Morphosyntactic annotation of the Georgian National Corpus – The diacronic dimension

Paul Meurer

Uni Research, Norway

Gudauri, September 23, 2013

Outline

Outline

The Georgian National Corpus (GNC) project

- Aim: building a balanced, richly annotated diacronic corpus of the Georgian language, accessible via the Internet (Web interface; web services)
- Time frame: five years, started late 2012
- Project partners: Uni Frankfurt (Project leader: Jost Gippert); four Georgian universities; Bergen
- Funding: The Volkswagen Foundation (Germany), additional funding by Rustaveli Foundation (Georgia)
- Hosting: In Norway (Bergen University Clarin Center), with mirrors in Tbilisi (National Library) and Frankfurt
- Corpus tool: Corpuscle (developed at Uni Research, Bergen)

The GNC: Content

- Old and Middle Georgian texts
  - text-critical editions with facs, dipl and norm levels of annotation
  - building on work done in Frankfurt in the Titus/Armazi project
  - recoded in XML according to TEI P5 guidelines
  - plus newly transcribed or digitized texts
- Modern Georgian texts (19. century – present)
  - literary texts from lib.ge and other sources, but proof-read anew
  - non-fictional texts: digitized modern newspapers, RFE/Radio Liberty etc.
  - 19th century newspapers
- Dialectal and spoken language texts
  - Georgian Dialect Corpus (Marina Beridze et al.)
  - SSGG – The Sociolinguistic Situation of Present-day Georgia
- … and more to come

Outline

The GNC: Text Annotation

- Annotation format: TEI-P5 XML
- Header: Detailed information (date, place, authorship, language variety etc.) about the original, the scholarly edition(s) the digital version is based on, and the digital version itself
- Levels of text annotation:
  - for digitized manuscripts: facsimile, diplomatic, normalized; critical apparatus and cross references
  - for all texts: structural coding (headings, paragraphs, lines of verse etc.)
- Grammatical annotation:
  - lemma, morphosyntactic features, verb frames; names/named entities
- All levels of annotation will be searchable

The GNC: Images, audio and video

Integration of images and audio/video

- for digitized manuscripts: scanned images of manuscripts
- for transcribed speech: if available audio or video

Text Annotation: Challenges

Transcribed manuscripts: There are line breaks inside words

- line breaks are kept in facsimile level
- words are units in diplomatic/normalized levels
- after tokenization, words should be coded as <w> elements across levels
  This is needed both for corpus building and morphosyntactic annotation
- Solution: use TEI <reg> (regularization) and <orig> (original form) elements

Text Annotation: Challenges

```
<choice>
   <gnc:facs> ⵇⵍⴱⵀⵊ, ⵞⵜⵀⵀⵛ ⵛⵍ ⵣⵛⵀⵛ </gnc:facs>
   <gnc:dipl> ⵊⵜⵜⵏⴰ ⵔ{ⵢⵙⵕ}ⵣ ⵠⵔⵏⴰ\ⵜⵒ </gnc:dipl>
</choice>
<lb type="Manuscript_line" n="6"/>
<choice>
   <gnc:facs> ⵑⵍ ⵜⵇⵣⵛⵖⵛⵓⵛ ⵜⵛⵙⵓⵛ : </gnc:facs>
   <gnc:dipl> მოწაფეთა მისთა. </gnc:dipl>
</choice>
```

