Computational Morphology: Introduction

Yulia Zinova

09 April 2014 – 16 July 2014

First part: Finite state techniques

1. 09 April – introduction of the course; What is computational morphology?
2. 16 April – finite state automata (FSA); finite state transducers (FST);
3. 23 April – weighted FSAs and FSTs; important algorithms;
4. 30 April – small test on FSAs and FSTs; morphological theory: overview of morphological operations;
5. 7 May – morphological operations (continuation); examples of FST application;

Computational issues, implementation and history

6. 14 May – computational issues; realizational vs. incremental morphology;
7. 21 May – computational implementation of fragments:
   - stem alternations in Sanskrit,
   - position classes in Swahili,
   - double plurals in Breton
8. 28 May – test?; history of the computational morphology: KIMMO

Machine learning, state of the art in Computational Morphology

9. 04 June – machine learning techniques for morphology;
10. 11 June – machine learning techniques for morphology;
11. 18 June – no class or replacement;
12. 25 June – discussing some paper?
13. 2 July – discussing some other paper?
14. 9 July – test or homework due;
15. 16 July – discussion of the test/homework, conclusion remarks.

Grading

- Attendance:
  - I don't care about attendance itself;
  - if you attend, please participate;
  - if you want to do something else – do it not in class;
- Active participation:
  - comments and questions during the class;
  - answering questions (even incorrectly, does not matter);
  - brings you 20 points;
- Tests/Assignments:
  - FSA and FST test: 20 points;
  - End of May: test or assignment on the analysis of some new data with methods learned; 20 points;
  - July: bigger test or assignment; 40 points.

Grades

- for a BN you need 50 points;
- AP:
  - 1.0: 95 – 100
  - 1.3: 91 – 94
  - 1.7: 87 – 90
  - 2.0: 83 – 86
  - 2.3: 80 – 82
  - 2.7: 75 – 79
  - 3.0: 70 – 74
  - 3.3: 65 – 69
  - 3.7: 60 – 65
  - 4.0: 50 – 59

Morphology

- Morphology: "study of shape" (Greek)
- Morphology in different fields:
  - Archaeology: study of the shapes or forms of artifacts;
  - Astronomy: study of the shape of astronomical objects such as nebulae, galaxies, or other extended objects;
  - Biology: the study of the form or shape of an organism or part thereof;
  - Folkloristics: the structure of narratives such as folk tales;
  - River morphology: the field of science dealing with changes of river platform;
  - Urban morphology: study of the form, structure, formation and transformation of human settlements;
  - Geomorphology: study of landforms

Morphology in linguistics

- The study of the internal structure and content of word forms;
- First linguists were studying morphology:
  - ancient Indian linguist Pānini formulated 3,959 rules of Sanskrit morphology in the text Astādhyāyī;
  - The Greco-Roman grammatical tradition was also engaged in morphological analysis.
  - Studies in Arabic morphology: Marāḥ al-arwāḥ and Ahmad b. ʻalī Masʻūd, end of XIII century;
  - Well-structured lists of morphological forms of Sumerian words: written on clay tablets from Ancient Mesopotamia; date from around 1600 BC.

An ancient example

- Well-structured lists of morphological forms of Sumerian words: written on clay tablets from Ancient Mesopotamia; date from around 1600 BC;

badu	'he goes away'	inĝen	'he went'
baddun	'I go away'	inĝenen	'I went'
bašidu	'he goes away to him'	inšiĝen	'he went to him'
bašiduun	'I go away to him'	inšiĝenen	'I went to him'

(see Jacobsen, 1974, 53-4)

Terminology

- **Word-form, form:** A concrete word as it occurs in real speech or text.
  For computational purposes, word is a string of characters separated by spaces in writing;
- **Lemma:** A distinguished form from a set of morphologically related forms, chosen by convention (e.g., nominative singular for nouns, infinitive for verbs) to represent that set.
  Also called the canonical/base/dictionary/citation form. For every form, there is a corresponding lemma.

