Лингвистические основы машинного перевода. Краткий курс

Linguistic Foundations of Machine Translation. An overview course

Л.Л.Иомдин

Лаборатория компьютерной лингвистики

Института проблем передачи информации РАН им. А.А. Харкевича

УНЦ компьютерной лингвистики РГГУ

Leonid Iomdin

Laboratory of Computational Linguistics,

A.A.Kharkevich Institute for Information Transmission Problems, RAS

Education and Research Centre of computational linguistics, RSUH

iomdin@iitp.ru, ***iomdin@gmail.com***

Морфологический анализ
Morphological Analysis

- **Поверхностный** морфологический анализ: свойства морфем
- Surface morphological analysis: properties of morphemes:

*ид**ущ**ий – лет**ящ**ий – кол**ющ**ий - молч**ащ**ий*

*брат – брат**ья**, отец – отц**ы**, мать – мат**ери**, сестра - с**ё**стры*

Морфологический анализ
Morphological Analysis

- **Глубинный** морфологический анализ – общие характеристики слов - Deep morphological analysis: lexeme features

*ид**ущ**ий – ИДТИ, прич, несов, непрош, ед, муж, им/вин,неод*

*брат**ья** - БРАТ, мн, им*

отцы - ОТЕЦ, мн, им

*мат**ери** – МАТЬ, мн, им*

Глубинный морфологический анализ
Deep Morphological Analysis

- Для задач автоматического анализа текста, выходящего за пределы морфологии, поверхностная морфология, как правило, не нужна. Поэтому анализ практически всегда ориентируется на глубинную морфологию.
  For automatic analysis of text going beyond morphology we hardly need surface morphology. Therefore, the analysis is almost always oriented at deep morphology.

Глубинный морфологический анализ
Deep Morphological Analysis

- Вход – предложение в обычной орфографической записи
  Input – a sentence in a conventional orthographic form
- Выход – морфологическая структура предложения
  Output – a morphological structure of a sentence

Глубинный морфологический анализ
Deep Morphological Analysis

- Вход – предложение в обычной орфографической записи
- Input – a sentence in a conventional orthographic form

Все эти типы стали есть в литейном цехе

Глубинный морфологический анализ
Deep Morphological Analysis

Выход – морфологическая структура предложения

Output – a morphological structure of a sentence

1.1	ВСЕ1	S,ЕД,СРЕД,ИМ,НЕОД,ALTJO
1.2	ВСЕ1	S,ЕД,СРЕД,ВИН,НЕОД,ALTJO
1.3	ВСЕ3	S,МН,ИМ,ОД
1.4	ВЕСЬ	A,ЕД,СРЕД,ИМ,ALTJO
1.5	ВЕСЬ	A,ЕД,СРЕД,ВИН,ALTJO
1.6	ВЕСЬ	A,МН,ИМ
1.7	ВЕСЬ	A,МН,ВИН,НЕОД
1.8	ВСЕ2	PART,ALTJO
2.1	ЭТОТ	A,ИМ,МН,CAPIT,CAP
2.2	ЭТОТ	A,ВИН,МН,НЕОД,CAPIT,CAP
3.1	ТИП1	S,ИМ,МН,МУЖ,НЕОД
3.2	ТИП1	S,ВИН,МН,МУЖ,НЕОД
3.3	ТИП2	S,ИМ,МН,МУЖ,ОД
4.1	СТАТЬ1	V,ПРОШ,МН,ИЗЪЯВ,СОВ
4.2	СТАНОВИТЬСЯ1	V,ПРОШ,МН,ИЗЪЯВ,СОВ
4.3	СТАНОВИТЬСЯ2	V,ПРОШ,МН,ИЗЪЯВ,СОВ
4.4	СТАЛЬ	S,РОД,ЕД,ЖЕН,НЕОД
4.5	СТАЛЬ	S,ДАТ,ЕД,ЖЕН,НЕОД
4.6	СТАЛЬ	S,ПР,ЕД,ЖЕН,НЕОД
4.7	СТАЛЬ	S,ИМ,МН,ЖЕН,НЕОД
4.8	СТАЛЬ	S,ВИН,МН,ЖЕН,НЕОД

