for nominals

| Feature | Comment |
|---|---|
| Number of nominals | |
| Number of different lemmas of nominals | |
| Number of different word-forms of nominals | |
| Proportion of nominals of all the words | |
| Frequency list of parts of speech of nominals | |
| Frequency list of possessive suffixes of nominals | |
| Frequency list of compound word lengths of nominals | |
| Frequency list of lemmas of nominals | |
| Frequency list of word-forms of nominals | |

In the verb report (see table 6), the verbs have been divided into three overlapping classes for calculations. The "grammatical" verbs are the ones carrying the morphological markers characteristic to the verbs; these include basic verbs (excluding infinitives), and auxiliary verbs in the temporal verb chains. The "semantic" verbs are the ones carrying the meaning in the verbal expressions, including the participle forms in the temporal and negative verb chains instead of the auxiliaries. Finally, the set of the finite verbs is based on the semantic verbs, choosing from there only the ones where the active or passive marker is present, and excluding the infinitive forms. The finite verbs carry special weight, the number of clauses in the material being the same as the number of the finite verbs.

Table 6: Standard features for verbs

| Feature | Comment |
|---|---|
| The number of grammatical verbs in the material | |
| The number of semantic verbs in the material | |
| The number of finite verbs (clauses) in the material | |
| The number of the different word-forms of the semantic verbs in the material | |
| The number of the different lemmas of the semantic verbs in the material | |
| The number of the different word-forms of the grammatical verbs in the material | |
| The number of the different lemmas of the grammatical verbs in the material | |
| The number of the different word-forms of the finite verbs in the material | |
| The number of the different lemmas of the finite verbs in the material | |
| Proportion of semantic verbs of total words | |
| Proportion of grammatical verbs of total words | |
| Proportion of finite verbs of total words | |

| | |
|---|---|
| Frequency list of voices of verbs | All the verbs having a voice |
| Frequency list of moods of verbs | Of finite verbs |
| Frequency list of tempora of verbs | Of finite verbs |
| Frequency list of infinitive types of verbs | All the words of part of speech "V" |
| Frequency list of types of participles | |
| Frequency list of parts of speech of first participles | |
| Frequency list of parts of speech of second participles | |
| Frequent word-forms of the semantic verbs | |
| Frequent lemmas of the semantic verbs | |
| Frequent word-forms of the grammatical verbs | |
| Frequent lemmas of the grammatical verbs | |
| Frequent word-forms of the finite verbs | |
| Frequent lemmas of the finite verbs | |

In addition to the three reports explained above, there is a fourth one that concentrates on the vocabulary of the text set. There, the most common lemmas and word-forms are arranged into frequency lists according to the parts of speech. Also, some basic numbers are calculated for each part of speech. The numbers are: number of words, different lemmas and different word-forms carrying the part of speech marker. Table 7 shows a fragment of a frequency list for adverbs in the presidents' New Year's speeches (English translations are added).

Table 7: Example of a lexical frequency list (adverbs in the speeches)

| **Word** | **Number** | **Proportion (%)** |
|---|---|---|
| Myös (also) | 365/4249 | 8.6 |
| Nyt (now) | 185 | 4.4 |
| Kuitenkin (however) | 137 | 3.2 |
| Vielä (yet) | 112 | 2.6 |
| Jo (already) | 106 | 2.5 |
| Vain (only) | 104 | 2.4 |
| Edelleen (still) | 96 | 2.3 |

## 5.2 SPECIAL FEATURES FOR THE CURRENT PROJECT

In this project, we have taken a special focus on the use of the first person singular, i.e. the manner in which the presidents refer to themselves, or refer to something else using the first person singular. For this, we have calculated certain special morphological features (see table 8). These features include (for ease of use) ones that are also included in the general reports above, and ones that are not included in them. The ones not included are the proportions of first person singular pronouns, first person singular verbs, and first person singular possessive suffixes, along with the frequency lists of word-forms of the first person singular verbs. In addition to the report, special searches have been performed, e.g., by searching for all the sentences with negative or temporal verb chains where the auxiliary is in the first person singular. They have been listed to identify the first person singular semantic verbs which are not marked with any person marker. For an exhaustive list of the features, see table 8.

Table 8: General standard features

| **Feature** | **Comment** |
|---|---|
| Number of texts | For both text sets |
| Number of sentences | For both text sets |
| Number of clauses | For both text sets |
| Number of words | For both text sets |
| Average text length in words | For both text sets |
| Average clause length in words | For both text sets |
| Average sentence length in words | For both text sets |
| Frequency list of parts of speech | For both text sets |
| Proportion of 1st person singular personal pronouns compared to all personal pronouns | For both text sets, and individual presidents. |
| Proportion of 1st person plural personal pronouns compared to all personal pronouns | For both text sets, and individual presidents. |
| Proportion of verbs in 1st person singular compared to all verbs with person indicators | For both text sets, and individual presidents. |
| Proportion of verbs in 1st person plural compared to all verbs with person indicators | For both text sets, and individual presidents. |
| Proportion of 1st person singular possessive suffix compared to all possessive suffixes | For both text sets, and individual presidents. |
| Proportion of 1st person plural possessive suffix compared to all possessive suffixes | For both text sets, and individual presidents. |
| Words relating directly to mental processes (to hope, to know, to believe) | For sets of passages according to topics |
| Frequency list of all the verbs in 1st person singular | For both text sets, individual presidents, and topics |

| | |
|---|---|
| Frequency list of all the words with 1st person singular possessive suffix | For both text sets and individual presidents |
| List of all the sentences with 1st person singular forms for negation, copula or temporal auxiliary | For individual presidents |

The basic reports have been calculated from the whole set of speeches and from the newspaper corpus that is used for comparison, as well as from the set of speeches by each individual president. The special features for the first person singular have been calculated for both corpora, for each president, and finally for each of the three text passage topics in the texts of the speech corpus.

## 6 SPEECHES AND NEWS ARTICLES

A significant strength of the Teko corpora and method lies in the fact that the corpora and the basic reports run on them are comparable. Thus, as the corpora accumulate, we can find out how a given text set may differ from other text sets as regards, e.g., its morphological features. Such an analysis also provides information about the linguistic differences between the presumed genres. In what way could the speeches be a distinctive set of texts as compared with another set of texts? In the following, we will demonstrate this by comparing two Teko corpora with each other – the New Year's speeches and news articles.

