Constructing an annotated corpus for Georgian – Tools and resources

Paul Meurer

Uni Computing; University of Bergen, Norway

Kutaisi, September 29, 2011

Outline

Outline

Designing an annotated corpus

Questions to consider:

- Scope: domain coverage, longitudinal (diachronic) coverage, dialectal, spoken material?
- Size: e.g. Russian National Corpus: 150 mio words, BNC, ANC: 100 mio, EANC: 110 mio
- Balancing vs. maximal size
- Sources: OCR-scanning, web harvesting, digital originals, others. A lot of work has to be invested into formatting and cleaning of the material (boilerplate removal, correction of scanning errors, duplicate removal etc.)
- Copyright matters

Designing an annotated corpus

Questions to consider:

- Grammatical annotation level: lemma, POS, morphosyntax, Named Entities, syntax, semantics, discourse, etc.; which frameworks to use?
- Meta annotation: Title, source, author, translator, year/date, genre, topic etc. (see e.g. EAGLES standard)
- An appropriate corpus tool to make the corpus accessible and searchable

A corpus for Georgian

My aim: a large annotated corpus of written standard modern Georgian, no balancing, for linguistic research

Sources: Texts from the Internet only

- www.open.ge (newspapers)
- www.civil.ge (news)
- www.tavisupleba.ge (news and background)
- lib.ge etc. (literature)

Size by now: 125 million words, growing

Grammatical annotation (subcorpus): lemma, morphosyntax; ambiguity: 1.2

Meta annotation: Title, source, author, translator, year/date

Search tool: Korpuskel

Outline

Morphology: Parsing model

First approach:

- Finite state transducer augmented with feature structure unification (general model of Georgian inflection)
- Disjunctive unification with a lexicon of existing forms to discard nonexisting verb analyses
- Implemented in Common Lisp, based on Parc Xerox's old fsa module

New implementation:

- based on **fst** (Xerox finite state tool, soon open source)
- automatically derived from old implementation
- flag diacritics mimic feature structure unification; compiled out at the end $\Rightarrow$ pure finite state
- lexicon compiled into the transducer
- interfaces well with LFG Grammar

Morphology

The lexicon

- Verb entries derived from Kita Tschenkélis ‘Georgisch-Deutsches Wörterbuch’ (52 000 entries, 3 823 verb entries)
- Other entries: Tschenkéli, Rayfield et al. (A Comprehensive English-Georgian Dictionary, 131 644 entries); material from Levan Chkhaidze.

Coverage

- Measured on 2/3 of “Data Tutašxia”: 1.3% (3.4%) unknown words, 3.5% (6.9%) unknown types (mainly names, Old Georgian, Russian, typos, missing adverbials)

Challenges

- foreign names
- named entities (difficult b/o missing case distinction)
- Old and Middle Georgian words and spellings
- dialect words

Verbal morphology: Superparadigms

Kita Tschenkéli's "Deutsch-Georgisches Wörterbuch":

- Verb forms are grouped hierarchically according to root, verbal class (transitive, unergative, unaccusative, indirect; causative, stative passive), and preverb, in that order
- Very valuable: detailed information about valency
- Participles are missing and are being added manually

The full set of paradigms deriveable from a given root I call a Superparadigm.

Morphology: Example analyses

‘wine’

ġvino →

ġvino N Nom Sg Full

Morphology: Example analyses

‘wine’

ġvino →

ġvino N Nom Sg Full

‘big’

did →

didi A Dat Reduced

didi A Adv Reduced

Morphology: Example analyses

‘wine’

ġvino →

ġvino N Nom Sg Full

‘big’

did →

didi A Dat Reduced

didi A Adv Reduced

‘it is for the girls, too, he said’

gogo-eb-isa-tvis-ac-aa-o →

gogo N Anim Gen Pl Full Tvis C Aux IndSpeech3

Morphology: Example analyses

‘wine’

ġvino →

ġvino N Nom Sg Full

‘big’

did →

didi A Dat Reduced

didi A Adv Reduced

‘it is for the girls, too, he said’

gogo-eb-isa-tvis-ac-aa-o →

gogo N Anim Gen Pl Full Tvis C Aux IndSpeech3

‘in childhood’

bavšvob-isa-s → bavšvoba N DGen DSg Dat Sg

Morphology: Example analyses

‘he apparently painted it’/‘he will paint it for her’/‘unpaintable’ (Dat)

da-u-xaṭ-av-s →

da-xaṭva V Trans Base Fut <S-DO3-OBen> <NomSubj> <DatObj> <DatObjBen> Subj3Sg Obj3 ObjBen3

da-xaṭva V Trans Base Perf <S-DO> <DatSubj> <NomObj> Obj3 Subj3Sg

da-xaṭva VPart NegPart Dat Sg Full

Morphology: Example analyses

‘he apparently painted it’/‘he will paint it for her’/‘unpaintable’ (Dat)

da-u-xaṭ-av-s →

da-xaṭva V Trans Base Fut <S-DO3-OBen> <NomSubj> <DatObj> <DatObjBen> Subj3Sg Obj3 ObjBen3

da-xaṭva V Trans Base Perf <S-DO> <DatSubj> <NomObj> Obj3 Subj3Sg

da-xaṭva VPart NegPart Dat Sg Full

Analysis output for verb forms is:

Masdar + Paradigm ID + features

In corpus annotation, the Paradigm ID is dropped.

Morphology: Open problems

Problem: Not all verb forms have an unambiguous masdar

- apasebs $\longrightarrow$ še-paseba / da-paseba / čamo-paseba
- but not: $\longrightarrow$ paseba

The correct one can at most be inferred from context.

