Mumbai, June 11-15

UNL Grammar Workshop

Indian Chapter

DAY #2

Summarizing

- Corpus
- NL-UNL (Analysis) Dictionary
  - From the word list to the NL-UNL dictionary
- UNL-NL (Generation) Dictionary
  - From the English dictionary to UNL-NL dictionary

Issues (Corpus)

- Corpus
  - How to translate:
    - Words (articles, prepositions, etc) that do not exist in the target language
      - DO NOT TRANSLATE THEM
        - the book = a book = book
    - Expressions that cannot be directly mapped onto the target language (book by John, book from Paris)
      - IF UNAVOIDABLE, INSERT THE MISSING WORDS
        - book by John => Book written by John
        - book from Paris => Book brought from Paris

Issues (NL-UNL dictionary)

- What is considered to be an "entry"?
  - Only the **words** appearing in the translated corpus
  - The English dictionary is only an EXAMPLE of the UNL-NL dictionary (the actual dictionary is to be made out of the word list, and not as a "translation" of the English dictionary)
  - Double-check:
    - Entries that **do not appear** in the corpus should be removed from the NL-UNL dictionary
    - Entries that **appear in the corpus** but are not present in the NL-UNL dictionary should be included there.

Issues (NL-UNL dictionary)

- Which features are necessary?
  - The ones required by your language:
    - If your language is number-inflective, NUMBER (NUM) is to be informed whenever the case
    - If your language is gender-inflective, GENDER (GEN) is to be informed whenever the case
    - If your language is case-inflective, CASE (CAS) is to be informed whenever the case
    - Etc.

Issues (NL-UNL dictionary)

- What to do with "complex words"?
  - To treat them as a single entry if this is not expensive; and to split them otherwise.
  - For instance
    - Postpositions, derivational morphemes and other non-inflectional markers are likely to be represented as separate entries when they are simply concatenated to the root, in order to avoid excessive proliferation of dictionary entries
    - However, inflectional morphemes that alter the structure of the root are to be represented as part of the entries, whenever they lead to the proliferation of roots.

Issues (UNL-NL Dictionary)

- The UNL-NL Dictionary must take the English Generation Dictionary (and not the word list) as the starting point
- The UNL-NL Dictionary must contain only **base forms** (and not word forms). The base form depends on the language. It is the "citation form" of the entry, i.e., the entry as it would appear in an ordinary monolingual dictionary:

LANG	WORD FORMS	BASE FORM
English	table, tables	table
English	die, dies, died, dying	die
English	am,are,is,was,were,...	be
French	beau, belle,beaux,belles	beau
Latin	amo, amas,amat,amamus,...	amo

Issues (UNL-NL Dictionary)

- The UNL-NL Dictionary must contain only the paradigm number in case of regular or quasi-regular entries
  - [table] {2883} "table" (POS=NOU,NUM=SNG,**PAR=M2**) <eng,0,0>;
  - In order to define the paradigms, group the words according to their morphological behavior and assign, to each group, an arbitrary number, larger than 1.
- The UNL-NL Dictionary must contain the inflectional rules in case of irregular entries
  - [foot] {2883} "10028466 5" (POS=NOU,PAR=M1,**FLX(PLR:="feet";)**) <eng,0,0>;
  - In order to avoid very long inflectional rules, include only the rules necessary to generated the inflected forms appearing in the corpus.

Warning

- We have been following the steps described in the workshop page at the Wiki.
- The best is the enemy of the good (Voltaire, "La Bégueule", 1772)
  - Our main goal during this workshop is not to build dictionaries but grammars.

Morphology

Inflectional Paradigms

- Inflectional paradigms (regular) x inflectional rules (irregular)
  - book, books => inflectional paradigm (absolutely regular)
  - man, men => inflectional paradigm (irregular, but very common in several lexemes: "service man", "gentleman", "superman", etc.)
  - foot, feet => inflectional rules (irregular and uncommon)

Rule types: simple rules

- **prefixation**
  CONDITION := "ADDED" < DELETED;
- **suffixation**
  CONDITION := DELETED > "ADDED";
- **infixation**
  CONDITION := [REFERENCE] > "ADDED";
  CONDITION := "ADDED" < [REFERENCE];
- **replacement**
  CONDITION := DELETED : "ADDED";
  CONDITION := [INTERVAL] : "ADDED";

Rule types: complex rules

- **circumfixation**
  - CONDITION := "ADDED" < DELETED , DELETED > "ADDED";
- **prefixation + infixation**
  - CONDITION := "ADDED" < DELETED , [REFERENCE] > "ADDED";
- **infixation + suffixation**
  - CONDITION := [REFERENCE] > "ADDED" , "DELETED" > "ADDED";

Examples of Paradigms (nouns)

- English
  - **M2**
    Add "s" to form the plural (*table>tables; boy>boys; computer>computers;*)
    SNG:=o>"";
    PLR:=o>"s";
  - **M3**
    Replace "y" by "ies" to form the plural (*baby>babies;*)
    SNG:=o>"";
    PLR:="y">"ies";
  - **M6**
    Replace "fe" by "ves" to form the plural (*life>lives; wife>wives;*)
    SNG:=o>"";
    PLR:="fe">"ves";

Examples of paradigms (nouns)

- French
  - **M20**
    Pluriel en -s, sans flexion de genre (*histoire, chaise*)
    SNG:=o>"";
    PLR:=o>"s";
  - **M235**
    Adjectifs en -teur (*consolateur, destructeur*)
    MCL&SNG:=o>"";
    FEM&SNG:=3>"rice";
    MCL&PLR:=o>"s";
    FEM&PLR:=3>"rices";

