Human Language Technologies – The Baltic Perspective
A. Utka et al. (Eds.)
© 2020 The authors and IOS Press.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA200617

Adding Compound Splitting and Analysis to a Semantic Tagger of Modern Standard Finnish – On the Way to FiSTComp

KIMMO KETTUNEN[1]
University of Eastern Finland, Joensuu, Finland

Abstract. This study continues a work in progress for implementing a full-text lexical semantic tagger for Finnish, FiST. The tagger is based on a 46,226 lexeme semantic lexicon of Finnish that was published in 2016 [1]. Kettunen [2], [3] describes the basic working version of FiST. FiST is based on freely available components: the first implementation uses Omorfi and FinnPos for morphological analysis and disambiguation of Finnish words. The current paper describes work with compound splitting for semantic tagging and its effects on the lexical coverage of the tagger. We try out two different approaches to morphological analysis and disambiguation of words for an improved version of FiST, FiSTComp: FinnPos [4], and Turku Dependency Parser [5], [6], UD1. Both these tools disambiguate morphological interpretations of words and provide boundary markings for compounds, but details and granularity of constituent decomposition vary. Our results with two-, three and four-part compounds show that analysis of compounds through their constituents with UD1 may improve the lexical coverage of the tagger with about 6.6 % units at best. Although we are able to proceed in basic problems of compound splitting, the results are still initial and further work is needed as compounds are a complex phenomenon.

Keywords. Semantic tagging, compounds, Finnish

1. Introduction

Kettunen [2], [3] has introduced the first version of a lexical semantic tagger of modern standard Finnish called FiST. Details of the tagger's implementation and first evaluation results are described in [2], and [3] continues with more evaluation. [7] have used the tagger for analysis of Finnish parliamentary speeches related to rights of everyman in three different decades. [1] describes the Finnish semantic lexicon and principles of its compilation in detail[2]. [1] also evaluates a now obsolete Finnish semantic tagger of Kielikone Ltd. that was the first semantic tagger for Finnish.

So far, the lexical coverage of FiST has been evaluated with about 30 different texts of various genres and sizes. Most of the texts are modern Finnish, but also texts older than 100 years have been analyzed successfully, e.g. the prose of several late 19th and early 20th century Finnish authors. The largest analyzed texts so far have been the Finnish Europarl documents v.6 with 28.6 million words and part of the Open Subtitle

[1] Corresponding Author: Kimmo Kettunen; University of Eastern Finland, Joensuu, Finland; E-mail: Kimmo.kettunen@uef.fi.

[2] The Finnish semantic lexicon is available at https://github.com/UCREL/Multilingual-USAS

collection with 45.2 million words. Both of these analyses achieve lexical coverage of 90.9 % [2], [3].

The first version of FiST analyzed only compounds that were included in its lexicon. As Finnish language uses compounding amply and formation of compounds is quite free, any Finnish lexicon is lacking a great part of compounds found in texts. In this study, we improve the compound handling of FiST by also using the constituents of compounds in semantic analysis. It is obvious that analysis of constituents of compounds should improve the coverage of compounds and the lexical coverage of the tagger, but it is not self-evident what the best practice for performing the analysis is and how much improvement can be achieved.

Our research topic in this paper is twofold: first, we need to find out what type of compound splitting with the available morphological analyzers is most beneficial for semantic tagging of texts. Second, we want to examine how much compound splitting improves the lexical coverage of our Finnish semantic tagger. The solutions we will offer for compound analysis are preliminary, but they are a step forward to a better lexical coverage of the tagger and more comprehensive than the basic compound engine introduced in [1]. [1] introduces a simple compound engine where the last constituent of the compound is separated from the beginning of the word and the two parts are given semantic tags, if possible.

Our test and development data consists of a corpus of speeches at the Finnish Parliament during 1991–2015. The data is part of the The ParlSpeech data set [8]. The data we use is a part of the whole Finnish corpus of 245,852 speeches, and it contains speeches where innovation has been mentioned in the speech. The size of the test and development data is 4,220 speeches and about 2.17 million word tokens. Parliamentary speeches contain probably more compounds than e.g. newspaper texts and thus these texts suit well for our analyses.

