Séminaire Délicortal

LE PROJET PHRASÉOROCHE

LIDILEM, UGA

Adam Renwick

QUI ?

- **M1 : FLE, University of Canterbury (NZ)**
- **M2 : LTMT, Lyon 2**
- **Thèse : *Recommandation et implantation : le cas des termes de la spatiologie***
- **ATERs**
- **Création de corpus *ad hoc* : tweets, *Les Simpson, CRTTMed,* Canards, -INGs, thèses, PubMed…**
- **Postdoc : Projet PhraséoRoche**

POURQUOI DES CORPUS ?

- **Preuves concrètes de l'emploi réel de la langue**
  - **Étudier, quantifier, développer + tester des hypothèses**
- **Scientificité !**
- **Point de vue contrastif avec la physique**

PHRASÉOROCHE

- **Partie de l’action « Phraséologie en diachronie »**
  - **PhraséoCorr**
  - **Phraséologie et genres textuels : le cas du roman médiéval**
  - **Phraséo 13-18 => PhraséoRoChe**

PHRASÉOROCHE

- Phraséologie dans les Romans de Chevalerie = PhraséoRoChe
  - But central comprendre le rôle que la phraséologie joue dans la structuration des genres textuels en utilisant les méthodes de la linguistique de corpus outillée
  - Unités préconstruites dans un type de texte
  - Comment faire ? Corpus !
  - Diachronie longue : 13e - 17e

ÉQUIPE

- O Kraif, J Sorba, A Renwick, I Fabry (LIDILEM)
- C Denoyelle, C Lignereux, P Mounier, M Basset (Litt&Arts)
- G Souvay (ATILF)

