HARALDUR BERNHARÐSSON
JÓHANNES BJARNI SIGTRYGGSSON

Menota Lemmatization Notes

Discussion notes for the MLA Colloquium
Oslo, May 29–30th, 2006

[May 26, 2006]

1. MLA—technical matters

1.1 Non-appearing pages

MLA sometimes fails to show pages due to an error involving entities. The error message is always of this type:

“Error: Entity “X” could not be resolved (1).”

Exx: CR06-Hrbl 27 and 29.

1.2 The “Mismatch problem”

There are still MLA pages where the lemmatization cannot be saved (written into the xml-file) due to an error. The error message is always of this type:

“Error: Mismatch: “X” and “Y” are not equal.”

Exx. from CR26-Akv:

– Page 2: Error: Mismatch: "ï„‰rÃ³tt megÄ±r" and "ï„‰rÃ³tt megÄ±r" are not equal.

– Page 13: Error: Mismatch: "vann Å¿tyÉ¢va" and "vann Å¿tyÉ¢va" are not equal.

– Page 39: Error: Mismatch: "ber harÃ¾a" and "ber harÃ¾a" are not equal.

1.3 Longer “pos” blocks shorter “pos”

In some instances a longer grammatical analysis (“pos”) appears to block a shorter analysis: once the longer “pos” has been entered into the database (in the right-hand window), the shorter “pos” becomes inactive and no longer appears in the left-hand window. Exx.:

– “Prep **governing-acc**” blocks “Prep”

– “Adv pos **enclit**” blocks “Adv pos”

– “Verb fin pres imp 2. sg act redupl **enclit**” blocks “Verb fin pres imp 2. sg act redupl”

1.4 Showing the <dipl>

Currently MLA shows the <dipl> version of the text without the italicization of expansions. In some instances it would be useful to be able to distinguish the expansions from the rest. For instance, in Skírnismál 25 “Ser. þ. þ. m. m*er* er e. h. h. h*er*.” is shown like this in MLA: “Ser þv þenna męki mer. er ek hefi hendi her.”

In Thorell’s 1977 word index of DG 11, the Uppsala manuscript of Snorri’s Edda, such abbreviations (initial followed by a period) appear to be left out of the index.

1.5 Erased or unclear forms

Frequently underpunctuation fails to appear in MLA. Also, forms in square brackets cause an error, for instance Hrbl. 18 “a[h]t” and Hrbl. 19 “þr[vð]moþga”. When one tries to click on them in the MLA there appears an error message.

1.6 Empty norm-forms

When something has been written in the normalization text-field in MLA it isn’t possible to erase it and leave nothing there, because the form appears agains when the page is saved. The problem seems to be that something has to be in the <norm> in the XML-file after it has been created. When one makes a space in the empty text field and then saves the page nothing appears in it after that.

1.7 The neutralization of irregular forms in the MLA data base

When irregular forms (misspellings or incomplete forms) are lemmatized they are entered into the MLA database, like all other forms; as a result they will keep appearing in MLA’s suggestions for analysis. For instance, a single 1st pers. pres. ind. “tekr” entered in the data base will cause MLA to suggest that all forms “tekr” could be 1st person (instead of only 2nd or 3rd person).

It would be helpful to be able to neutralize such irregular forms and prevent them from reappearing.

1.8 Expelling incorrectly analyzed forms

Mistakes can be made: an incorrectly analyzed form, even if marked with “ut”, keeps sitting in the data base (haunting the “lemmateur”!). It would be comforting to be able to expell those embarrassing forms.

1.9 Notes

It would be nice if notes (by the editor) in the xml-text would appear in the MLA.

1.10 Adding <norm> fields?

Frequently more <norm> fields are needed, especially for punctuation. Can that be solved in the MLA?

1.11 A simple xml-editor within MLA?

Every time a minor correction needs to be made to an xml-file in the MLA, the user has to download the file, open it up in a editor, fix the file and finally upload it again to MLA.

This is a somewhat time consuming process and therefore the following question has come up: Is it possible to equip the MLA with a window that would allow the user to edit the underlying xml-encoded text (without having to download and upload)? Perhaps it would be sensible to have it display the text by line groups, as the MLA already does on the right-hand screen.

This is no doubt possible, but it is more a question of how much work it would require to make it happen.

2. Lemma names

As discussed earlier (cf. also the Menota Handbook §8.2), we have decided to use the ONP word list as a standard for the lemma names. Three points on this subject:

(a) Non-standard lemma names in MLA.
Currently only a part of the lemma names in the MLA corpus conform to the ONP standard. The deviations are mostly of two sorts:

(i) instead of "ǽ", many of the lemma names have plain "æ";

(ii) instead of "j" (for the semivowel) many (most) of the lemma names have "i"; exx. "telia", "velia", "iór" for the expected "telja", "velja", "jór".

(b) The ONP standard has changed.
What was initially written "aptr", "eptir", "opt" is now "aftr", "eftir", "oft", etc.

What is the most sensible way to correct this? Do we change the lemma names manually in the xml-file itself by series of find-and-replace operations?

(c) Unfortunate ONP lemma names.
The ONP lemma names are not always suitable as a standard. At the ONP the general principle is to have the lemma name in singular, even in words that only occur in plural. This principle yields lemma names like "ørlag", "rak", "skap" instead of "ørlǫg", "rǫk", "skǫp".

These ONP lemma names are unfortunate, and we find in hard to accept them. Therefore we have used the more common plural lemma names.