Terminology

- **Lexeme:** An abstract entity, a dictionary word; it can be thought of as a set of word-forms. Every form belongs to one lexeme, referred to by its lemma.
  For example, in English, steal, stole, steals, stealing are forms of the same lexeme steal; steal is traditionally used as the lemma denoting this lexeme.
- **Paradigm:** The set of word-forms that belong to a single lexeme.

Example

- The paradigm of the Latin lexeme INSULA 'island'

	singular	plural
nominative	insula	insulae
accusative	insulam	insulas
genitive	insulae	insularum
dative	insulae	insulis
ablative	insula	insulis

Terminology: Complications

- The terminology is not universally accepted, for example:
  - lemma and lexeme are often used interchangeably (and so will we use it too);
  - sometimes lemma is used to denote all forms related by derivation;
  - paradigm can stand for the following:
    1. set of forms of one lexeme;
    2. a particular way of inflecting a class of lexemes (e.g. plural is formed by adding -s);
    3. a mixture of the previous two: set of forms of an arbitrarily chosen lexeme, showing the way a certain set of lexemes is inected (language textbooks).

Morpheme

- Morphemes are the smallest meaningful constituents of words;
- e.g., in *books*, both the suffix *-s* and the root *book* represent a morpheme;
- words are composed of morphemes (one or more).

Morpheme

- Morphemes are the smallest meaningful constituents of words;
- e.g., in *books*, both the suffix *-s* and the root *book* represent a morpheme;
- words are composed of morphemes (one or more).
- Your examples?
    1. a word with 1 morpheme?

Morpheme

- Morphemes are the smallest meaningful constituents of words;
- e.g., in *books*, both the suffix *-s* and the root *book* represent a morpheme;
- words are composed of morphemes (one or more).
- Your examples?
    1. a word with 1 morpheme?
    2. 2 morphemes?

Morpheme

- Morphemes are the smallest meaningful constituents of words;
- e.g., in *books*, both the suffix *-s* and the root *book* represent a morpheme;
- words are composed of morphemes (one or more).
- Your examples?
    1. a word with 1 morpheme?
    2. 2 morphemes?
    3. 3 morphemes?

Morpheme

- Morphemes are the smallest meaningful constituents of words;
- e.g., in *books*, both the suffix *-s* and the root *book* represent a morpheme;
- words are composed of morphemes (one or more).
- Your examples?
    1. a word with 1 morpheme?
    2. 2 morphemes?
    3. 3 morphemes?
    4. 4 morphemes?

Morpheme

- Morphemes are the smallest meaningful constituents of words;
- e.g., in *books*, both the suffix *-s* and the root *book* represent a morpheme;
- words are composed of morphemes (one or more).
- Your examples?
    1. a word with 1 morpheme?
    2. 2 morphemes?
    3. 3 morphemes?
    4. 4 morphemes?
    5. 5 and more morphemes?

Morphs and allomorphs

- The term *morpheme* is used both to refer to an abstract entity and its concrete realization(s) in speech or writing.
- When there is a need to make a distinction, the term *morph* is used to refer to the concrete entity, while the term *morpheme* is reserved for the abstract entity only.
- Allomorphs are variants of the same morpheme, i.e., morphs corresponding to the same morpheme;
- Allomorphs have the same function but dierent forms. Unlike the synonyms they usually cannot be replaced one by the other.
- Examples?

Examples of allomorphs

(1) a. indenite article: **an** orange - **a** building
b. plural morpheme: cat-**s** [s] - dog-**s** [z] - judg-**es** [@z]
c. opposite: **un**-happy - **in**-comprehensive - **im**-possible - **ir**-rational

Morphemes

- The order of morphemes/morphs matters:

  (2) a. talk-ed $\neq$ *ed-talk
  b. re-write $\neq$ *write-re
  c. un-kind-ly $\neq$ *kind-un-ly

- Complications: how would you decompose *cranberry* into morphemes?

Morphemes

- The order of morphemes/morphs matters:

  (2) a. talk-ed $\neq$ *ed-talk
  b. re-write $\neq$ *write-re
  c. un-kind-ly $\neq$ *kind-un-ly

- Complications: how would you decompose *cranberry* into morphemes?
- The *cran* is unrelated to the etymology of the word *cranberry* (crane (the bird) + berry).

