Gap Analysis Report for the Wordweb

Þórður Arnar Árnason

The Árni Magnússon Institute for Icelandic Studies

At the start of our gap analysis, a plan was made on how to prioritise changes to the Wordweb's database. The initial focus would be on errors that affected large groups of Lemmas, and ones that could be easily fixed, before moving on to smaller sets and more time-consuming changes. We then analysed the database, getting to know the data and identifying the most discernible patterns in its grammatical markings. The results from this analysis guided our choices in which particular errors to start correcting. Our work on some of these subsequently prompted the discovery of even more errors that were also in need of correction.

As a primary source, we relied on the database table which held the original Wordweb's data on Lemmas, in particular the following columns: "flid", the unique identifier of each entry; "fletta", the Lemma's raw text; "ofl", the grammatical mark for the controlling, or the only, word in the Lemma; and "mark", a list of grammatical marks for the words in each polylexical Lemma. This was not the full extent of available data on the Lemmas, but the remainder was inapplicable for error correction.

The focus of the gap analysis was on the texts in "fletta". Many changes were made to that column in accordance with the contents of "mark" and "ofl". These changes can be put into three categories: Firstly, too many or unnecessary words. Secondly, incorrectly marked Lemmas, or ones not in accordance with "ofl" or "mark". Thirdly, unknown symbols, often resulting in information that was incorrect or incomplete. We list the results of these changes below, after which we also discuss alterations based on a separate, preexisting datatable of automatically compiled potential errors.

Changes to lemmas with too many or unnecessary words

In this category, the errors were mainly caused by polylexical lemmas containing angle brackets (‘<’, ‘>’). The idea was to focus on two different problems pertaining to these lemmas.

In the first type of problem, word clusters of two or more words were inside angle brackets without being separated with a semicolon. In general, this resulted in an error as the format usually expects only a single word in that kind of context. Our approach was to create a regular expression which might be used to identify, and subsequently eliminate, these instances of multiple words within the brackets, e.g.

<word1; word2 word3; word4>

where "word2 word3" would result in an error. The following regular expression was applied to identify and eliminate a total of 1,165 instances of these word phrases:

```
(<[a-zA-Z0-9À-ž ]* +[a-zA-Z0-9À-ž]*)|(;+ *([a-zA-Z0-9À-ž]+ +[a-zA-Z0-9À-ž ]+))
```

When applied to the example provided above, this regular expression would deliver the results *<word1; word4>*, thus eliminating the error consisting of a word cluster where the format only accounts for single words e.g. *<word1; word2; word3>*.

There are pros and cons to this method. On occasion, the format would rope in valid word clusters. In those cases, the valid word cluster would be deleted, sometimes leaving empty brackets with no information. In most cases the format anticipated single words, making the few examples of the valid word cluster being deleted worth the information lost (these particular instances would be caught by other processes later on in the development cycle).

In the second type of problem, unnecessary data was found inside the angle brackets, complicating the Lemma instead of simplifying it. Let us look at an example:

```
"144392        ásækja <hann, hana>    so <fn-p-a>       so
```

The words in the angle brackets show which case the word uses (in this instance the accusative case), but as can be seen in the angle brackets of the corresponding mark (<fn-p-a>), only one word was expected. For this kind of error, we would identify the words in question, i.e. "hann, hana" (him, her), and replace it with a valid entry, such as "einhvern" (someone). These instances were isolated, and all 6,163 instances fixed.