3. Internal consistency of the grammatical analysis (pos)

We have noticed, for instance, that there is some discrepancy in how past participles are analyzed, as the encoding for Voice appears in two different places in the morphological analysis, and frequently it is missing altogether (cf. the chapter 8.5.8.2 in the Menota Handbook where Voice and Inflectional class are missing in the example of the past participle). At some point it will therefore be necessary to look at all past participles.

There is also discrepancy in MLA as to whether present participles are marked as Indef or Def. All present participles need to be checked with this in mind—but perhaps present participles should not be marked for Species at all (see below).

These problems are probably best fixed in the xml-file at the end, right?

4. Combined word-class tags?

(a) Neutralizing the distinction of prepositions and adverbs. It does not seem practical to retain in all instances the distinction between adverbs and prepositions; in such cases it would be convenient to have a combined tag: Prep/Adv.

Also, would it be sensible to have a separate tag for particles, such as *of/um* (the "füllwort")?

(b) Also, *einn* can be a numeral, a pronoun, and an adjective. The distinction can be very difficult, and it is questionable if it is at all profitable to try to make this distinction. Therefore, a simple solution suggests itself: a combined word-class tag: Num/Pron/Adj.

5. Implementing an "X or Y analysis"

In section 8.4 of the Menota Handbook there is a discussion of homography and zero values, cf. especially 8.4.2 where the encoder can choose to analyze a given form as either Acc or Dat. How can this be implemented in MLA?

6. Adjectives used adverbially

Adjectives in acc. sing. neuter, such as *brátt*, *fljótt*, *hátt* of *bráðr*, *fljótr*, *hárr*, frequently are used adverbially, that is from a syntactic point of view they are adverbs, not adjectives. Similarly, the dat. plur. *stórum* frequently appears as an adverb.

These can be treated in two ways:
(a) as adverbs under the lemma names *brátt*, *fljótt*, *hátt*, *stórum*.
(b) as "adverbially used" adjectives under the lemmas *bráðr*, *fljótr*, *hár*, *stórr*, analyzed either as

  (i) adjective n. sg. acc. or perhaps
  (ii) adv.

We have been practicing the method undir (b) and (i). What is the "party line" on this subject?

7. Genitival compounds and other compounds

In the xml-file of the Codex Regius, genitival compounds are most often—but not always—in two separate <w> tags. It can be very difficult to decide what to treat as a genitival compound and what not; the distinction is bound to be somewhat arbitrary. Consider, for instance: *jöfra brúðr* (Grp 40), *þjóðar þengill* (Grp 41), *hers oddviti* (Grp 41).

Words that are left in two <w>-tags pose no particular problems; thus both elements of the compound are easily retrievable, each under its own lemma. As this will not be a word index, but a lemmatized concordance, the user will have no difficulty seeing when a particular word is part of a genitival compound and when not.

By contrast, compounds in a single <w>-tag, call for special measures in order for the second member to be retrievable. The usual solution to this is to include a reference under the lemma under which the non-initial member of the compound would belong or even print a separate list of non-initial members of compounds.

What is the best way to deal with this in MLA-generated concordances? How do we, to take an example, create a reference to *allmikill* under the lemma *mikill*?

8. Redundancy in the grammatical analysis

8.1 Number

It seems superflous to mark cardinal numbers higher than *einn* as plural.

8.2 Species

The Species category (definiteness) is really only important in the following instances:

(a) nouns: the presence (+) vs. absence (÷) of the suffixed definite article;
(b) adjectives and past participles: “weak” (+) vs. “strong” (÷) inflection (in the positive and superlative).

In the following, the marking of Species (definiteness) appears superflous:

(a) the definite article itself should not be marked for definiteness (in MLA *inn* is currently labelled “Art m sg nom indef”).
(b) proper names: inherently definite (semantically) and thus **very rarely** appear with the definite article (note exceptions like *Esjan*)
(c) adjectives in the comparative: appear **only** in the “weak” declension;
(d) present participles: appear **only** in the “weak” declension.
(e) numerals never appear with the definite article.

9. Enclitics

In the Menota Handbook (§8.3.2.11), the importance of encoding the cliticization of a personal prononun to a verb is discussed.

In such forms the division between the “host” and the enclitic is bound to be somewhat arbitrary. We have opted for a division where the verbal form is as close to being “intact” as possible, even if such division is at the expense of the cliticized pronominal form. Exx.: “att-v” (*eiga*), “ert-v” (*vera*), “gazt-v” (*geta*), “knatt-v” (*knega*), “kyst-v” (*kjósa*), “lezt-v” (*láta*: *lét-st* + *þú*), “sátt-v” (*sjá*), “skalt-v” (*skulu*), “vart-v” (*vera*), veizt-v” (*vita*), “þott-v” (*þykkja*).

Also in instances where assimilatory effect appears in the orthography of the verbal form, as in “mvnd-v” (*munu*: *munt* + *þú*), “vild-o” (*vilja*: *vilt* + *þú*). — Thus we will have a number of instances where the 2nd person prounoun *þú* appears in an enclitic form which only has the orthographic representation “v”.

It seems sensible to encode the cliticization of other forms as well:

(a) the pronoun *es*
– “þatz” < *þat* + *es*—NB “þaz” (HHv 2)!
– “þanns” < *þann* + *es*

(b) the negative particle *-a(t)*
– “erat”

Note also series of clitics:
– “grátt-at-v” of *gráta*: grátt-at-u
– “var-c-a” of *vera*: var-k-a