Analyse syntaxique de l'ancien français: quelles propriétés de la langue influent le plus sur la qualité de l'apprentissage?

Gaël Guibon (1), Isabelle Tellier (1,2)
Sophie Prévost (1), Matthieu Constant (3) and Kim Gerdes (2,4)

(1) Lattice CNRS
(2) université Paris 3 - Sorbonne Nouvelle
(3) université Paris-Est, LIGM
(4) LPP CNRS
E-mails: gael.guibon@gmail.com, isabelle.tellier@univ-paris3.fr, matthieu.constant@u-pem.fr, sophie.prevost@ens.fr, kim@gerdes.fr

23 juin 2015

Objectifs

1. L'exploration de corpus peut-elle se faire à base d'apprentissage automatique?
2. Quels sont les résultats de l'analyse syntaxique en dépendance sur un corpus très hétérogène?
3. Quelles sont les propriétés qui importent quand on entraine un programme sur des textes et qu'on les teste sur d'autres ?

Syntactic Reference Corpus of Medieval French (SRCMF)

- Premier **corpus arboré de l'ancien français**
- Issu d'un **projet ANR et DGF** de mars 2009 à février 2012
- **15 textes** issus de la BFM et du NCA
- **Étiquetés** en morpho-syntaxe et syntaxe
- **Différentes variations** (morphologie, genre, époque, etc.)

Caractéristiques des textes utilisés

- **10** textes
- Relativement **petits** : de 1388 à 41 305 mots
- En **vers** ou/et en **prose**
- **Espacés dans le temps** : Du 10ème au 13ème siècle
- **4 dialectes** : normand, anglo-normand, champenois, picard
- **4 domaines** : littéraire, historique, religieux, didactique

Textes utilisés

Texte	Date	Mots	Clauses	Forme
Vie Saint **Légier**	Fin 10e s.	1388	192	vers
Vie de Saint **Alexis**	1050	4804	562	vers
Chanson de **Roland**	1100	28 766	3857	vers
Lapidaire *en prose*	Milieu 12e s.	4708	468	prose
Yvain, Chretien de Troyes	1177-1181	41 305	3788	vers
La **Conqueste** *de Constantinople* de Robert de Clari	$\geq$ 1205	33 534	2308	prose
Queste del Saint **Graal**	1220	40 417	3078	prose
Aucassin *et Nicolete*	Fin 12e s.-début 13e s.	9844	1101	**vers** & **prose**
Miracles de Gautier de **Coinci**	1218-1227	17 360	1422	vers
Roman de la **Rose** de Jean de Meun	1269-1278	19 339	1449	vers

Table: Textes du SRCMF utilisés dans nos expériences

Particularités de l'ancien français

- Importante variabilité des formes ...
- ... mais évaluable sur un seul texte
- Normalisation des formes écartée

POS \ Corpus	French Treebank	Yvain
Nom propre	1	**1.25**
Nom commun	1.31	1.31
Verbe (infinitif)	1	**1.10**
Verbe (conjugué)	2.48	3.15
Déterminant	1.06	2.21
Adjectif	1.63	1.68
Adverbe	1.01	1.40
Nombre moyen de formes par lemme	1.57	2.25

