XV UNL School, 21-25 July 2014, Geneva

Exercise #4 – Dictionary

- Goal: To prepare the dictionary for the training corpus
- Deliverable:
  - *tgrammar_<ID>.txt*
- Activities:

1. Create a dictionary *dic_<ID>.txt* comprising all and only the entries appearing in your word list according to the instructions in the Annex.
2. Upload the segmented corpus to the UTK (UNLWEB>UNLDEV>UTK>SENTENCE>SENTENCES>ADD)
3. Check the results of the tokenization:
   a. UNLWEB>UNLDEV>UTK>CORPUS>CORPU>LOAD corpus
   b. UNLWEB>UNLDEV>UTK>CORPUS>DICTIONARY>LOAD dictionaries (default + your language)
   c. UNLWEB>UNLDEV>UTK>SENTENCE>PROCESS>TOKENIZE
4. Do the necessary changes until the tokenization is correct. Do not worry about the correct categorization (i.e., whether the string has been correctly categorized). This will be handled later.

ANNEX – INSTRUCTIONS FOR THE DICTIONARY

a. The dictionary entries must be provided, one per line, in a plain text (.txt) file in UTF-8 encoding, in the format below:

> [word form] {} "uw" (list of features) <language code,frequency,priority>;

b. You should include, in the dictionary, all and only the word forms appearing in your corpus:

> ~~[human being]{} "human being" (LEMMA=human being,LEX=N,POS=NOU,NUM=SNG) <eng,0,0>;~~ (this does not appear)
> [human beings]{} "human being" (LEMMA=human being,LEX=N,POS=NOU,NUM=PLR) <eng,0,0>;

c. Do not include punctuation signs and the blank space (they are already provided in the Default Dictionary)
d. Do not include temporary entries (such as digits)
e. Leave the field {} empty.
f. Use English lemmas (nouns in singular, verbs in infinitive) as UW's

> [homines]{} "~~homines~~" (LEMMA=homo,LEX=N,POS=NOU,GEN=MCL,NUM=PLR,CAS=NOM) <lat,0,0>;
> [homines]{} "~~human beings~~" (LEMMA=homo,LEX=N,POS=NOU,GEN=MCL,NUM=PLR,CAS=NOM) <lat,0,0>;
> [homines]{} "human being" (LEMMA=homo,LEX=N,POS=NOU,GEN=MCL,NUM=PLR,CAS=NOM) <lat,0,0>;

g. Only open class words (nouns, adjectives, adverbs and verbs) are mapped to UW's. Determiners, prepositions, conjunctions and auxiliary verbs are mapped to attributes (att=@def, att=@indef, att=@paucal, etc.) and/or to relations (rel=and, rel=tim, rel=plc, etc.). The list of attributes and relations may be found at www.unlweb.net/wiki/Tagset.

> [the]{} "~~the~~" (LEMMA=the,LEX=D,POS=ART) <eng,0,0>;
> ~~[the]{} "" (LEMMA=the,LEX=D,POS=ART) <eng,0,0>;~~ (att=@def is missing)
> [the]{} "" (LEMMA=the,LEX=D,POS=ART,att=@def) <eng,0,0>;

h. The list of features must be provided in the format ATTRIBUTE=VALUE and must contain all and only the morphological attributes that are relevant to your language.

> ~~[homines]{} "human beings" (LEMMA=homo,LEX=N,POS=NOU,GEN=MCL,NUM=PLR) <lat,0,0>;~~ (case is missing)
> [human beings]{} "human being" (LEMMA=human being,LEX=N,POS=NOU,~~GEN=MCL~~) <eng,0,0>;
> [άνθρωποι]{} "human being" (LEMMA=άνθρωπος,LEX=N,POS=NOU,GEN=MCL,NUM=PLR,CAS=NOM) <ell,0,0>;
> [esseri umani]{} "human being" (LEMMA=essere umano,LEX=N,POS=NOU,GEN=MCL,NUM=PLR) <ita,0,0>;

i. Include the lemma as a feature in the feature list.

> [human beings]{} "human being" (LEMMA=human being,LEX=N,POS=NOU) <eng,0,0>;

j. Use only the tags provided in the Tagset (www.unlweb.net/wiki/Tagset)

> [the]{} "" (~~determiner, article~~) <eng,0,0>; (these are not tags from the Tagset)

k. Duplicate ambiguous word forms (if the ambiguity appears in the corpus):

> [this]{} "" (LEMMA=this,LEX=D,POS=DET,att=@proximal) <eng,0,0>; (*this book*: this = determiner)
> [this]{}"00"(LEMMA=this,LEX=R,POS=DEP,att=@proximal)<eng,0,0>; (*this is the book*: this = pronoun)
> [is] {}""(LEMMA=be,LEX=V,POS=COP,ATE=PRS,PER=3PS)<eng,0,0>; (*John is a student*: is = copula)
> [is] {}""(LEMMA=be,LEX=I,POS=AUX,ATE=PRS,PER=3PS)<eng,0,0>; (*John is coming*: is = auxiliary)

l. Use the ISO 639-2 language codes (available at http://www.loc.gov/standards/iso639-2/php/code_list.php)

m. Frequency is used to disambiguate in UNLization (IAN,SEAN); priority is used to disambiguate in NLization (EUGENE). In both cases, the higher prevail over the lower. The values can be set from 0 to 255. Since this is a very small corpus, you may set frequency = 0 and priority = 0.