Our newspaper corpus contains 583 brief news articles from eight Finnish provincial newspapers: Aamulehti, Etelä-Saimaa, Keskisuomalainen, Kouvolan Sanomat, Lapin Kansa, Länsi-Suomi, Savon Sanomat and Turun Sanomat. Some of the texts have been taken from the ordinary news pages of the provincial newspapers (i.e. not from the front page, the main news page, nor from the sections of the newspaper covering economy, culture, or any other special topic). The newspapers were published in March 2002.

In table 9, we can see the sizes of the text sets along with information on the mean lengths of certain structural units (texts, sentences, and clauses) in the sets. We can notice that the units in the speeches are in all cases somewhat longer than the corresponding units in the news articles.

Table 9: Certain general features in the speeches and news articles

| | **Speeches** | **News** |
|---|---|---|
| Texts | 73 | 583 |
| Sentences | 4446 | 7625 |
| Words | 60485 | 89400 |
| Text length (in words) | 828.6 | 153.3 |
| Sentence length | 13.6 | 11.0 |
| Clause length | 7.6 | 6.9 |

However, the comparison of the average length of clauses and sentences does not necessarily shed light on, e.g., what the New Year's speeches are like as compared with the news articles in this sense. In fact, it has been shown in other contexts that, e.g., the sentences in the New Year's speeches have been radically shortened over the years. While the average sentence length in the speeches given by President Ryti in the 1930s exceeded twenty words, that of the speeches given by Ahtisaari and Halonen in the 1990s and 2000s remains around ten words. It seems that the average total length of the speeches is established as something between 800 and 900 words. The speeches were at their shortest in the early years of the tradition, Svinhufvud giving speeches of around 200 words. The longest speeches were given by President Ryti, amounting to around one thousand words. We could interpret this by concluding that the speeches have been affected by "mediatization", i.e. the fact that the special nature and requirements of the media through which the speeches are broadcast and presented are recognized (Heikkinen 2006, 176.) Further, we must note that there is considerable variation in, e.g., the length of the texts even within the news articles. Therefore, it is worthwhile to investigate these variables not only by comparing whole corpora but also looking at parts of them (see Heikkinen & Lehtinen & Lounela 2005).

One of our research questions is whether we can talk about a genre. What are the New Year's speeches like essentially, what is the linguistic nature of the genre? Do the speeches differ from, e.g., news articles in terms of the distribution of parts of speech? As we know, the distribution of the parts of speech has often been used as a corpus linguistic variable and in the comparison between different genres.

Table 10: Distribution of parts of speech in the speeches and the news articles

| **Part of speech** | **Speeches** | **News** |
|---|---|---|
| Nouns | 35.9 | 44.7 |
| Verbs | 17.5 | 17.8 |
| Adjectives | 11.9 | 6.5 |
| Adverbs | 7.0 | 6.7 |
| Conjunctions | 7.6 | 5.7 |
| Pronouns | 8.1 | 4.8 |
| Numerals | 1.6 | 3.6 |

The most striking feature in table 10 is that the speeches have considerably fewer nouns and more adjectives and pronouns than the newspapers. These calculations lead us to qualitative questions. Why are there, for example, clearly fewer nouns and clearly more adjectives in the speeches than in the news articles in relative terms? To answer this question, we need to revert to the qualitative analysis of the text.

The wealth of nouns in the news articles can be explained by, e.g., the fact that the participants and circumstances in the processes presented are often named in the news articles in a detailed manner, cf. the following passage from the Länsi Suomi newspaper (12 March 2002): *The city is committed to the following projects: Puupelletti project by the energy office of Satakunta province; Leather Centre Fennica at the adult education centre of the municipality*

*of Huittinen; and the Four parts of nature – land, water and myself at the Christian Institute of Eurajoki municipality.* As for the large number of adjectives found in the speeches, they can be explained by the need to classify and determine issues by relational clauses in speeches. The typical clauses in the speeches include relational clauses, many of which are attributive, such as *Competition is important; We need to be persistent and look into the future, Finland's relations to our neighbouring countries are excellent; We are more and more dependent on each other.* These examples are from President Halonen's speech of 2001.

We can also study the differences and similarities between the text sets by looking at the differences between the lists of the most frequent words. Table 11 shows the lists of the most frequent words in two text sets. The common grammatical words are frequent in these corpora. We could compare them with the Parole corpus of the Finnish language, which includes millions of words: the same words are the most frequent (with the example of the noun 'year'). It seems that we will not find any significant generic differences by looking at these variables. It can be noted that there are only 6 894 different lemmas in the speech corpus, whereas the number of the different lemmas in the newspaper corpus is practically twice as big; it is 14 705. Even though the sizes of the text sets (in words) are much closer to each other; the newspaper corpus is one third bigger than the speech corpus. So, the vocabulary of the speeches seems to be more restricted than the vocabulary of the newspaper text. The fact that the proportions of the common words are bigger in the speech corpus than in the newspaper corpus illustrates this; cf. table 11.

Table 11: Most frequent words, proportions of all the words in the texts

| **Speeches** | **News** |
|---|---|
| Olla (to be) 7.3 (4388) | Olla 4.9 (4412) |
| Ja (and) 4.0 | Ja 2.8 |
| Se (it) 1.7 | Ei 1.2 |
| Vuosi (year) 1.6 | Se 1.0 |
| Että (that as a SUB) 1.3 | Vuosi 0.9 |
| Ei (negation) 1.1 | Että 0.8 |
| Tämä (this) 1.0 | Joka 0.7 |
| Me (we) 1.0 | Myös (also) 0.6 |
| Joka (which) 1.0 | Saada (to get) 0.5 |

Table 12 shows the most frequent nominals in the two text sets. Both lists include – somewhat surprisingly – the words 'year', 'Finnish/Finland', and 'time'. Both the speeches and the news articles reflect a need to anchor events to time and place.