Possible solution:

- apasebs $\longrightarrow$ $*$-paseba

The tagset I

- POS: N Prop Pron Pp A AInt Q ALLQ Det Adv Cj Card Ord V VPart Neg Period ExclPoint IntMark Ellipsis Comma Semicolon Dash Quote LParen RParen
- Subclass: Poss Pers Rel Refl Interr Digits Alphabetic Pot
- Case: Nom Erg Dat Gen Inst Adv Voc
- Number: Sg Pl OldPl
- Person: 1 2 3 Poss1sg Poss2sg Poss3sg Poss1pl Poss2pl Poss3pl
- Declension type: Full Reduced Bound Free
- Double declension tags: DGen DSg DPl
- Postpositions: Dan Dmi Ebr Ebriv Cin Gan Si Tan Ken Mde Mdis Mebr Tvis Vit Ze Iani

The tagset II

- Tense: Pres Impf ConjPres Fut Cond ConjFut Aor Opt Perf PluPerf ConjPerf Imp
- Participle: Masdar PresPart PastPart FutPart NegPart
- Verb class: Trans Unacc Unerg Inv Caus
- Verb valency: <S> <S-DO> <S-DO3-OBen> ...
- Verb agreement: Subj1 Subj1Pl Subj1Sg Subj2 Subj2Pl Subj2Sg Subj3 Subj3Pl Obj1 Obj1Pl Obj1Sg Obj2 Obj2Pl Obj2Sg Obj3
- Argument case: <NomSubj> <ErgSubj> <DatSubj> <NomObj> <DatObj> <GenObj> <DatObjTh> ...
- Clitics: IndSpeech1 IndSpeech2 IndSpeech3 C Ve Ga Long Aux
- Semantics: Title Meas Mass Coll Temp Anim Inanim
- Style: Old Subnorm Dialect Rare Bracket
- NE tags: Name FirstName LastName City Institution Geo River Sea Area

Tools for morphology development

- Lexicon stored in database, rules stored in files
- Web interface to the morphology
- Paradigm and superparadigm display
  - used for editing
  - helps in detecting missing forms and overgeneration
- Regression testing on a large corpus

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
- can be used as a preprocessing tagger for LFG analysis

Constraint Grammar

Constraint Grammar (CG)

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

- Rules operate on morphosyntactically analyzed text
- Rules either REMOVE, SELECT, ADD or REPLACE a tag or a set of grammatical tags in a given sentence context
- The last reading is never discarded, this makes CG a very robust formalism
- Levels of analysis: morphosyntax, syntactic functions, dependencies
- Fast open source implementation: vislcg3

Constraint Grammar: Example rules I

Removing rare forms

```
REMOVE ("xoli" N Voc) ;

REMOVE ("<xart>" "xari") ;
REMOVE ("<ert>" "eri") ;
REMOVE ("<an>" "ani") ;

REMOVE (Voc) IF (0 (Nom)) ;

SELECT ("da" Cj) IF (NOT -1 (Poss Nom));
```

Constraint Grammar: Example rules II

Negation

```
SELECT  V    IF (-1 Neg) (NOT -1 (Aux)) ;
SELECT (Neg) IF (1 V) ;
SELECT (Neg) IF (NOT 1 V) ;
REMOVE  V    IF (1C Neg) ;
```

Modal “unda”

```
SELECT (V Modal) IF (1 OPT) ;
SELECT (V Modal) IF (1 Neg) (2 OPT) ;
SELECT (V Modal) IF (-1 OPT) ;
SELECT  OPT      IF (-1 (V Modal)) ;
```

Constraint Grammar: Example rules III

Agreement

```
SELECT (Pron Pers Erg 1 Sg)
    IF (0* ErgSubj + Subj1Sg
           BARRIER CLB | (V) - (Modal))

SELECT (V Subj1Sg)
    IF (0C ErgSubj)
       (0* (Pron Pers Erg 1 Sg)
           BARRIER CLB | (V) - (Modal)) ;

REMOVE (V Subj1Sg <NomSubj>)
    IF (0* (Nom Full) BARRIER CLB) ;
```

Constraint Grammar: Example rules III

Case disambiguation

```
TEMPLATE NPGen =
        (? NA + Gen)
    OR (? NA + Gen + Reduced LINK 1 T:NPGen) ;

REMOVE (Gen)
    IF (NOT 1 (Pp))
       (NOT 1 Gen)
       (NOT 1 NACProp)
       (NEGATE 1 ("da" Cj) | (",") | (Adv)
                 LINK 1 Gen)
       (NEGATE 1 T:NPGen LINK 1 Gen)
       (NOT 0* GenArg BARRIER CLB) ;
```

Constraint Grammar: Challenges

Ambiguous Adjective attachment

gardacvlili “deceased”

- gardacvlili A Nom Reduced
- gardacvlili A Gen Reduced

mdguris “staying traveller”

- mdguri N Gen Sg Full

koneba “belongings”

- koneba N Nom Sg Full

Constraint Grammar: Challenges

Forms that are difficult to disambiguate

daiçqo

- daçqoba V Unacc Aor <S> Subj3Sg
- daçqoba V Trans Opt <S-DO-R> Subj2Sg Obj3
- daçqoba V Trans Aor <S-DO-R> Subj3Sg Obj3
- daçqeba V Unacc Aor <S> Subj3Sg
- daçqeba V Trans Opt <S-DO-R> Subj2Sg Obj3
- daçqeba V Trans Aor <S-DO-R> Subj3Sg Obj3

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:

biċebi (biċa/biċi)
common forms: 3017
biċi only: 10232
biċa only: 9

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