STO-LMF

1. BASIC INFORMATION

1.1 Lexicon type

Monolingual lexicon, morphological part, nouns, adjectives and verbs

1.2 Representation of the lexicon

One XML file

1.3 Character encoding

The characters have been encoded in UTF8

2. ADMINISTRATIVE INFORMATION

2.1 Contact person

Name: Sussi Olsen

E-mail: saolsen@cst.dk

2.2 Copyright statement and information on IPR

3. TECHNICAL INFORMATION

3.1 Data structure of an entry

The following data structure holds for the data in the first batch. The structure will be expanded with more features for batch 2 and with a syntactic layer for batch 3.

A lexical entry contains part of speech, lemma and word forms.

For the lemma, decomposition specifies if the word is a compound. It is also specified whether the lemma is in accordance with the official Danish orthography.

For each word form, all the relevant grammatical features are specified varying according to the part of speech.

Example of entry

```
<LexicalEntry>
    <feat att="partOfSpeech" val="NOUN_COMMON"/>
    <feat att="mu_id" val="HUKOMMELSE"/>
    <feat att="gmu_id" val="GMU_HUKOMMELSE_1"/>
    <feat att="origin" val="EDB-KORPUS"/>
    <feat att="autonomy" val="YES"/>
    <Lemma>
      <FormRepresentation>
        <feat att="spelling" val="hukommelse"/>
        <feat att="ro_approved" val="YES"/>
        <feat att="fugeResultat" val="hukommelse"/>
        <feat att="decomposition" val=""/>
      </FormRepresentation>
    </Lemma>
    <WordForm>
      <feat att="gender" val="COMMON"/>
      <feat att="grammaticalNumber" val="singular"/>
      <feat att="definiteness" val="indefinite"/>
      <feat att="case" val="unmarked"/>
        <FormRepresentation>
          <feat att="writtenform" val="hukommelse"/>
          <feat att="inflectionalParadigm" val="MFG0076"/>
            </FormRepresentation>
    </WordForm>
```

3.2 Lexicon size

86,935 morphological entries of the STO lexicon: 70305 nouns, 10572 adjectives and 6055 verbs.

4. CONTENT INFORMATION

4.1 The natural language(s) of the lexicon

Danish

4. 2 Entry Type

Lexical entries in LMF

4.3 Attributes

The key attributes of the lexicon are the lexical entry, the lemma and the word forms.

4.4 Coverage of the lexicon

The entire STO lexicon covers about 88,000 morphological entries, 43,000 syntactic entries and 10,000 semantic entries.

In this batch 86,000 morphological entries (nouns, verbs and adjectives) are included – the entries of the remaining part of speech will be included in the next batch.

4.5 Intended application of the lexicon

Intended applications are all kinds of language technology applications that need morphological information.

4.6 Reliability (automatically/manually constructed)

The original STO lexicon has been 100 % manually validated. The update to LMF has been automatically validated against the LMF DTD (ISO 24613).

5. RELEVANT REFERENCES AND OTHER INFORMATION

ISO 24613, International Standard: *Language resource management - Lexical Markup Framework (LMF)*, ISO 2008

Braasch A. et al.: *STO Sprogteknologisk Ordbase, monolingual lexicon, Documentation, version 2*, 2008. CST, KU, Copenhagen.

Braasch, Anna & Olsen, Sussi: STO: A Danish Lexicon Resource - Ready for Applications, 2004 In: *Fourth International Conference on Language Resources and Evaluation, Proceedings,* Vol. IV. Lisbon, pp. 1079-1082.