Table 12: Most frequent nominals, proportions of all nominals

| **Speeches** | **News** |
|---|---|
| Se (it) 19.1 (1031) | Se 7.3 (885) |
| Vuosi (year) 17.4 | Vuosi 6.3 |

| | |
|---|---|
| Tämä (this) 11.8 | Joka 5.3 |
| Me (we) 11.5 | Hän (she/he) 3.6 |
| Joka (which) 10.9 | Tämä 3.1 |
| Suomi/suomi (Finland/Finnish) 7.9 | Aika 2.4 |
| Maa (country) 7.3 | Uusi (new) 2.3 |
| Aika (time) 7.3 | Mies (man) 2.3 |
| Kaikki (every/all) 6.0 | Muu (other) 2.3 |
| Hyvä (good) 5.5 | Suomi/suomi 2.2 |

As we look at the words on this list, we can see that the speeches deal with topical issues in the republic and the world, whereas the news articles cover topics that are relevant to certain, rather restricted areas and to the people living in them. It also seems that the speeches describe reality in a more abstract way than the news articles. The news topics are more versatile and the descriptions in them are more concrete, referring to persons, times, and places (see Heikkinen & Lehtinen & Lounela 2005). There are also word lists available for each president and, indeed, differences can be found between the different presidents; e.g., President Kekkonen made frequent use of vocabulary belonging to the sphere of economy (Heikkinen 2006, 183).

Earlier we saw that the speeches include many more pronouns than the news articles. We could make the assumption that from a textual point of view, it is important what type of an author (and reader) is constructed through the speeches. A significant linguistic choice in such construction is that of person references. Our analysis yields information on various person references. Table 13 displays features that refer to the first person singular and plural: the person forms of the verbs, first person pronouns, and the person suffixes of nouns. The speeches refer to 'me' and 'us' clearly more often than the newspapers, and in different ways. This could be expected in many senses, and not least because the New Year's speeches were originally meant as texts spoken by one person to other people, whereas authorship is less evident and visible in the news articles.

Table 13: Features referring to first person in the speeches and news articles

| **Feature** | **Speeches** | **News** |
|---|---|---|
| Verbs in 1st person singular | 5.6% of verbs with person inflection | 1.5% |
| Verbs in 1st person plural | 8.8% of verbs with person inflection | 1.0% |
| Minä (I, me) | 7.8% of personal pronouns | 4.3% |
| Me (we, us) | 75.7% of personal pronouns | 13.6% |
| 1st person singular possessive suffix | 8.2% of possessive suffixes | 3.3% |
| 1st person plural possessive suffix | 61.8% of possessive suffixes | 4.4% |

## 7 HOW TO INVESTIGATE THE UNIFORMITY OF THE TEXTS SETS?

### 7.1 DIFFERENCES BETWEEN PRESIDENTS

Before, we have shown that certain differences between two text sets can be identified through quantitative comparisons. We have also shown that it is fruitful to revert from a whole corpus to an individual text, when looking for an explanation for quantitative data.

On the basis of our demonstration it seems that the New Year's speeches share a set of features that make them distinct from news articles. Yet, as we have already suggested above, the New Year's speeches are not a uniform group of texts, either. Thus, we need to look into how heterogeneous the set of texts we are considering here is. We will begin testing this by an analysis of the differences between the New Year's speeches, first for each president and then for each topic. As a basis for this demonstration we have chosen one feature or series of features, i.e. the open references to the first person singular and plural.

Table 14: Features referring to first person in the speeches of three presidents

| **Feature** | **Svinhufvud** | **Kekkonen** | **Halonen** |
|---|---|---|---|
| Verbs in 1st person singular | 3.8% of verbs with person inflection | 7.1% | 7.7% |
| Verbs in 1st person plural | 27.8% of verbs with person inflection | 6.6% | 12.3% |
| Minä (I, me) | 6.6% of personal pronouns | 13.3% | 1.5% |
| Me (we, us) | 93.3% of personal pronouns | 70.9% | 79.2% |
| 1st person singular possessive suffix | 0.0% of possessive suffixes | 11.2% | 11.0% |
| 1st person plural possessive suffix | 81.0% of possessive suffixes | 62.6% | 57.6% |

In table 14 we can see that there are great differences between the presidents. We can also see the use of certain first person features by Svinhufvud who was president in the 1930s; Kekkonen whose presidency has been the longest in Finland so far, and our current President Halonen. By way of generalization we can say that Svinhufvud talked about 'us' more than the others, and only rarely about himself; whereas Kekkonen made frequent references to both himself and us. As for the current President Halonen, she avoids using the pronouns 'I' and 'me', and prefers 'we' and 'us'. Methodically, we should, of course, note that these references generate many different meanings in the texts. For example, President Kekkonen often used I/me references for intertextual and metatextual purposes, or for "the management of the textual structure", i.e. whilst referring to other texts he had read, explaining the origins of his own speech, or commenting on its contents (see Heikkinen & Lounela forthcoming.)

References to the first person do not seem to be a self-evident feature joining a text set or a "generic feature". The speech genre offers the speakers a freedom of choice in terms of how often they make references to me and us. But could there be another factor behind these references, for example, the topic that is being spoken about?

### 7.2 DIFFERENCES BETWEEN TOPICS

We understand speech texts as semantic units and units comprising different parts or moves or stages (e.g. Eggins 1994). We assume that the different parts of the texts reflect different linguistic choices that have to do with, for example, what the texts are about. We could imagine that Russia or the Soviet Union is talked about in a different way than the US or the European Union. But how could this be studied?

It is typical of the speech corpus that the texts display fairly clear topics and boundaries between topics. Thus, we have decided to encode the topics in the corpus for each text passage. A text passage is a typographic element in the original written text. By topic, we refer to speech topic, i.e. what the text passages are about. We have divided the topics into three categories at the most general level of abstraction: home country, world, and general.