2. Compounds in Finnish

2.1. Basics

Creation of compounds, words which are formed by concatenating two or more words without a space between them[3], is a very productive means of making up new words in Finnish and many other languages [9], [10]. Finnish compounds are most often formed from nouns, but other parts of speech can also appear in compounds [10], [13]. Most common are Noun+Noun and Adjective+Noun compounds. According to [14], about 89 % of compounds in *Nykysuomen sanakirja* (Dictionary of Modern Finnish), are nouns. Typical examples of Finnish compounds are e.g. *puutalo* (puu+talo, 'wooden house', literally wood+house), and *ihmisoikeus* (ihmis+oikeus, 'human right'). By adding more words to the beginning or end of two-part compounds, new more complex compounds can be formed: *puutalorakentaminen* (puu+talo+rakentaminen, 'building of wooden houses'), *ihmisoikeusloukkaus* (ihmis+oikeus+loukkaus, 'violation of human right(s)').

There is no clear upper limit to recursive concatenation of constituents in compound creation, but compounds with five constituents are already on the upper limit of

[3] A hyphen is used to separate compound constituents in several cases for clarity This happens, for example, when the constituents in a compound have adjacent same vowels, i.e. a hiatus between, e.g. *kilpa-auto ('racing car')* [10:401]. Also abbreviations, numbers and special signs are written with hyphen in a compound.

concatenation in frequency [10: 405]. [15] has analyzed compounds that have four or more constituents and in her newspaper data of ca. 13,000 tokens about 84 % of the long compounds consist of four constituents and ca. 12 % have five constituents. In Tyysteri's [12] data of new compounds from years 2000–2009 (over 28,000 tokens), two constituent compounds are the norm: 83.6 %; three constituent compounds form 15.5 % of the data and four constituent compounds only 0.9 %. Longer compounds are almost negligent in the data [12].

2.2. Types of Compounds

The largest modern Finnish grammar, *Iso suomen kielioppi* [10], describes compound forming of Finnish in detail. Here, we concentrate only on the basics of compounds for our purpose and do not try to cover all the varieties, as part of the compound classes are rare[4]. In addition to [10], [11] and [12] have been useful sources in details of Finnish compound forming.

The most common type of compound is a determinative compound (aka. a subordinate compound). In a determinative compound, the last constituent of the compound specifies the basic meaning of the word and is the head of the whole construction, whereas the first constituent modifies the whole. The meaning of the whole is more or less the sum of the constituents, i.e. the meaning is transparent and compositional. *Puutalo* is a type of a house, where *puu* ('wood') modifies the basic meaning of *talo* ('house'). In determinative compounds, constituents have thus a semantically non-symmetrical relationship with each other. The compound denotes a subordinate concept to the head of the compound [11].

In some determinative compounds, the meaning of the compound cannot easily be deduced from the sum of the meanings of the compound constituents. Examples of such compounds are e.g. *tietokone* (tieto+kone, 'computer', literally 'knowledge machine') and *potkuhousut* (potku+housut, 'playsuit' (for a baby), literally 'kick trousers') [1]. Such items are many times referred to as "lexicalized compounds", and their meaning is non-transparent.

Copulative or co-ordinate compounds are the second main compound type [10], [11]. They consist of two or more compound constituents, which are in a symmetrical relationship with each other. Constituents of copulative compounds represent the same part of speech and their relationship is semantically additive. A hyphen is often used to separate the constituents. Examples of copulative compounds are e.g., *kanttori-urkuri* ('cantor and organist') and *parturi-kampaamo* ('barber and hairdresser').