Forte variabilité

Variation d'un nom propre

`Yvain Yvein` + la marque du sujet principal : `Yvains Yveins`

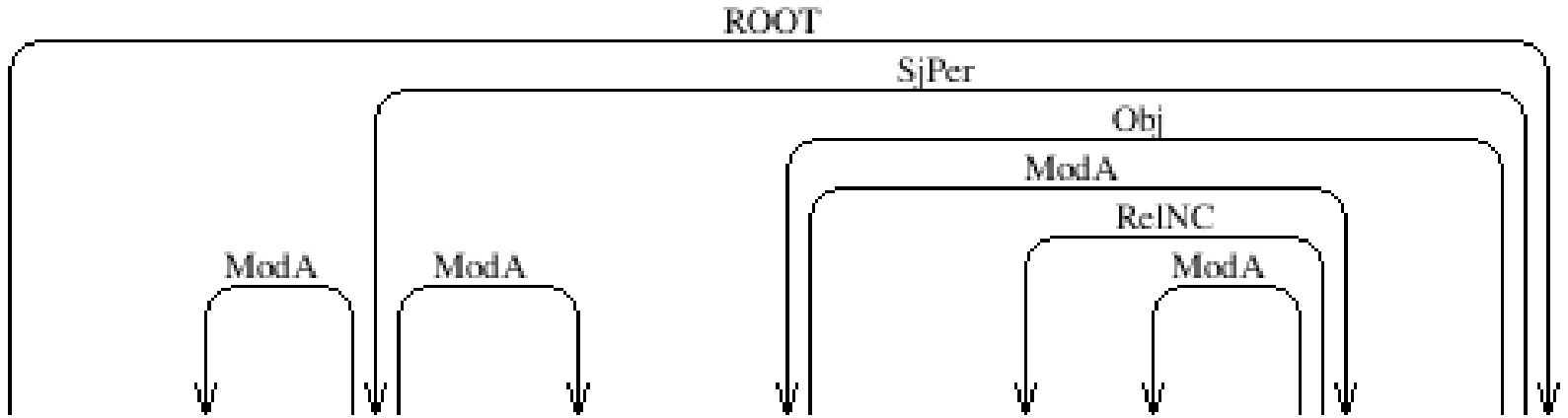
Variation d'un nom commun

`vilains vileins vilainne`

Souplesse dans l'ordre des mots

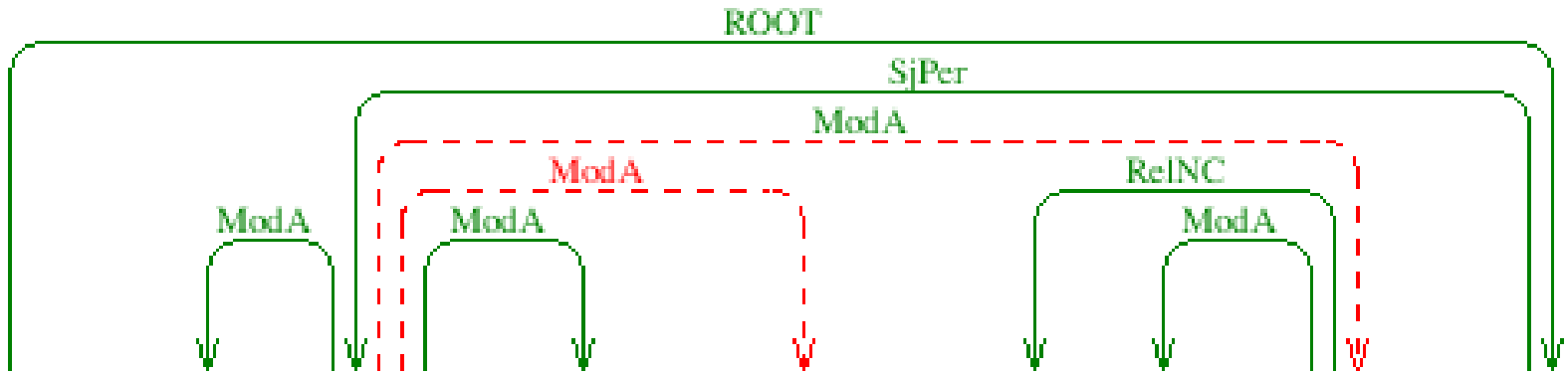
Ordre SOV

[Li quens Rollant]Subj [Gualter de l'Hum]Obj [apelet]Verb
Le comte Roland appelle Gautier de l'homme



	Li	quens	Rollant	Gualter	de	l'	Hum	apelet
0	1	2	3	4	5	6	7	8
<ROOT>	_	_	_	_	_	_	_	_
	DETdet	NOMcom	NOMpro	NOMpro	PRE	DETdet	NOMcom	VERcjg

Difficultés de l'analyse syntaxique



	Li	quens	Rollant	Gualter	de	l'	Hum	apelet
0	1	2	3	4	5	6	7	8
<ROOT>	_	_	_	_	_	_	_	_
	DETdet	NOMcom	NOMpro	NOMpro	PRE	DETdet	NOMcom	VERcjg

Pour le contexte actuel...

- **Très peu d'apprentissage automatique** sur l'ancien français
- Corpus bien **plus hétérogène et variable** que le français contemporain

...et l'intérêt de l'analyse en dépendance

- **Complète** les connaissances linguistiques
- **Compare** l'efficacité des méthodes actuelles sur un corpus de langue ancienne
- **Évalue** la difficulté et l'impact de chaque caractéristique
- **Mesure** l'influence des relations linguistiques entre textes sur les résultats des expériences