Table 15: Distribution of topics (of text passages) by presidents

| **President** | **Home country** | **World** | **General** |
|---|---|---|---|
| Svinhufvud | 11 | 0 | 3 |
| Kallio | 80 | 17 | 7 |
| Ryti | 23 | 1 | 5 |
| Paasikivi | 147 | 25 | 24 |
| Kekkonen | 226 | 59 | 48 |
| Koivisto | 236 | 165 | 21 |
| Ahtisaari | 116 | 54 | 9 |
| Halonen | 62 | 46 | 13 |
| Others | 59 | 9 | 2 |

The most popular topic in the entire corpus has clearly been 'the home country'. The presidents have used a total of 42 342 words to deal with matters having to do with that particular topic, whereas the topic 'world' has only required a total of 13 550 words, and that of general matters just 2 601 words. Table 15 allows us to make interpretations about the types of topics each of the presidents has favoured. We can complete these calculations by looking at words, rather than paragraphs, per topics. For example, the paragraphs in President Koivisto's speeches were generally very short, which explains the large number of topics as compared with, e.g., Kekkonen, who gave many more speeches than Koivisto. Thus, the features analysed are relative in many ways.

The overall picture can also be specified in other ways, e.g., by analysing what types of topics have been popular in the openings or endings of the speeches. The most popular

topics in the first passages of the speeches represent the category "general". However, there are nearly as many topics belonging to the category "home country". The opening topics display interesting differences per different presidents. For example, Koivisto had a distinctive habit of opening his speeches with topics from the "world" category. The endings are clearly dominated by the topic "general" with each president. The last paragraph typically involves best wishes for the New Year. (Heikkinen & Lounela forthcoming.)

We can combine the categorization of the topics with all the information we have gained by analyzing the speeches. For example, we can try to find out whether the topic has an influence on the way in which the speakers refer to themselves (see table 16). It would seem that topic does matter. The references to the first person singular are the most common in the passages with a 'general' topic. This could be compared with, e.g., the available information on the topics that favour references to the first person plural. They are the most popular in the topics belonging to the category "home country".

Table 16: Features referring to first person singular according to the text passage topics

| Feature | Home country | World | General |
|---|---|---|---|
| Verbs in 1st person singular | 4.4% of verbs with person inflection | 3.3% | 33.1% |
| Minä (I, me) | 6.7% of personal pronouns | 6.9% | 31.8% |
| 1st person singular possessive suffix | 7.0% of possessive suffixes | 6.7% | 26.6% |

Table 17 shows the outcome of an analysis of the most frequent verbs by topics. Here we are looking at the first person singular only, i.e. the choices that the speakers make to refer to themselves. The first column ("all topics") shows that the first person singular is used more often with verbs that express mental and relational processes. Let us now revert to the corpus and look at where such cases are found. It does seem that this combination of features (first person singular plus the most popular verbs) also varies according to topic.

Table 17: Verbs in first person singular according to the text passage topics

| All topics | Home country | World | General |
|---|---|---|---|
| To be 21.8% (84) | To be 27.3% (59) | To be 19.6% (10) | To wish 47.5% (48) |
| To wish (smbd sth) 13.7% | Not 8.8% | To like 11.8% | To thank 9.9% |
| To want 6.7% | To hope 7.9% | To believe 9.8% | To be 7.9% |
| To hope 6.2% | To want 5.1% | Not 7.8% | To want 7.9% |
| Not 6.0% | To believe 4.2% | To want 7.8% | To hope 4.0% |
| To believe 3.9% | To say 3.2% | To state or note 7.8% | To express 3.0% |

It is interesting even in a more general sense what types of verbs are favoured in the different topics. Thus, we decided to count how the most general verbs expressing different mental processes are distributed between different topics. Table 18 illustrates the occurrences of a couple of common verbs that express mental processes. We can see that different verbs expressing mental processes are preferred in different topics. When the speakers are talking about the world, they often talk about hoping and believing, and when they are talking about the home country, they frequently talk about knowing.

Table 18: Words referring to mental processes in speeches, according to topic

| **All** | **Home country** | **World** | **General** |
|---|---|---|---|
| To hope 56 | To hope 34 | To hope 18 | To hope 4 |
| To know 33 | To know 24 | To believe 10 | To believe 4 |
| To believe 32 | To believe 18 | To know 7 | To know 2 |

## 8 PROCESS ANALYSIS AND ADDING IT TO THE CORPORA

Above we have considered the opportunities and challenges of the analysis and interpretation of different data available on a corpus. One significant advantage in such a combination of qualitative and quantitative analysis is that it generates constantly more precise research questions – at the same time creating a need to analyse the data further and supplement the corpora with new analyses, thus making it possible to take them into account in different calculations.

The analysis of the first person singular references in the New Year's speeches activates the question of what types of verbal processes "I/me" and "we/us" participate in, together with the more general question of what types of verbal processes are constructed in the New Year's speeches in general. At this stage, we aim to expand the method with a process analysis complying with the Systemic Functional Grammar. It is not an easy task for many reasons, but we see it as worthwhile, especially since similar ventures have already been taken by other researchers, albeit concerning the English language (e.g. Neale 2006).

Our starting point in the categorization of the processes of the New Year's speeches was that of the Systemic Functional Grammar (Halliday & Matthiessen 2004), paying special attention to the characteristics of the Finnish language (Shore 1992; 1996; 2005; also ISK 2004). First, we had to answer the question of what a process is and how processes are generally expressed in Finnish. Whilst marking the processes in the New Year's speech corpus, we decided to omit to encode as processes modal verbs or verbal structures that have clearly specific interpersonal functions in the texts: *voida (can/be able to/be allowed to)*, *saada (be allowed to)*, *saattaa (may/might)*, *sopia*, *taitaa (seem)*, *mahtaa (can)*, *pystyä(can/be able to)*, *kyetä (be able to)*, *päästä (be allowed to/get)*, *pitää (need to)*, *täytyä* (*must/have to) tulla (be/come)*, *joutua (have to)*, *tarvita (need to)*; *on -(t)tAvA* (see Kangasniemi 1992). We also decided to settle with dividing the processes into three main categories

at this stage: material, mental and relational. What has been especially challenging in this work is that there is no thorough description of the Finnish language in compliance with the SF Grammar available.

Whilst talking about the processes, we refer roughly to the verbs carrying the core ideational contents to the individual clauses. The processes have been encoded manually on the finite verbs; this means that we are dealing with the core process of the clause. Compared to a morphological analysis, this method is probably more subjective – we should perhaps consider whether a set of 'watertight' encoding guidelines could be established.