<hann, hún	=	65	Nominative case
<hún, hann	=	0	Nominative case
<hann, hana	=	1729	Accusative case
<hana, hann	=	1	Accusative case
<honum, henni	=	3134	Dative case
<henni, honum	=	1	Dative case
<hans, hennar	=	1233	Genitive case
<hennar, hans	=	0	Genitive case

Incorrect marking, not in accordance with ofl or mark

After the changes listed here above had been made, we still had problems with unnecessary information in the angle brackets which were incorrectly marked. Where "ofl" indicated that the angle brackets should only contain a pronoun ("<fn->"), in 47 cases the Lemma either showed both a pronoun and a noun, or a singular noun (<hann; fiskistofninn>, where "hann" is the correct pronoun but "fiskistofninn" is a noun). These instances were isolated, and the unwanted words removed as listed:

<einhvern; word1> = 4	Accusative case
<einhverjum; word2> = 1	Dative case
<einhvers; word3> = 42	Genitive case

In other cases the gender of the lemma was incorrectly marked. This included 257 instances where more than one gender was specified, i.e. "kk/kvk" (male/female). These instances were individually isolated, analysed, and changed as appropriate. For example, the word "blikksmíði" had been marked "hk/kvk" (neutral/female) and the marking was changed to female as the neutral marking is either out-dated or simply wrong.

Unknown symbols found in mark or ofl, primarily ‘?’ and ‘/’.

In keeping with our intent of focusing on large-scale errors, we identified multiple instances of the symbols ‘?’ or ‘/’ in the "mark" column, which should have shown the word class of the lemma. In general, these instances were a result of an incorrect gender, or no gender at all, being assigned to their corresponding Lemmas. These 260 instances were isolated, and their correct gender assigned.

Automated error search database

As mentioned earlier, one of the Wordweb's database tables included data from what appeared to be an automatic consistency check that had been run sometime before the start of this project. It should be noted that this consistency check was not run on the same parameters as our search. It had clearly cast a much wider net, and some of its errors concerned aspects of the Wordweb that were deprecated in the new version. The Lemma counts accompanying each error code should thus not necessarily be taken to indicate actual faults or errors in the system, but rather as indications that some aspects of the original Wordweb's numerous database tables weren't aligning as well as they should.

The table contained 18 different error types. Of these 18 types, 7 contained no Lemmas, while the 103,760 errors were split rather unevenly between the remaining 11 as follows:

2 – missing mark - 372

3 – marking contains fewer words than the lemma - 800

4 – angle brackets open too soon in the marking - 681

5 – too few words within angle/square brackets in the lemma - 115

6 – lemma not found in BÍN database - 6644

7 – no analysis method for word class - 3581

9 – lemma found in the BÍN database but not in the Wordweb - 923

10 – more than one monolexical Wordweb lemma possible - 31.836

11 – number analysis not possible - 362

12 – mismatch in marks between tables 89

18 – pronoun-; analysis postponed - 58.357

There are four main errors that account for approximately 97% of all errors in the file. These are error types 6, 7, 10 and 18. With error types 6, 7 and 10, we made no changes. Type 6, "Lemma not found in the BÍN database", was likely a result of incorrect markings. Type 7 is a more complicated error as it involved Lemmas with unknown words, such as ones simply written as "S". These two errors could not be fixed en masse, and were left unaltered during the gap analysis. Type 10, which accounts for approximately 30% of the errors, was usually not an actual error. It states that the words could be associated with more than one Lemma in the Wordweb, which is normal and not an error as far as the gap analysis was concerned. It should be noted that as part of the Wordweb's conversion to the new format, our algorithms would automatically seek out any potential links between words in polylexical Lemmas, and their corresponding monolexical Lemmas.

The final major type, 18, contains over 56% of all errors. This error type states ‘pronoun- analysis postponed’. This error was often a result of one of the problems mentioned here above, such as where <hann, hana> was in the angle brackets rather than the simplified <einhver>. By making these changes we got rid of over 6,000 errors. In other cases, the reason was not so obvious and would need further analysis for correction.

The gaps in the database are understandably many since the Wordweb has been updated over many years by many different people. These gaps, or errors, are sometimes intertwined but others are completely isolated from others. The total number of changes made on the problems named in this report are around 8,000. Further work would undoubtedly result in even more changes and hopefully strengthening and simplifying the Wordweb for future use.