
абвгдеёжзийклмнопрстуфхцчшщъыьэюя
Slavic Grammar Matrix
DELPH-IN

Developing a Russian HPSG based on the Russian National Corpus

DELPH-IN Summit 2009, Barcelona

The SlaviGraM Project

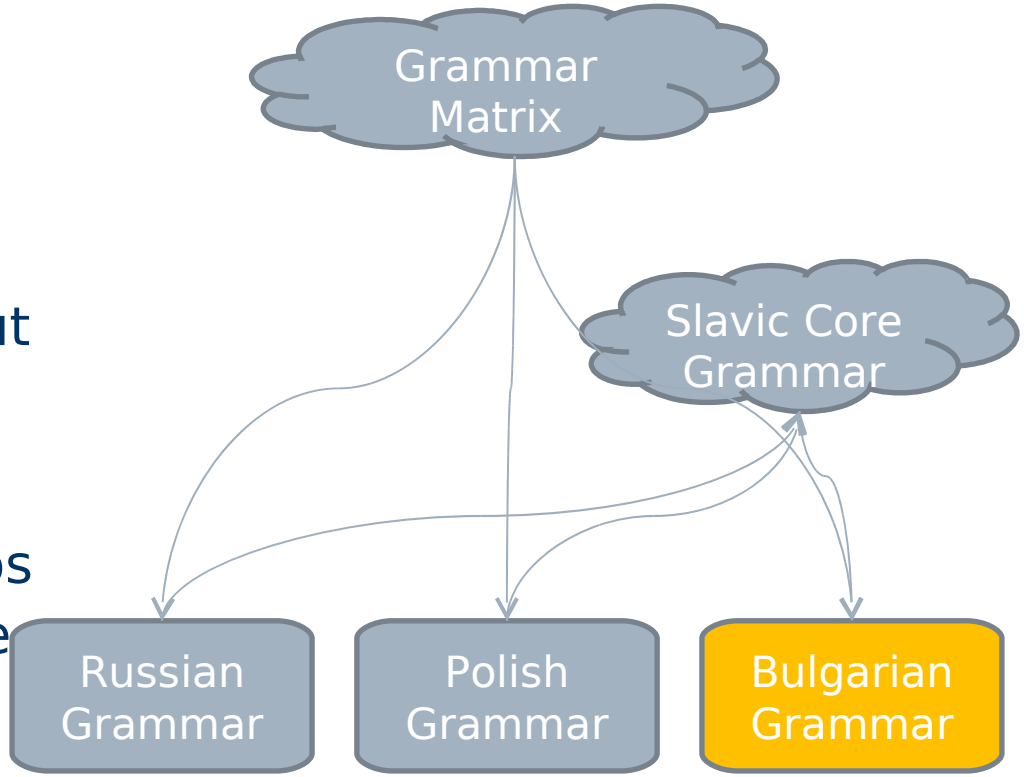
- Proof-of-Concept Implementation
  - strategies for constructing a cross-linguistic resource based on concepts of shared and non-shared grammar in modelling Slavic morphosyntax
- Focus on Russian Resource Grammar
  - construction of Russian HPSG in the DELPH-IN environment combining the Grammar Matrix and a Slavic Core Grammar with corpus-based grammar elaboration; exploiting the Russian National Corpus (RNC) in Russian Resource Grammar (RRG) engineering.
- Cross-Slavic Validation (two showcases)
  - construction of Bulgarian HPSG in the DELPH-IN environment combining the Grammar Matrix and a Slavic Core Grammar with corpus-based grammar elaboration exploiting the BulTreeBank
  - construction of Polish HPSG in the DELPH-IN environment combining the Grammar Matrix and a Slavic Core Grammar with corpus-based grammar elaboration exploiting the IPPI PAN Corpus

Idea

- construction of Russian HPSG
- in the DELPH-IN environment
- combining the Grammar Matrix
- and a Slavic Core Grammar
- with corpus-based grammar elaboration
  - exploiting the Russian National Corpus (RNC) as structured grammatical knowledge resource

Top-down vs. Bottom-up

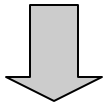
- Use Grammar Matrix to quickly build small grammars for individual languages
- Shared analyses from individual languages are put into the Slavic Core
- When new language is added, the Slavic Core helps to more efficiently build the new grammar, and potentially receives cross-slavic validation