We have based our categorization on what has been said about process analysis in Halliday's grammar and other sources of the SF Grammar. We have paid special attention to the specific characteristics of the Finnish language by looking at the syntactic and other conditions that have to do with the possibilities of the Finnish language to express different processes. Table 19 displays some of the most obvious of them (see also Shore 1996).

Table 19: Grammatical features of the process categories (+ = valid, - = not valid, + / B = valid with restrictions)

| **Feature** | **Material** | **Mental** | **Relational** |
|---|---|---|---|
| Aspect | + | - | - |
| *olla* + *-mAssA* *['be (in the process of) doing something/ be about to do something']* | + | +/ - | - |
| Projection | - | + | - |
| *Be* verb central | - | - | + |
| Two congruent participants | - | - | + |

Grammatically, the (transitive) material processes are distinct from other processes especially in terms of aspect, i.e. whether they allow the description of a finalized process or not; in other words, whether the aspect is limited (Shore 1996, 252; see also ISK 2004, 1437). This feature is strong with the material processes, and less strong with other processes: (aspect underlined in the examples): *Asi<u>aa</u> (~ asia) on valmisteltu perusteellisesti. [The issue has been prepared thoroughly.]* [Halonen 2002] Mental processes are limited in their aspect. (Shore 1996, 255).

A typical material process may be related to the '*be (in the process of) doing something/ be about to do something'* structure, where the temporary nature of the situation is in focus: the event is attached to a given moment or period. Often, there is also an implication that the author foresees the event as being about to take place. (Shore 1996, 253–254;

ISK 2004, 1446–1447.) – – *kuntien selkä on murtumassa taakan alle. [the backs of the municipalities are about be broken'* [Kekkonen 1959] The *Olla + -mAssA* structure *['be (in the process of) doing something/ be about to do something']* is not generally used in connection with verbs describing mental processes, since the mental processes have to do with space more than events. A special characteristic of the mental processes is that of projection (in the wide sense of the word). An issue is presented as being distant or detached from reality, i.e. as a description of the second degree. Thus, in this sense, projection is a strongly mental feature. (Shore 1996, 257; 2005.) *Pienen kansan jäseninä me sydämestämme* ***toivomme****, että tämä työ tulee menestymään.[As members of a small nation, we* ***hope****, from the bottoms of our hearts, that this project will succeed.]* [Kekkonen 1957]

It has been suggested that the process types are "crypto types" (Whorf), i.e. types of a covert system of transitivity. They generally lack overt markers on the textual surface. (See, e.g. Shore 1996, 239.) For example, the ends of the clauses do not include particles expressing transitivity which would instantly show which process type is in question. Nor are there such unambiguous differences in the morphology of the verbs that would reveal the type of the process. Here, in fact, lies one of the challenges of the Teko method: a large part of the quantitative data obtained by using the method is based on morphology in particular.

On the basis of what has been summarized above, we have tentatively encoded the processes in six New Year's speeches (three speeches by Kekkonen and three speeches by Halonen). So far, we have only made preliminary calculations. What we are after is testing a method rather than conducting actual research. The major problem here, again, is that there is no basic description of the Finnish language based on the Systemic Functional Grammar.

Table 20: Clause processes in the speeches by Kekkonen and Halonen

| **Process** | **Kekkonen** | **Halonen** |
|---|---|---|
| Material | 42% | 31% |
| Mental | 15% | 15% |
| Relational | 40% | 50% |

Judging by this preliminary analysis it would seem that there are differences between the different presidents as regards the processes they describe. We could generalize and say that Kekkonen preferred material processes, whereas Halonen seems to prefer relational processes.

We can also combine process information with topic classification. Table 21 could be used to analyse, for example, whether the choice of the process has to do more with the topic or the author of the speech.

Table 21: Clause processes by topic in the speeches by Kekkonen and Halonen

| Topic | Kekkonen | | Halonen | |
|---|---|---|---|---|
| Home country | Rel | 45% | Rel | 53% |
| | Mat | 42% | Mat | 33% |
| | Ment | 13% | Ment | 14% |
| World | Rel | 39% | Rel | 53% |
| | Mat | 39% | Mat | 33% |
| | Ment | 21% | Ment | 14% |
| General | Rel | 33% | Rel | 28% |
| | Mat | 42% | Mat | 38% |
| | Ment | 24% | Ment | 34% |

## 9 DISCUSSION

Above, we have been dealing with the methodical questions that we have been faced with whilst trying to combine qualitative and quantitative analysis in the research into the New Year's speeches. We have also presented the Teko project, together with the method we have developed for it, in a wider scope. The Teko project aims at compiling small well encoded textual corpora which are mutually comparable.

The corpora have been analysed morphologically. They can be supplemented with grammatical analysis of a higher level or other markings, and these markings can be used together with morphological analysis in the quantitative part of the research. We have also given some thought, both in this paper and our earlier studies, to how qualitative and quantitative research approaches could work dynamically together. This research method is typically based on a complete morphological analysis and a set of qualitative working hypotheses. The morphological analysis either confirms or challenges the hypotheses, on the basis of which we can focus on certain features; define textual passages that are to be analysed; and add and combine several levels of quantitative analysis. This process can be repeated several times. "The final word" in our research method always belongs to qualitative analysis, our fundamental purpose being the analysis of the meanings and the (grammatical) meaning potential of the texts. The meanings hidden behind the figures only start unfolding after a meticulous textual analysis.

In the light of this project, this kind of corpus linguistic text analysis seems to be a promising approach for getting in touch with interesting features in texts or text sets – or genres –, and interpreting them, even if it has its problems. Important questions for future research include the following: what kind of syntactic and semantic information can we (and should) encode in the corpus? How we can do it? How do we combine it with readily available morphological information? And how do we interpret the combinations?

We have started to supplement the New Year's speech corpus with data on the clause processes analysed according to the SF Grammar, but there are still many challenges we need to tackle in that venture. For example, we have not added analyses of the participants and circumstances of the processes (even if they have, of course, been dealt with in the process analysis), nor have we added any other type of dependency information. Further, we have yet to start adding interpersonal and textual analysis to the corpus. Yet, we have already carried out some experiments towards the analysis of the thematic structure by looking at the themes and rhemes in the textual clauses and, e.g., the distributions of the parts of speech in them, together with the word frequencies as far as the morphological analysis allows (Heikkinen & Lehtinen & Lounela 2004).