5.1	ЕСТЬ1	V,ИНФ,НЕСОВ
5.2	БЫТЬ	V,НАСТ,ЕД,ИЗЪЯВ,1-Л,НЕСОВ
5.3	БЫТЬ	V,НАСТ,ЕД,ИЗЪЯВ,2-Л,НЕСОВ
5.4	БЫТЬ	V,НАСТ,ЕД,ИЗЪЯВ,3-Л,НЕСОВ
5.5	БЫТЬ	V,НАСТ,МН,ИЗЪЯВ,1-Л,НЕСОВ
5.6	БЫТЬ	V,НАСТ,МН,ИЗЪЯВ,2-Л,НЕСОВ
5.7	БЫТЬ	V,НАСТ,МН,ИЗЪЯВ,3-Л,НЕСОВ
5.8	ЕСТЬ2	INTJ
6.1	В1	PR
6.2	В2	PR
6.3	В3	PR
6.4	В (ФИКТ-КОМПОЗИТ)	COM,CAP-MIX,STRICT_ABBR
7.1	ЛИТЕЙНЫЙ	A,ПР,ЕД,МУЖ
7.2	ЛИТЕЙНЫЙ	A,ПР,ЕД,СРЕД
8.1	ЦЕХ1	S,ПР,ЕД,МУЖ,НЕОД
8.2	ЦЕХ2	S,ПР,ЕД,МУЖ,НЕОД

Морфологическая структура предложения
Morphological Structure of a Sentence

- МС предложения – последовательность МС всех входящих в него слов
  MorphS of a sentence is a sequence of MorphS's of all words belonging to the sentence.
- МС слова – совокупность МС всех омонимов данного слова
  MorphS of a word is the set of all MorphS of all homonyms of this word
- МС омонима – имя лексемы (лемма) плюс часть речи плюс набор словоизменительных морфологических характеристик
  MorphS of a homonym is the lexeme name (lemma) plus part of speech plus a set of all inflectional features

Морфологическая структура предложения
Morphological Structure of a Sentence

MorphS of a word is the set of all MorphS of all homonyms of this word

MorphS of a homonym is the lexeme name (lemma) plus part of speech plus a set of all inflectional features (each of lines 4.1-4.8

4.1	***СТАТЬ1***	**V,ПРОШ,МН,ИЗЪЯВ,СОВ**
4.2	***СТАНОВИТЬСЯ1***	**V,ПРОШ,МН,ИЗЪЯВ,СОВ**
4.3	***СТАНОВИТЬСЯ2***	**V,ПРОШ,МН,ИЗЪЯВ,СОВ**
4.4	***СТАЛЬ***	**S,РОД,ЕД,ЖЕН,НЕОД**
4.5	***СТАЛЬ***	**S,ДАТ,ЕД,ЖЕН,НЕОД**
4.6	***СТАЛЬ***	**S,ПР,ЕД,ЖЕН,НЕОД**
4.7	***СТАЛЬ***	**S,ИМ,МН,ЖЕН,НЕОД**
4.8	***СТАЛЬ***	**S,ВИН,МН,ЖЕН,НЕОД**

Морфологическая структура предложения
Morphological Structure of a Sentence

- Морфологические характеристики - это значения (values) морфологических категорий

  Morphological features are values of morphological categories

Морфологические категории
Morphological Categories

Морфологические категории разные у разных частей речи

Different parts of speech have different morphological categories

Морфологические категории
Morphological Categories

- Словоизменительные морфологические категории – Inflectional categories

(например, число и падеж русского существительного)

- Классифицирующие морфологические категории – Classifying categories

(род и одушевленность русского существительного – в русском языке других таких нет)

Части речи в русском языке
Parts of Speech in Russian

Существительное	S
Прилагательное	A
Числительное	Num
Глагол	V
Наречие	Adv
Союз	Conj
Предлог	Pr
Частица	Part
Междометие	Intj

Части речи в английском языке
Parts of Speech in English

Noun	S
Adjective	A
Numeral	Num
Verb	V
Adverb	Adv
Conjunction	Conj
Preposition	Pr
Particle	Part
Interjection	Intj

Article	**Art**

Части речи – Parts of Speech

Местоимения – не особая часть речи. Pronouns form no specific part of speech.