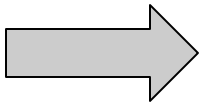
Integrating External Morphological Analyzers

- Mystem (http://company.yandex.ru/technology/mystem/)
- Perl-based configurable SPPP wrapper

```
профессор{профессор=S,муж,од=им,ед}
читает{читать=V,несов=непрош,ед,изъяв,3-л}
книгу{книга=S,жен,неод=вин,ед}
```



```
S+им := noun-nom-irule.
S+вин := noun-acc-irule.
S+ед := noun-sg-irule.
V+непрош := verb-nonpast-irule.
V+изъяв := verb-indic-irule.
V+3-л := verb-3p-irule.
```



```
<?xml version="1.0" encoding="utf-8"?>
<segment>
 <token form="профессор" from="0" to="9">
  <analysis stem="профессор">
   <rule id="noun-nom-irule" form="профессор"/>
   <rule id="noun-sg-irule" form="профессор"/>
  </analysis>
 </token>
 <token form="читает" from="10" to="16">
  <analysis stem="читать">
   <rule id="verb-nonpast-irule" form="читает"/>
   <rule id="verb-indic-irule" form="читает"/>
   <rule id="verb-3p-irule" form="читает"/>
  </analysis>
 </token>
 <token form="книгу" from="17" to="22">
  <analysis stem="книга">
   <rule id="noun-acc-irule" form="книгу"/>
   <rule id="noun-sg-irule" form="книгу"/>
  </analysis>
 </token>
</segment>
```

Morphological information in RNC

Part of speech
- ☐ noun
- ☐ adjective
- ☐ numeral
- ☐ numeral adjective
- ☐ verb
- ☐ adverb
- ☐ predicative
- ☐ parenthesis
- ☐ pronoun
- ☐ adjective pronoun
- ☐ predicative pronoun
- ☐ adverbial pronoun
- ☐ preposition
- ☐ conjunction
- ☐ particle
- ☐ interjection

Antroponymic
- ☐ family name
- ☐ first name
- ☐ patronymic

Case
- ☐ nominative
- ☐ vocative*
- ☐ genitive
- ☐ genitive 2
- ☐ dative
- ☐ accusative
- ☐ accusative 2*
- ☐ instrumental
- ☐ locative
- ☐ locative 2
- ☐ adnumerative

Number
- ☐ singular
- ☐ plural

Gender
- ☐ masculine
- ☐ feminine
- ☐ neuter
- ☐ common*

Animacy
- ☐ animate
- ☐ inanimate

Mood / Verb form
- ☐ indicative
- ☐ imperative
- ☐ imperative 2
- ☐ infinitive
- ☐ participle
- ☐ gerund

Tense
- ☐ present
- ☐ future
- ☐ past

Person
- ☐ first
- ☐ second
- ☐ third

Voice
- ☐ active
- ☐ passive
- ☐ middle

Aspect
- ☐ perfective
- ☐ imperfective

Degree / Adj. form
- ☐ comparative
- ☐ comparative 2*
- ☐ superlative
- ☐ full form
- ☐ short form

Transitivity
- ☐ transitive*
- ☐ intransitive*

Other features
- ☐ dictionary form
- ☐ numeral recording
- ☐ anomalous form*
- ☐ distorted form*
- ☐ non-dictionary form**
- ☐ initials*
- ☐ abbreviation*
- ☐ indeclinable*