It could, of course, be asked why we did not start from the SF Theory's conception of language in the first place, building the corpora on the basis of a detailed encoding according to the SF Grammar. There are many reasons for this, one of the most prominent being that the linguistic basic theory in the SF tradition on the Finnish language has so far been rather restricted. In this sense, every research project based on the SF Grammar is, in fact, basic research. Further, we have had to construct our corpora and method gradually, with scarce resources. We have aimed at making use of the methodical resources that have been available and that we have been able to improve with moderate effort. On the one hand, we have had access to semi-automatic morphological analysis, and on the other hand, to the qualitative analysis of individual texts. We have also had access to interesting materials and research questions. Based on them, it has seemed natural to develop the Teko method in a way that has enabled us to gradually compile and partly even analyse mutually comparable corpora that are interesting from the perspectives of corpus linguistics, text analysis, general linguistics, culture, and society.

Although we recognize the benefits of the SF analysis in the study of the social meanings of language use, the use of basic morphological analysis which is carried out semi-automatically also has its advantages: the basic results of several studies are mutually comparable; combining different types of analyses is methodically interesting; and the process of analysis is relatively similar, irrespective of the conception of language held by the researchers. The background theory of the research does, of course, affect the analysis in the sense that the automatic program of analysis brings with it a certain theory of language. Thus, a core challenge in this development work is that the morphological analysis does not self-evidently support a grammatical analysis following the SF Grammar. Our examples above included expressions referring to the speaker. The automatic analyser does recognize reliably explicit I/me references, but the resulting automatic analysis does not shed light on what type of I/me is being referred to, what processes the I/me in question participates in, what kind of participator role it has, what thematic position it is presented in, etc. However, certain morphological phenomena of the word level in Finnish can be connected with interpretation complying with the SF Grammar rather directly. The morphological analyser can, e.g., recognize the forms of the fourth infinitive (*teke-minen* [do-ing]), which are often classified as grammatical metaphors in the tradition of the SF Grammar.

In our methodical considerations, we have been faced with a special need to reconsider some of the basic concepts of textual analysis. We have, for example, decided that it may be too daring to talk about genre analysis whilst combining qualitative and quantitative methods. Therefore, we have ended up by using rather more cautious concepts, such as text set or presumed genre. Of these two, *text set* aims at being a pretheoretical concept, as "neutral" as possible, providing a basis for the exploration of the concept of *genre*. The *presumed genre,* in turn, is a tool based on intuitive and collective knowledge of the world, which we use to refer to the presumptions we have about genres, their system, and the genre potential. This way, we can form a chain of concepts for research purposes: *text* > *presumed genre* > *text set* > *genre*. Our starting point is that of individual texts, which we use to make presumptions about the genre, using these presumptions, in turn, to choose the set of texts to be studied (which form a specialised corpus). In the qualitative-quantitative analysis of the text sets, we may even attain interpretations about the linguistic nature of the genre.

Many text sets materialize as genres when contemplated as specific linguistic action or when considering, e.g., the layout of the texts. Yet, we have to carry out qualitative-quantitative analysis to obtain information about how uniform the texts and genres are linguistically, and what type of linguistic heterogeneity the genre allows, as it were; and what that heterogeneity is based on. For example, based on this paper, the New Year's speech as a genre allows wide variation. The linguistic choices and meanings in the texts are affected, e.g., by who is speaking and what topics are spoken about.

## REFERENCES


Beaugrande, Robert de. 2001. Large corpora, small corpora, and the learning of "language". In Mohsen Ghadessy, Alex Henry & Robert L. Roseberry (eds.), *Small Corpus Studies and ELT*, 3–28. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Butt, David. 1996. Theories, Maps and Descriptions: An Introduction. In Ruqaiya Hasan, Carmel Cloran & David G. Butt (eds.), *Functional Descriptions. Theory in Practice*, xv–xxxv. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Eggins, Suzanne. 1994. *An Introduction to Systemic Functional Linguistics*. London: Pinter Publishers.

Hakulinen, Auli & Fred Karlsson & Maria Vilkuna. 1980. *Suomen tekstilauseiden piirteitä: kvantitatiivinen tutkimus [Features of Finnish textual clauses: a quantitative study]*. Helsinki: University of Helsinki.

Halliday, M. A. K. 1996. On Grammar and Grammatics. In Ruqaiya Hasan, Carmel Cloran & David G. Butt (eds.), *Functional Descriptions. Theory in Practice*, 1–38. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Halliday, M. A. K. 2006. Afterwords. In Geoff Thompson & Susan Hunston (eds.), *System and Corpus. Exploring Connections*, 293–299. London: Equinox Publishing Ltd.

Halliday, M. A. K. & Christian M. I. M. Matthiessen. 2004. *An Introduction to Functional Grammar*. Third edition. London: Arnold.

Heikkinen, Vesa. 1999. *Ideologinen merkitys kriittisen tekstintutkimuksen teoriassa ja käytännössä [Ideological meaning in the theory and practice of critical text analysis]*. Helsinki: SKS.

Heikkinen, Vesa. 2006. Uudenvuodenpuheiden piirteitä 1935-2006 ja näkymiä vallan medioitumiseen [Features of New Year's speeches 1935-2006 and perspectives on the mediatization of power]. In Tuija Nikko & Pekka Pälli (eds.), *Kieli ja teknologia [Language and Technology]*, 173–194. Helsinki: Helsinki School of Economics.

Heikkinen, Vesa & Outi Lehtinen & Mikko Lounela. 2004. Piirre tekstissä, teksti korpuksessa. Kohti dynaamisia analyyseja [Feature in text, text in corpus. Towards dynamic analyses]. In Marja Nenonen (ed.), *Papers from the 30th Finnish conference of linguistics*, 49–54. Joensuu: University of Joensuu.

Heikkinen, Vesa & Outi Lehtinen & Mikko Lounela. 2005. Lappeenrantalaismies löi toista nenään baarissa, Uutisia ja uutisia [A man from Lappeenranta punched another man in the nose in a bar. News and news]. In Vesa Heikkinen (ed.), *Tekstien arki. Tutkimusmatkoja jokapäiväisiin merkityksiimme [Everyday in texts. Explorations to our everyday meanings]*, 231–258. Helsinki: Gaudeamus.

