LatMor: A Latin Finite-State Morphology

Uwe Springmann, Helmut Schmid, Dietmar Najock

[1]US, HS: Center für Informations- und Sprachverarbeitung (CIS),
Ludwig-Maximilians-Universität München

[2]DN: Institut für Griechische und Lateinische Philologie,
Freie Universität Berlin

Workshop
"Treebanking Ancient Languages - Current and Prospective Research",
December 14-15, 2015, Universität Leipzig

2015-12-15

Computer assisted morphological analysis

- presupposes electronic text collections
- enables lemma-based search
- linguistic research: vocabulary studies
- groundwork for further study (syntax)

History of ancient language morphological analyzers

- Fr. Roberto Busa: Index Thomisticus (Latin, 1949-1980)
- David Packard (Greek, 1973) (Packard 1973)
- Joseph Denooz (LASLA, 1973) (Denooz 1973)
- Bozzi/Marinone (LEMLAT, Pisa, 1982) (Passarotti 2004)
- Najock/Morgenroth (LatLem, mid-80s)
- Greg Crane (Morpheus: Greek, later also Latin, 1984) (Crane 1991; Crane 1998)
- many systems since:
  - Whitaker's Words, Collatinus, PROIEL, ...
  - G. Crane: "*Anybody can write a Latin analyzer over a weekend* …"

So why yet another Latin morphological analyzer?

We need open data (including code sources) to enable incremental progress:

- *availability*: older implementations mostly unavailable, copyrighted, patented ...
- *documentation*: lacking documentation makes it hard to install, run and adapt older systems even when they are available
- *speed*: analyze 100,000 wordforms per second rather than hundreds ...

Lexical sources

- hand-compiled Berlin Latin Lexicon (70,000 lemmata) by group of D. Najock at Freie Universität Berlin (80s):
  - main source: Georges Handwörterbuch (Georges 1913)
  - proper names: Lewis & Short (Lewis and Short 1907)
  - vowel quantities checked and added from Menge (Menge, Güthling, and Pertsch 1983)
  - further vocabulary additions from concordance work of Najock et al.
- lexical entries with pseudo-stems:

```
vt audio: 4          (transitive verb, 4th conj.)
su serv/us, i: m
```

```
POS: su; pseudo-stem: serv; ending: us; genitive: i: (long i); gender: masculine
```

Finite State Analyzer

- lexical entries get converted into unambiguous input form for transducer:

```
<Stem>audïre<V><base><Verb-i>
<Stem>servus<N><base><NMasc-o>
```

- transducer is implemented with Stuttgart Finite State Tools (SFST) (Schmid 2006)
- inflection rules generate correct surface forms (with triggers such as `<delete>`,`<shorten>`,`<ins-u>`):

```
$Verb-X-active$ = \
{<active><sg><1>}:{<delete><shorten>ö} |\
    $Verb-X-active-2$

$Verb-X-active-2$ = \
{<active><sg><2>}:{s} |\
    {<active><sg><3>}:{<shorten>t} |\
    {<active><pl><1>}:{mus} |\
    {<active><pl><2>}:{tis} |\
    {<active><pl><3>}:{<shorten><ins-u>nt}
```

```
Example 3rd pl:
audiunt (shorten i, insert u)
laudant (shorten a)
```

Demonstration: Speed

Analyze Caesar's Commentarii de bello Gallico (11,420 tokens)

- web service: http://services.perseids.org/bsp/morphologyservice: ca. **20 min**
- local Morpheus installation: **6 sec** (i5-3320M CPU @ 2.60GHz)

```
$ time MORPHLIB=stemlib  cruncher -L < caesar.txt 2>/dev/null > crunched

real 0m5.836s
user 0m5.134s
sys  0m0.691s
```

- LatMor: **0.1 sec**

```
$ time fst-infl2 latmor.ca caesar.txt >/dev/null
reading transducer from file "latmor.ca"...
finished.
11400

real 0m0.111s
user 0m0.110s
sys  0m0.000s
```

