Challenges of Language Change and Variation: towards an Extended Treebank of Medieval French

Mathilde Regnault, Sophie Prévost and Eric Villemonte de la Clergerie

Syntaxfest, TLT, the 29th of August 2019

Laboratoire Lattice (CNRS, Ecole normale supérieure / PSL, Université Sorbonne nouvelle)
Inria Paris, Almanach

ANR Projet Profiterole

PRocessing Old French Instrumented TExts for the Representation Of Language Evolution

Goal: develop and adapt **resources and tools** to extend the SRCMF *treebank* (Prévost and Stein 2013), annotated with dependency syntax, with new texts of Old French and Middle French

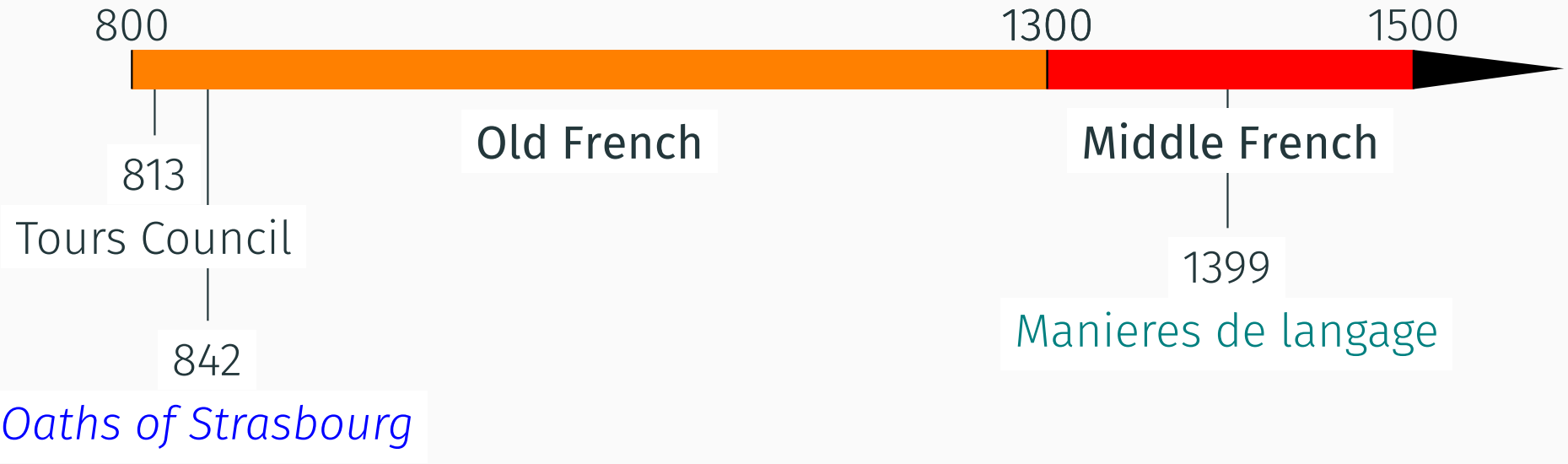
1st treebank containing Medieval French texts: MCVF (Martineau 2008), constituency syntax

	SRCMF	**extension**		MCVF	
	OFr	**OFr**	**MFr**	OFr	MFr
nb texts	16	**17**	**29**	11	14
size (nb words)	251,435	**244,906**	**501,990**	518,090	655,228

OFr: Old French 9th-13th c.

MFr: Middle French 14th-15th c.