SURVOL...

```
<?xml version='1.0' encoding='utf-8'?>
<corpus><doc><meta>authors  (c) Équipes SRCMF and BFM
bibl    (c) Équipes SRCMF and BFM, CNRS/ENS-LYON 2009-2015
date
discipline  littérature
firstPage
genre   roman
language    fr
languageId  fr
lastPage
monogr.title    SRCMF ArtusdeBretagne
publisher
title   Artus de Bretagne
volume
year
</meta><text><p id="p1"><s id="s1" newdoc_id="ArtusdeBretagne" sent_id="1">1   Apréz   APRÈS   ADP _   _   3   case    _   _
2   la  LE  DET _   _   3   det _   _
3   mort    MORT1   NOUN    _   _   45  obl _   _
4   le  LE  DET _   _   6   det _   _
5   bon BON ADJ _   _   6   amod    _   _
6   roy ROI1    NOUN    _   _   3   appos   _   _
7   Artus   ARTHUR  PROPN   _   _   6   flat    _   _   ,
8   qui QUI1    PRON    _   _   12  nsubj   _   _
9   tant    TANT    ADV _   _   12  advmod  _   _
10  fu  ÊTRE1   AUX _   _   12  cop _   _
11  nobles  NOBLE1  ADJ _   _   6   amod    _   _
12  roys    RAI1|ROI1   NOUN    _   _   3   acl:relcl   _   _
13  et  ET  CCONJ   _   _   14  cc  _   _
14  gentilz GENTIL1 ADJ _   _   12  conj    _   _   ,
15  et  ET  CCONJ   _   _   16  cc  _   _
16  entour  ENTOUR  ADV _   _   12  conj    _   _
17  qui QUI1    PRON    _   _   18  nsubj   _   _
18  fu  ÊTRE1   VERB    _   _   3   acl:relcl   _   _
19  et  ET  CCONJ   _   _   20  cc  _   _
20  regna   RÉGNER  VERB    _   _   18  conj    _   _
21  toute   TX2 DET _   _   23  det _   _
22  la  LE  DET _   _   23  det _   _
23  noblece NOBLESSE    NOUN    _   _   18  obj _   _
24  de  DE  ADP _   _   27  case    _   _
25  toute   TX2 DET _   _   27  det _   _
26  la  LE  DET _   _   27  det _   _
27  chevalerie  CHEVALERIE  NOUN    _   _   23  nmod    _   _
28  de  DE  ADP _   _   31  case    _   _
29  tout    TX2 DET _   _   31  det _   _
30  le  LE  DET _   _   31  det _   _
31  monde   MONDE1  NOUN    _   _   27  nmod    _   _   ,
32  si  SI4 ADV _   _   55  advmod  _   _
33  comme   COMME   SCONJ   _   _   34  mark    _   _
34  furent  ÊTRE1   VERB    _   _   55  advcl   _   _
35  Gauwain GAUVAIN PROPN   _   _   55  nsubj   _   _   ,
36  Lancelot    LANCELOT    PROPN   _   _   35  flat    _   _
37  et  ET  CCONJ   _   _   40  cc  _   _
38  maint   MAINT   DET _   _   40  det _   _
39  autre   AUTRE   ADJ _   _   40  amod    _   _
40  chevalier   CHEVALIER   NOUN    _   _   23  conj    _   _
41  preu    ADPUX   ADJ _   _   35  amod    _   _
42  et  ET  CCONJ   _   _   43  cc  _   _
43  bon BON ADJ _   _   41  conj    _   _   ,
44  il  IL  PRON    _   _   45  expl    _   _
45  ot  AVOIR1  VERB    _   _   0   root    _   _
46  en  EN1 ADP _   _   47  case    _   _
47  Bretagne    BRETAGNE    PROPN   _   _   45  obl _   _
48  un  UN  DET _   _   49  det _   _
49  duc DUC1    NOUN    _   _   45  obj _   _   :
50  preudomme   PRUDHOMME   ADJ _   _   49  amod    _   _
51  fu  ÊTRE1   AUX _   _   55  cop _   _
52  sus SUS ADP _   _   54  case    _   _
53  tous    TX2 DET _   _   54  det _   _
54  autres  AUTRE   PRON    _   _   55  obl _   _   ,
55  vertueus    VERTUEUX    ADJ _   _   49  amod    _   _
56  et  ET  CCONJ   _   _   57  cc  _   _
57  puissans    PUISSANT    VERB    _   _   55  conj    _   _   ,
58  riches  RICHE   ADJ _   _   18  conj    _   _
59  et  ET  CCONJ   _   _   74  cc  _   _
60  fors    FORS    ADJ _   _   55  conj    _   _
```

TEXTES

- **Genre en diachronie longue : Roman de Chevalerie**
  - **Littré [article chevalerie] Romans de chevalerie : romans où sont décrits les exploits, les caractères, les mœurs, les amours des chevaliers tels que l'imagination les avaient idéalisés.**
  - **Mais c'est pas aussi simple que ça : Vielliard, F. 2007.**
    - **Lien avec des chansons de geste, romans courtois en vers**
  - **On prend des textes en *prose***

TEXTES

- **Identification, via bibliographies, recueils…**
- **SRCMF, BFM, Agrégation**
- **FranText : 37 *romans* du XVIIe**
- **XVIIe : 89 ouvrages**
- **=> Délimitation : place des réécritures (Vers -> Prose), éditions ?**

TEXTES

- Manuscrit, manuscrit scanné ? Livre scanné ? GoogleBooks, BNF, Gallica…
- Rendre de nouveaux textes disponibles
- Délimitation en trois phases et équilibre des sous-corpus
  - AF (avant 1300) : 9 textes ; 0,8 M mots
  - MF (1300-1550) : 18 textes ; 1,9 M mots
  - FPC (1550-1650) : 13 textes ; 2,1 M mots

TEXTES

- Manuscrit, manuscrit scanné ? Livre scanné ? GoogleBooks, BNF, Gallica…
- Rendre de nouveaux textes disponibles
- Délimitation en trois phases et équilibre des sous-corpus
  - AF (avant 1300) : 9 textes ; 0,8 M mots => 400k mots
  - MF (1300-1550) : 18 textes ; 1,9 M mots => 820k mots
  - FPC (1550-1650) : 13 textes ; 2,1 M mots => 615k mots
- Carottages de 50k mots OU totalité du texte