OK | Clear | Cancel

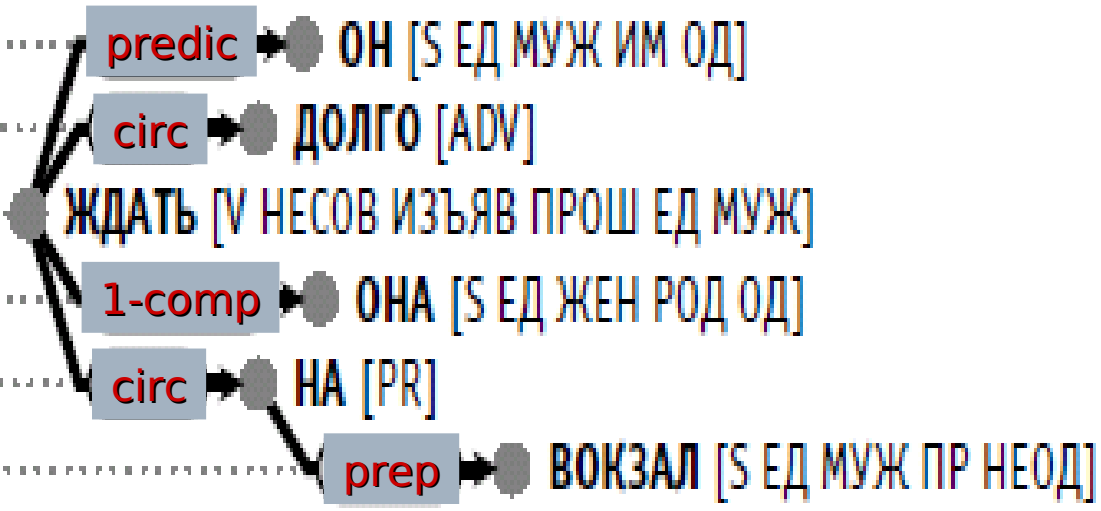
* - only in the corpus with resolved homonymy
** - only in the corpus with unresolved homonymy

Making the most of the syntactic Information in RNC

- In the deeply annotated sub-corpus of the RNC, every sentence is marked up with a dependency syntactic structure, with nodes corresponding to the words of the sentence, and labeled edges encoding the syntactic relations.
- The syntactic formalism originates in the Meaning-Text Theory (Mel'cuk 1995), but the inventory of syntactic relations has been extended for the purposes of corpus annotation, incorporating a number of specific linguistic decisions (Boguslavsky et al. 2002; Apresjan et al. 2006).
- Unlike the morphologically annotated portion of the RNC, the deeply annotated sub-corpus only contains fully disambiguated annotations (i.e. both morphological and syntactic ambiguity is resolved).

A sample structure

- ‘he’ *Он*
- ‘long’ *долго*
- ‘waited’ *ждал*
- ‘her’ *ее*
- ‘at’ *на*
- ‘station’ *вокзале*



Grammatical features used in the syntactic sub-corpus

Part of speech ☐ nominal ☐ adjective ☐ numeral ☐ verb ☐ adverb ☐ preposition ☐ conjunction ☐ particle ☐ interjection ☐ compound word ☐ word-sentence ☐ foreign word, non-lexical formula	**Case** ☐ nominative ☐ genitive ☐ partitive ☐ dative ☐ accusative ☐ instrumental ☐ prepositive ☐ locative ☐ vocative	**Aspect** ☐ perfective ☐ imperfective **Tense** ☐ present ☐ non-past ☐ past
	Grade ☐ comparative ☐ comparative 2 ☐ superlative	**Person** ☐ first ☐ second ☐ third
Animacy ☐ animate ☐ inanimate	**Form** ☐ short form	
Gender ☐ masculine ☐ feminine ☐ neuter	**Representation** ☐ finite verb ☐ infinitive ☐ participle ☐ gerund	**Voice** ☐ passive
Number ☐ singular ☐ plural	**Mood** ☐ indicative ☐ imperative	**Other** ☐ part of a compound word

Access to the syntactic sub-corpus

Main corpus | Syntactic corpus | Spoken corpus

русская версия

Search by exact form

Word or phrase

search clear

Lexico-grammatical search

Word

Gramm. features select

Distance from parent: from to

Syntactic relationship select

Word

Gramm. features select

search clear

Russian National Corpus
© 2003–2008

Search provided by Яndex.Server

Syntactic relations in RNC

Actantial relationships

- ☐ predicative

- ☐ dative subjective
- ☐ agentive
- ☐ quasi-agentive
- ☐ non-intrinsic agentive

- ☐ I completive
- ☐ II completive
- ☐ III completive
- ☐ IV completive
- ☐ V completive

- ☐ copula

- ☐ I non-intrinsic completive
- ☐ II non-intrinsic completive
- ☐ III non-intrinsic completive

- ☐ non-actantial completive
- ☐ completive appositive

- ☐ prepositional
- ☐ subordinating conjunctional
- ☐ comparative
- ☐ comparative conjunctional
- ☐ elective

Attributive

determinative
- ☐ (proper) determinative
- ☐ descriptive determinative
- ☐ approximative ordinal
- ☐ relative