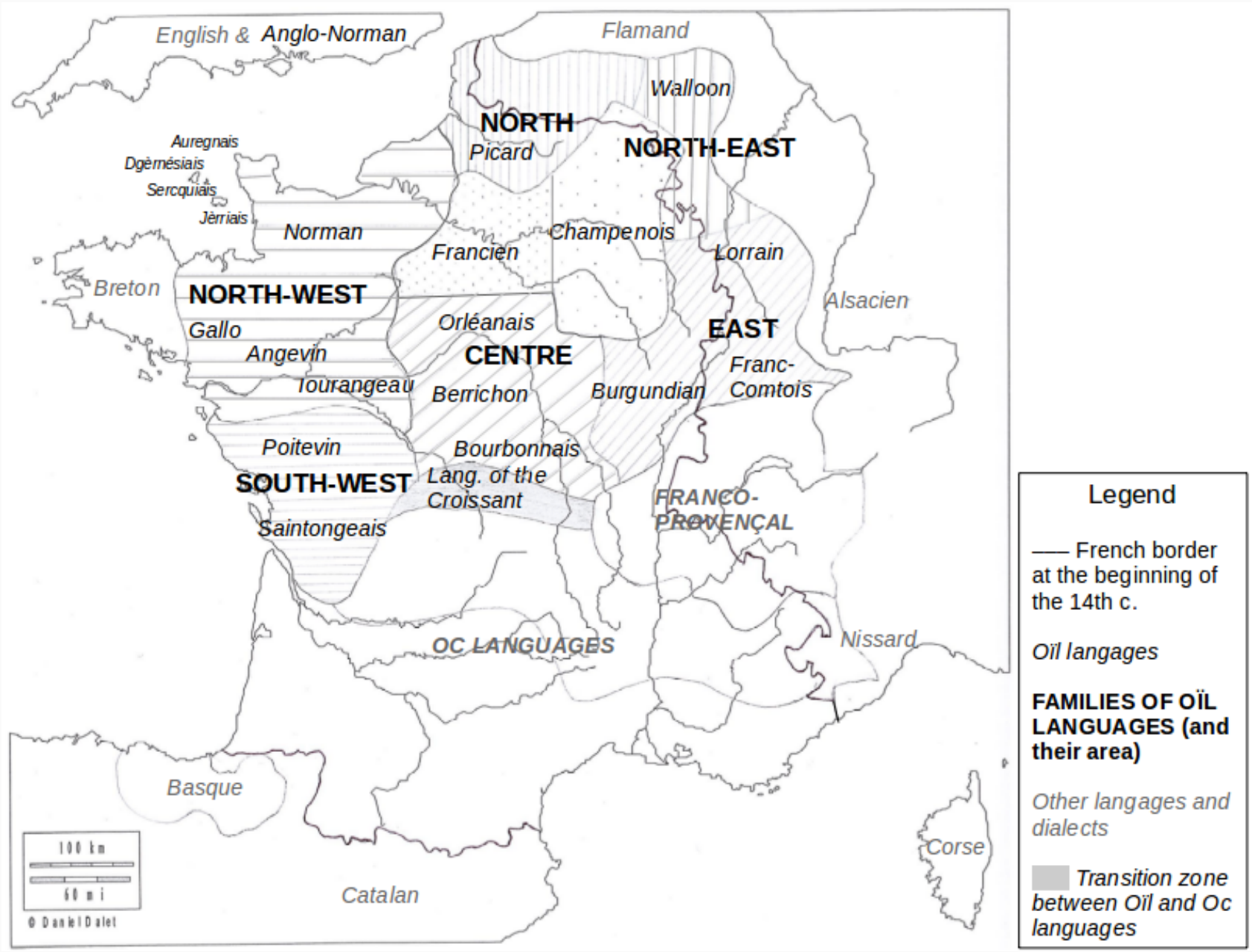
Evolution of Medieval French



[...] si salvarai eo cist meon fradre **Karlo** et **in** aiudha et **in** cadhuna cosa, si **cum** om **per** dreit son fradra salvar dift

Cy comence un petit livre pour enseigner les enfantz de leur entreparler comun françois.

Extending the SRCMF treebank



Parsing Historical Texts

Medieval French

Free word-order (Buridant 2000)

Prevalent order: SOV, then SVO from the 11th c. on

The 5 other orders appear too: `SOV, OSV, OVS, VSO, VOS`

> ex. SOV: 'Franc et paien merveilus colps i rendent',
> *Chanson de Roland* (around 1100)
> trad. *Franks and pagans throw fierce punches (at each other)*

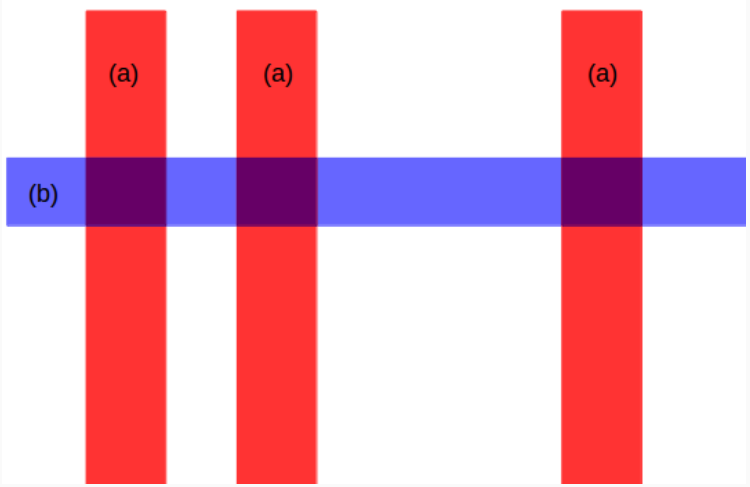
Hyp. main differences between states of Medieval French:
frequencies of linguistic phenomena and word-orders

Decline of nominal declension and no fixed spelling

Variation in Medieval French

Crossing of two perspectives

- **Diachrony**:
  stages of French over 7 c.
- **Synchrony**:
  - dialects
  - domains (genres)
  - form (verses/prose)



Interaction between synchronic axes (a) and the diachronic axis (b)

Consequences on...

- syntax (ex. word-order)
- spelling
- frequencies of constructions

Related work

On the SRCMF treebank, Guibon et al. (2015)

- Mate parser (Bohnet 2010)
- search for discriminating characteristics

Work on ancient languages

- **normalisation**, Bollmann and Søgaard (2016): adding an annotation layer with normalised forms of words
- **transfer**, Scrivner and Kübler (2012): training on treebanks of Old French and Old Catalan to annotate a corpus of Old Occitan

Related work: adapting grammars

State of the art

- Rocio et al. (2003) adapted a pipeline for Contemporary Portuguese to Medieval Portuguese (12^{th}-13^{th} c.)
- adaptation of a grammar:
  - XLE experience (Kaplan, King, and Maxwell III 2002)
  - LinGO Redwoods treebank (Oepen et al. 2004; Toutanova et al. 2005)

Goal: adapt a grammar for Contemporary French in order to make a diachronic grammar for former states of French (9^{th}-15^{th} c.)

Adapting a French Metagrammar

Tree Adjoining Grammars (TAGs)

- Joshi, Levy, and Takahashi (1975), Abeillé (1993)

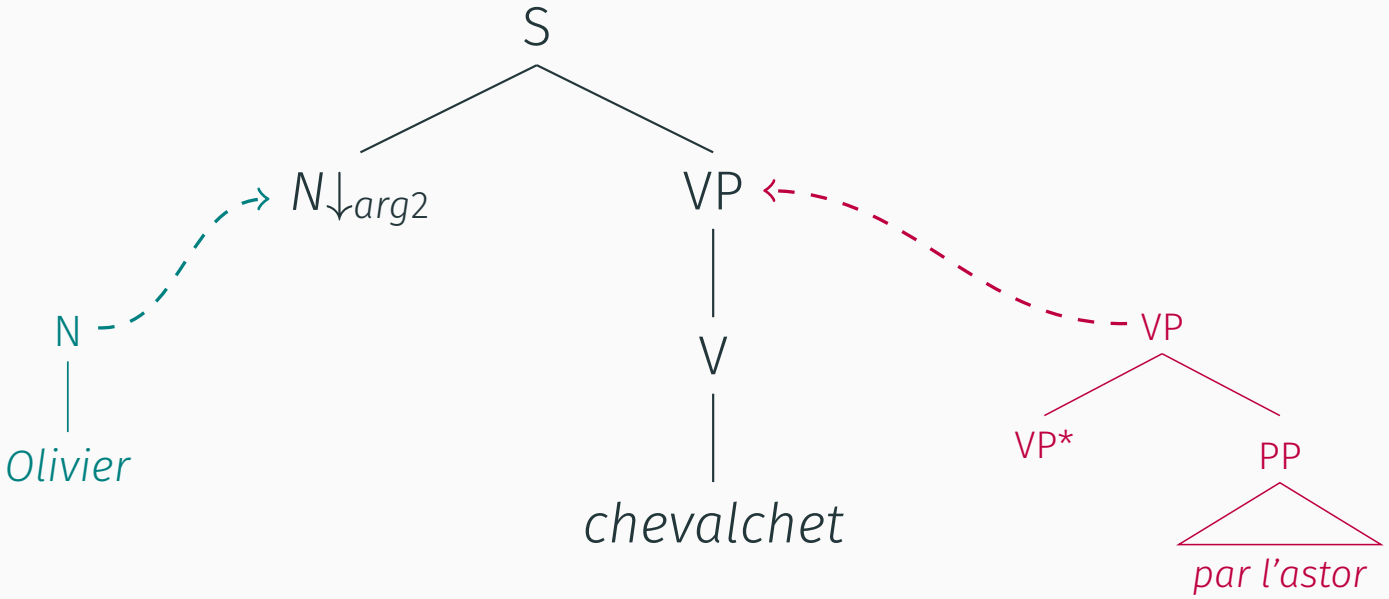


Figure 1: "Olivier chevalchet par l'astor, *Chanson de Roland*, transl. *Oliver rides through the melee*

