25

Exploring Morphologically Analysed Text Material

MIKKO LOUNELA

In text linguistics, it is possible to carry out research by carefully analysing a small set of texts, sometimes just a few. For some examples of this kind of text analysis, see Heikkinen (2005). Such a methodological choice is not easily combined with the idea of using corpus-based methods and quantitative analysis as an essential part of research. In text linguistics, however, the text type constitutes an important research problem, and some work has been done in classifying texts according to their quantitative morphological and syntactic characteristics. Such work has been going on for a few decades now, see e.g. Biber (1988). For a related approach to Finnish texts, see Saukkonen (2001).

In the Research Institute for the Languages of Finland (RILF), we are aiming at a fruitful combination of quantitative morpho-syntactic analysis and deep text analysis based on Lexical Functional Grammar, i.e. LFG (2004). This work includes providing a morpho-syntactic analysis (in due form) of a selected group of texts, and calculating a "morphological fingerprint" of the text group. One group of texts forms a text material, usually of moderate size (consisting of fewer than 100 texts, with approximately 100,000 words). This article focuses on the problems and choices in adding the morpho-syntactic annotation to the text material, and in defining intuitive linguistic categories such as part-of-speech, verb, finite verb, and tense using semi-automatic word-level morpho-syntactic analysis.

The language of our texts is Finnish. The design of our materials is based on the XML (1996-2004) language, using modified TEI (2001-2003) P4 structure. The morphological annotation is based on the analysis provided

Inquiries into Words, Constraints and Contexts.
Antti Arppe et al. (Eds.)

by FINTWOL (1995-2000), a morphological analyser developed at Lingsoft (version 1998/03/02), based on the Two-level model introduced by Koskenniemi (1983). An overview of FINTWOL's tag set is presented in Fintwol/tags (2001). The morphological analysis goes through a careful hand-made disambiguation and augmentation. Our model for text materials is described in Lehtinen and Lounela (2004). The exploration of the text material is carried out using the Xquery (2000-2005) language.

25.1 Morphological Analysis and Text Structure

The FINTWOL morphological analyser provides each word of the material with morphological information. This information includes the base form (lemma) of the word, and an unordered set of tags, expressing morphological features of the word. If the word can represent more than one word-form, FINTWOL will list all its possible readings. In the case of compound words, the word-internal boundaries are marked in the lemma. At RILF, we use our own pre-processor to enhance FINTWOL's capabilities in processing Finnish abbreviations and numerical expressions.

The following illustrates FINTWOL analysis and the ambiguity it may produce. The Finnish word-form *alustamassa* may be interpreted either as a compound noun it alustamassa ("platform mass"), or a third infinitive or a deverbalised derivation of the verb *alustaa*, ("knead" or "format"):

```
"<alustamassa>"
      "alusta#massa"  N NOM SG
      "alustaa"  V INF3 INE
      "alustaa"  DV-MA INE SG
```

In order that the FINTWOL analysis would be usable in the TEI-format, the information it gives has to be split and embedded in the XML-element. The element for a text word in TEI P4 definition is "w", and it has such attributes as "lemma" for the base-form and "type" for the part-of-speech information. Following the Corpus Encoding Standard, i.e. CES (2000), we have added an attribute, "msd", to include the morpho-syntactic description in the word element. The following (simplified) example shows how the Fintwol analysis is embedded in XML-encoding:

```
<w lemma="alusta#massa" type="N" msd="NOM SG">alustamassa</w>
<w lemma="alustaa" type="V" msd="INF3 INE">alustamassa</w>
<w lemma="alustaa" type="DV-MA" msd="INE SG">alustamassa</w>
```

The XML-type word elements are then disambiguated by hand (only the most likely analysis is retained), and some information concerning multi-word features (e.g., perfect tense) is added. The words are then included in the general text structure, e.g., in text chapters, headers, legends, etc. This type of text material can be analysed quantitavely according to its morpho-syntactic

features.

25.2 Quantitative Lexical Analysis

The morpho-syntactic fingerprint of a group of texts consists mainly of figures and frequency lists of the morpho-syntactic features of the words in the material. At present, the fingerprint that we have designed in RILF consists of four different parts: (1) the general part, (2) the verbal part, (3) the nominal part, and (4) the lexical part.