General attributive
- ☐ (proper) attributive
- ☐ compound

appositive
- ☐ (proper) appositive
- ☐ dangling appositive
- ☐ nominative appositive
- ☐ numerative appositive

quantitative
- ☐ (proper) quantitative
- ☐ approximative quantitative
- ☐ approximative co-predicative
- ☐ approximative delimitative
- ☐ distributive
- ☐ additive

circumstantial
- ☐ (proper) circumstantial
- ☐ durative
- ☐ multiple durative
- ☐ distantional
- ☐ circumstantial tautological
- ☐ subjective circumstantial
- ☐ objective circumstantial
- ☐ subjective co-predicative
- ☐ objective co-predicative
- ☐ delimitative
- ☐ parenthetic
- ☐ complement clause
- ☐ expository
- ☐ adjunctive
- ☐ precising

Coordinative

- ☐ coordinative
- ☐ sentential coordinative
- ☐ conjunctional coordinative
- ☐ communicative coordinative
- ☐ multiple

Syncategorematic

- ☐ analytical
- ☐ passive analytical
- ☐ auxiliary
- ☐ quantitative auxilary
- ☐ correlative
- ☐ expletive

- ☐ proleptic
- ☐ elliptic

Mapping the components of a RNC dependency relation to HPSG categories

- Observe the following straightforward convention:
  - in a given syntactic dependency relation,
  - the “governor X” corresponds to the lexical head of the head daughter, while
  - the “dependent Y” corresponds to the lexical head of the non-head daughter.

E.g. predicative syntactic relation predicate(X) – first argument(Y)

Head–Subject Schema (trivial cases)

Профессор[Y] читает[X].
professor.$_{\text{NOMINATIVE}}$ read.$_{\text{FINITE.ACTIVE}}$

‘The professor reads’

Заявка[Y] изучается[X].
proposal.$_{\text{NOMINATIVE}}$ study.$_{\text{FINITE.REFLEXIVE.PASSIVE}}$

‘The proposal is being studied.’

vs.

Комитет[Y] изучает[X] заявку
committee.$_{\text{NOMINATIVE}}$ study.$_{\text{FINITE.ACTIVE}}$ proposal.$_{\text{ACCUSATIVE}}$

‘The committee studies the proposal.’

E.g. predicative syntactic relation predicate(X) – first argument(Y)

Head–Subject Schema (non-verbal predication)

- noun(X): Москва[Y] — столица[X] России.
  Moscow.$_{\text{NOMINATIVE}}$ capital.$_{\text{NOMINATIVE}}$ Russia.$_{\text{GENITIVE}}$
  ‘Moscow is the capital of Russia.’

- adjective(X): Профессор[Y] какой-то странный[X].
  professor.$_{\text{NOMINATIVE}}$ somewhat strange.$_{\text{FULL-ADJECTIVE.NOMINATIVE}}$
  ‘The professor is kind of strange.’

  Профессор[Y] очень добр[X].
  professor.$_{\text{NOMINATIVE}}$ very kind.$_{\text{SHORT-ADJECTIVE}}$
  ‘The professor is very kind.’

  Профессор[Y] должен[X] уходить.
  professor.$_{\text{NOMINATIVE}}$ obliged.$_{\text{PREDICATIVE-ADJECTIVE}}$ leave.$_{\text{INFINITIVE}}$
  ‘The professor must leave.’

1.1. predicative syntactic relation predicate(X) – first argument(Y)

Head–Subject Schema (non-nominative subject)

- Y is a noun in genitive or partitive (also known as “second genitive”)

  Хлеба[Y] не осталось[X]
  bread.$_{\text{GENITIVE}}$ NEG remain.$_{\text{IMPERSONAL.REFLEXIVE.PAST}}$

  ‘No bread left.’

  Сахару[Y] хватит[X] на всех
  sugar.$_{\text{PARTITIVE}}$ suffice.$_{\text{IMPERSONAL.NON-PAST}}$ on all

  ‘There will be enough sugar for everybody.’

- Y is a distributive or approximative PP

  Пришло[X] до[Y] десяти детей.
  come.$_{\text{IMPERSONAL.PAST}}$ to ten.$_{\text{GENITIVE}}$ children

  ‘Up to ten children came.’