BUT

- Analyser les textes via le Lexicoscope (Kraif & Diwersy (2012)
- http://phraseotext.univ-grenoble-alpes.fr/lexicoscope_2.0/
- Textes XML-TEI
  - Lemmatisation
  - Étiquetage morphosyntaxique
  - Analyse des dépendances

OUTILS

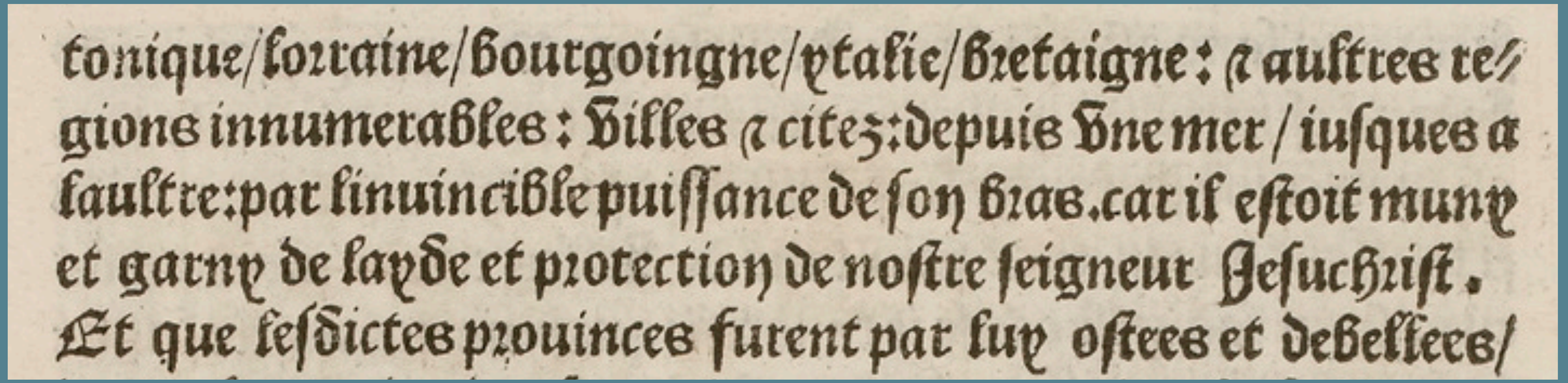
- 1. Transkribus : Reconnaissance optique de caractères (ROC /OCR)
- https://readcoop.eu/transkribus/?sc=Transkribus
- Gratuit
- OCR sur serveur, téléchargement d’un dossier contenant les résultats

TRANSKRIBUS

- 14 modèles français
  - Manuscrit vs imprimé
  - Pré-entraînés sur des corpus hétérogènes
    - Annuaires des propriétaires 1898-1923
    - Recueil des édits, déclarations, lettres-patentes etc enregistrés au parlement de Flandres : XVIIIe
    - Journaux français - fin XVIIe - XXe