- 2 possible operations on elementary trees:
  1. substitution
  2. adjunction

Metagrammars

Metagrammars (Candito 1996; Candito 1999)

- **modular description** of a language (making extension and maintenance easier)
- hierarchy of classes (inheritance mechanism)
- together with a lexicon, produces a **Lexicalised TAG (LTAG)**

French Metagrammar (FRMG)

- Villemonte de la Clergerie (2005)
- compact description: 451 classes, 381 factorised trees
- uses the *Lefff* lexicon (Sagot 2010) (compatible *hypertags*)
- online: `http://alpage.inria.fr/frmgwiki/`

Adaption in order to parse Medieval French

- similarities between Contemporary French an Medieval French
- **goal**: modular system capable of handling variation
- lexicon for Medieval French: *OFrLex* (Sagot 2019)

Adapting FRMG

Main steps

1. change some descriptions:
   - loosen some contraints (ex. word-order, verbal agreement)
   - delete or add descriptions to build a large coverage grammar

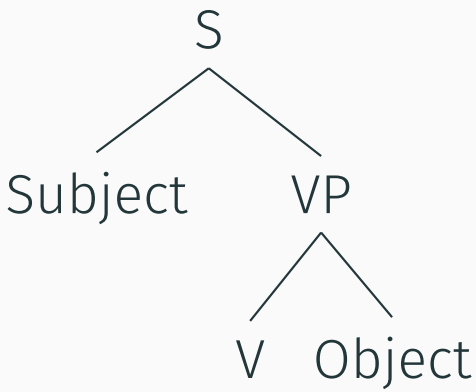


Figure 2: Main constituents in FRMG

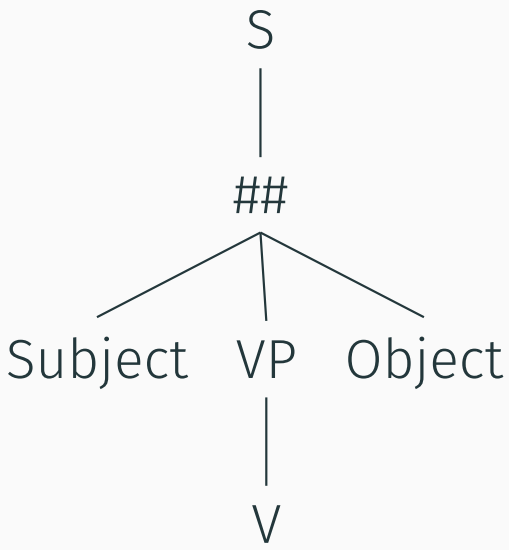


Figure 3: Main constituents in our metagrammar, following Abeillé (2002)

2. add mechanisms in order to handle variation

Free word-order of main constituents

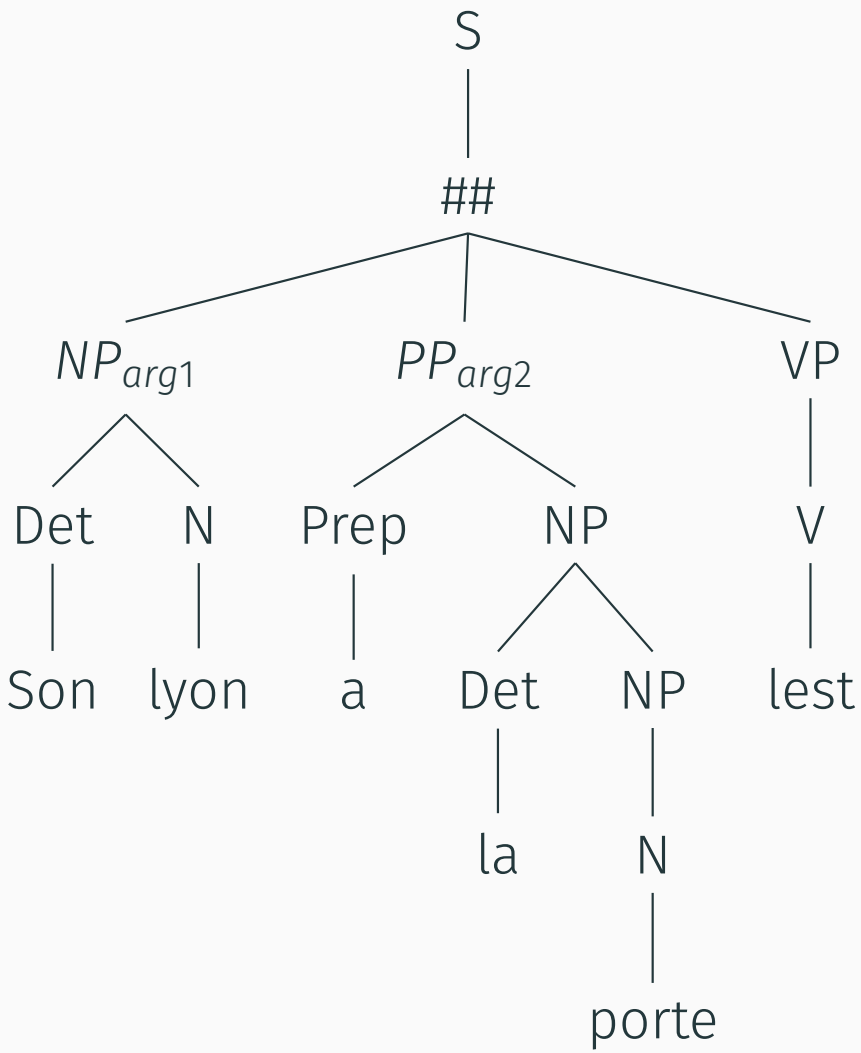


Figure 4: ”Son lyon a la porte lest”, *Yvain*, Chrétien de Troyes, transl. *[He] leaves his lion at the door*