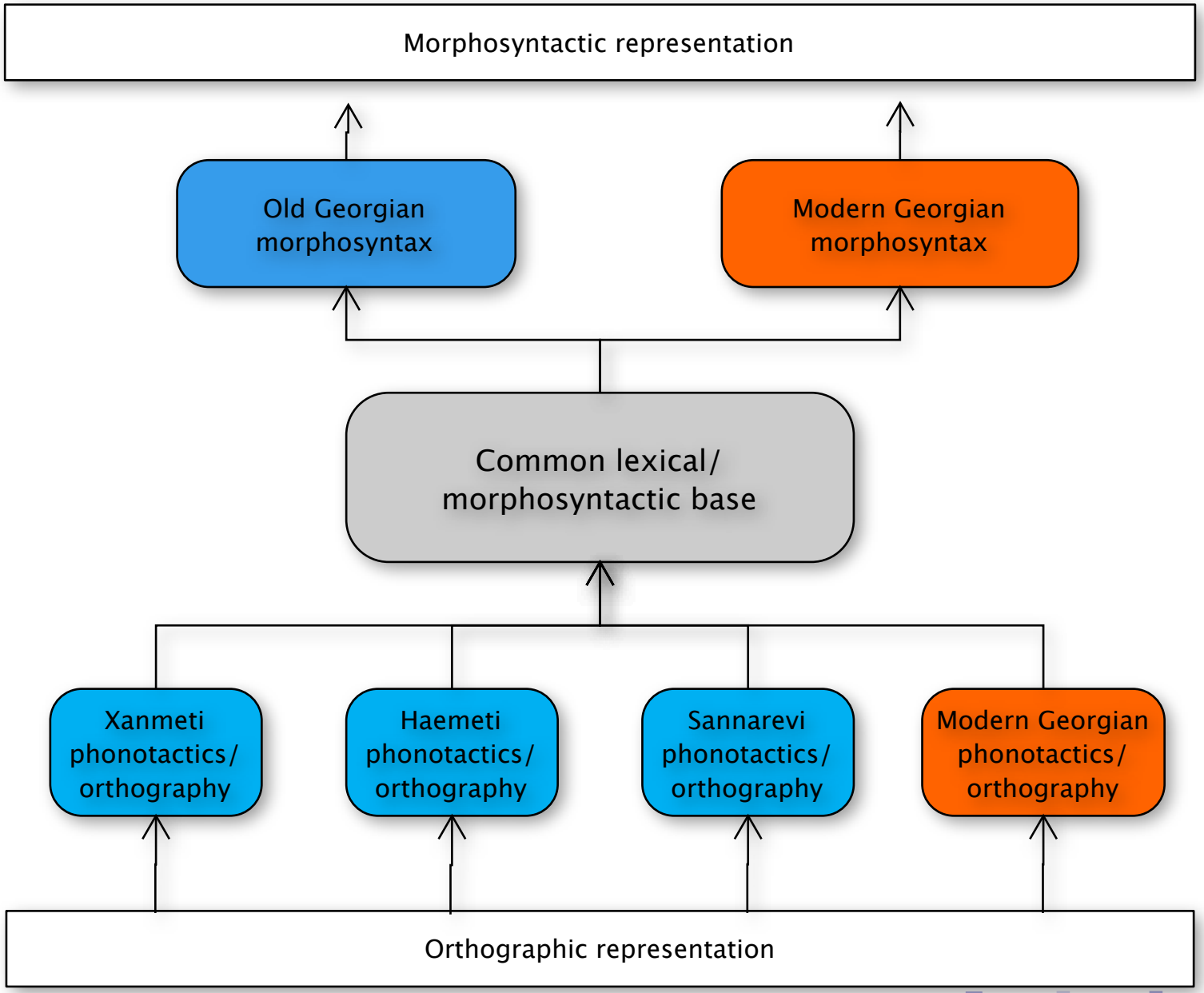
Text Annotation: Challenges

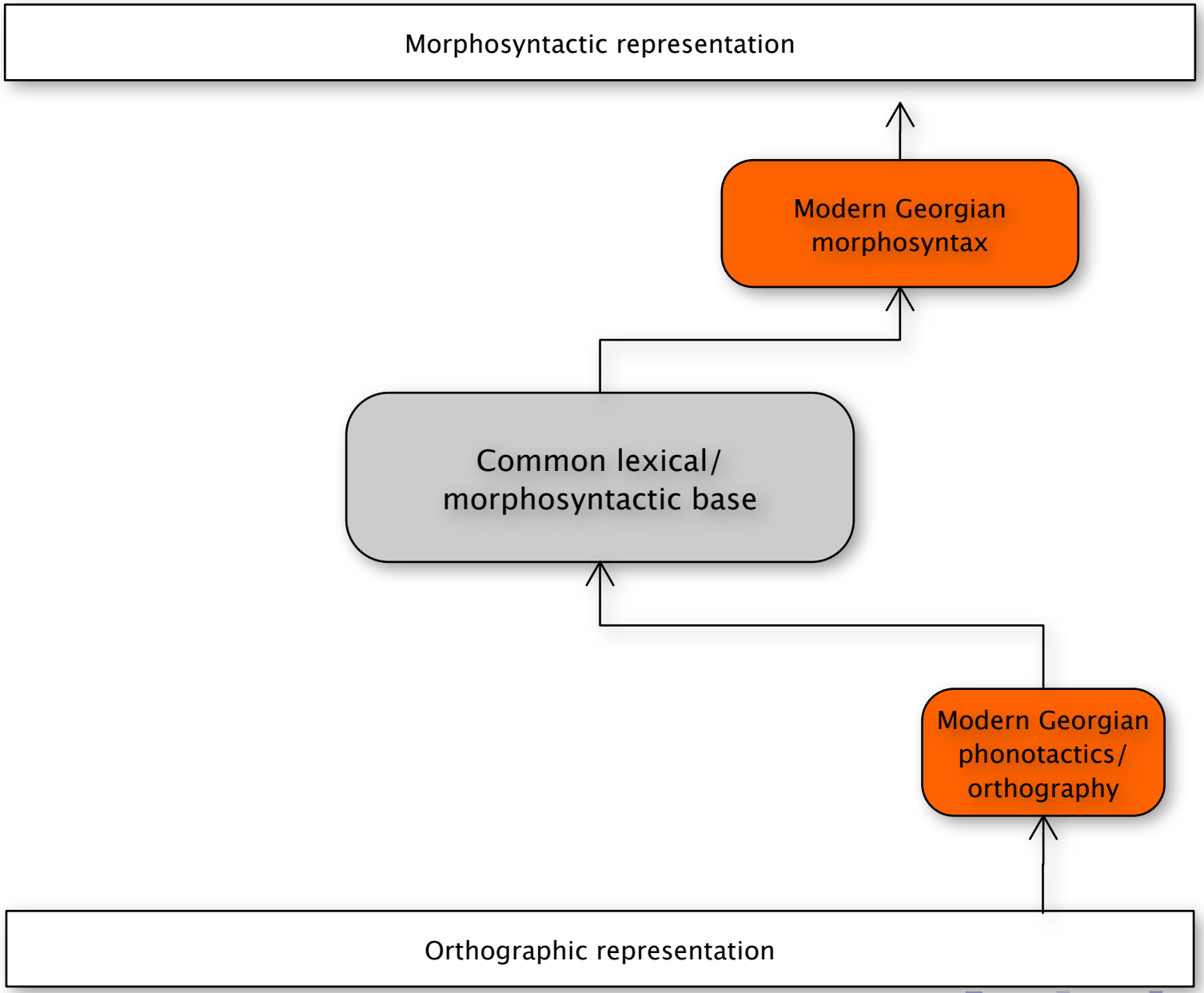
```
<w id="2076">
  <choice>
    <gnc:facs>ⴒⵊⵀⵛ<reg type="ws">ⵖⵊ</reg></gnc:facs>
    <gnc:dipl>წინა\მე</gnc:dipl>
  </choice>
</w>
<lb type="Manuscript_line" n="6"/>
<choice>
  <gnc:facs><orig type="ws" idref="2076">ⵖⵊ</orig></gnc:facs>
</choice>
<w id="2077">
  <choice>
    <gnc:facs>ⵒⵆⴒⵛⵕⵊⵜⵛ</gnc:facs>
    <gnc:dipl>მოწაფეთა</gnc:dipl>
  </choice>
</w>
```