Exactitude - Accuracy

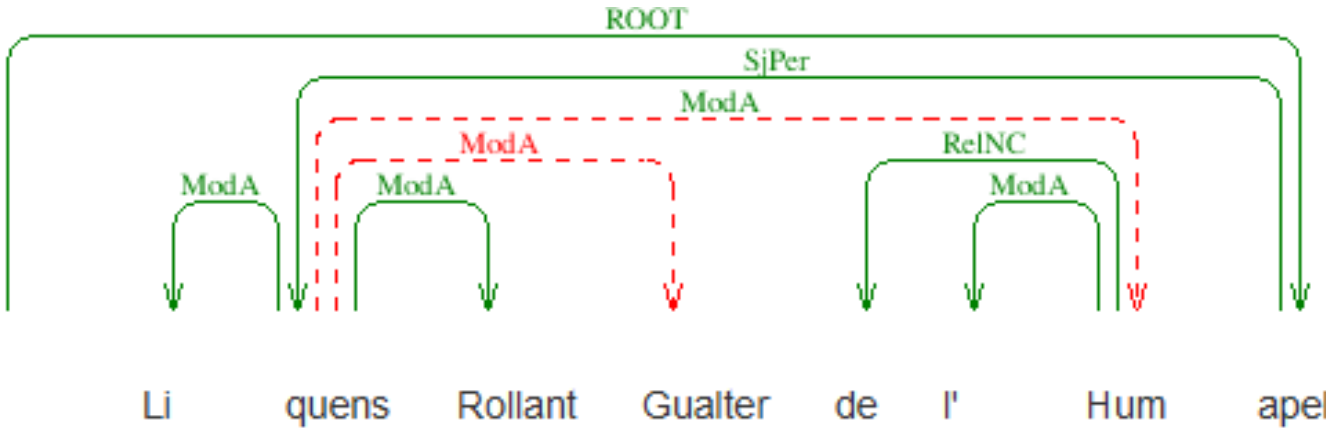
Pourcentage d'**étiquettes correctes**

UAS - Unlabelled Attachment Score

Pourcentage de **gouverneurs corrects**

LAS - Labelled Attachment Score

Pourcentage de gouverneurs corrects **bien étiquetés syntaxiquement**



Précédents travaux

- LREC 2014 (Achim Stein)[4] et TLT 2014[2]
- Méthode principale : **apprentissage croisé**
- Nous proposons un autre angle d'analyse du corpus : **les méta-données**

Protocole

1. **Lemmatisation** : *TreeTagger* avec fichiers de paramètres (Achim Stein)
2. **Étiquetage morpho-syntaxique** : CRF sous *Wapiti* (Thomas Lavergne [3])
3. **Analyse syntaxique en dépendance** : *Mate* (Bernd Bohnet [1])

Approche lexicale: méthode

1. **5 textes sélectionnés** pour leur taille et particularité (Aucassin)
2. Chaque corpus est **divisé en sous-corpus** d'environ 16 000 mots (sauf Aucassin)
3. Chaque modèle de sous-corpus est **appliqué sur les autres** (apprentissage croisé)

Approche lexicale
Objectifs

Question

La variabilité lexicale est-elle **déterminante lors de l'apprentissage?**

Corpus	unités distinctes	mots	mots/ unités	répétitions par lemme	taille moy. des clauses
Aucassin	1962	9844	5,02	2.38	9
Conq.	3424	33 534	9,79	**2.9**	**14**
Graal	3874	40 417	**10,43**	2.56	12
Roland	4304	28 766	6,68	2.00	7
Yvain	**5040**	**41 305**	8,19	2.17	10

Approche lexicale
Résultats

- **Corrélation insuffisante** entre mots connus et taux d'exactitude
- Le vocabulaire partagé **ne permet pas de prédire la qualité** de l'étiquetage morpho-syntaxique
- **Taux de répétition des mots** n'explique pas tous les résultats

Train (Roland) \ Test	Graal	Conq.
Exactitude	82.66	**84.13**
Mots connus	**60.10**	54.14
Vocabulaire partagé	**9.07%**	6.23%

Leave One Out
Objectifs et méthode

Méthode

Apprentissage sur les 9 textes; test sur le dernier

Buts

- Disposer de **corpus d'apprentissage plus grands**
- Repérer les **sources d'hétérogénéité** (*fractures*)
- **Identifier les lexiques et syntaxes** éloignées des autres

Leave One Out
Résultats

	Mots connus	Exactitude	UAS	LAS
Moy. 9 autres	84.97%	90.33%	84.78%	74.70%
9 sur Legier	40.56%	66.64%	61.74%	46.04%

- *Vie Saint Légier* se démarque des autres
- Jusque **15.7% de baisse** entre UAS et LAS
- **Le plus petit texte** (1388 mots) avec le moins d'unités distinctes (578)
- Le texte **le plus ancien**
- Un grand corpus d'entrainement ne signifie pas une bonne reconnaissance
- Confirme la nécessité de **sous-corpus d'apprentissage orientés** autrement

La forme du texte
Objectifs

Questions

- Y a-t-il une forme plus difficile à reconnaître?
- Ces deux formes sont-elles proches?
- Les résultats représentent-ils ces formes?

