Technical Reports on Language Technology
Report No 77, 2021
http://www.njas.helsinki.fi/salama

Enhanced method for describing compound words[1]

Arvi Hurskainen
Department of Languages
FIN-00014 University of Helsinki, Finland
arvi.hurskainen@helsinki.fi

Abstract

In Technical Report No. 75 I proposed a method for describing compound words in Finnish. The aim in that report was to combine the coverage of the compound words and the limited capacity of the analyser. The number of compound words is almost limitless, which is why the compound words cannot be simply listed into the lexicon. In the report I described the method, where each compound words, regardless its length, was split into two segments. The split point was in front of the last segment. By so doing it was possible to handle compound words without needing to bother about the inflection class of each compound word. The inflection instructions were in the lexicon already in the list of basic words, that is, the non-compound words.
The method reduced the size of the lexicon, but it turned out that the reduction was not sufficient when I added the whole set of compound words in the corpus of one million word types. I made two decisions. First, I moved the description of verbs into another lexicon, so that the size of the initial lexicon became smaller. Second, I split the compound words so that each segment of the compound was isolated and moved into the corresponding sub-lexicon. The maximum number of such segments in a compound was defined to be four. If a compound had more than four segments, one or more segments were joined together, so that the maximum number became four.
Using these measures, it was possible to describe all compound words in the corpus. The report describes the method in detail.

Key Words: *compound words, morphological description..*

1 Introduction

One can consider various approaches for describing compound words. Because the number of compounds is in theory almost limitless, it is not feasible to try to list compounds as such into the lexicon. Exceptions are common two-segment compounds, which have been lexicalised and which often are considered as single words.

Because compounds are most often composed of nouns, it is tempting to construct such a lexicon, where each noun is allowed to combine with any other noun, and the last member of the compound is directed to its inflection lexicon.

This approach has several problems. First, the individual nouns as parts of compound are not always in base form. They are often in genitive form, and also other forms occur. Second, there are also other part-of-speech categories as segments of compounds, and it is not feasible to allow all of them to combine with nouns in the lexicon. Third, using this method, the lexicon would become excessively large.

This report describes the method, where the starting point is the corpus of 38 million words of news text from the year 2017. All unique word-forms were extracted, and the resulting corpus had one million words. I went manually through the corpus of unique word forms and separated compound segments from each other in all compound word types.

The task seemed almost impossible to carry out. However, the process became gradually easier, because I made the work in bunches and the words of each bunch were added to the lexicon. When a new bunch from the corpus was taken for manual handling, it was first analysed with the enhanced version of the lexicon, and only the non-analysed words were extracted. The method left an increasingly smaller percentage of the original list for manual work.

2 The structure of the lexicon for describing compound words

In the technical Report No. 75[2] I described the system, where all compound words were described so, that only the rightmost segment was removed, and the string on the left was listed in a separate sub-lexicon regardless its length. The result was that many sub-segments of the compound were listed several times, as far as the total string differed from other strings.

In this new implementation, the rightmost segment is removed, and the left part is split into up to three segments, depending on the length of the compound. Each segment is then located into its sub-lexicon, depending on its location in the compound.

As a result, all first segments of the compound are located into one sub-lexicon. All second segments are located into another sub-lexicon, and all duplicates are removed. All third segments are located yet into the third sub-lexicon, and all duplicates are removed. As a result, we have a condensed lexicon structure.

There is still one feature to be taken into consideration. That is the arbitrary use of the dash between segments. In Finnish there is the rule that the dash is only used between segments of a compound, when the left segment ends in the same vowel as the next segment begins (e.g. *ulko-ovi*). In practice, people tend to use the dash also elsewhere, where language rules would not expect it. Strictly speaking, this habit violates language rules. But because people behave as they do, we must cope with this problem, too.

I have solved the problem so that after each segment of the compound there is optionally a dash. Using this solution, we need to list each segment only once, regardless of whether it appears with a dash or without it.

The structure of the lexicon on the part compound words is in (1).