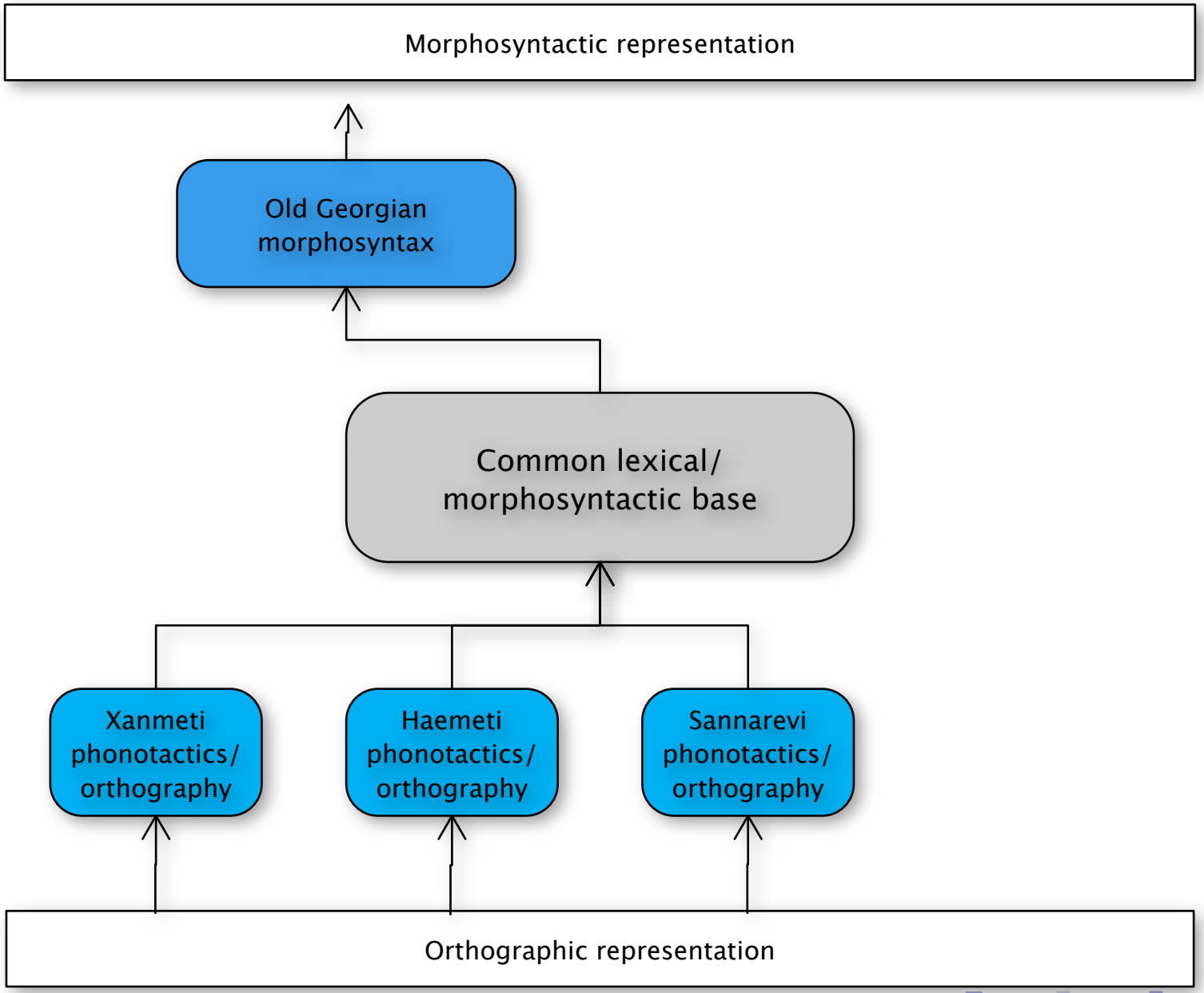
Outline

The GNC: Grammatical annotation

- Based on: Morphosyntactic analyser for Modern Georgian
- Implemented in: FST (Xerox Finite State Tool)
- Modularized to ease extension to other varieties of Georgian
- Core module: Finite state transducer operating on the concatenative level
  - Lexical basis: derived from major dictionaries of Old and Modern Georgian and various word lists
  - Lower side: intermediate morphophonemic representation
  - Upper side: lemma form plus morphosyntactic features
  - Mostly common for Old and Modern Georgian, but certain affixes and forms are marked for language variety
- The core transducer is composed both on lower and upper side with transducers that are specific for each language variety
- Lower side transducers: phonotactic and orthographic pecularities
- Upper side transducers: morphosyntactic pecularities


Morphosyntactic representation
Old Georgian morphosyntax
Modern Georgian morphosyntax
Common lexical/ morphosyntactic base
Xanmeti phonotactics/ orthography
Haemeti phonotactics/ orthography
Sannarevi phonotactics/ orthography
Modern Georgian phonotactics/ orthography
Orthographic representation


Morphosyntactic representation
Modern Georgian morphosyntax
Common lexical/ morphosyntactic base
Modern Georgian phonotactics/ orthography
Orthographic representation


Morphosyntactic representation
Old Georgian morphosyntax
