

Morphological modelling

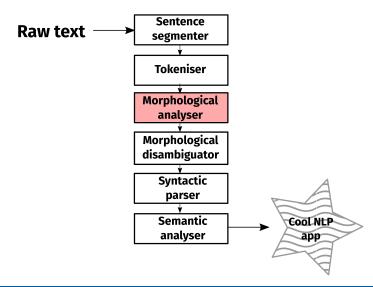
Francis M. Tyers

ftyers@hse.ru https://www.hse.ru/org/persons/209454856

Национальный исследовательский университет «Высшая школа экономики» (Москва)

28 октября 2018 г.





The story so far



В 1942—1945 годах профессором Г. С. Петровым и сотрудниками была разработана серия клеев БФ^[1]. Советский учёный-химик Петров знаменит также «контактом Петрова» и работами в области химии и технологии карболита (бакелита, фенолформальдегидных пластмасс)^[2].



В 1942—1945 годах профессором [[Петров, Григорий Семёнович|Г. С. Петровым]] и сотрудниками была разработана серия клеев БФ-геf>[http://chem21.info
/раде/034120176225149200221127252239157188201019105199/ Справочник по пластическим массам Том 2
(1969) стр.149.]</ri>
/геf>. Советский учёный-химик Петров знаменит также «контактом Петрова» и работами в области химии и технологии [[карболит]]а ([[бакелит]]а, фенолформальдегидных пластмасс)-геf>[http://www.planet-of-people.org/htmls/rus/nadezhdin/plastmassa.htm Надеждин Н. Я. История науки и техники. Пластмасса!-- Заголовок добавлен ботом -->]{{Недоступная ссылка|date=Июль 2018 |bot=InternetArchiveBot }}{{битая ссылка}}



В 1942—1945 годах профессором Г. С. Петровым и сотрудниками была разработана серия клеев БФ. Советский учёный-химик Петров знаменит также «контактом Петрова» и работами в области химии и технологии карболита (бакелита, фенолформальдегидных пластмасс).



В 1942 — 1945 годах профессором Г. С. Петровым и сотрудниками была разработана серия клеев БФ . Советский учёный-химик Петров знаменит также « контактом Петрова » и работами в области химии и технологии карболита (бакелита , фенолформальдегидных пластмасс) .

The story so far



В 1942—1945 годах профессором Г. С. Петровым и сотрудниками была разработана серия клеев БФ^[1]. Советский учёный-химик Петров знаменит также «контактом Петрова» и работами в области химии и технологии карболита (бакелита, фенолформальдегидных пластмасс)^[2].



В 1942—1945 годах профессором [[Петров, Григорий Семёнович|Г. С. Петровым]] и сотрудниками была разработана серия клеев БФ-геf>[http://chem21.info
/раде/034120176225149200221127252239157188201019105199/ Справочник по пластическим массам Том 2
(1969) стр.149.]</ri>
/геf>. Советский учёный-химик Петров знаменит также «контактом Петрова» и работами в области химии и технологии [[карболит]]а ([[бакелит]]а, фенолформальдегидных пластмасс)-геf>[http://www.planet-of-people.org/htmls/rus/nadezhdin/plastmassa.htm Надеждин Н. Я. История науки и техники. Пластмасса!-- Заголовок добавлен ботом -->]{{Недоступная ссылка|date=Июль 2018 |bot=InternetArchiveBot }}{{битая ссылка}}



В 1942—1945 годах профессором Г. С. Петровым и сотрудниками была разработана серия клеев БФ. Советский учёный-химик Петров знаменит также «контактом Петрова» и работами в области химии и технологии карболита (бакелита, фенолформальдегидных пластмасс).



В 1942 — 1945 годах профессором Г. С. Петровым и сотрудниками была разработана серия клеев БФ . Советский учёный-химик Петров знаменит также « контактом Петрова » и работами в области химии и технологии карболита (бакелита , фенолформальдегидных пластмасс) .

Overview



- Morphology: What is it? Why should we care?
- Modelling morphology: With finite-state machines

•

Morphology

What is morphology?



Morphology is:

« the branch of linguistics that studies patterns of word formation within and across languages, and attempts to formulate rules that model the knowledge of the speakers of those languages. »

This is a big field, here we are interested in practical models.

Why produce models?



