English Morphology for Computational Linguistics[1]

Isabel Moskowich – University of A Coruña

imoskowich@udc.es

www.udc.es/grupos/muste

Course outline

UNIT 1. Language taxonomy

- Language diversity
- Criteria for classification
- The typological classification

UNIT 2. Morphology basic concepts

- What is Morphology?
- The branches of morphology

  Derivational Morphology

  - derivational morphology
  - inflectional morphology
- Other key concepts
  - Affixes (prefixes, infixes, suffixes)
  - Vowel change
- What is a grammatical category?
- Inflectional Morphology
  - Lexeme
  - Grammatical category
  - Morpheme
  - Morph
  - Allomorph

UNIT 3. Types of languages

- Inflectional languages
- Agglutinative langugages
- Isolating languages

[1] This module on Morphology will be taught in two sessions. Each sesión consists of a four-hour contact sesión with the teacher in class plus three hours of self-study.

UNIT 4. Inflection and derivation

- English inflectional morphology
  - Affixation
  - Vowel change
  - Paradigms
- English derivational morphology
  - affixation
  - Word classes
- Complex words
- Morphology and other levels of analysis

UNIT 5. Computers and grammar

- A definition of Computational morphology
- Encoding
- Mark-up
- Annotation
- Types of annotation
- Formats

Supplementaru (web) materials

Glossary of linguistic terms

On language typology

- The classification of languages (very basic): http://lingvo.info/en/babylon/typology
- What is Morphology? https://www.youtube.com/watch?v=b-1PT4ZwwsM&feature=youtu.be

Some useful definitions

- Britannica's definition of inflection: https://www.britannica.com/topic/inflection
- .Britannica's definition of agglutinating language: https://www.britannica.com/topic/agglutination-grammar
- Turkic languages: https://www.britannica.com/topic/Turkic-languages
- Britannica's definition of isolating languages: https://www.britannica.com/topic/isolating-language

Supplementary material

- Difference between inflection and derivation: https://www.youtube.com/watch?v=GdrgdVgmX28&feature=youtu.be
- Some notes on inflectional morphology: https://www.thoughtco.com/inflectional-morphology-words-1691065
- English Inflectional Morphology: https://www.thoughtco.com/what-is-an-inflectional-morpheme-1691064
- On infixation: http://www.viviancook.uk/Words/infixes.htm

Some notes on derivational morphology: https://www.thoughtco.com/derivational-morpheme-words-1690381

Some exercises

- https://www.calpoly.edu/~jrubba/Morph.html
- https://benjamins.com/sites/z.156/exercise/c4q4

For computers and Morphology

- Leech's maxims of annotation:
https://www.lancaster.ac.uk/fss/courses/ling/corpus/Corpus2/2MAXIMS.HTM

  Geoffrey Leech was a pioneer corpus linguist who developed a set of rules for "good practice" in text annotation.

  Annotation has changed very much since he first published this (as it is now often automatic rather than manual), but his guidelines are still applicable.

  Read this maxims and think how to apply them to your own languages.

- Techniques in computational morphology:
http://ccl.pku.edu.cn/doubtfire/nlp/Lexical_Analysis/Word_Lemmatization/Introduction/Computational%20Morphology.htm

  Computational morphology deals with the processing of word forms, in both their graphemic (written) and phonemic (spoken) form. It is around us as it has many different applications (such as Spell-checkers or automatic hyphenation in word processors).

  These tasks imply hard problems for a computer programme. This reading provides some insights into why this is so and what techniques are available to tackle these tasks.

Links to corpora

A corpus is a colletion of texts that has been compiled based on some **principles** and encoded to become machine-readable. Corpora are used for many different aims nowadays and with many immediate applications (for instance, oral corpora are used to produce voice recognition devices).

- Applications of oral corpora:
http://www.oddcast.com/home/demos/tts/tts_example.php
  - BNC: The British National Corpus: http://www.natcorp.ox.ac.uk/

  One of the first and most important corpus of the English language

- Archivo de textos hispánicos de la Universidad de Santiago (ARTHUS):
http://adesse.uvigo.es/data/corpus.php

  The Archivo de textos hispánicos de la Universidad de Santiago (ARTHUS) contains texts in Spanish both from Spain and America (1,450,000 tokens). All the texts have nbeen syntactically analysed and the corresponding data are recorded in the databeses Base de datos sintácticos del español actual (BDS) andBase de verbos, alternancias de diátesis y esquemas sintáctico-semánticos del español (ADESSE).

- The Coruña Corpus of English Scientific Writing:
http://www.udc.es/grupos/muste/corunacorpus/index.html

  The Coruña Corpus of English Scientific Writing is being compiledat the Universidade da Coruña by the Research Group for Multidimensional Corpus-based Studies in English (MUStE). The corpus is a tool for the

study of the history of English during the late Modern English period (18th and 19th c.).

- The Helsinki Corpus of English Texts: http://www.helsinki.fi/varieng/CoRD/corpora/HelsinkiCorpus/
  First diachronic (historical) corpus ever compiled. It contains raw text marked-up in COCOA format. It gatjhers texts from 850 to 1710 with more than 1500000 words.Directed by Matti Rissanen and published in 1991.
- The Brown Corpus
  One of the first corpora ever compiled (1960s)
- Cambridge English Corpus
  A corpus designed for teaching English
  - The Corpus of the Tatar Language: http://www.corpus.tatar/en

  One of the first tries to compile and annotate a corpus in Tatar

Web-based software for annotation

- CLAWS: http://ucrel.lancs.ac.uk/claws/
  This is a part of speech (POS) automatic tagger
- USAS: http://ucrel.lancs.ac.uk/usas/
  A tool for semantic annotation
- ELAN: http://tla.mpi.nl/tools/tla-tools/elan/
  Tool for audio and video corpora annotation

Web-based corpus software

Sometimes, the computerised collections of texts (corpora) are not independently released but built-in into some kind of web that includes the software required for analysis. You have a couple of cases here.

- CQPWeb: https://cqpweb.lancs.ac.uk/
  This web-based sortware may give you access to different corpora. There are different servers around the world hosting CQPWeb. The one here is set at Lancaster University.
- The Sketch Engine: http://www.sketchengine.co.uk/
  This web-based software offers you the opportunity to upload and analyse yoir own corpus.

FOR SELF-STUDY (Morphology Module)

In this section you will find different resources that you can use on your own for non-contact hours with the teacher.

- Exercises and quizes: https://www.cs.bham.ac.uk/~pxc/nlp/InteractiveNLP/
- To know more (English Language and Linguistics)… http://www.ello.uos.de/field.php

As we have seen Uzbekh and Kazakh are languages belonging to the Tukic group inisde the Altaic family. As such, they are agglutinative language, very different from Englis (in the Germanic group ofthe Indo-European family)

- More on agglutinative languages: https://en.wikipedia.org/wiki/Agglutination

This may be useful as a starting point for annotating word structure in Kazakh nad Uzbekh

- University Centre for Computer Corpus Research on Language (Lancaster U.): http://ucrel.lancs.ac.uk/
- Another perspective