Out of these two basic compound types, determinative compounds are far more common than copulative compounds. On the basis of the 94,110 word basic lexicon of Finnish, the Kotus lexicon[5], we would estimate that whereas determinative compounds are counted in tens of thousands in a basic lexicon, copulative compounds are counted in about 30–50 in the same lexicon. Lantee's [15] compound analysis data consists of ca. 13,000 compounds. About 85 % of tokens in the data are determinative compounds**.** The rest 15 % are either copulative compounds or determinative compounds that have as a determinative part either a phrase or a copulative compound.

[4] Usually three types are distinguished: determinative, copulative and appositive [10-11, 13].
[5] http://kaino.kotus.fi/sanat/nykysuomi/

2.3. Share of Compounds

The total share of compounds has been counted for the largest Finnish dictionaries. [14] states that the still largest but nowadays slightly outdated Finnish dictionary, *Nykysuomen sanakirja,* has about 65 % of compounds out of its ca. 201,000 lexemes. The more modern dictionary, *Perussanakirja*, has 94,100 lexemes out of which 52,269 (55.5 %) are compounds according to [16].

The number of compounds in dictionaries is one aspect of productivity of compounds in Finnish, another is their frequency in texts and speech. We estimated this with analyses of large corpora with an automatic morphological analyzer. The largest data we had available were Europarl's Finnish data v.7[6] with ca. 31.95 million words, Open subtitle corpus [7] with ca. 144.48 million Finnish words, and The Finnish Parliamentary data with ca. 57.32 million words [8]. Out of these, Open subtitles represents spoken language data, although it is slightly artificial.

We ran the texts through morphological analyzer [17]. Europarl v7 had 4,125,947 (12.9 %) unique compounds in Omorfi's [17] analysis and Open subtitles 5,845,351 (4.1 %). The Finnish Parliamentary data had 7,692,148 (13.4 %) unique compounds. In the analysis of [15], the ca. 31,270,992 million token Helsingin Sanomat 2000–2001 newspaper corpus contained about 2.5 million compounds, which is 8 % out of the total words. These figures are similar to the older data of [18]: they had a 3.8 % share of compounds in speech and 14.6 % in texts.

2.4. Structure of Compounds

Compounding is based on the concatenation of two or more words together. A two-part simple determinative compound consists of two simple words, and its structure is straightforward: *kivi+talo*. However, more complex compounds can consist of other compounds or word combinations. These complex compounds have a layered structure where relations of the constituents are hierarchic. According to [10], multiple constituents are more common for the first constituents of a compound. As examples, [10] list the multipart compounds shown in Table 1.

Table 1. Multipart compounds with hierarchic structural analysis: Det and Cop refer to determinative and copulative compounds

1)	[isän+maan]+rakkaus	'love for homeland'	Det
2)	ala+[ikä+raja]	'minimum age limit'	Det
3)	[maa+talous]+[oppi+laitos]	'rural institute'	Det
4)	[[aika+kaus]+lehti]+katsaus	'survey of periodicals'	Det
5)	sähkö+[[parran+ajo]+kone]	'(electric) razor'	Det
6)	[palo+päällikkö]+[[väestön+suojelu]+ohjaaja]	'fire chief and civil defense instructor'	Cop

The problems that multipart determinative compounds bring to automatic analysis can be further illuminated with examples from [15]. Three-constituent compounds can in principle be decomposed in two ways:

[6] https://www.statmt.org/europarl/
[7] http://opus.nlpl.eu/OpenSubtitles.php

7) [koira+valjakko]+kilpailut ‘dog sled race’
8) kirjasto+[tieto+kanta] ‘library database’

It seems that the first type is more common, but the latter one is also frequent in our data. Four constituent compounds are still more complex, as they can be decomposed in three ways. Lantee [15: 30] gives the following examples:

9) [arvo+paperi]+markkina]+laki ‘securities market law’
10) [mäki+hyppy]+[viikon+loppu] ‘ ski-jump weekend’
11) kesä+[kauppa+[korkea+koulu]] ‘ summer school of economics’

These decompositions can still be decomposed further, which increases the number of possible combinatorial analyses to five. If the compound has five or more parts, possibilities for analyses would increase.