[2] http://www.njas.helsinki.fi/salama/describing-compound-words-in-finnish.pdf

(1)

```
LEXICON Start
 #  NORM;

LEXICON NORM
 PreNoun;
 MN;
 LN;
 Noun;
 Adj;

LEXICON PreNoun
aakkos D1 "= ";
aallon D1 "= ";
aalto D1 "= ";
aamiais D1 "= ";
aamu D1 "= ";
aamun D1 "= ";

LEXICON D1
- MN "= C1";
- Noun "= C1";
- Adj "= C1";
  MN " C1";
  Noun "= C1";
  Adj "= C1";

LEXICON MN
aallon D2 "= ";
aalto D2 "= ";
aamiais D2 "= ";
aamu D2 "= ";
aate D2 "= ";
aaton D2 "= ";
aatto D2 "= ";

LEXICON D2
- LN "= C2";
- Noun "= C2";
- Adj "= C2";
  LN "= C2";
  Noun "= C2";
  Adj "= C2";

LEXICON LN
aalto D3 "= ";
```

```
aihe D3 "= ";
aika D3 "= ";
aine D3 "= ";
ajo D3 "= ";
akku D3 "= ";
ala D3 "= ";
alainen D3 "= ";

LEXICON D3
- Noun "= C3";
- Adj "= C3";
  Noun "= C3";
  Adj "= C3";

LEXICON Noun
NounStem " N";

LEXICON NounStem
alh N1o "alho";
alkupotk N1u "alkupotku";
alm N1u "almu";
amm N1u "ammu";
antipast N1o "antipasto";
apartament N1o "apartamento";
apr N1o "apro";
arh N1o "arho";
arm N1o "armo";
ar N1o "aro";
arv N1o "arvo";
astiast N1o "astiasto";
as N1u "asu";

LEXICON Adj
AdjStem " A";

LEXICON AdjStem
omai N38 "omainen";
velt N1-Co "veltto";
soundi A38 "soundinen";
ahtimi A38 "ahtiminen";
ruuhkai A38 "ruuhkainen";
luiskai A38 "luiskainen";
teemai A38 "teemainen";
täytei A38-f "täyteinen";
eräi A38-f "eräinen";
```

```
LEXICON N1o
! vuon:o
o ### " SG NOM";
o POS " SG NOM/GEN/ACC";
on ### " SG GEN/ACC";
oa POS-an " SG PAR";
o ### " SG ACC-N";
ona POS-an " SG ESS";
ossa POS-an " SG INE";
osta POS-an " SG ELA";
oo POS " SG ILL";
oon ### " SG ILL";
olla POS-an " SG ADE";
olta POS-an " SG ABL";
olle POS-en " SG ALL";
otta POS-an " SG ABE";
okse POS-en " SG TRA";
oksi ### " SG TRA";
ot ### " PL NOM/GEN/ACC";
ot HanKo " PL NOM/ACC/ACC-N";
oje ## " PL GEN";
ojen ### " PL GEN";
oja POS-an " PL PAR";
oina POS-an " PL ESS";
oissa POS-an " PL INE";
oista POS-an " PL ELA";
oihi ## " PL ILL";
oihin ### " PL ILL";
oilla POS-an " PL ADE";
oilta POS-an " PL ABL";
oille POS-en " PL ALL";
oitta ## " PL ABE";
oikse POS-en " PL TRA";
oiksi ### " PL TRA";
oin ## " PL INS";
oine POS-en " PL KOM";
! comp, sup
om N16-H " CMP";
oi N51 " SUP";

LEXICON N38
! toi:nen
nen ### " SG NOM";
se POS " SG NOM/GEN/ACC";
sen ### " SG GEN/ACC";
sta  POS-an " SG PAR";
```

```
ne ### " SG ACC-N";
sena  POS-an " SG ESS";
sessa  POS-an " SG INE";
sesta POS-an " SG ELA";
see POS " SG ILL";
seen ### " SG ILL";
sella POS-an " SG ADE";
selta POS-an " SG ABL";
selle POS-en " SG ALL";
setta POS-an " SG ABE";
sekse POS-en " SG TRA";
seksi ### " SG TRA";
set HanKo " PL NOM/ACC/ACC-N";
se POS " PL NOM/GEN/ACC";
sie ## " PL GEN";
sien ### " PL GEN";
ste ## " PL GEN";
sten ### " PL GEN";
sia POS-an " PL PAR";
sina POS-an " PL ESS";
sissa POS-an " PL INE";
sista POS-an " PL ELA";
sii ## " PL ILL";
siin ### " PL ILL";
silla POS-an " PL ADE";
silta POS-an " PL ABL";
sille POS-en " PL ALL";
sitta ## " PL ABE";
sikse POS-en " PL TRA";
siksi ### " PL TRA";
sin ## " PL INS";
sine POS-en " PL KOM";
!
sem N16-H " CMP";
si N51 " SUP";
s ESTI " ADV";
```