’Free’ word-order of modifiers

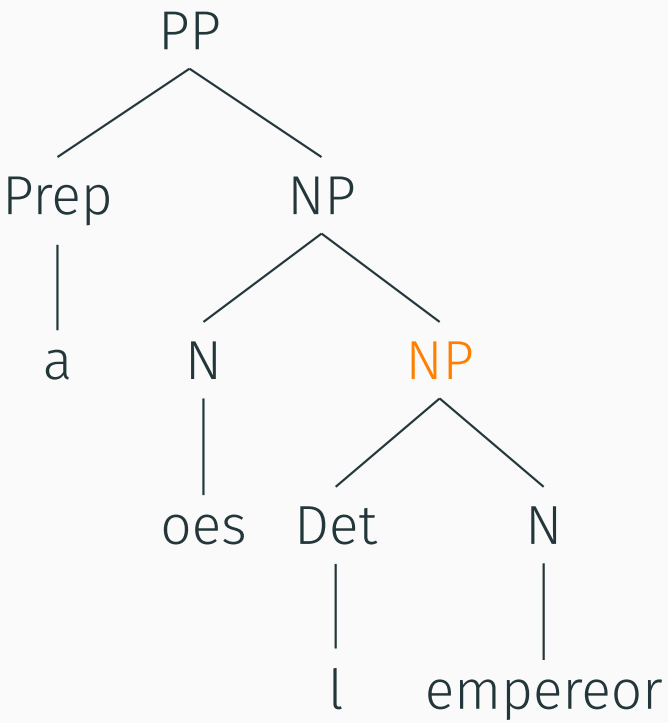


Figure 5: Modifier on the right: ”a oes l’empereor”, *Chanson de Roland*, transl. *in the emperor’s name*

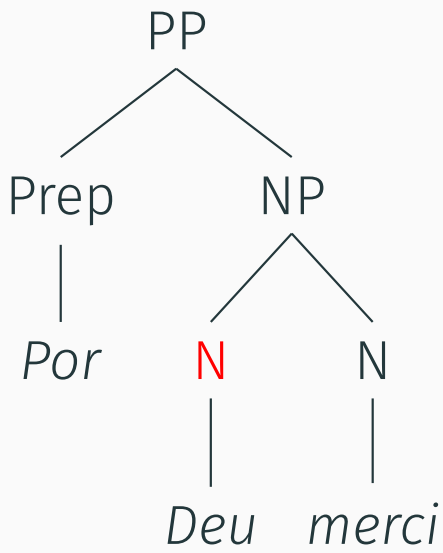


Figure 6: Modifier on the left: ”Por Deu merci”, *Melion*, transl. *For God’s pity*

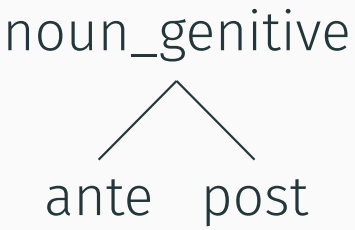


Figure 7: Hierarchy of classes

Conclusion

Conclusion

Work in progress: adaptation of FRMG

A first grammar enables the analysis of sentences

Next steps:

- complete the metagrammar
- train and evaluate the disambiguation
- add facets (filters) to build a diachronic grammar capable of handling several dialects

Thank you!

References i


Anne Abeillé. *Une grammaire électronique du français*. CNRS Editions, 2002.


Anne Abeillé. *Les Nouvelles Syntaxes*. Paris: Armand Colin, 1993.


Bernd Bohnet. "Very High Accuracy and Fast Dependency Parsing is not a Contradiction". In: *Proceedings of the 23rd international conference on computational linguistics* (2010). Ed. by Association for Computational Linguistics.


Marcel Bollmann and Anders Søgaard. "Improving historical spelling normalization with bi-directional LSTMs and multi-task learning". In: *COLING* (2016).


Claude Buridant. *Nouvelle Grammaire de l'ancien français*. Paris: Sedes, 2000.

References ii

Marie-Hélène Candito. “A principle-based hierarchical representation of LTAGs”. In: *Proceedings of the 16th conference on Computational linguistics* (1996).

Marie-Hélène Candito. “Organisation modulaire et paramétrable de grammaires électroniques léxicalisees application au francais et à l’italien”. In: (1999).

Antoine Destemberg. *Atlas de la France médiévale*. Paris: Autrement, 2017.

Gaël Guibon et al. “Searching for Discriminative Metadata of Heterogenous Corpora”. In: *Proceedings of the Fourteenth International Workshop on Treebanks and Linguistic Theories (TLT14)* (2015).

References iii

Aravind K. Joshi, Leon S. Levy, and Masako Takahashi. “Tree adjunct grammars”. In: *Journal of Computer and System Sciences* (1975).

Ronald M. Kaplan, Tracy Holloway King, and John T. Maxwell III. “Adapting existing grammars: the XLE experience”. In: *COLING-02 on Grammar engineering and evaluation* (2002).

France Martineau. “Un corpus pour l’analyse de la variation et du changement linguistique”. In: *Corpus [en ligne]* (2008).

Stephan Oepen et al. “LinGO Redwoods: A Rich and Dynamic Treebank for HPSG”. In: *Research on Language and Computation* (2004).

References iv

Sophie Prévost and Achim Stein. “Syntactic Reference Corpus of Medieval French”. In: (2013).

Vitor Rocio et al. “Automated creation of a medieval portuguese partial treebank”. In: *Treebanks: Building and Using Parsed Corpora*. Anne Abeillé, Kluwer Academic Publishers, 2003.

Benoît Sagot. “The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French”. In: *7th international conference on Language Resources and Evaluation (LREC 2010)* (2010).

Benoît Sagot. “Développement d’un lexique morphologique et syntaxique de l’ancien français”. In: *26ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN)* (2019).

References v

Lene Schøsler. *La déclinaison bicasuelle de l'ancien français : son rôle dans la syntaxe de la phrase, les causes de sa disparition*. Odense University Press, 1984.