3. Marking of Compounds in FiSTComp

3.1. An Initial Strategy

We have seen so far that compound structures may be complicated and a certain type of compound, i.e. the determinative compound, is the most frequent one. The number of constituents in a determinative compound is in theory unlimited, but two- and three-constituent determinative compounds are the most frequent ones. Four-constituent compounds occur to some extent too, but from five constituents on the frequencies are negligible [10], [12], [15]. Thus, we will concentrate only on compounds that have maximally four constituents in our compound tagging strategy.

In this paper, we use two different approaches to morphological analysis of words for FiSTComp: FinnPos [4], and Turku Dependency Parser [5–6], UD1. Both these tools disambiguate multiple morphological interpretations of words and provide word boundary markings for compounds, but the details and granularity of constituent decomposition vary. FinnPos’s style in compound splitting could be called cautious whereas UD1 is more prolific in splitting.

Most of the compounds – easily up to 85 % in different data – consist of two constituents, and these are easy to handle: FiSTComp tries first to analyze all split compounds as wholes, and if the whole is found in the lexicon, the program stops analysis and returns the result found in the lexicon. If the whole is not found in the lexicon, the two constituents are sought for in the lexicon and tagged, if possible.

As example analyses, we use words *pää+ministeri* (‘prime minister’ sg. nom.) and *oppositio+puolue* (‘opposition party’, sg. nom.). FiSTComp tries first to find the two-constituent word as a whole in the lexicon, and only after failure of that, the constituents *pää* and *ministeri* or *oppositio* and *puolue* would be sought for. Results of the analysis after FiSTComp look like this:

12) pääministeri Noun G1.1/S2
13) puolue Noun G1.2/S5+ oppositio Noun G1.2/S5+ COMP1

The first compound has been found in the semantic lexicon, and thus its meaning is one tag for the whole; the slash in the tag shows that the word belongs to two semantic categories. The second compound was not in the lexicon, and it is given the meaning of its constituents, *party* and *opposition*. The main constituent is presented first in the output of FiSTComp to mark its saliency for the meaning of the whole. Tag COMP1 is also attached to analyses where constituents of a two-part compound have been sought for in the lexicon.

In our test data of 2.17 million tokens, FinnPos analyses 91,139 tokens as compounds. Out of these 83,419 (91.5 %) are split to two constituents. In the same data, UD1 analyses 265,437 tokens as compounds, and out of these 227,549 (85.7 %) have two constituents. Analyses of UD1 seem to be far more useful for FiSTComp: 121,965 (45.9 %) of the marked compounds could be analyzed as wholes by FiSTComp, and the rest, 143,472 (54.1 %) were given a constituent analysis. This implies that UD1's compound splitting performs well. In comparison, out of FinnPos's compound analyses only about 3 % could be analyzed as wholes by FiSTComp.

3.2. A Refined Strategy

For two constituent compounds, the analysis is straightforward, but for more complex compounds other solutions are needed. The simple solution, treating the last constituent of the compound as the main part of the compound, works in many cases, but there are also lots of cases where the main internal word boundary should be set differently. The following examples depict this. In examples 14–15, the main constituent of the compound consists of the last two constituents, and the first constituent is a modifier for the whole.

14) aalto+[sulku+merkki] ('curly bracket')
15) aamu+[jumalan+palvelus] ('morning worship')

In example 16, however, the first two constituents should be kept together:

16) [aika+kaus]+julkaisu ('magazine')

This applies to four-constituent compounds, too. The last constituent may be sometimes the main part as in example 17, but many times, the last two constituents form the main part of the compound: this is the case with the examples 18–19.