  (3) cranberry = crane + berry $\neq$ cran + berry

- Zero-morphemes, empty morphemes.

Morphemes: types

- Bound morphemes cannot appear as a word by itself.
- Examples?

Morphemes: types

- Bound morphemes cannot appear as a word by itself.
- Examples?
- -s (dog-s), -ly (quick-ly), -ed (walk-ed)
- Free morphemes can appear as a word by itself; often can combine with other morphemes too.
- Examples?

Morphemes: types

- Bound morphemes cannot appear as a word by itself.
- Examples?
- -s (dog-s), -ly (quick-ly), -ed (walk-ed)
- Free morphemes can appear as a word by itself; often can combine with other morphemes too.
- Examples?
- house (house-s), walk (walk-ed), of, the, or

Morphemes: types

- The property of being bound or free is language-dependent: past tense morpheme is a bound morpheme in English (-ed) but a free morpheme in Mandarine Chinese (le)

(4) a. Ta chi le fan.
He eat past meal.
'He ate the meal.'

b. Ta chi fan le.
He eat meal past.
'He ate the meal.'

Morphemes: types

- **Content** morphemes carry some semantic content;
- **Functional** morphemes provide grammatical information;
- Examples?

Morphemes: Root

- Root is the nucleus of the word that axes attach too.
- In English, most of the roots are free.
- In some languages that is less common: in Russian, noun and verbal roots are bound morphemes, sometimes with zero affixes;
- Some words (compounds) contain more than one root: *homework*.

Morphemes: Affixes

- **Affix** is a morpheme that is not a root; it is always bound;
- **Suffix** follows the root;
- Suffixes in English: -ful in event-ful, talk-ing, quick-ly, neighbor-hood
- **Prefix** precedes the root;
- Prefixes in English: un- in unhappy, pre-existing, re-view;
- **Infix** occurs inside the root;
- Infixes in Khmer: -b- in lbeun 'speed' from leun 'fast';
- Infixes in Tagalog: -um- in s-um-ulat 'write'
- **Circumfix** occurs on both sides of the root
- Circumfixes in Tuwali Ifugao: baddang 'help', ka-baddang-an 'helpfulness', *ka-baddang, *baddang-an;
- Circumfixes in Dutch:
  - berg 'mountain' – ge-berg-te 'mountains', *geberg, *bergte;
  - vogel 'bird', ge-vogel-te 'poultry', *gevogel, *vogelte

Typology of affixation

- Suffixing is more frequent than prexing;
- Infixing/circumfixing are very rare (Greenberg, 1957);
- Postpositional and head-final languages use suffixes and no prefixes;
- Prepositional and head-initial languages use not only prefixes, as expected, but also suffixes.
- Many languages use exclusively suffixes and no prefixes (e.g., Basque, Finnish).
- Very few languages use only prefixes and no suffixes (e.g., Thai, but in derivation, not in inflection).

Example problem

(5) I am swimming

- There is a lexeme 'to swim'
- The +ing portion tells us that this event is taking place at the time the utterance is referring to.
- Why there is an extra **m**?

Computational Morphology

- **Computational morphology** deals with developing techniques and theories for computational analysis and synthesis of word forms.
- Applications?

Computational Morphology

- **Computational morphology** deals with developing techniques and theories for computational analysis and synthesis of word forms.
- Applications?
- Spelling correction
- Search engines
- Machine translation
- Text generation
- Text-to-speech

Next class

- In the next session we will discuss regular languages, finite state automata, regular expressions and finite state transducers.
- A task for those of you who know something about regular languages and/or finite state automata: bring an expression or an automaton you like (and can draw on the board!).

References:

Greenberg, J. H. (1957). The nature and uses of linguistic typologies. *International journal of American linguistics*, pages 68–77.

Jacobsen, T. (1974). Very ancient texts: Babylonian grammatical texts. *Studies in the history of linguistics: traditions and paradigms*, page 41.