Introducing PaLaFra

A Project on the Creation and Analysis of an Electronic Corpus of Historical Texts of Old French and Late Latin

Lars Döhling
Lehrstuhl für Medieninformatik

FAKULTÄT FÜR SPRACH-, LITERATUR- UND KULTURWISSENSCHAFTEN

Universität Regensburg

“Le *pas*sage du *la*tin au *fran*çais”

- Romance languages originate from spoken Latin
  - Co-existence in space and time
- No written records of *Vulgar Latin* available, but of Late Latin and Old French
- Hopefully, comparing them will give insights on the transition

- Interdisciplinary project
  - Romance linguistics
  - Computer linguistics
  - Media informatics
- Cofunded by ANR and DFG
- Located at Lille, Lyon, Ratisbon, Tübingen

Goals of PaLaFra

Providing a **diachronic corpus** of **Late Latin** and **Old French** helping to analyze the **transition** from Latin to French

- Creation of a corpus of Late Latin
- Corpus annotation
- Connect French and Latin corpora
- Create and investigate a small parallel corpus

- Open access, e.g. via TXM

Corpora

North France (Gaul); religious, historical, juridical

- Late Latin corpus @ Regensburg/Tübingen
  - ≈200 texts, 6th—8th century
  - Sourced from *digital Monumenta Germaniae Historica* (dMGH)

- French corpus @ Lyon
  - 47 texts, 9th—14th century
  - Morphosyntactically annotated with Cattex09
  - Part of *Base de Français Médiéval* (BFM)

- Parallel corpus @ Lille
  - 3+ texts

Annotating Late Latin

- Morphosyntax + Lemmata
- Based on previous work by CompHistSem (Eger et al., 2015)

- Tools: Winner + more recent ones
- Texts: 7 hagiographic texts, cross-evaluated

	Hagiographics	Eger et al.
Sentences	≈1k	≈15.5k (0.5k hagiographical)
Tokens	≈21k	

Late Latin — Morphosyntax

CompHistSem tagset

	LaPOS	MarMoT	Lexicon	Eger et al. (LaPOS)
Accuracy	72,6%	77,3%	53,5%	85,0%
Runtime	4min	25min	<1min	
Error			Missing 27,3% Ambigue 16,6%	

Only tokens with mophology (62%)

	LaPOS	MarMoT	Lexicon
Accuracy	59,8%	66,4%	34,1%
Error			Missing 42,0% Ambigue 19,9%

Late Latin — Lemmatization

- Word tokens only (82%)
- Case-insensitive

	Base-line	**Lemma Gen**	**BTagger**	**Lexicon**	**Eger et al. (LaPOS+ Lemma Gen)**
Accuracy	37,4%	82,6%	83,8%	74,1%	95,3%
Runtime	<1min	<1min	11h	<1min	
Error				Missing 27,4% Ambigue 4,2%	

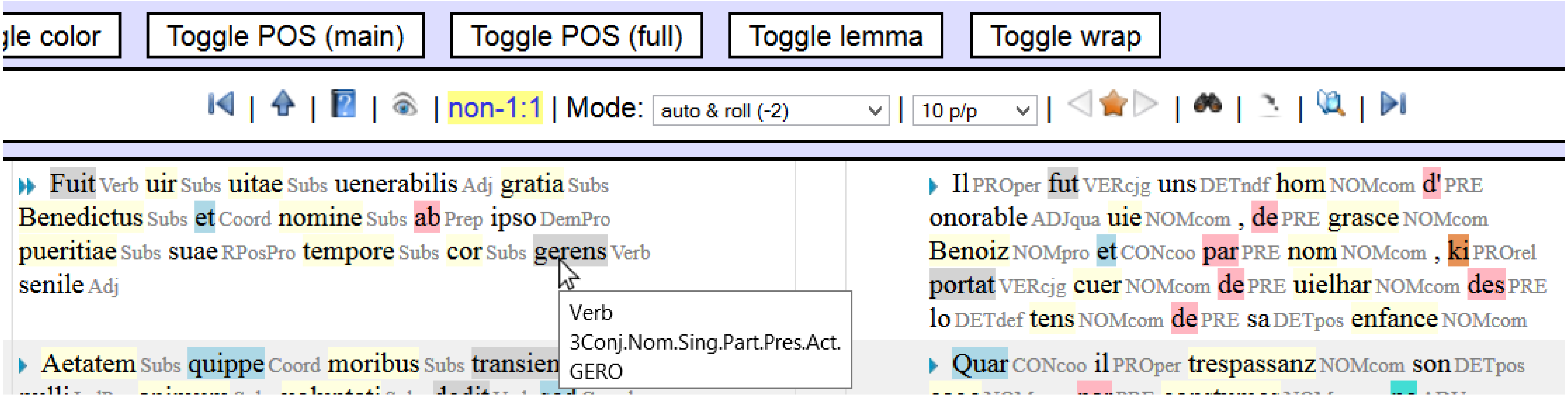
Late Latin — Discurs-pragmatic Labeling

- Typically, hagiographic texts are divided into distinguishable parts, e.g. miracles or virtues
- Different linguistic facies per part
  - Forewords: very elaborate Latin (authors try to impress the reader)
  - Narrative parts: plain style marked by **proto-romance features**

- Automatically labeling following the multi-dimensional approach of Biber (1995)
    - Linguistic features, e.g. usage of personal pronouns

Connecting Corpora — PaLaFra Tagset

- Common tagset for Late Latin / Old French
  - Enables bilingual analyses
  - (Automatic?) mapping between CompHistSem / Cattex09
- Visual tagset comparison via adapted InterText



Fuit Verb uir Subs uitae Subs uenerabilis Adj gratia Subs Benedictus Subs et Coord nomine Subs ab Prep ipso DemPro pueritiae Subs suae RPosPro tempore Subs cor Subs gerens Verb senile Adj

Verb
3Conj.Nom.Sing.Part.Pres.Act.
GERO

Aetatem Subs quippe Coord moribus Subs transien

Il PROper fut VERcjg uns DETndf hom NOMcom d' PRE onorable ADJqua uie NOMcom , de PRE grasce NOMcom Benoiz NOMpro et CONcoo par PRE nom NOMcom , ki PROrel portat VERcjg cuer NOMcom de PRE uielhar NOMcom des PRE lo DETdef tens NOMcom de PRE sa DETpos enfance NOMcom

Quar CONcoo il PROper trespassanz NOMcom son DETpos

Vita Benedicti (LASLA) / Vie de saint Benoit (Cattex09)

Outlook

- Late Latin corpus creation
- Corpus annotation
- Discurs-pragmatic labeling
- Usability of annotation tools

Workshop October 11+12, 2016 @ Lille(Lyon)

www.palafra.org

References

Steffen Eger, Tim vor der Brück, and Alexander Mehler. 2015. *Lexicon-assisted tagging and lemmatization in latin: A comparison of six taggers and two lemmatization methods. LaTeCH 2015*, page 105.

Douglas Biber. 1995. *Dimensions of register variation: A cross-linguistic comparison*. Cambridge University Press.