17) [aika+kaus+lehti]+artikkeli ('magazine article')
18) [asian+ajo][valta+kirja] ('power of attorney')
19) [elo+hopea][lämpö+mittari]('mercury thermometer')

For three- and four-part compounds, a more elaborate initial strategy could be like this: as earlier, the whole word is first sought for in the lexicon. If it is not found, then two splittings are tried in this order for three-part compounds, keeping in mind that Finnish compounds are right-headed [13]:

1/2+3	try to find the longest possible end match first
1+2/3	if the longer end match does not succeed, try to find the last constituent first and then the initial combined part

If these do not bring results, then all the constituents need to be sought for separately in the lexicon. The same kind of strategy applies to four-part compounds, although this does not cover all the possibilities [15].

1+2/3+4 longest plausible match
1+2+3/4 the last constituent and the beginning as a whole

3.3. Results and Problems

We saw earlier that compound splitting with FinnPos was not very useful for FiSTComp even with two constituent compounds, and thus we use only results of UD1's compound analyses with FiSTComp in the analysis of the 2.17 million token corpus.

We compared FiSTComp's analysis with basic FiST which has no elaborated compound handling. FiSTComp achieved lexical coverage of 93.4 % with the corpus, whereas FiST achieved lexical coverage of 86.8 %. The gain was thus quite clear: 6.6 % units. As the only difference between the tagger versions is compound handling, splitting of compounds improves lexical coverage of the tagger significantly if the morphological analysis phase performs well.

UD1 marked 265,437 (12.23 %) words as compounds in the data. 85.7 % of these were marked as two-constituent compounds, 12.9 % as three-constituent compounds and 1.3 % had four constituents. The data had also 106 five-constituent and eight six constituent compounds in UD1's analysis.

There are some clear problems in compound analysis that rely on a morphological analyzer which is not integrated with the semantic tagger. First and foremost is the case when the morphological component does not produce good enough boundary markings for compounds, which seemed to be the case with FinnPos. Another problem is that morphological analyzers may, e.g., produce inaccurate analyses for some parts of the compounds. These include category changes in word class, e.g. from deverbal nouns to verbs: noun *tuottavuus+ohjelma* ('productivity program') is analyzed as *tuottaa+ohjelma* ('to produce + program'). Many times, the tags in semantic categories of the constituents are right even in these cases, but, anyhow, a whole word analysis would be better. Base forming of compound constituents may also make the analyzed word impossible to find in the lexicon as a whole. *Veron+kevennys (*'tax cut', the first constituent in sg. gen.*),* e.g., is analyzed as *vero+kevennys,* where the first constituent is lemmatized to sg. nom, and thus the word could not be found as a whole in the lexicon even if it is there. The form of the first constituent of a Finnish compound is most of the times sg. nom., but also genitive forms are common. Also, clear misanalyses occur in the morphological analysis phase: *tulevaisuus+valiokunta* ('future committee') becomes *tulla+valiokunta* ('come + committee') which blurs the meaning of the compound.

4. Conclusion

This paper has described an initial version of FiSTComp, a semantic tagger for Finnish with advanced compound analysis via compound constituents. As was shown, constituent analysis improves the lexical coverage of the tagger markedly – with 6.6 % units – in comparison to a tagger version with no compound constituent analysis. In a corpus with about 227,000 found compounds, FiSTComp was able to give some level of

constituent analysis to 54 % of the compounds which otherwise would have been left unanalyzed. FiSTComp thus improves compound handling, but further improvement is needed. Compounds are a multifaceted phenomenon, and so far we have scratched the surface of their structural composition. The question of representing the analysis results from a lexical semantic point of view, for example, would need separate discussion, which needs to be left for later development.