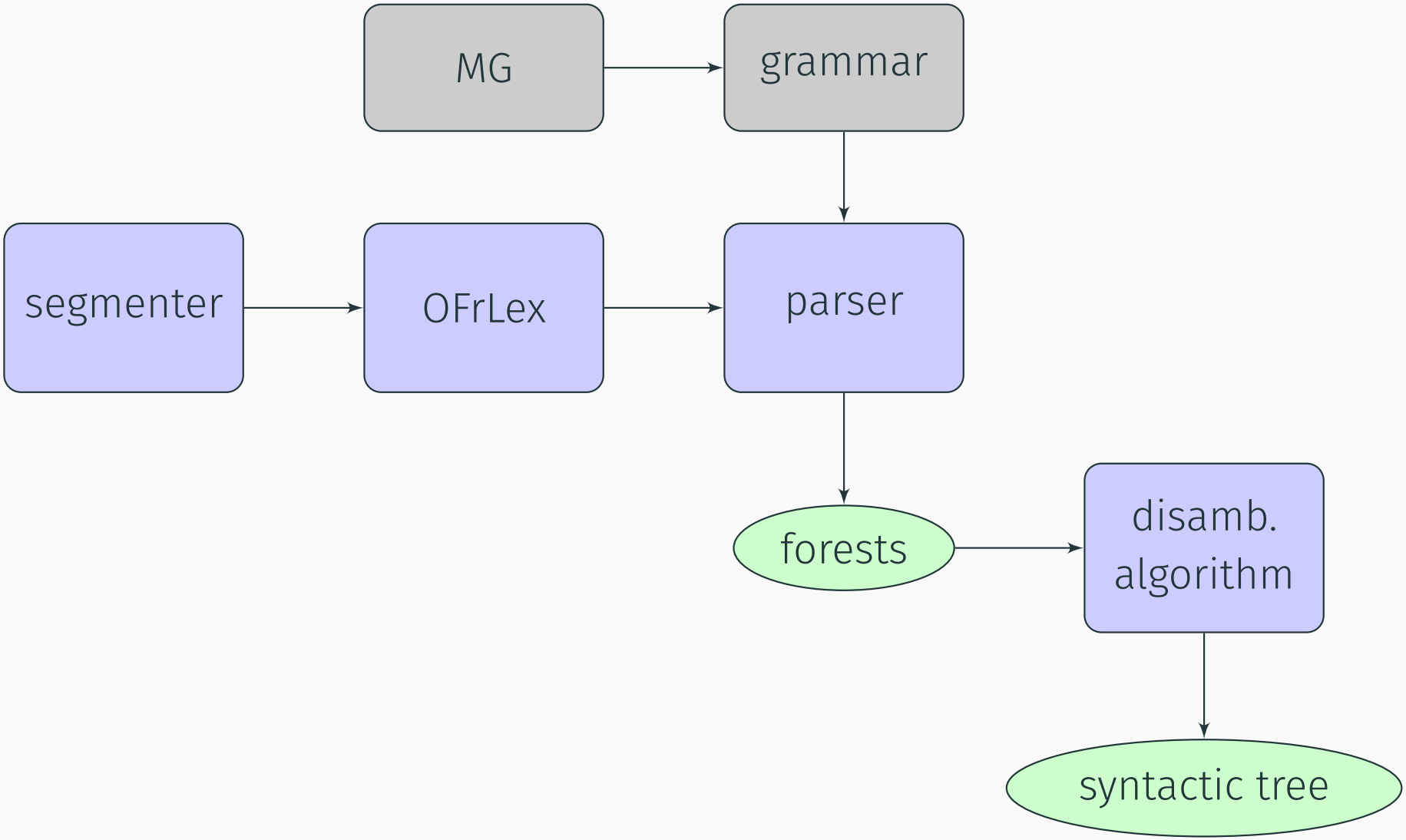
Olga Scrivner and Sandra Kübler. "Building an old Occitan corpus via cross-Language transfer". In: *KONVENS* (2012).

Kristina Toutanova et al. "Stochastic HPSG Parse Disambiguation using the Redwoods Corpus". In: *Research on Language and Computation* (2005).

Eric Villemonte de la Clergerie. "From metagrammars to factorized TAG/TIG parsers". In: *Proceedings of the Ninth International Workshop on Parsing Technology - Parsing '05* (2005).

Eric Villemonte de la Clergerie. “Building factorized TAGs with meta-grammars”. In: *The 10th International Conference on Tree Adjoining Grammars and Related Formalisms - TAG+10* (2010).