Examples of paradigms (nouns)

- Latin
  - **M2**
    Nouns of the Second Declension which end in -us (nominative singular) and -i (genitive singular). (*equus, dominus, hortus*)
    SNG&NOM:=o>"";
    SNG&GNT:="us">"i";
    SNG&DAT:="us">"o";
    SNG&ACC:="us">"um";
    SNG&VOC:="us">"e";
    ABL&SNG:="us">"o";
    NOM&PLR:="us">"i";
    GNT&PLR:="us">"orum";
    DAT&PLR:="us">"is";
    ACC&PLR:="us">"os";
    VOC&PLR:="us">"i";
    ABL&PLR:="us">"is";

Examples of paradigms (adjectives)

- French
  - **M3**
    Adjectifs en -n (*ancien, breton*)
    MCL&SNG:=o>"";
    FEM&SNG:=o>"ne";
    MCL&PLR:=o>"s";
    FEM&PLR:=o>"nes";
  - **M4**
    Adjectifs en -au (*nouveau, beau*)
    MCL&SNG:=o>"";
    FEM&SNG:=2>"lle";
    MCL&PLR:=o>"x";
    FEM&PLR:=2>"lles";

Examples of paradigms (adjectives)

- Latin
  - **M12**
    Adjectives that the Masculine is declined like second declension (i.e.annus, -i), the Feminine is declined like first declension (i.e. aba, -ae), and the Neuter is declined like second declension (i.e.bellum, -i) (*"malus","malam","eximius","eximia";*)
    NOM&SNG&MCL="us">"us";
    NOM&SNG&FEM="us">"a";
    NOM&SNG&NEU="us">"um";
    NOM&PLR&MCL="us">"i";
    NOM&PLR&FEM="us">"ae";
    NOM&PLR&NEU="us">"a";
    VOC&SNG&MCL="us">"e";
    VOC&SNG&FEM="us">"a";
    VOC&SNG&NEU="us">"um";
    VOC&PLR&MCL="us">"i";
    VOC&PLR&FEM="us">"ae";
    VOC&PLR&NEU="us">"a";
    ACC&SNG&MCL="us">"um";
    ACC&SNG&FEM="us">"am";
    ACC&SNG&NEU="us">"um";
    ACC&PLR&MCL="us">"os";
    ACC&PLR&FEM="us">"as";
    ACC&PLR&NEU="us">"a";
    GNT&SNG&MCL="us">"i";
    GNT&SNG&FEM="us">"ae";
    GNT&SNG&NEU="us">"i";
    GNT&PLR&MCL="us">"orum";
    GNT&PLR&FEM="us">"arum";
    GNT&PLR&NEU="us">"orum";
    DAT&SNG&MCL="us">"o";
    DAT&SNG&FEM="us">"ae";
    DAT&SNG&NEU="us">"o";
    DAT&PLR&MCL="us">"is";
    DAT&PLR&FEM="us">"is";
    DAT&PLR&NEU="us">"is";
    ABL&SNG&MCL="us">"o";
    ABL&SNG&FEM="us">"a";
    ABL&SNG&NEU="us">"o";
    ABL&PLR&MCL="us">"is";
    ABL&PLR&FEM="us">"is";
    ABL&PLR&NEU="us">"is";

Examples of paradigms (verbs)

- English
  - **M17**
    Regular verbs ending in -e (*vote > voted, voted, voting, votes*)
    INF:=0>"";
    PAS:=0>"d";
    PTP:=0>"d";
    3PS&PRS:=0>"s";
    GER:=1>"ing";
  - **M18**
    Regular verbs ending in -y (*unify > unified, unified, unifying, unifies*)
    INF:=0>"";
    PAS:="y">"ied";
    PTP:="y">"ied";
    GER:=0>"ing";
    3PS&PRS:="y">"ies";

Examples of paradigms (verbs)

- French
  - **M15**

Examples of paradigms (verbs)

- **M11 (234 rules)**
  The First Conjugation has a distinguishing vowel "a", and infinitive termination in "-are".Each Conjugation of a fully inflected verb has three different stems: Present Stem, Perfect Stem, Participial Stem. In the First Conjugation Perfect Stems are categorized in -avi, -ui, -i, Perfect reduplicated, and Deponents. (*poto;dono;indico.*)

How to create an inflectional paradigm?

- Paradigms apply only to inflectional categories
- Paradigms must include only simple forms generated through affixation (prefixation, suffixation and infixation)
- Paradigms target only regular or quasi-regular words.

Creating an inflectional paradigm

1. Define the inflectional values
2. Create the rule to each inflectional value

Practical example

Task #4

- Create the inflectional paradigms necessary to generate the inflected forms appearing in the corpus
- Convert the resulting inflectional paradigms to the grammar syntax as follows:
  - (%x,MX):=(%x,-MX,+FLX(RULES));
    - WHERE X is the paradigm number
    - RULES are the RULES of the paradigm
  - For instance:

(%x,M2):=(%x,-M2,+FLX(SNG:=0>""; PLR:=0>"s";));

- Save the grammar in a plain text (.txt) file with UTF-8 encoding and upload it to UNLWEB>UNLDEV>PROJECTS>EUGENE>GRAMMAR

Observation

- Those having CLEA700 may follow this steps inside the UNLarium and export the inflectional grammar