Они распадаются на – They are classed into
местоименные существительные – pronominal nouns (*я, ты, он, что, кто* и др., *I, he, who, mine, yours* etc);

- местоименные прилагательные – pronominal adjectives (*мой, твой, свой, чей, каковой* и др., *my, your, whose* etc);
- местоименные наречия - pronominal adverbs (*где, там, тут, откуда, оттуда, почему* и др., *where, there, whence, why* etc).

Morphological Features in English

Cases of Nouns

Main (= common) case comm
Possessive case poss

Morphological Features in English

Cases of Personal Pronouns

Main (=nominative) case nom
Objective case obj

Morphological Features in English

Number of Nouns and Verbs

Singular Number	sg
Plural Number	pl

Morphological Features in English

Degrees of Comparison of Adjectives and Adverbs

Positive	posit
Comparative	comp
Superlative	sup

Morphological Features in English

Representation of Verbs

Main Form	mf
Active participle	ing
Passive participle	pp

Morphological Features in English

Tense of Verbs

Nonpast	prs
Past	pst

Morphological Features in English

Person of Verbs

First	FP
Second	SP
Third	TP

(Словоизменительные) морфологические категории в русском языке
(Inflectional) Morphological Categories in Russian

- Существительные: число и падеж
  - nouns: number and case
- Прилагательные: число, падеж, род, одушевленность, краткость, степени сравнения
  - adjectives: number, case, gender, animacy, brevity, degrees of comparison
- Числительные: падеж, род, одушевленность, число
  - numerals: number, case, gender, animacy

Морфологические категории в русском языке
Morphological Categories in Russian

- Глаголы: репрезентация, наклонение, время, вид, залог, лицо, число, род, падеж, одушевленность, краткость
  - Verbs: representation, mood, tense, aspect, voice, person, number, gender, case, animacy, brevity (the latter three are only relevant for participles)

Морфологические категории в русском языке
Morphological Categories in Russian

- Наречия: степени сравнения
  - Adverbs: degrees of comparison
- Союзы: нет категорий
  - Conjunctions: no categories
- Предлоги: нет категорий
  - Prepositions: no categories
- Частицы: нет категорий
  - Particles: no categories
- Междометия: нет категорий
  - Interjections: no categories

Морфологические категории в русском языке
Morphological Categories in Russian

- Дополнительная категория:

смягчение сравнительной степени прилагательных и наречий

Additional category:

Attenuation of the comparative degree of adjectives and adverbs

Морфологические характеристики в русском языке
Morphological Features of Russian

Падеж существительного – Case of Noun:

Именительный - nominative
Родительный - genitive
Партитивный (2-й родительный) - partitive
Дательный - dative
Винительный - accusative
Творительный - instrumental
Предложный - prepositional
Местный (2-й предложный) - locative
Звательный - vocative
Счетная форма – count form

Морфологические характеристики в русском языке
Morphological Features of Russian

Падеж прилагательного, числительного, причастия – Case of Adjective, Numeral, Participle:

Именительный - Nominative
Родительный - Genitive
Дательный - Dative
Винительный - Accusative
Творительный - Instrumental
Предложный - Prepositional

Морфологические характеристики в русском языке
Morphological Features of Russian

Род прилагательного, числительного, причастия – Gender of Adjective, Numeral, Participle:

Мужской - masculine
Женский - feminine
Средний - neuter

Морфологические характеристики в русском языке
Morphological Features of Russian

Число - Number:

Единственное - singular
Множественное - plural

Морфологические характеристики в русском языке
Morphological Features of Russian

Одушевленность прилагательного, числительного, причастия
– Animacy of adjective, numeral, participle