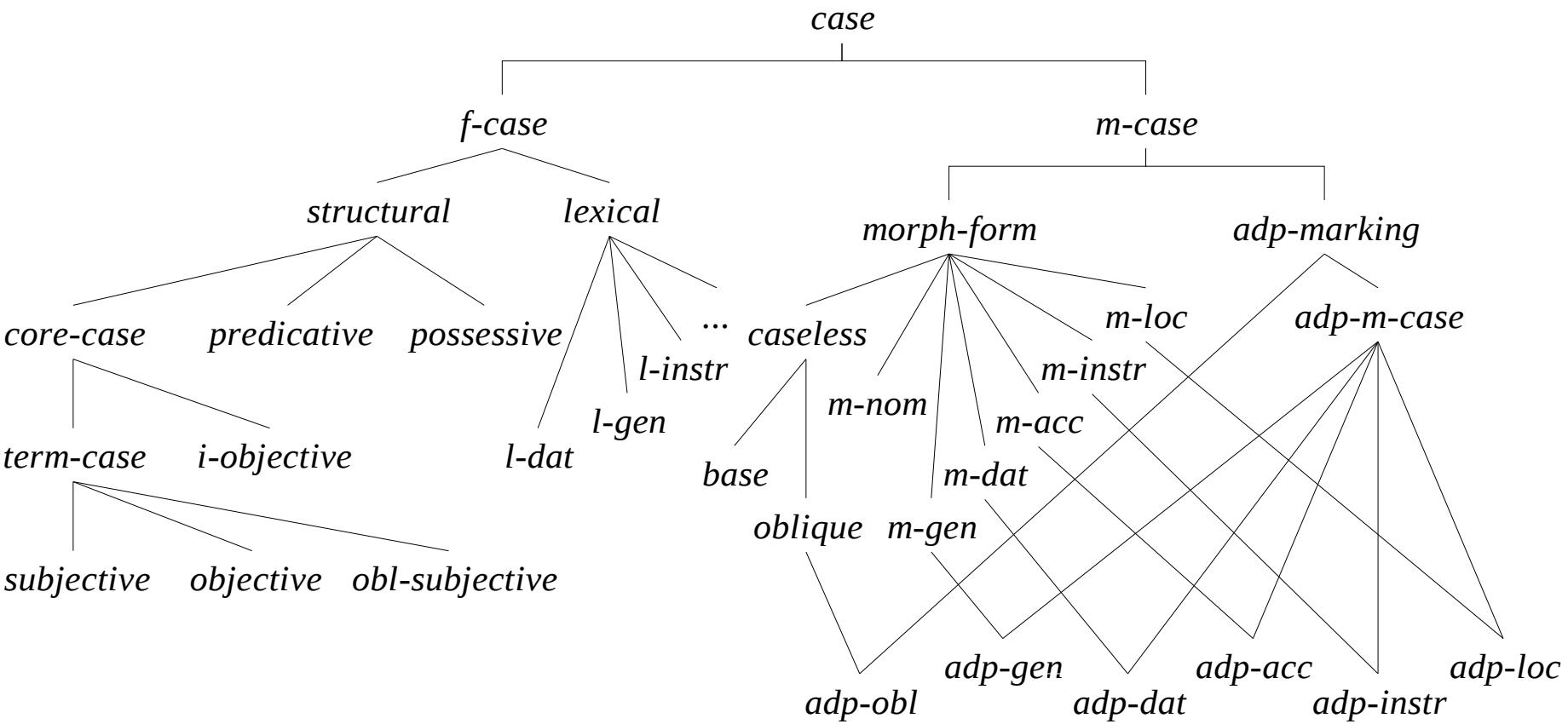
  Нам досталось[X] по [Y] груше.
  we.$_{\text{DATIVE}}$ recieve.$_{\text{IMPERSONAL.REFLEXIVE.PAST}}$ of pear.$_{\text{DATIVE}}$

  ‘We’ve got a pear each.’