Common lexical/ morphosyntactic base
Xanmeti phonotactics/ orthography
Haemeti phonotactics/ orthography
Sannarevi phonotactics/ orthography
Orthographic representation

Phonotactics

- da#H<ɕer>s, mi#vH<ɕer>, H<ɕer>. Hi<tku>mis

```
define phtXanmeti
      [ {vH}  ->  H    || _ u "<" ]
  .o. [ {vH}  -> {Hu} || _ (?) "<"]
  .o. [ H -> x ]

define phtHaemeti
      [ {vH}  ->  H    || _ u "<" ]
  .o. [ {vH}  -> {Hu} || _ (?) "<"]
  .o. [ H -> h ]

define phtSannarevi
      [ H   -> s  || _ "<" sRoot ]
  .o. [ H   -> 0  || _ ("<") Vowel ]
  .o. [ H   -> h ]
```

Phonotactics: Examples

Function	Intermediate rep.	Xanmeti	Haemeti	Sannarevi	Modern G.
DO3	da#H<ɕer>s	daxɕers	dahɕers	dasɕers	daɕers
IO3	mi#vH<ɕer>	mixuɕer	mihuɕer	mivsɕer	mivsɕer
S2	H<ɕer>	xɕer	hɕer	sɕer	ɕer
i-Passive	Hi<tku>mis	xitkumis	hitkumis	itkumis	itkmis

DO plural marking

- da-xaṭ-av-s →
  da-xaṭ-va V Pres <S-DO> S:3Sg DO:3
- da-xaṭ-a →
  da-xaṭ-va V Aor <S-DO> S:3Sg DO:3Sg
- da-xaṭ-n-a →
  da-xaṭ-va V Aor <S-DO> S:3Sg DO:3Pl
- da-gu-xaṭ-a →
  da-xaṭ-va V Aor <S-DO> S:3Sg DO:1 [Incl]
- da-gu-xaṭ-n-a →
  da-xaṭ-va V Aor <S-DO> S:3Sg DO:1Pl [Incl]
- da-gu-i-xaṭ-n-a →
  da-xaṭ-va V Aor OV <S-DO-IO> S:3Sg DO:3Pl IO:1 [Incl]
  da-xaṭ-va V Aor SV <S-DO> S:3Sg DO:1Pl [Incl]
- da-m-i-xaṭ-n-a →
  da-xaṭ-va V Aor OV <S-DO-IO> S:3Sg DO:3Pl IO:1 [Excl]
  da-xaṭ-va V Aor OV <S-DO-IO> S:3Sg DO:3Pl IO:1Sg
  da-xaṭ-va V Aor SV <S-DO> S:3Sg DO:1Pl [Excl]