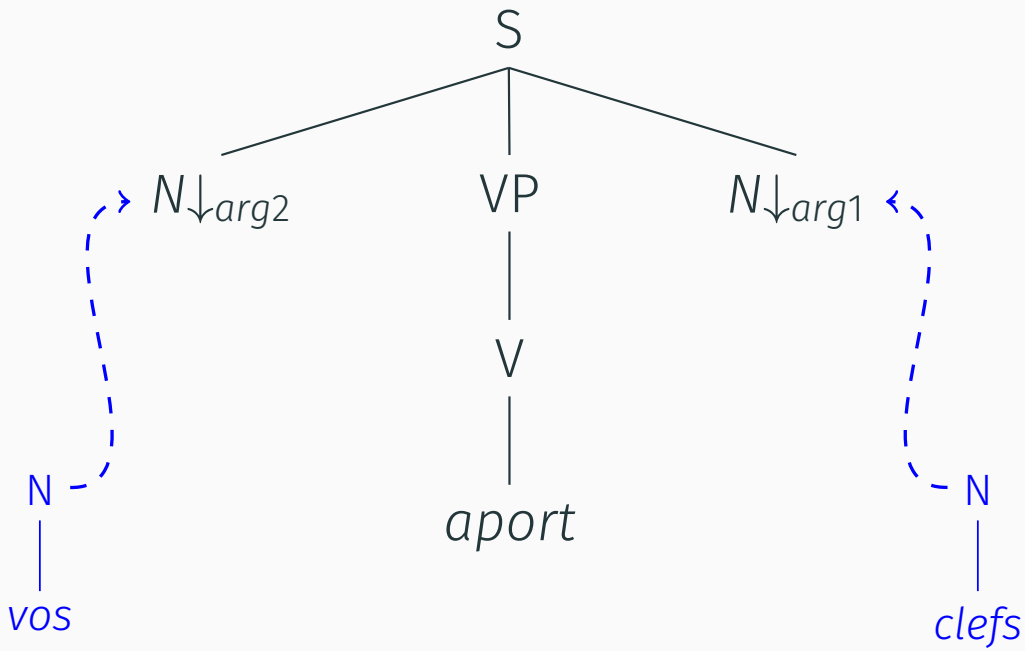
Architecture of the Alpi system

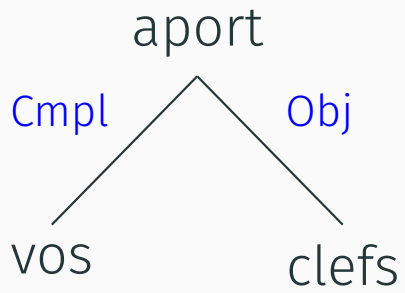


Substitution

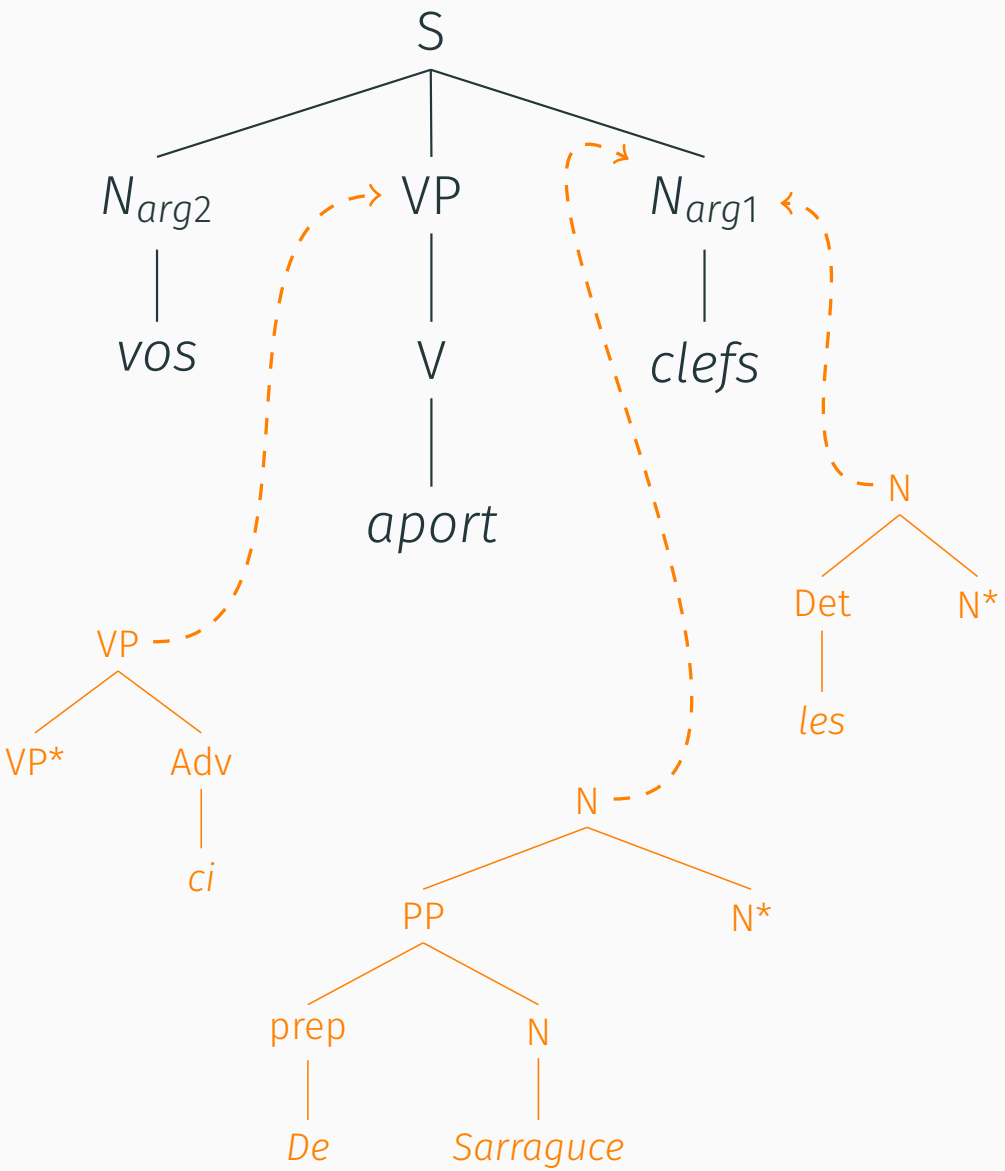
Example of a derivation

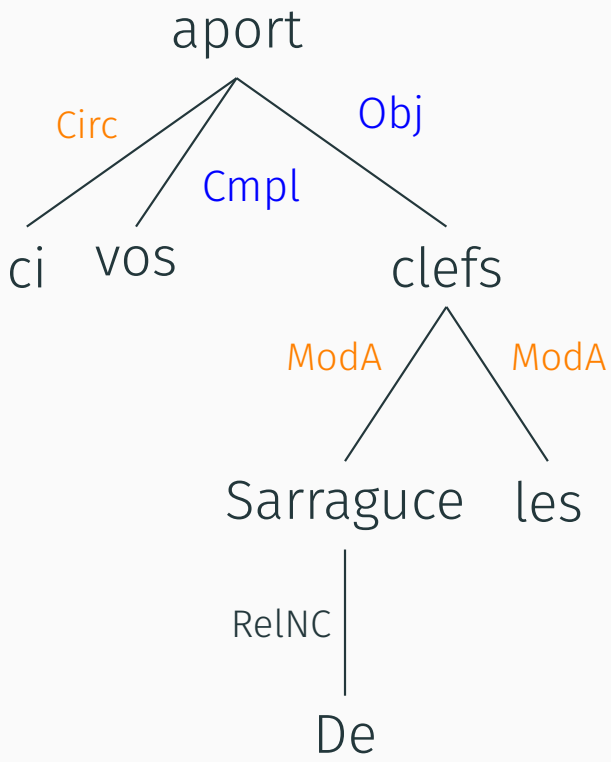
”De Sarraguce ci vos aport les clefs”, *Chanson de Roland*, trad. *[I] bring you the keys of Zaragoza here.*