La forme du texte

Informations générales

- 2 formes: **vers** et **prose**
- **Majorité de vers** en ancien français jusqu'au 13ème siècle
- Moins de diversité lexicale pour la prose

Corpus	Nb de mots	Nb unités
Prose [entrainement]	41 910	4320
Vers [entrainement]	41 907	6840
Prose [test]	36 749	4370
Vers [test]	34 478	4417

La forme du texte
Résultats

Train \ Test		Prose [test]	Vers [test]
prose	UAS	**85.47%**	76.33%
	LAS	**74.96%**	62.96%
	ACC	**91.36%**	83.61%
vers	UAS	**83.12%**	82.79%
	LAS	**71.52%**	71.40%
	ACC	90.06%	**90.78%**

Il ressort que:

- Il semble préférable d'**utiliser le vers en entrainement**
- La prose semble **plus simple à analyser**
- Les **tests sur Aucassin** sont meilleurs avec un modèle de prose
- La forme ne se reflète pas sur l'analyse en dépendance

Évolution langagière
Objectifs

Questions

- Peut-on remarquer l'évolution langagière sur:
  - La variabilité des formes
  - La diversité lexicale
  - La complexité syntaxique

Informations

- Uniquement sur des texte du **12ème et 13ème siècle**
- **Clauses plus longues** au 13ème siècle (12 mots par clause en moyenne)

Évolution langagière

Résultats

Train \ Test		12ème Siècle [test]	13ème Siècle [test]
12ème Siècle	UAS	66.07%	**66.72%**
	LAS	50.71%	**52.89%**
	ACC	71.88%	**75.19%**
13ème Siècle	UAS	72.31%	**81.15%**
	LAS	57.57 %	**69.52%**
	ACC	78.28%	**87.32%**

- **Proximité** des sous-corpus en nombre de mots et lexiques communs.
- **Pas de relation** entre morpho-syntaxe et date de parution
- Varier les corpus d'entrainement de siècles différents **n'améliore pas l'analyse en dépendance**

Conclusions

- Apprentissage automatique **utilisable pour de l'exploration de corpus** hétérogène...
- ...mais à condition de **choisir les bonnes méta-données**
- **Pas de propriété dominante**
- Nécessité de prendre ensemble variété lexicale et taux de répétition
- **Taille en entrainement pas déterminante**

Perspectives

- **Vérification** de la variabilité des formes plus élevée pour le vers
- Explorer davantage **en fonction des méta-données** : le dialecte et le domaine donnent des résultats plus pertinents
- Quantification exacte de l'**impact de la longueur des clauses**


Bernd Bohnet.
Very high accuracy and fast dependency parsing is not a contradiction.
In *The 23rd International Conference on Computational Linguistics (COLING 2010)*, Beijing, China, 2010.


Gaël Guibon, Isabelle Tellier, Matthieu Constant, Sophie Prévost, and Kim Gerdes.
Parsing poorly standardized language dependency on old french.
In *13th Treebank and Language Theory (TLT)*, 2014.


Thomas Lavergne, Olivier Cappé, and François Yvon.
Practical very large scale CRFs.
In *Proceedings the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 504–513. Association for Computational Linguistics, July 2010.

Achim Stein.
Parsing heterogeneous corpora with a rich dependency grammar.
In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may 2014. European Language Resources Association (ELRA).

Corpus de test	UAS moyen	LAS moyen	UAS mots inconnus / connus	LAS mots inconnus / connus
Aucassin	75.71 %	61.92 %	67.73% — 78.61%	44.41% — 63.62%
Roland	76.54 %	58.77 %	63.76% — 77.28%	44.10% — 60.36%
Graal	77.62 %	63.94 %	71.53% — 82.74%	49.37% — 69.07%
Yvain	71.14 %	56.03 %	70.53% — 80.75%	46.41% — 65.69%
Conq.	**77.67 %**	**65.67 %**	62.98% — 79.08%	41.24% — 66.49%