Одуш - Anim
Неод - Inanim

Морфологические характеристики в русском языке
Morphological Features of Russian

Степени сравнения прилагательного – Degrees of comparison of adjective

Положительная - positive
Сравнительная - comparative
Превосходная - superlative

Морфологические характеристики в русском языке
Morphological Features of Russian

Степени сравнения наречия - Degrees of Comparison of Adverb

положительная - positive

сравнительная - comparative

превосходная – superlative:

только потенциальные и устаревшие формы – *покорнейше прошу, нижайше вам кланяюсь, презабавнейше, тщательнейше, деликатнейше*

Морфологические характеристики в русском языке
Morphological Features of Russian

Краткость прилагательного и причастия - Brevity

полное - full

краткое - short

усеченное – truncated:

красна девица, средь бела дня, на босу ногу, лиха беда начало

Морфологические характеристики в русском языке
Morphological Features of Russian

Репрезентация глагола – Representation of verb

Личная форма - finite
Инфинитив - infinitive
Причастие - participle
Деепричастие - gerund

Морфологические характеристики в русском языке
Morphological Features of Russian

Наклонение глагола - Mood

Изъявительное - indicative

Повелительное - imperative

Сослагательного нет – оно только аналитическое

No conjunctive mood as it only appears as analytical

Морфологические характеристики в русском языке
Morphological Features of Russian

Время глагола - Tense of Verb

Непрошедшее - nonpast
Прошедшее - past
Настоящее (для глагола *быть*) – present (for the verb *быть* only)

Морфологические характеристики в русском языке
Morphological Features of Russian

Вид глагола – Aspect of Verb

Несовершенный - imperfective
Совершенный - perfective

Морфологические характеристики в русском языке
Morphological Features of Russian

Залог глагола – Voice of Verb

Действительный - active
Страдательный - passive

Морфологические характеристики в русском языке
Morphological Features of Russian

Лицо глагола – Person of Verb

Первое - first
Второе - second
Третье - third

Аналитические формы слов
Analytical forms of words

Будущее время: *Буду работать*

1.1 БЫТЬ V,НЕПРОШ,ЕД,ИЗЪЯВ,1-Л, НЕСОВ

2.1 РАБОТАТЬ V,ИНФ,НЕСОВ

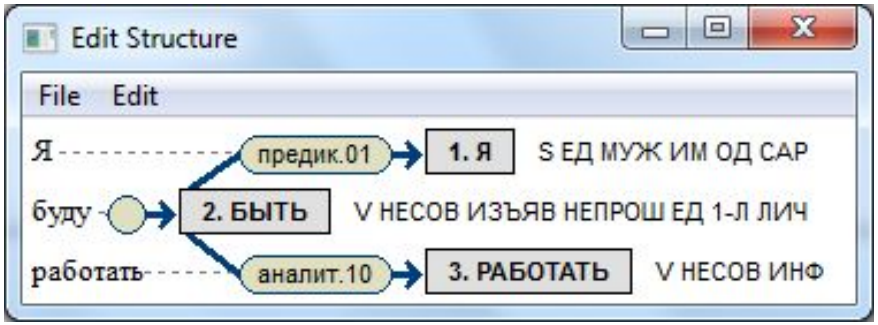
Страдательный залог: *Был отвергнут*

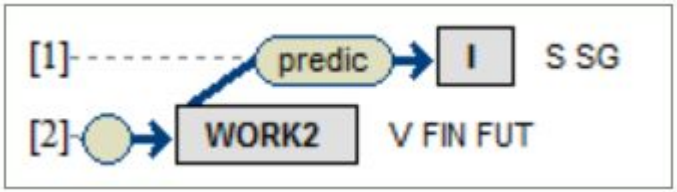
1.1 БЫТЬ V, ПРОШ, ЕД, ЛИЧ, ИЗЪЯВ, МУЖ, НЕСОВ