Adjonction


S
N_{arg2}
VP
N_{arg1}
vos
V
clefs
aport
VP
VP*
Adv
ci
N
Det
N*
les
N
PP
N*
prep
N
De
Sarraguce


aport
Circ
Cmpl
Obj
ci
vos
clefs
ModA
ModA
Sarraguce
les
RelNC
De

Classes of the metagrammar

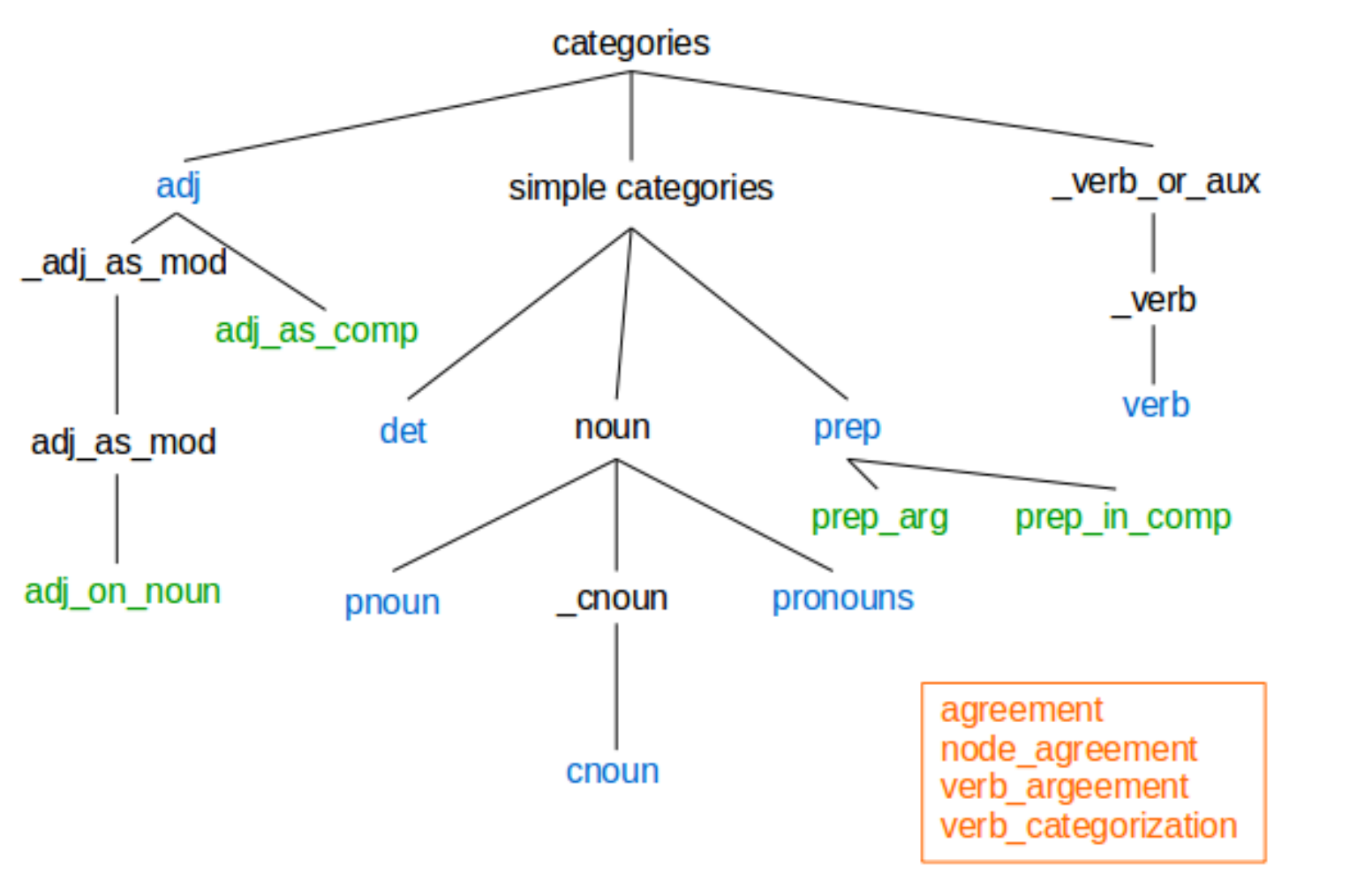


Figure 8: Some classes of the metagrammar

Classes of the metagrammar

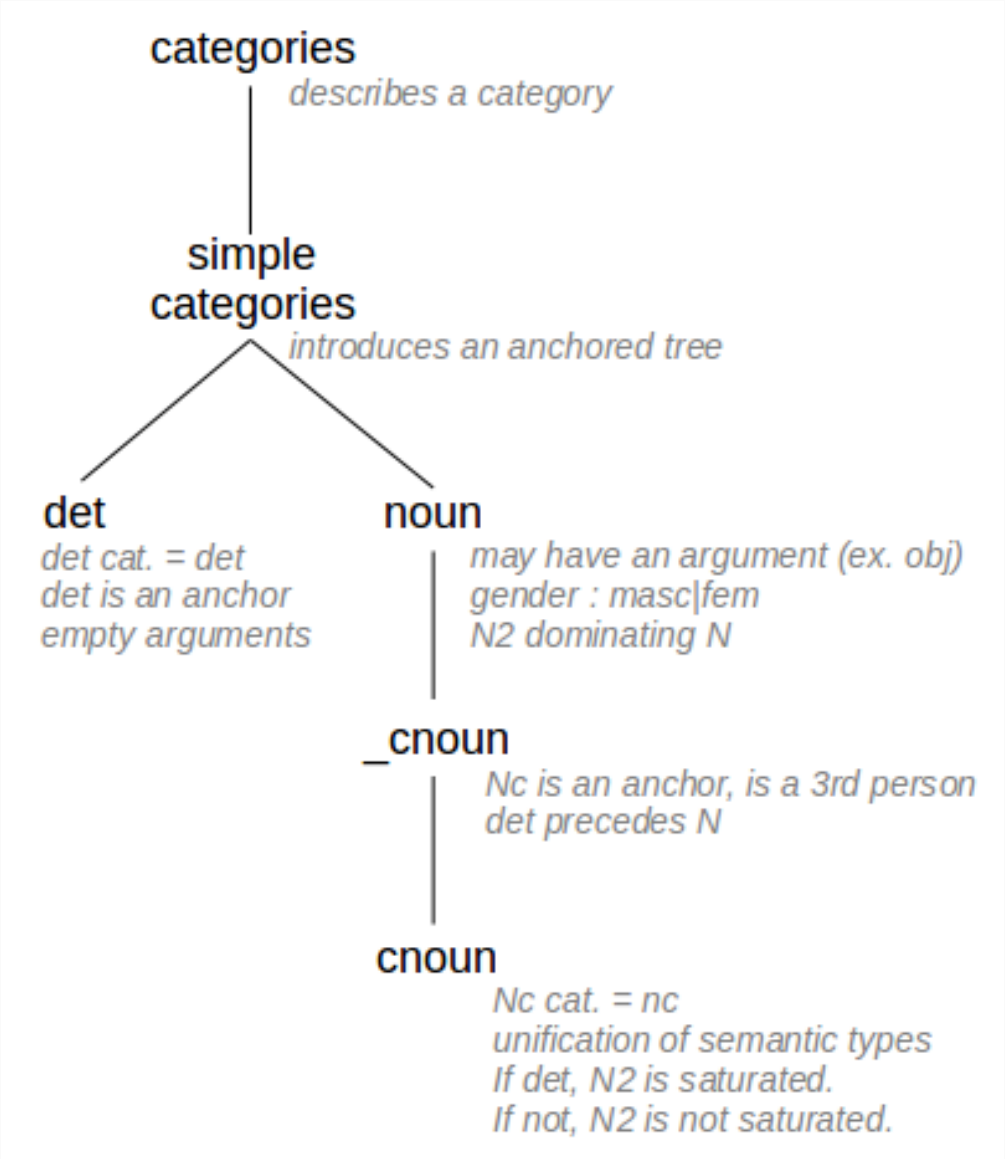
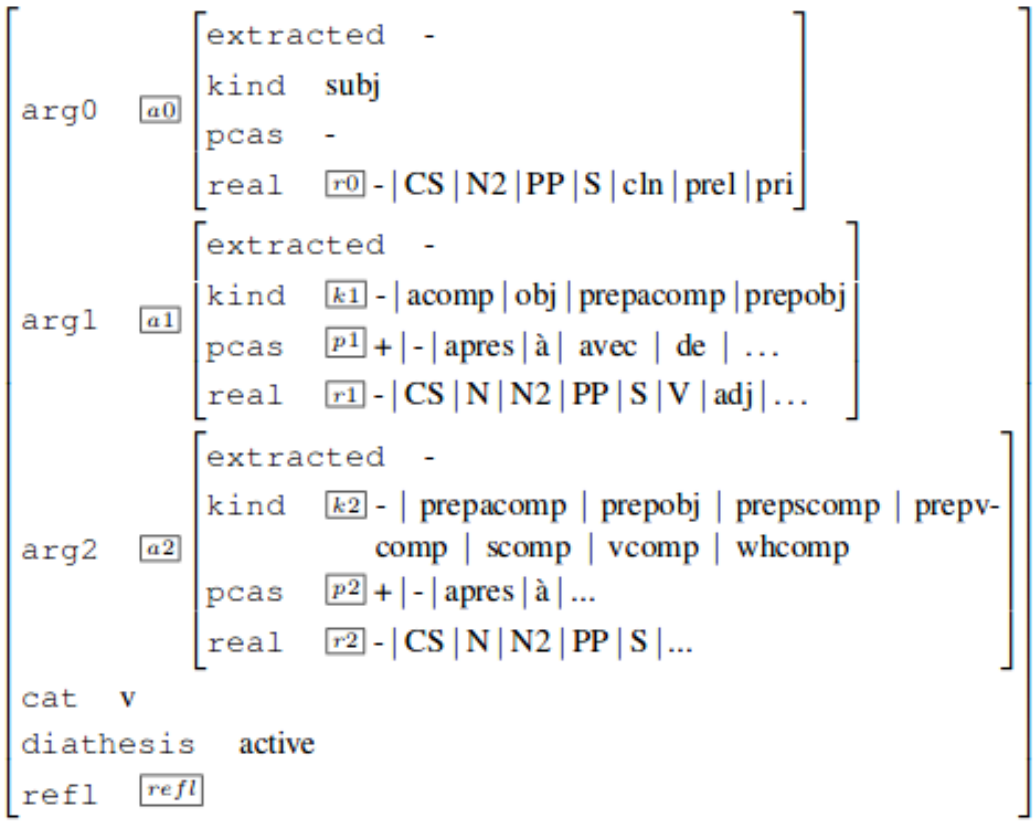
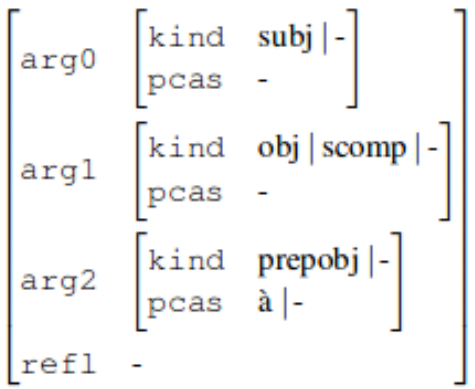


Figure 9: Determiner and common noun

Hypertag



(a) for tree #198



(b) for "to promise"

Figure 2: Grammar and lexicon hypertags

Figure 10: from Villemonte de la Clergerie (2010)

Facets I

Adding facets to filter analysis according to external characteristics of texts

ex. ”Deu Samuel apela”, which is the subject?

Example of a facet

- Criteria: **date**
- until the 13th c., OSV order is only possible with a subject pronoun (Schøsler 1984)

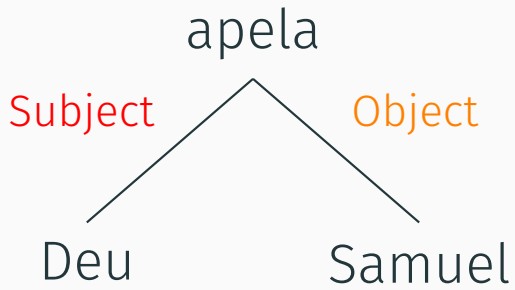


Figure 11: SOV analysis

Facets II

Example of a facet

- Criteria: **dialect**
- object clitics are usually before the verb, but not necessarily in Picard

ex. from *Escouffe* (v. 4 954-55), as cited by Buridant (2000)

Et,	se viaus non,	prestés **me**	huimais	L'ostel
(And,)	otherwise,	give me	for today	shelter

Cleft sentences

	SRCMF	BFM
12th c.	3 (50%)	22 (45%)
14th c.	-	15 (68%)

Figure 12: Texts with cleft sentences