Table: Moyenne des résultats de l'analyse en dépendance par corpus de test

XP	Mots inconnus / Mots connus	Exactitude	UAS	LAS
9 sur Alexis	20.05% — 79.95% /	85.91 % 79.01% — 87.12%	81.10 % 72.17% — 83.34%	69.86% 55.56% — 73.44%
9 sur Aucassin	13.92% — 86.08% /	91.21 % 83.58% — 92.45%	86.87 % 74.74% — 88.84%	77.20% 59.56% — 80.06%
9 sur Clari	15.92% — 84.08% /	92.55 % 88.40% — 93.33%	87.12 % 79.52% — 88.63%	78.35% 66.12% — 80.66%
9 sur Coinci	12.96% — 87.04% /	89.72 % 75.51% — 91.84%	79.91 % 66.89% — 81.85%	69.27% 49.64% — 72.20%
9 sur Lapidaire	17.99% — 82.01% /	88.89 % 77.69% — 91.35%	84.88 % 74.26% — 87.21%	75.57% 55.61% — 79.95%
9 sur Legier	59.44% — 40.56% /	66.64 % 52.58% — 76.24%	61.74 % 54.53% — 66.67%	46.04% 33.93% — 54.30%
9 sur Graal	7.17% — 92.83% /	93.58 % 85.51% — 94.19%	89.51 % 80.79% — 90.19%	80.82% 66.06% — 81.97%
9 sur Roland	22.70% — 77.30% /	90.74 % 85.36% — 92.32%	87.91 82.39% — 89.54%	76.23 % 63.82% — 79.88%
9 sur Rose	14.04% — 85.96% /	90.74 % 80.63% — 92.96%	81.56 % 68.89% — 83.64%	70.94% 52.36% — 73.98%
9 sur Yvain	10.55% — 89.45% /	89.61 % 86.11% — 90.02%	84.19 73.76% — 85.42%	74.08% 58.08% — 75.97%

Table: Resultats des tests sur un corpus par le modèle appris sur les neuf autres

Train \ Test		Prose [test]	Vers réduit [test]
prose	**UAS**	**85.47%**	76.33%
	LAS	**74.96%**	62.96%
	ACC	**91.36%**	83.61%
	Mots inconnus / Mots connus	16.49% — 83.51%	21.26%— 78.74%
	Lexique différent / commun	57.02% — 42.98%	77.05% — 22.95%
	UAS Mots inconnus / Mots connus	73.76% — 87.78%	65.87% — 79.15%
	LAS Mots inconnus / Mots connus	55.48% — 78.81%	46.37% — 67.44%
	ACC Mots inconnus / Mots connus	77.33% — 94.14%	76.78% — 85.46%
vers réduit	**UAS**	**83.12%**	82.79%
	LAS	**71.52%**	71.40%
	ACC	90.06%	**90.78%**
	Mots inconnus / Mots connus	18.81% — 81.19%	14.03% —85.97%
	Lexique différent / commun	66,47% — 33,53%	42.52% — 57.48%
	UAS Mots inconnus / Mots connus	73.43% — 85.37%	72.39% — 84.49%
	LAS Mots inconnus / Mots connus	55.45% — 75.24%	55.62% — 73.98%
	ACC Mots inconnus / Mots connus	81.02% — 92.15%	84.13% — 91.86%

Train \ Test		12ème Siècle [test]	13ème Siècle [test]
12ème Siècle	**UAS**	66.07%	**66.72%**
	LAS	50.71%	**52.89%**
	ACC	71.88%	**75.19%**
	Mots inconnus / Mots connus	30.88% — 69.12%	32.33% — 67.67%
	Lexique différent / commun	58.13% — 41.87%	62.52% — 37.48%
	UAS Mots inconnus / Mots connus	51.41% — 72.62%	52.00% — 73.75%
	LAS Mots inconnus / Mots connus	33.24% — 58.52%	33.82% — 62.00%
	ACC Mots inconnus / Mots connus	55.83% — 79.05%	57.88% — 83.46%
13ème Siècle	**UAS**	72.31%	**81.15%**
	LAS	57.57 %	**69.52%**
	ACC	78.28%	**87.32%**
	Mots inconnus / Mots connus	22.52% — 77.48%	17.65% — 82.35%
	Lexique différent / commun	70.17% — 29.83%	52% — 48%
	UAS Mots inconnus / Mots connus	59.44% — 76.06%	68.51% — 83.86%
	LAS Mots inconnus / Mots connus	39.32% — 62.87%	50.11% — 73.69%
	ACC Mots inconnus / Mots connus	62.20% — 82.95%	67.78% — 91.51%