More Than Just a Russian Grammar
E.g. Shared Slavic Case Hierarchy

functional dimension

marking dimension



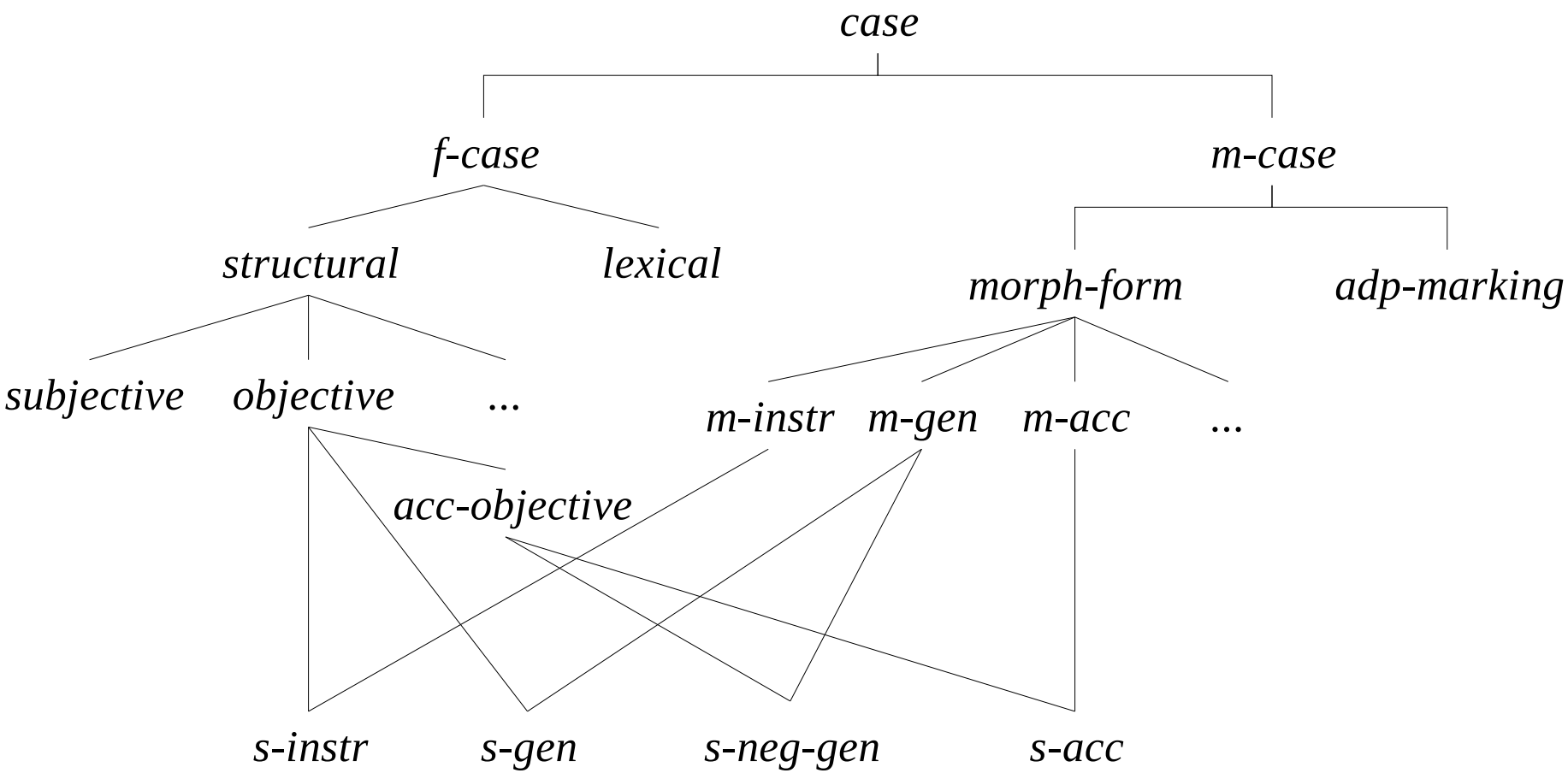
Parametric constraints on the instantiation of functional case

- Language-specific: the particular encoding of functional case, e.g., the instrumental / prepositional marking of the agentive phrase in passive.

- Idiosyncratic: lexically determined subclasses of structural arguments, e.g. accusative, genitive and instrumental direct objects in Polish.

- Context-sensitive: case alternations triggered by syntactic or semantic context, e.g., genitive of negation.

Structural case in Polish: the case of direct object NPs



Future Work

- Extend the Russian grammar
- Extend the Slavic Core Grammar
- Start grammars for Polish, Bulgarian, Czech, … …