The recognition of words starts with the strings listed in LEXICON Start. Here we have only the hash #, which marks the beginning of the wordform. From it you are allowed to proceed to LEXICON NORM. Here are listed the names of the lexicons with normal words. We see that, in addition to Noun and Adj, there are three other lexicon names, such as PreNoun, MN and LN. These are the names of the three sub-lexicons, which contain various segments of compound words. PreNoun contains the first segment of each compound word regardless its length. MN (medium noun) contains the second segment of the compound word, if such one exists in the compound. LN (last noun)

contains the third segment of the compound word. Note that all these three groups contain such segments of the compound, where the last part, the inflecting part, is excluded.

The lexicon is constructed so that you can start the word recognition from any of these five sub-lexicons. You can start from the first segment, or the second segment, or the third segment of the compound word. You can also go directly to the noun or adjective lexicon.

It is assumed that compound words are hierarchically constructed in the way as shown in (2).

(2)
ohjemisto(palvelu(liike(toiminta)))

It means that such words as *toiminta*, *liiketoiminta*, *palveluliiketoiminta* and *ohjelmistopalveluliiketoiminta* are valid words. The lexicon structure recognises all these forms.

From the lexicon with the name PreNoun, there is access to the lexicon with the name D1. This is a kind of switchboard, where you can either add a dash into the string or leave it out. From here is access to MN (medium noun), the second member of the compound. There is also access directly to the sub-lexicons of nouns and adjectives. Here we also add the tag C1 (compound segment 1), so that we can check, which segments of the compound lexicons were used in each reading.

Note that there is no access to the third section of the compound. This precaution is taken for avoiding unnecessary extra readings that the analyser produces.

Another switchboard is after the lexicon with the name MN. From it there is access to the nouns and adjectives, but also to the lexicon LN, which contains the segments that are in the third position of the compound.

The third switchboard lexicon D3 has access to nouns and adjectives.

We go to these lexicons via a separate lexicon, where we add the POS mark, N or A, to the reading. This method simplifies the listing of the actual words. Each noun and adjective has its own inflection code. This code is the name of the corresponding inflection lexicon. Only two examples are given above, N1o and N38.

The structure of the lexicon on the part of compound words is graphically described in Appendix 1. Only compound nouns and adjectives are included.

Next we analyse the compound word in (3).

(3)
"<ohjelmistopalveluliiketoiminta>"
"ohjelmistopalveluliiketoiminta" C1 C2 N SG NOM/ACC-N
"ohjelmistopalveluliiketoiminta" C1 C2 C3 N SG NOM/ACC-N
"ohjelmistopalveluliiketoiminta" C2 C3 N SG NOM/ACC-N

The word is recognised through three different routes. These are described in (4).

(4)

1	2	3	4
ohjelmisto	palvelu		liiketoiminta
ohjelmisto	palvelu	liike	toiminta
	ohjelmisto	palvelu	liiketoiminta

We see that the last slot has the segments *tominta* and *liiketoiminta*. The noun lexicon has the simple segment *toiminta* and the compound segment *liiketoiminta*. In theory, the noun lexicon should have only simple segments and compound words should be described by splitting the segments and placing them into separate lexicons. The implementation of the compound words was done on the top of the basic lexicon, which includes also a large number of compounds. If we remove the description of *liiketoiminta* as a single word from the lexicon, we get only one interpretation (5).

(5)
"<ohjelmistopalveluliiketoiminta>"
"ohjelmistopalveluliiketoiminta" C1 C2 C3 N SG NOM/ACC-N

Next we test how we can cope with the system, where part of compound nouns are described as single strings, and part is described by splitting the parts into separate lexicons (6).

(6)
"<riistamaaliammunta>"
"riistamaaliammunta" N SG NOM/ACC-N
"riistamaaliammunta" C1 N SG NOM/ACC-N
"riistamaaliammunta" C1 C2 N SG NOM/ACC-N
"riistamaaliammunta" C2 N SG NOM/ACC-N
"riistamaaliammunta" C2 C3 N SG NOM/ACC-N

The construction of the readings is described in (7).

(7)

1	2	3	4
			riistamaaliammunta
riista			maaliammunta
riista	maali		ammunta
	riista		maaliammunta
	riista	maali	ammunta

The morphological description of each of the five interpretations is precisely the same. It means that from the viewpoint of further processing it does not matter which interpretation we choose. However, we can assume that the whole-word analysis is more preferable than the analysis based on split segment combination.