2.1 ОТВЕРГАТЬ V, НЕПРОШ, МН, ЛИЧ, ИЗЪЯВ, 3-Л, СОВ

2.2 ОТВЕРГАТЬ V, ПРОШ, ЕД, ПРИЧ, КР, МУЖ, СОВ, СТРАД

Аналитические формы слов – дальнейшая судьба
Analytical forms of words – Future Fate



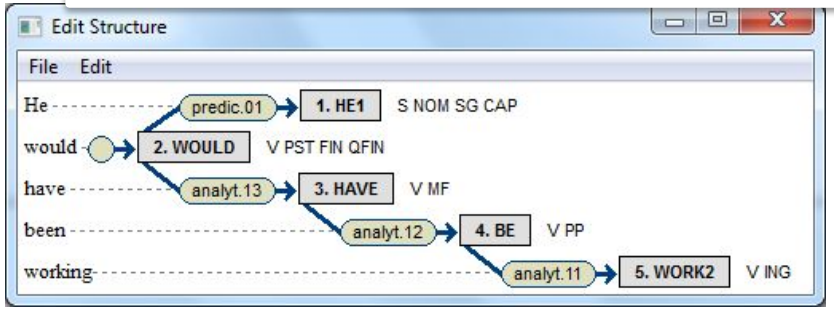


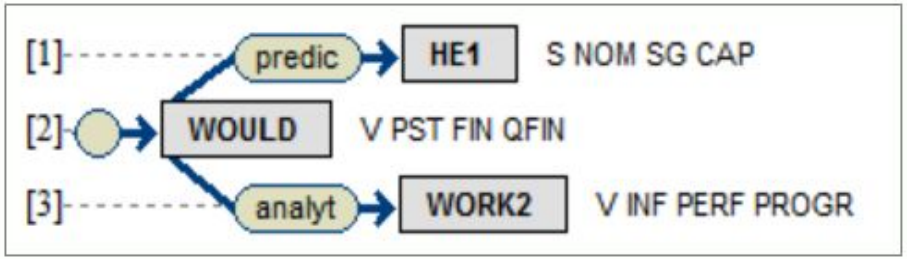
Аналитические формы слов
Analytical forms of words

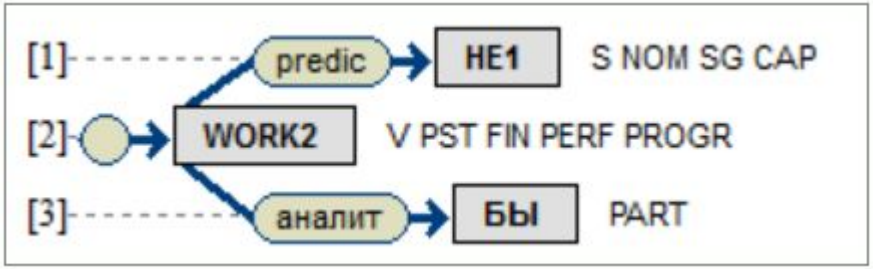
Future in the Past Perfect Continuous:
Would have been working

1.1 WOULD V,PST
2.1 HAVE V,MF
3.1 BE V,PP
4.1 WORK V,ING

Аналитические формы слов – дальнейшая судьба
Analytical forms of words – Future Fate







Элементы морфологического анализатора
Elements of Morphological Analyzer

- Морфологический словарь
  – Morphological dictionary
- Стандартные морфологические объекты
  – Standard morphological objects
- Алгоритм морфологического анализа
  – Algorithm of Analysis
- Процедуры и программы
  – Procedures and programs