Evaluation: Coverage

	Caesar		Nepos		Godfrey	
all	type	token	type	token	type	token
PROIEL	70.0	51.6	69.4	47.9	63.1	50.6
Parsley	89.5	95.2	90.0	94.3	86.7	91.7
Words	90.5	96.6	88.1	93.3	93.0	95.4
Morpheus	92.5	93.8	89.0	92.7	87.6	92.7
LEMLAT	92.8	94.1	89.2	92.8	88.1	93.1
LatLem	92.4	97.1	94.3	97.2	88.1	93.0
LatLem+que	**97.8**	**99.0**	97.8	98.8	89.4	93.8
LatMor	97.3	98.8	**97.9**	**99.1**	**96.0**	**97.2**
lowercase only	type	token	type	token	type	token
PROIEL	73.4	51.9	73.2	49.3	70.2	54.4
Parsley	90.7	96.0	90.5	94.6	93.1	95.5
Words	93.7	97.8	95.2	97.6	96.4	97.9
Morpheus	**99.6**	**99.8**	**99.5**	**99.8**	98.1	99.0
LEMLAT	99.5	99.7	99.3	99.7	**98.4**	**99.2**
LatLem	93.5	97.9	95.4	97.9	96.1	97.8
LatLem+que	99.3	**99.8**	99.3	99.7	97.7	98.6
LatMor	98.2	99.2	98.7	99.4	97.7	98.7

Table 1: Coverage of different morphological analyzers on three Latin texts.

Future work

- Fix errors:
  - non-standard numerals (XIIII XXXX)
  - first name initials (Q. Cn.)
  - typos such as XIlI (third letter is wrong)
  - unknown proper names (Commius Lucterius)
  - conversion of the lexicon entry had failed (*progredi iureiurando sese totidem*)
  - missing lexicon entries (*conloquium*)
  - forms not generated by the inflectional paradigm (*oportere veniri*)
  - inflection errors (*vehementiter* instead of *vehementer, meridieo* instead of *meridie*)
  - prefixed forms missing in the lexicon (*perspici* failed, but spici was analyzed)
  - missing alternative forms (*orereturr* as an alternative form of *oriretur*)
- Remaining errors in the lexicon conversion program and the inflection tables need to be identified and fixed.
- The list of proper names in the lexicon needs to be extended.
- Derivation rules should be implemented to analyze e.g. prefix verbs.

References I

Crane, Gregory. 1991. “Generating and Parsing Classical Greek.” *Literary and Linguistic Computing* 6 (4). ALLC: 243–45.

———. 1998. “New Technologies for Reading: The Lexicon and the Digital Library.” *The Classical World*. JSTOR, 471–501.

Denooz, Joseph. 1973. “Recherches Sur Le Traitement Automatique de La Langue Latine.” *Revue de L’Organisation Internationale Pour L’Etude Des Langues Anciennes Par Ordinateur*, no. 1. LASLA.

Georges, Karl Ernst. 1913. *Ausführliches Lateinisch-Deutsches Handwörterbuch*. Hannover u. Leipzig: Hahnsche Buchhandlung.

Lewis, Charlton T, and Charles Short. 1907. *A New Latin Dictionary. Founded on the Translation of Freund’s Latin German Lexicon Edited by E. a. Andrews, LL d.* Clarendon Press.

References II

Menge, Hermann, Otto Güthling, and Erich Pertsch. 1983. *Langenscheidts Großes Schulwörterbuch Lateinisch-Deutsch*. Langenscheidt.

Packard, David W. 1973. "Computer-Assisted Morphological Analysis of Ancient Greek." In *Proceedings of the 5th Conference on Computational Linguistics - Volume 2*, 343–55. COLING '73. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.3115/992567.992595.

Passarotti, Marco Carlo. 2004. "Development and perspectives of the Latin morphological analyser LEMLAT." *Linguistica Computazionale* 20 (A). Italy: 397–414.

Schmid, Helmut. 2006. "A Programming Language for Finite State Transducers." In *Finite-State Methods and Natural Language Processing: 5th International Workshop (FSMNLP 2005)*, edited by Anssi Yli-Jyrä, 4002:308–9. Lecture Notes in Artificial Intelligence. Springer, Heidelberg, Germany.