We can implement this choice by adding to the disambiguator the following simple rules (8).

(8)
REMOVE (C1);
REMOVE (C2);
REMOVE (C3);

These rules remove all the readings, where the tags C1, C2 or C3 occur. After disambiguation we get the reading as in (9).

(9)
"<riistamaaliammunta>"
"riistamaaliammunta" N SG NOM/ACC-N

Next we test what happens, when the system produces several interpretations, but none of them is a single-word analysis (10).

(10)
"<lauluääni>"
"lauluääni" C1 N SG NOM/ACC-N
"lauluääni" C2 N SG NOM/ACC-N
"lauluääni" C3 N SG NOM/ACC-N

This structure can be described as in (11).

(11)

1	2	3	4
laulu			ääni
	laulu		ääni
		laulu	ääni

After disambiguation we get the result as in (12).

(12)
"<lauluääni>"
"lauluääni" C3 N SG NOM/ACC-N

We see that the rules in (8) removed readings one by one, first the reading with the tag C1, and then the reading with the tag C2. The last rule could not apply, because after its application there would be no reading left. The system does not allow the removal of the last reading. Therefore, the reading with the tag C3 was left. This is also a natural choice considering the hierarchical structure of the segments in compounds, as described in (2).

Finally, we test how the system analyses structures, which have alternative orthographies. Especially we point out the use of the dash for separating parts in compound nouns. In the corpus we find constructions such as in (13).

(13)
```
"<olympia-asiassa>"
        "olympia-asia"   C1 N SG INE
        "olympia-asia"   C2 N SG INE

"<olympia-asukohussa>"
        "olympia-asukohu"   C1  C2 N SG INE
        "olympia-asukohu"   C2  C3 N SG INE

"<olympia-asut>"
        "olympia-asu"   C1 N PL NOM/ACC/ACC-N
        "olympia-asu"   C2 N PL NOM/ACC/ACC-N

"<olympia-trampoliinilta>"
        "olympia-trampoliini"   C1 N SG ABL
        "olympia-trampoliini"   C2 N SG ABL

"<olympia-turnauksien>"
        "olympia-turnaus"   C1 N PL GEN
        "olympia-turnaus"   C2 N PL GEN

"<olympiaboikotista>"
        "olympiaboikotti"   C1 N SG ELA
        "olympiaboikotti"   C2 N SG ELA

"<olympiaboikottia>"
        "olympiaboikotti"   C1 N SG PAR
        "olympiaboikotti"   C2 N SG PAR
```

We see that every compound was analysed, although *olympia-trampoliini* and *olympia-turnaus* are written with a dash, although they should not. The recognition was made possible by allowing alternative continuations after each segment, one with the dash and another without.

After disambiguation the result is as in (14).

(14)
```
"<olympia-asiassa>"
        "olympia-asia"   C2 N SG INE
"<olympia-asukohussa>"
        "olympia-asukohu"   C2  C3 N SG INE
"<olympia-asut>"
        "olympia-asu"   C2 N PL NOM/ACC/ACC-N
"<olympia-trampoliinilta>"
        "olympia-trampoliini"   C2 N SG ABL
"<olympia-turnauksien>"
        "olympia-turnaus"   C2 N PL GEN
```

```
"<olympiaboikotista>"
        "olympiaboikotti"   C2 N SG ELA
"<olympiaboikottia>"
        "olympiaboikotti"   C2 N SG PAR
```

3 Recall and precision of the system

When the recognition system of compound words is constructed as described above, it allows for almost a limitless number of compound words. The first section of the compound can be combined with any noun or adjective. Also any section in the second and third slot can be combined with any noun and adjective. Therefore, the system recognises also such structures, which are grammatically correct, but which do not occur in real life.

In order to test the system with previously unknown text material, I collected a list of 250,000 unique words from a corpus of news texts of the year 2019 and extracted manually all compound words and formed a compound word corpus. This corpus has a total of 28,441 words.

The analysis of the corpus resulted in 2,230 unknown compound words. This means that 7.8 percent of the compound words were not analysed. This is quite a high percentage, and we need to take a closer look at the non-analysed compound words.

The first observation is that the corpus includes a large portion of such text, which deals with inter-galactical universes with a large number of such invented vocabulary that does not appear in normal text.

Another group of unknown words were typos.

The third group were compound verbs and adverbs, which were in advance excluded from the analysis system.