English or Chinese:

- A full form list is a possibility
- Few or no inflectional forms
 - e.g. 5 forms per English verb {see, sees, saw, seen, seeing}

Other languages:

- Difficult or impossible to enumerate all forms
- Very productive inflection and derivation
 - Russian verbs: over 150 forms (maximally)
 - Turkish verbs: thousands of forms



```
PÝCNO, a, g. pl. annel; pýcno, a, g. pl. annel; river. Triver) bed, channel; course of a river. Change the course of a river. Change the course of the cours
```

A morphological lexicon consists of entries:

- Lemma: The citation form of a word (cf. headword)
- Stem: The part of a word affixes attach to
- Paradigm: A description of how the word inflects:



Add additional meaning or change the meaning of a lexical stem:

- **Suffixes:** hus 'house' huset 'the house'
- Prefixes: kjent 'known' ukjent 'unknown'
- Infixes: ktieb 'book' kotba 'books'
- Circumfixes: nagy 'big' legnagyobb 'biggest'

Morphological processes



- Inflection: Inflectional morphemes carry grammatical information, such as number, case, tense, etc., but do not change the word category
- Derivation: Derivational morphemes change the basic semantic meaning of a word, and can also change word category.
- Compounding: A process where two or more words are joined together to form one, typically of the same category or supertype.
- Clitics: Syntactically independent word that functions phonologically as an affix of another word.
- Incorporation: Where a nominal (e.g. direct object) or adverbial is included into a verb form.

Inflection



Examples of inflection categories:

- Case: дом·у 'house-LOC', ev·de 'house-LOC', talo·ssa 'house-INE'
- Possession: ev-im 'house-1sg', talo-ni 'house-1sg'
- Number: дом·a 'house-PL, ev·ler 'house-PL', talo·t 'house-PL'
- Tense, aspect, mood: говори·ла 'say-PAST.F, söyle·di 'say-PAST', puhu·i 'say-PAST'
- Comparison: больше 'big-comp', пысак рах 'big-comp', iso mpi 'big-comp'

Высшая школа экономики

In general: Change in meaning is regular.

Derivation



Examples of derivational affixes:

- Actor: diş·çi /tooth-er/ 'dentist'
- State: boş·luk 'emptiness', nycm·oma 'emptiness'
- Diminutive: dog·gie, kedi·cik /cat-DIM/ 'kitten'

Can often be stacked:

- temizlikçi /temiz-lik-çi/ clean-ness-er = cleaner
- поверхностный /по-верх-ность-ный/ on-surface-ness-ly = superficial

Высшая школа экономики

Change in meaning may be irregular, compare:

- cooker /cook-er/ 'machine that cooks'
- cleaner /clean-er/ 'person who cleans'
- looker /look-er/ 'person that looks good'

May be limited to particular stems.

Compounding



New words are formed from morphologically/syntactically independent words:

- This may be indicated in the writing system or not.
 - infrastruktuurontwikkelingsplan, or
 - infrastructure development plan
- tri-noun compounds, but different orthographical treatment

Note: a given compound word may be split different ways, or a given word may appear as a compound, but not be one:

- Freitag = Friday (not "Frei" + "tag" = free day)
- kulturforskeren = the ethnographer, and not
 - kultur+forskeren = "culture researcher"
 - kultur+forske+ren = "culture research clean"



Clitics are syntactically separate words that are phonologically conditioned by another unit (word, phrase).

Pronominal:

- Spanish: me lo das me it you.give 'You give it to me'
- Spanish: dámelo! give-me-it 'Give it to me!'

• Verb forms:

- Serbo-Croatian: govorit ću vs. govoriću 'I will speak'
- English: I'm 'I am', gonna 'going to'

Other:

- Question words (e.g. Finnish onko? is-QST? 'Is there?')
- Tense markers (e.g. Kurdish -ê)

Should these be tokenised prior to analysis?

Incorporation



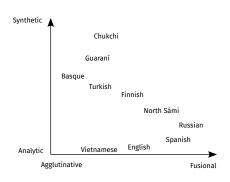
Гақорапэнратлэн Сыкванақай рэмкык "Cıkwaŋaqaj chased after the reindeer in the other encampment."

```
га-қора-пэнр-ат-лэн Сыкванақай рэмк-ык 
PERF-reindeer-chase-s3sG Cıkwaŋaqaj folk-LOC
```

- Syntactically determined (not lexically!)
- Can be valency changing, e.g.
 - DOBJ + V.TR \rightarrow V.INTR
- ...

Morphological typology





- Analytic—Synthetic:
 - Morphemes per word
- Agglutinative—Fusional:
 - Meanings per morpheme

Modelling

Finite-state morphology



Representing the lexicon



Analysis and generation



Two-level rules



Rule application



Rewrite rules



Development

24/29

Development guidelines



Applications



What we have not covered



- Templatic morphology:
- Machine learning approaches:



Go through the following practical:

https://ftyers.github.io/2017-КЛ_МКЛ/hfst.html

This will take you through all of the main steps to build a transducer.