The general part includes information such as the average lengths of texts, sentences and clauses, and the frequencies of punctuation marks, lemmas, and most common word-forms and parts-of-speech in the material. The nominal part concerns words of the types "N" (noun), "A" (adjective), "PRON" (pronoun) and "NUM" (numeral). It includes the frequencies of cases, comparatives, numera, word-forms and lemmas as well as the frequency lists of the most common word-forms and parts-of-speech of the nominals in the material. The verbal part of the fingerprint includes the frequencies of features such as voice, mood and tense as well as frequencies of infinitive forms, participles and the most common verbal lemmas and word forms. The lexical part of the fingerprint consists of frequency lists of the most common lemmas and word-forms of each of the parts-of-speech (values of the "type"-attribute) in the material.

Some of these features can be obtained directly from the FINTWOL analysis, while some of them require combining the FINTWOL tags and interpreting the combinations. In the remainder of this article I will consider defining more or less problematic features such as part-of-speech, verb, finite verb and tense. A sample verbal fingerprint analysis along with the xquery code used to produce it can be seen on the web site of RILF, at <http://www.kotus.fi/julkaisut/2005-ml-1/>.

25.3 Part-of-speech

The transformation from FINTWOL to XML includes recognising the part-of-speech tag in the FINTWOL analysis. In the previous example, the most obvious candidate is the first tag of the analysis, but this is not always the case. In the following FINTWOL analysis of the word-form *kaavoittaja* ("planner"), the obvious part-of-speech tag "N" (noun) is preceded by the tag providing information about the derivation of the word-form ("DV-JA"):

```
"<kaavoittaja>"
      "kaavoittaja"  DV-JA N NOM SG
```

There are also FINTWOL analyses where no obvious part-of-speech tag is present, and those in which we have to choose between more than one good candidate. In the last line of the first example (*alustamassa*), the best

candidate for the part-of-speech is in this particular case the derivation tag "DV-MA", as other possibilities are in practice less appropriate. More about the part-of-speech problematics concerning current morphological analysers for Finnish can be found in Heikkinen and Lounela (forthcoming).

At RILF, we have developed a simple algorithm for automatically finding the best part-of-speech candidate in the FINTWOL analysis. The algorithm divides the FINTWOL tags into four classes, of which we choose the most likely part-of-speech in the following manner.

1. If the analysis contains one or more of the tags "A", "ABBR", "AD-A", "ADV", "C", "INTJ", "N", "NUM", "PP", "PREP", "PRON", "PSP", or "V", choose the one that appears last in the tag sequence.
2. If the analysis does not contain any of the tags mentioned above, choose the last of "Q", "PCP1", "PCP2", or "A/N".
3. If the analysis does not contain any of the tags mentioned above, choose the last tag indicating the derivative properties of the words (any tag beginning "D?-", where "?" denotes any character).
4. If none of the above applies, choose the first tag in the analysis.

This algorithm gives us the following list of part-of-speech tags, when applied to a 20,000-word sample of material from the Finnish newspaper Aamulehti after analysis by FINTWOL and without any subsequent disambiguation.

TAG	PART-OF-SPEECH
UNKNOWN	Unrecognised word-form
A	Adjective
A/N	Adjective or noun
ABBR	Abbreviation
AD-A	Ad-adjective
ADV	Adverb
C	Conjunction
DV-MA	Deverbal derivation with ending "ma"
FORGN	Foreign word
INTJ	Interjection
N	Noun
NUM	Numeral
PCP1	First participle
PCP2	Second participle
PP	Post- or preposition
PRON	Pronoun
PSP	Postposition
V	Verb

25.4 Verb

As the part-of-speech has its FINTWOL-based operational definition, we can start to define other linguistic features. Here, I will focus on the verbs and the morpho-syntactic properties that are closely related to them.

25.4.1 Verb

The definition of the verb itself might seem quite unproblematic, since it is a primary part-of-speech category, as seen earlier. However, when we take a look at analyses of a few text samples, the picture changes. When analysing verb chains, such as those containing negation *ei juossut*, "[he/she/it] did not run", or the perfect tense, *on juossut*, "[he/she/it] has run", we notice that the number of the finite verb forms in the negative construction is two, while the perfect construction has only one finite verb form, as the *juossut* is defined as a participle form in the construction. The infinitive form *saamme juosta* "[we] may run" consists, again, of two verbs. The analyses for the verb form *juossut* in the example have been selected from the three alternatives given by the FINTWOL analyser, the third interpretation being an adjective. The analysis for the form *juosta* is selected from two analyses, the other possibility being present negative passive. All the following examples will be manually disambiguated:

```
<w lemma="ei" type="V" msd="NEGV SG3">ei</w>
<w lemma="juosta" type="V" msd="PAST ACT NEG SG">juossut</w>

<w lemma="olla" type="V" msd="COP PRES ACT SG3">on</w>
<w lemma="juosta" type="PCP2" msd="ACT POS NOM SG">juossut</w>

<w lemma="saada" type="V" msd="PRES ACT PL1">saamme</w>
<w lemma="juosta" type="V" msd="INF1 NOM">juosta</w>
```