In (15) is shown the breakdown of the failure categories after the analysis of the corpus of 28,441 compound words.

(15)

Unknown compounds	**To be excluded**	**Real compounds**	**Total corpus**
Galactical words	551		
English	11		
Typos	268		
Verbs	132		
Adverbs	10		
Total unknown, incl. unreal	2230		
Real unknown compounds		1258	
Total corpus			**28441**

We see that there were various factors that distorted the result. There were 551 compounds from the very strange galactical language that does not occur in normal text. Also there were English words, typos, verbs and adverbs, which were not the subject of our study. While the total number of unknown compounds was 2,230, the number of true unknown compounds was 1,258. When we compare this number with the original corpus

size, we see that 4.4 percent of the compound words were not recognised by the analysis system.

When we take a closer look at these unknown compounds, we see that most of them were not analysed, because the last segment of the compound, the inflecting part, was not listed in the lexicon. Many of them were listed as part of compounds but not as single words. For example, there are such compounds as *grahamkeksi*, *koirankeksi*, *suolakeksi*, *täytekeksi*, *vohvelikeksi*, and *voileipäkeksi* in the lexicon, but not the simple word *keksi*. Therefore, the word *suklaakeksi* was not recognised.

The results give strong motivation for reconstructing the lexicon so that all such compounds that are listed directly as single words will be split into segments, and all preceding segments except for the last segment, will be moved to separate already existing lexicons.

As a conclusion we can say that out of the 4.4 percent of unrecognised compound words many were such which either did not have the base word in the lexicon or which were listed only as part of a compound word. The recall should improve when we reconstruct the lexicon.

The question on precision is less relevant in this system, because it allows for a large number of segment combinations, many of which do not occur in real life. Tests show, however, that typos are normally identified, although no prise typo recognition cannot be guaranteed.

4 Reconstructing the lexicon

The original lexicon had a total of 5,550 such compound nouns, which also could be treated in the way that I have shown above. In other words, they could be split into component parts without losing the possibility of translating them using the translation of individual segments. In addition, there were also such compounds, for which splitting is not feasible, because they cannot be translated using the translation of individual segments. In other words, they are multi-word expressions.

When the 5,550 compound words were split into segments, and each segment was counted only once, the total number of unique segments was reduced to 4,491.

The reduction of lexical entries was even more drastic, when the segments were added to the existing sub-lexicons and all duplicates were removed. The number of single-word nouns was originally 30,987. When this new arrangement was done, the number of single-word nouns was reduced to 27,720.

Also the number of left-side segments in slot 1 was reduced from 13,516 to 11,749.

In all, the size of the lexicon was reduced considerably, while at the same time it recognises even more compound words than before.

We can test the performance of the new lexicon using the example in (3) above (16).

(16)
```
"<ohjelmistopalveluliiketoiminta>"
        "ohjelmistopalveluliiketoiminta"   C1  C2  C3 N SG NOM/ACC-N
```

Now we have only one interpretation, because the compound *liiketoiminta* as a single word was removed from the lexicon.

We can also test the analysis of the example (6) above (17).

(17)
"<riistamaaliammunta>"
"riistamaaliammunta" C1 C2 N SG NOM/ACC-N
"riistamaaliammunta" C2 C3 N SG NOM/ACC-N

Now we have only two readings, one composed of segments 1 and 2 plus basic word, and another composed of segments 2 and 3 plus basic word.