DO plural marking

```
[ "+DO:1Sg"  <- "+DO:1",
  "+DO:2Sg"  <- "+DO:2",
  "+DO:3Sg"  <- "+DO:3"
  || 2ndSeries S=DO ?* _ ?* "+DirObjSg" ]
.o.
[ "+DO:1Pl"  <- "+DO:1",
  "+DO:2Pl"  <- "+DO:2",
  "+DO:3Pl"  <- "+DO:3"
  || 2ndSeries S=DO ?* _ ?* "+DirObjPl" ]
.o.
[ "+DO:3Pl" <- "+DO:3"
  || 2ndSeries S=DO=IO ?* _ ?* "+DirObjPl" ]
.o.
[ "+DO:3Sg" <- "+DO:3"
  || 2ndSeries S=DO=IO ?* _ ?* "+DirObjSg" ]
```

Lemma, feature set

Morphosyntactic analysis and features are being revised to be more in line with the Georgian linguistic tradition and expectations (work together with Lela Cixelašvili)

Lemma:

- For nominals: nominative form, with marking of syncopation
- For verbs: masdar + verbal root

Types of features:

- word class (POS), subclass
- morphological features (case, number, tense, version, verbal class,…)
- verb frame with arguments (S, DO, IO)
- (morpho)syntactic features: argument case marking; argument person/number

Morphosyntactic analysis: Examples

Noun:

- jm-isa-ta-y (the one of those of the brother) →
  jm[a]-y N Hum Sg Gen DPl DOldPl DGen DDSg DDNom

Verb:

- da-v-i-ȼer-e-n-i-t (we were written; we wrote them for ourselves) →
  da-ȼer[a]/ȼer V Pass Aor Pv <S> <S:Nom> S:1Pl
  da-ȼer[a]/ȼer V Act Aor Pv SV <S-DO> <S:Erg> <DO:Nom> S:1Pl DO:3Pl

Tmesis:

- mo-vinme-vides (somebody will come) →
  mo-slv[a]/vid V MedPass Conj-II Pv <S> <S:Nom> S:3Sg vinme Pron Indet Anim Nom

Morphosyntactic analysis: Examples

Tmesis:

- mo-vinme-vides (somebody will come) $\rightarrow$
  mo-slv[a]/vid V MedPass Conj-II Pv <S> <S:Nom> S:3Sg vinme
  Pron Indet Anim Nom

word	lemma	features
da	da	Cj
mo-vinme-vides	mo-slv[a]/vid	V MedPass Conj-II Pv <S> <S:Nom> S:3Sg
	vinme	Pron Indet Anim Nom
kalakad	kalak-i	N Sg Advb

Outline

Disambiguation

Morphosyntactic annotation is ambiguous and can (partially) be disambiguated based on context.

Possible approaches: statistical and rule-based taggers.

Advantages of a rule-based approach:

- better suited for rich tagsets
- ambiguity/precision ratio can be controlled
- residual ambiguity can be eliminated manually or with a statistical module

Chosen rule-based formalism: Constraint Grammar (CG)

Fred Karlsson v. 1 (1990), Eckhard Bick v. 3

Disambiguation

Initial procedure:

- CG will only be sparingly used
- disambiguation will be done manually using a Web-based tool

Probably, most or all Old Georgian and Middle Georgian material will be treated this way. For Modern Georgian, we will have to rely on automatic disambiguation because of the material's large size.

Plan for year 2:

- Small amount of fully disambiguated, manually corrected and virutally error-free texts for all language stages (gold standard)

Plans: Statistical disambiguation

Idea: A word might be ambiguous as a common form of two nominal or verbal paradigms.

If one of the paradigms is much commoner than the other, this can be computed by counting the occurrences (in a large corpus) of those forms that are not common to both paradigms.

Examples:

biče̩bi (biče̩a/biče̩i)
common forms: 3017
biče̩i only: 10232
biče̩a only: 9

ča̩moviġe (Trans/Unacc)
common forms: 701
Trans only: 5031
Unacc only: 0