When counting verbs, we divide the verb category into two: (1) semantic verbs and (2) grammatical verbs. The semantic verbs include the participle forms of the temporal verb chains forming perfects and pluperfects, as well as all the words of the type "V", except for the auxiliaries in the negative and temporal verb constructions.

The grammatical verbs include the same set of words as the semantic verbs, with some exceptions. The infinitive verb forms (marked with "INF1", "INF2", "INF3", etc.) are excluded, and the auxiliary is selected from the temporal chains. In the temporal chains, the participle has a part-of-speech marker "PCP2", so this can be achieved by just counting the "V"-tags, and excluding the infinitives.

To make the operational definitions for all the possible tenses, we have to mark the perfect and the pluperfect tenses in the material. An extra attribute ("function") is added to the data model for this purpose.

```
<w ... type="V" msd="COP PRES ACT SG3" function="P">on</w>
<w ... type="PCP2" msd="ACT POS NOM SG" function="P">juossut</w>
```

For the sake of consistency, it would probably be necessary to function-mark modal verb chains, such as *saamme juosta* (see above), but this is not done at present.

25.4.2 Finite Verb and Tense

According to the latest authoritative and quite comprehensive grammar of Finnish, *Iso Suomen Kielioppi*, by Hakulinen et al. (2005), a finite verb in Finnish is a verb that is inflected in tense, mood and person. A finite verb functions as the nucleus of a clause. Identifying the finite verbs is essential for obtaining figures related to clauses, which we consider very important. Mapping this definition of finiteness to FINTWOL analysis is, however, problematic, for at least two reasons.

Firstly, in the FINTWOL analysis the indicative mood is provided as the default value for all verb forms. In order to follow the definition given in Hakulinen et al. (2005), we should know which verbs inflect in mood, in order to to be able to identify the finite verbs. This information is, however, not available.

Second, while the FINTWOL analysis does not include a tag for person inflection in the negative verb forms, it includes one in the negative auxiliaries in negative verb chains, eg. *emme juokse* ("[we] do not run"), below. This means that in negative forms finiteness is divided between the semantic verb and the negative auxiliary. Identifying it would require information about word dependencies, but that kind of information is not available in the FINTWOL analysis:

```
<w lemma="ei" type="V" msd="NEGV PL1">emme</w>
<w lemma="juosta" type="V" msd="PRES ACT NEG">juokse</w>
```

At present, the fingerprint analysis defines finite verbs as a set of semantic verbs, where the active or passive marker is present, with the infinitive forms excluded.

We use the number of finite verbs as an indicator of the number of clauses in the text materials. Concerning the problems and uses of this type of work, see Heikkinen et al. (2000). The following sentence (*älä juokse ja huuda*, "do not run and shout") is interpreted as having two finite verbs, and thus clauses, even though it has only one word with inflection markings for person, none for tense, and three for mood:

```
<w lemma="ei" type="V" msd="NEGV IMPV ACT SG2">Älä</w>
<w lemma="juosta" type="V" msd="IMPV ACT NEG SG">juokse</w>
<w lemma="ja" type="C" msd="COORD">ja</w>
<w lemma="huutaa" type="V" msd="IMPV ACT NEG SG">huuda</w>
<w lemma="." type="PUNCT" msd="FULLSTOP">.</w>
```

Having all the above definitions, defining the tenses of the verbs is quite straightforward. We use the finite verbs as the base set of words expressing temporal information of the texts. As the perfect and pluperfect are explicity marked, we can identify the tenses directly, using the FINTWOL tags "PAST" and "PRES" combined with the information provided by the function attribute.

25.5 Conclusion

In this article, I have presented proposals for operational definitions for some linguistic categories for Finnish. The proposals are based on hand-augmented morphological analysis of Finnish texts, the analysis being provided by the FINTWOL morphological analyser. The defined categories include part-of-speech, verb, finite verb, and tense.