Heikkinen, Vesa & Mikko Lounela. 2006. Sanaluokka automaattisen analyysin kategoriana [Part of speech as a category of automatic analysis]. In Krista Kerge & Maria-Maren Sepper (ed.), *Finest linguistics. Proceedings of the annual Finnish and Estonian conference of linguistics*, 42–58. Tallinn: Tallinn University Press.

Heikkinen, Vesa & Mikko Lounela. Forthcoming. Uudenvuodenpuheiden minä-viittaukset. Ihminen, instituutio, valintojensa valtias [Me-references in New Year's speeches. Man, institution, master of one's choices]. In Vesa Heikkinen (ed.), *Piirteen paino [The weight of a feature]*. Helsinki.

Hunston, Susan. 2002. *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

ISK 2004 = *Iso suomen kielioppi [Comprehensive Grammar of the Finnish Language]*. Auli Hakulinen et al. Helsinki: SKS.

KAINO 2008 = *Kaino – Kotuksen aineistopalvelu [Kaino. Material service of the Research Institute for the Languages of Finland]*. http://kaino/ (1 Apr 2008).

Kangasniemi, Heikki. 1992. *Modal expressions in Finnish*. Helsinki: SKS.

Kankaanpää, Salli. 2006. *Hallinnon lehdistötiedotteiden kieli [Language of Administrative Press Releases]*. Helsinki: SKS.

KOTUS 2006 = *Kotimaisten kielten tutkimuskeskus [Research Institute for the Languages of Finland]*. Theory and methods of genre analysis. http://www.kotus.fi/index.phtml?l=en&s=257 (1 May 2008).

KOTUS 2008 = *Kotimaisten kielten tutkimuskeskus [Research Institute for the Languages of Finland]*. Archives and collections held by the Research Institute. http://www.kotus.fi/index.phtml?l=en&s=54 (1 May 2008)

Lehtinen, Outi & Mikko Lounela. 2004. A model for composing and (re-)using text materials for linguistic research. In Marja Nenonen (ed.), *Papers from the 30th Finnish Conference of Linguistics*. Joensuu : University of Joensuu.73-78.

Lounela, Mikko. 2005. Exploring morphologically analysed text material. In Antti Arppe & al. (eds.), *Inquiries into words, constraints and contexts. Festschrift in the honour of Kimmo Koskenniemi on his 60th birthday*, 359-267. Helsinki: Gummerus.

Lounela, Mikko. 2007. Anatomy of an XML-based Text Corpus Server. In Joakim Nivre & Heiki-Jaan Kaalep & Kadri Muischnek & Mare Koit (eds.), *Proceedings of the 16th Nordic Conference of Computational Linguistics NODALIDA-2007*, 337–344. Tartu: University of Tartu.

Lounela, Mikko. Forthcoming. A Process Model for Composing High-quality Text Corpora. *6th International Conference on Language resources and Evaluation Proceedings*. European Language Resources Association, 2008.

Matthiessen, Christian & Christopher Nesbitt. 1996. On the Idea of Theory-Neutral Descriptions. In Ruqaiya Hasan, Carmel Cloran & David G. Butt (eds.), *Functional Descriptions. Theory in Practice*, 39–84. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Neale, Amy. 2006. 'Matching' corpus data and system networks: using corpora to modify and extend the system networks for TRANSITIVITY in English. In Geoff Thompson & Susan Hunston (eds.), System and Corpus. Exploring Connections, 143–163. London: Equinox Publishing Ltd.

Saukkonen, Pauli. 1984. Mistä tyyli syntyy? [How is a style born?] Helsinki: WSOY.

Saukkonen, Pauli. 2001. *Maailman hahmottaminen teksteinä. Tekstirakenteen ja tekstilajien teoriaa ja analyysia [Perceiving the world as texts. Theory and analysis of textual structure and genres]*. Helsinki: Helsinki University Press.

Shore, Susanna. 1992. Aspects of a Systemic Functional Grammar of Finnish. Unpublished Ph.D. Thesis. Sydney: Macquarie University.

Shore, Susanna. 1996. Process Types in Finnish: Implicate Order, Covert Categories, and Prototypes. In Ruqaiya Hasan, Carmel Cloran & David G. Butt (eds.), *Functional Descriptions. Theory in Practice*, 237–264. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Shore, Susanna. 2005. Referoinnista projektioon ja metarepresentaatioon [From referring to projection and metapresentation]. In Markku Haakana & Jyrki Kalliokoski (eds.), Referointi ja moniäänisyys [Referring and multiple voices], 44–82. Helsinki: SKS.

Sinclair, John. 2001. Preface. In Mohsen Ghadessy & Alex Henry & Robert L. Roseberry (eds.), *Small Corpus Studies and ELT*, xii–xv. Amsterdam/Philadelphia: John Benjamins Publishing Company.

TEI 2004 = The XML version of the TEI guidelines. The TEI Consortium 2004. http://www.tei-c.org/release/doc/tei-p4-doc/html (1 May 2008).

Thompson, Geoff & Susan Hunston. 2006. Introduction. System and Corpus: two traditions with a common ground. In Geoff Thompson & Susan Hunston (eds.), *System and Corpus. Exploring Connections*, 1–14. London: Equinox Publishing Ltd.

Tiililä, Ulla. 2007. *Tekstit viraston työssä [Textts and text production in a city department]*. Helsinki: SKS.

TWOL = *Fintwol. Finnish morphological analyser.* http://www2.lingsoft.fi/doc/fintwol (1 May 2008).

APPENDIX

Listed below are the morphological markers in use as presented at the web page of the manufacturer of the analyser (Lingsoft 2008), slightly modified for presentation purposes.