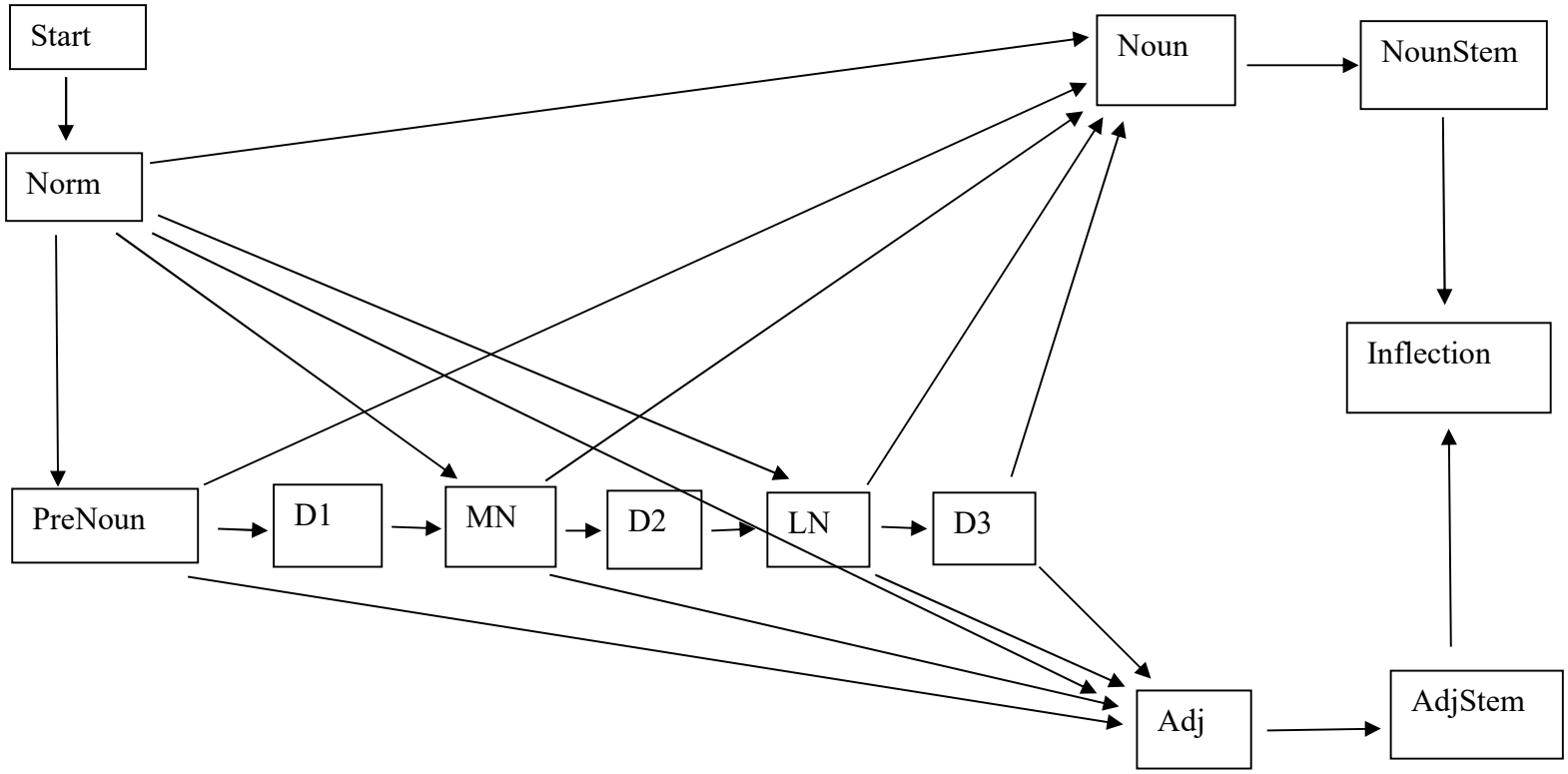
We see that the new lexicon structure gives results with less alternative readings.

I made a test on the recall of the enhanced lexicon using the corpus of 10,000 compound nouns. It had 123 unrecognised words. In other words, only 1,2 percent of the compound words were not recognised. These were mostly names of very rare chemical compounds and such rare words that were not listed in the basic list of nouns.

5 Conclusion

The enhanced structure of the morphological lexicon reduces considerably the size of the lexicon itself. It also recognises compound words in unknown text with the recall of 98,8 percent. The unrecognised words were such, for which the left sections were listed in the lexicon, but the listing of the base word was missing. Such additions can be easily done.

Appendix 1.


Start
Norm
PreNoun
D1
MN
D2
LN
D3
Noun
NounStem
Inflection
Adj
AdjStem

Key to Appendix 1.

Start – The lexicon for starting any wordform.
Norm – The lexicon, from which all normal wordforms start. Only those arcs that are relevant for the study are described.
PreNoun – The lexicon for the leftmost segments of compound words.
D1 – The lexicon, which allows continuation with or without a dash.
MN – The lexicon for the second segments of compound words.
D2 – The lexicon, which allows continuation with or without a dash
LN – The lexicon for the second segments of compound words.
D3 – The lexicon, which allows continuation with or without a dash
Noun – The lexicon for adding the identification code N to the reading.
NounStem – The lexicon, where base forms of nouns are listed.
Inflection – This is the collective name for more than 200 inflection lexicons.
Adj – The lexicon for adding the identification code A to the reading.
AdjStem – The lexicon, where base forms of adjectives are listed.