References

[1] Löfberg. L. Creating large semantic lexical resources for the Finnish language. Lancaster University, 2017. 422 pages. https://doi.org/10.17635/lancaster/thesis/3
[2] Kettunen, K. FiST – towards a Free Semantic Tagger of Modern Standard Finnish. In: Proceedings of the Fifth International Workshop on Computational Linguistics for Uralic Languages, 2019. p. 66-76.
[3] Kettunen, K. Nykysuomen automaattisesta semanttisesta merkitsemisestä. In: Jantunen, JH, Brunni, S, Kunnas, N, Palviainen, S, Västi, K, editors. Proceedings of The Research Data And Humanities (Rdhum) 2019 Conference: Data, Methods And Tools. Studia humaniora ouluensia. p. 215–228.
[4] Silfverberg, M, Ruokolainen, T, Lindén, K, Kurimo, M. FinnPos: an open-source morphological tagging and lemmatization toolkit for Finnish. Lang Resources & Evaluation 2016 50: 863–878.
[5] Haverinen, K, Nyblom, J, Viljanen, T, Laippala, V, Kohonen, S, Missilä, A, Ojala, S, Salakoski, T, Ginter, F. Building the essential resources for Finnish: the Turku Dependency Treebank. Lang Resources & Evaluation 2014 48: 493–531.
[6] Pyysalo, S, Kanerva, J, Missilä, A, Laippala, V, Ginter, F. Universal Dependencies for Finnish. In: Proceedings of the 20th Nordic Conference of Computational Linguistics, NODALIDA 2015, May 11-13, 2015, Vilnius, Lithuania. p. 163-172.
[7] Kettunen, K, LaMela M. Digging Deeper into the Finnish Parliamentary Protocols – Using a Lexical Semantic Tagger for Studying Meaning Change of Everyman's Rights (allemansrätten). In: Reinsone, S, Skadiņa, I, Baklāne, A, Daugavietis, J, editors. Proceedings of the Digital Humanities in the Nordic Countries 5th Conference Riga, Latvia, October 21-23, 2020. p. 63-80. http://ceur-ws.org/Vol-2612/
[8] Rauh, C, De Wilde, Pieter, Schwalbach, J. The ParlSpeech data set: Annotated full-text vectors of 3.9 million plenary speeches in the key legislative chambers of seven European states. 2017. https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/E4RSP9
[9] Finkbeiner R, Schlücker B. Compounds and multi-word expressions in the languages of Europe. In: Schlücker, B, editor. Compounds and Multi-Word Expressions. De Gruyter. 2018. p. 1–43.
[10] Hakulinen, A et al. Iso suomen kielioppi. Helsinki: Kotimaisten kielten tutkimuskeskus, 2008. http://scripta.kotus.fi/visk/etusivu.php
[11] Hyvärinen, I. Compounds and multi-word expressions in Finnish. In Schlücker, B, editor, Compounds and Multi-Word Expressions. De Gruyter. 2018. p. 307–337.
[12] Tyysteri, L. Aamiaiskahvilasta ötökkätarjontaan. Suomen kirjoitetun yleiskielen morfosyntaktisten yhdyssanarakenteiden produktiivisuus. 2015. Annales Universitatis Turkuensis C 408.
[13] Niemi, J. Compounds in Finnish. Lingue e linguaggio 2009 VIII2: 237-256.
[14] Saukkonen, P. Suomen kielen yhdyssanojen rakenne. In: Commentationes Fenno-Ugricae in honorem Erkki Itkonen sexagenarii die XXVI mensis aprilis anno MCMLXXIII: Erkki Itkonen 60 v. 1973. SUST. 150. Helsinki: SUS. p. 332–339.
[15] Lantee, A. Pitkät yhdyssanat Helsingin Sanomissa. M.A. thesis, Kieli-ja käännöstieteiden laitos - School of Modern Languages and Translation Studies. 2010. University of Tampere.
[16] Nikolaev, A, Niemi, J. Suomen nominien taivutusjärjestelmän produktiivisuuden indekseistä. Virittäjä 2008 112(4): 518–544.
[17] Pirinen, TA. Development and Use of Computational Morphology of Finnish in the Open Source and Open Science Era: Notes on Experiences with Omorfi Development. SKY Journal of Linguistics 2015 28: 381–393.
[18] Pajunen, A, Palomäki, U. Tilastotietoja suomen kielen rakenteesta, 1 / Frequency Analysis of Spoken and Written Discourse in Finnish. 1984. Helsinki: KOTUS (Research Institute for the Languages of Finland).