Морфологический словарь – Morphological dictionary

ENTRY:КРУЖКА acct:a

осн:кру`ж(е)к чер:3 т:4 \\кру`жка ж 3*a

trs:mug

ENTRY:КРУЖОК acct:b

осн:круж(о`)к чер:2 т:9 \\в

trs:circle

Морфологический словарь – Morphological dictionary

ENTRY:КРУГЛЫЙ acct:adj_ac1

осн:кругл т:211 ф:51 \\кру`глый п 1а/с'

trs:round

ENTRY:КРУГЛО acct:adv_c

осн:кругл ф:38 ф:15 ф:52

trs:roundly

Морфологический словарь – Morphological dictionary

ENTRY:КРАСИВЫЙ acct:adj_a

осн:краси`в т:211 ф:51 осн:кра`ше хар:А, срав осн:краси`вше хар:А,срав

trs:beautiful

ENTRY:КРАСИВО acct:adv_a

осн:краси`в ф:15 ф:38 ф:52 осн:кра`ше хар:ADV,срав осн:краси`вше хар:ADV,срав

trs:beautifully

Морфологический словарь – Morphological dictionary

ENTRY:КРУЖИТЬ1 [НЕСОВ!] acct:vn_ct
  хар:V,осн:круж т:353 т:411
|| acct:vn_bt*2 хар:V,осн:круж т:353 т:411 \\c
  trs:whirl_about

ENTRY:КРУЖИТЬ2 [НЕСОВ!] acct:vn_ct
  хар:V,осн:круж нет:страд т:353 т:411
|| acct:vn_bt*2 хар:V,осн:круж нет:страд т:353 т:411
  trs:circle

Стандартные морфологические объекты – Standard Morphological Objects

- Списки окончаний – Lists of Endings
- Форматы - Formats
- Трафареты - Templates
- Маски - Masks
- Чередования - Alternations

Списки окончаний – Lists of Endings

ок:001 [вода]
'а'ед,им,'ы'ед,род,'е'ед,дат,'у'ед,вин,'ой'ед,твор, * 'е'ед, пр,'ы'мн,им,'#'мн,род,'ам'мн,дат, * 'ы'мн,вин,'ами'мн, твор,'ах'мн,пр,'ою'ед,твор,'о'сл

ок:011 [здание]
'е'ед,им,'я'ед,род,'ю'ед,дат,'е'ед,вин,'ем'ед,твор,'и'ед,пр, * 'я'мн,им,'й'мн,род,'ям'мн,дат,'я'мн,вин,'ями'мн,твор, * 'ях'мн, пр,'е'сл

Форматы - Formats

ф:001

хар:S,муж,неод

ф:002

хар:S,жен,неод

Трафареты и маски – Templates and Masks

т:001 [вода]

ф:2,ок:1

т:101 [жена]

хар:S,од,жен,ок:1/24,ок:'#'мн,вин

Чередования - Alternations

чер:001 [стрелок]

ед,им

стрелок, стрелка,стрелку,…,стрелки,…

чер:002 [лесок]

ед,им/ед,вин

лесок, леска, леску, лесок, …,лески,…

чер:003 [сосен]

мн,род

сосна, сосны, сосне,…, сосны, сосен, соснам…

Алгоритм морфологического анализа – Algorithm of morphological analysis

- Морфологические позиции (обойма) - ordered list positions
- Русский язык: 6 позиций

префикс основа тема суффикс окончание частица

Prefix	Base	Theme Suffix		Ending	Particle
1	2	3	4	5	6

писать написанный поинтереснее рассматривающийся

- Просмотр слева направо или справа налево – Scanning from left to right or from right to left
- **Конечный автомат! – finite-state automaton**

Алгоритм морфологического анализа – Algorithm of morphological analysis

Обработка сложных слов - Processing of Composita

нефтепереработка = нефте (=НЕФТЬ,сл) +переработка

пятитомный = пяти (=ПЯТЬ,сл) + томный2

полкруга = пол (=ПОЛ3,сл) + круг

минсвязи = мин (=МИНИСТЕРСТВО,им,сл-верш) + связь

бизнес-проект = бизнес (БИЗНЕС,им,ед) + проект

наркоманка = нарко (НАРКОТИЧЕСКИЙ,сл) + манка

(чтобы избежать этого, надо ввести слово в словарь: in order to avoid such parse one needs to add the word to the dictionary)

Алгоритм морфологического анализа – Algorithm of morphological analysis

Обработка неопознанных слов

Processing of Unidentified Words

- Технические приемы - Technicalities
- Guesser