1. Part-of-speech marker can be selected from the FINTWOL analysis as being
   (a) the last tag indicating primary word category (adjective, abbreviation, ad-adjective, adverb, conjunction, interjection, noun, numeral, post/preposition, pronoun, postposition, or verb), or
   (b) the last tag indicating secondary word category (quantifier, first or second participle, or adjective/noun), if the above does not apply, or
   (c) the last tag indicating derivative information, if none of the above applies, or
   (d) the first tag of the analysis, if none of the above applies.
2. A semantic verb is a word of part-of-speech "V", with the temporal and negative auxiliaries excluded, and with the participle forms ("PCP2") of the perfective and pluperfective verb chains included.
3. A grammatical verb is a word of part-of-speech "V", with the infinitive verb forms excluded.
4. A finite verb is a semantic verb with voice, active ("ACT"), or passive ("PSS"), with the infinitive forms excluded.
5. The tenses are counted on the basis of finite verbs. The markers "PRES" and "PAST" indicate present and past tense, and a special attribute ("function", with values "P" for perfect and "PL" for pluperfect) is added to words in the temporal verb chains to indicate corresponding tenses.

The overall process of giving functional morphological definitions to the general grammatical categories raises some issues concerning the design principles of morphological analysers for the Finnish language. Taking these things into account would greatly enhance the usability of such analysers.

First, the major linguistic categories, such as part-of-speech, should be consistently included in the analysis of each word. Second, no categories should be left as the default, such as the indicative mood is in the present FINTWOL analysis. The fundamental set may not be clear, which makes deducing the set of the words with the default value hard, or even impossible. Third, the ordering of the tags does matter. Tabular or XML-style representation of the analysis would help the user to identify the features behind the markers, and to see what information may be missing. Finally, a complete documentation of all the categories and markers used by the analyser is essential for its scientific use.

Acknowledgments

The views expressed in this article are based on practical work on annotating, analyzing and exploring text materials at the Research Institute for the Languages of Finland in 2000-2005. Part of the work was funded by the Academy of Finland in 2003-2004. I am grateful to Vesa Heikkinen and Outi Lehtinen for fruitful cooperation. I would also like to thank Vesa Heikkinen for his comments regarding this article.

References

Biber, Douglas. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.

CES. 2000. Corpus encoding standard. <http://www.cs.vassar.edu/CES/>. Vassar College. Visited March 2005.

FINTWOL. 1995-2000. Fintwol. <http://www.lingsoft.fi/cgi-bin/fintwol/>. Lingsoft, inc. Visited March 2005.

Fintwol/tags. 2001. Tags (partial list). <http://www.lingsoft.fi/doc/fintwol/intro/tags.html>. Lingsoft, inc. Visited March 2005.

Hakulinen, Auli, Maria Vilkuna, Riitta Korhonen, Vesa Koivisto, Tarja Heinonen, and Irja Alho. 2005. *Iso suomen kielioppi*. Helsinki: SKS.

Heikkinen, Vesa, ed. 2005. *Tekstien arki*. Helsinki: Gaudeamus.

Heikkinen, Vesa, Outi Lehtinen, and Mikko Lounela. 2000. Ihminen ja kone tekstiä mankeloimassa, kuusikohtauksinen keskustelu. In H. Sulkala and L. Nissilä, eds., *XXVII Kielitieteen päivät Oulussa 19. - 20.5.2000*. Oulu: University of Oulu.

Heikkinen, Vesa and Mikko Lounela. forthcoming. Sanaluokka morfologisen analyysin kategoriana. In K. Kerge and M.-M. Sepper, eds., *Finest Linguistics. Proceedings of the Annual Finnish and Estonian Conference of Linguistics. Tallinn, May 6-7, 2004*. Tallinn: TPÜ.

Koskenniemi, Kimmo. 1983. *Two-Level Morphology: A General Computational Model for Word-Form Recognition and Production*. Helsinki: University of Helsinki, Department of General Linguistics.

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

LFG. 2004. Lexical Functional Grammar (Stanford Web Site). <http://www-lfg.stanford.edu/lfg/>. University of Stanford. Visited March 2005.

Saukkonen, Pauli. 2001. *Maailman hahmottaminen teksteinä*. Helsinki: Helsinki University Press.

TEI. 2001-2003. Text encoding initiative. <http://www.tei-c.org/>. TEI Consortium. Visited March 2005.

XML. 1996-2004. Extensible markup language (xml). <http://www.w3.org/XML/>. World Wide Web Consortium. Visited March 2005.

Xquery. 2000-2005. Xml query (xquery). <http://www.w3.org/XML/Query/>. World Wide Web Consortium. Visited March 2005.