PART OF SPEECH

| | |
|---|---|
| **A** | adjective (*pieni, small*) |
| **ABBR** | abbreviation (*esim, e.g.*) |
| **AD-A** | ad-adjective (*melkein, almost*) |
| **ADV** | adverb (*hitaasti, slowly*) |
| **ART** | foreign article (*das*) |
| **C** | conjunction (*ja, and*) |
| **INTJ** | interjection (*hui, wow*) |
| **N** | noun (*koira, dog*) |
| **NUM** | numeral (*kaksi, two*) |
| **PP** | post- or preposition (*jälkeen*, *after*) |
| **PREP** | foreign preposition (*de*) |
| **PRON** | pronoun (*sinä, you*) |
| **PSP** | postposition (*vieressä*) |
| **Q** | quantifier (*moni, many*) |
| **V** | verb (*tulla, to come*) |

COMPARATION

| | |
|---|---|
| **POS** | positive (*kuuma*, *hot*) |
| **CMP** | comparative (*kuumempi*, *hotter*) |
| **SUP** | superlative (*paras*, best) |

CASE

| | |
|---|---|
| **NOM** | nominative (*koira, dog*) |
| **GEN** | genitive (*koiran*) |
| **PTV** | partitive (*koiraa*) |
| **ESS** | essive (*koirana*) |
| **TRA** | translative (*koiraksi*) |
| **INE** | inessive (*koirassa*) |
| **ELA** | elative (*koirasta*) |
| **ILL** | illative (*koiraan*) |
| **ADE** | adessive (*koiralla*) |
| **ABL** | ablative (*koiralta*) |
| **ALL** | allative (*koirallе*) |
| **ABE** | abessive (*koiratta*) |
| **CMT** | comitative (*koirineen*) |
| **INS** | instructive (*koirin*) |

NUMBER

| | |
|---|---|
| **SG** | singular (*pöytä, table*) |
| **PL** | plural (*pöydät, tables*) |

POSSESSIVE SUFFIXES

| | |
|---|---|
| **1SG** | 1st person singular (*tyttäreni, my daughter*) |
| **2SG** | 2nd person singular (*tyttäresi, your daughter*) |
| **3** | 3rd person singular or plural (*tyttärensä, her/his daughter*) |
| **1PL** | 1st person plural (*tyttäremme, our daughter*) |
| **2PL** | 2st person plural (*tyttärenne, your daughter*) |

MOOD

| | |
|---|---|
| **IMPV** | imperative (*mene!, go!*) |
| **COND** | conditional (*lukisi*, would read) |
| **POTN** | potential (*lukenee*, may read) |

There is no feature for indicative forms (*lukee*, *menee*).

TENSE

| | |
|---|---|
| **PRES** | present tense (*haluan, I want*) |
| **PAST** | past tense (*halusin, I wanted*) |

Perfect and pluperfect tenses are interpreted as participle forms.

VOICE

| | |
|---|---|
| **ACT** | active (*uin, I swim*) |
| **PSS** | passive (*uidaan, people swim*) |

PERSON

| | |
|---|---|
| **SG1** | 1st person singular (*menen, I go*) |
| **SG2** | 2nd person singular (*menet, you go*) |
| **SG3** | 3rd person singular (*menee, (s)he goes*) |
| **PL1** | 1st person plural (*menemme, we go*) |
| **PL2** | 2nd person plural (*menette, you go*) |
| **PL3** | 3rd person plural (*menevät, they go*) |
| **PE4** | passive ending (*mennään, people go*) |

NEGATIVE

| | |
|---|---|
| **NEGV** | negative verb (*en, not*) |
| **NEG** | negative form (en *tehnyt, I did not*) |

INFINITIVES

| | |
|---|---|
| **INF1** | 1st infinitive (*tulla*, *to come*) |
| **INF2** | 2nd infinitive (*tullessaan*, *while coming*) |
| **INF3** | 3rd infinitive (*tulemaan, to come*) |
| **INF5** | 5th infinitive (*tulemaisillaan, about to come*) |

The 4th infinitive (*tuleminen*) is interpreted as a noun.

PARTICIPLES

| | |
|---|---|
| **PCP1** | 1st participle (*lentävä*, *flying*) |
| **PCP2** | 2nd participle (*lentänyt*, *flown*) |

CLITICS

| | |
|---|---|
| **hAn** | *han/hän* (*poikahan*) |
| **kA** | *ka/kä* (*eikä*) |
| **kAAn** | *kaan/kään* (*poikakaan*) |
| **kin** | *kin* (*poikakin*) |
| **kO** | *ko/kö* (*oletko*) |
| **pA** | *pa/pä* (*oletpa*) |
| **S** | *s* (*onpas*) |

OTHER

| | |
|---|---|
| **FORGN** | foreign word (*British*) |
| **PROP** | proper noun (*Mikko*) |
| **pi** | *-pi* (*ompi*) |

ADDITIONAL MARKERS (NOT IN THE LINGSOFT LIST)

| | |
|---|---|
| **COORD** | coordinating conjunction |
| **COP** | copula |
| **DA-UUS** | deadjectival, -UUs-affix (*rikollisuus*, *criminality*) |
| **DEM** | demonstrative pronoun |
| **DV-ILLINEN** | deverbal -illinen-suffix (*ruumiillinen*, *bodily*) |
| **DN-INEN** | denominal -inen-suffix (*osainen*) |
| **DN-ITTAIN** | denominal -ittain-suffix (*osittain*, *partly*) |

| | |
|---|---|
| **DV-MA** | deverbal -ma-suffix (*luoma*) |
| **DV-MATON** | deverbal -maton-suffix (*murtumaton*, *unbreakable*) |
| **DV-NTAA** | deverbal -ntaa-suffix (*vähentää*, *minimize*) |
| **DV-TTA** | deverbal -tta-suffix (*huolestuttaa*, *to make worried*) |
| **DV-U** | deverbal -u-suffix (*rajoitu*) |
| **INTG** | interrogative pronoun (*mitä*, *what*) |
| **INTERR** | interrogative |
| **MAN** | adverb class manner |
| **PERS** | personal pronoun |
| **REF** | referative non-finite clause |
| **REL** | relative pronoun |
| **SUB** | subordinating conjunction |
| **TEMP** | temporal non-finite clause |

In addition to the Lingsoft markers, the lists include some markers that express properties of multi-word expressions, such as perfect (P) and pluperfect (PL), or lexicalized participles that function as nouns (function="N").

Vesa Heikkinen & Mikko Lounela
Research Institute for the Languages of Finland
Sörnäisten rantatie 25
FI-00500 HELSINKI
vesa.heikkinen@kotus.fi
mikko.lounela@kotus.fi