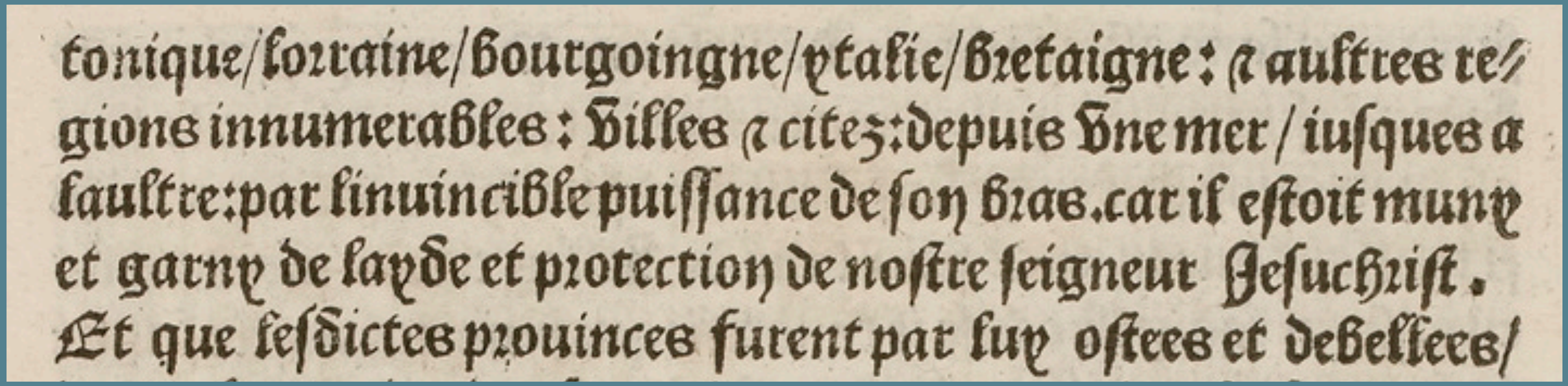
TRANSKRIBUS

- **Plusieurs modèles pré-entraînés pour le français, ancien français, français classique… mais le(s)quel(s) choisir ?**


tonique/lorraine/bourgoingne/ytalie/bretaigne: & aultres re/
gions innumerables: villes & citez: depuis vne mer/iusques a
faultre: par linuincible puissance de son bras. car il estoit muny
et garny de layde et protection de nostre seigneur Jesuchrist.
Et que lesdictes prouinces furent par luy ostees et debellees/

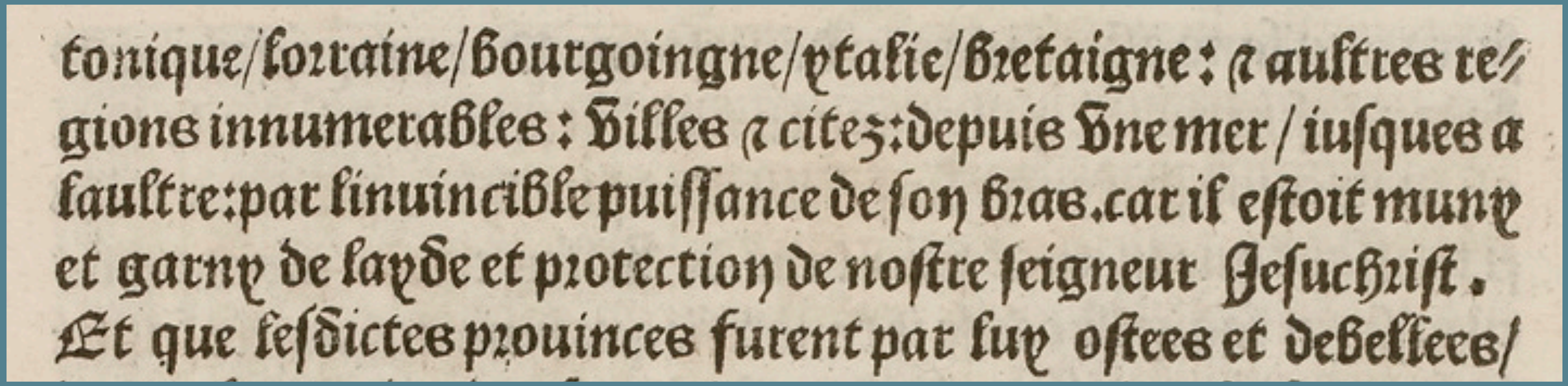
Anonyme, 1527. *Cronique et histoire faicte et composée par reverend pere en dieu Turpin*. Paris, P. Vidoue. f°ii.r.

tonique/ forraine/ bourgoingne/ ptalie/ bretaigne: s auftres te-
gions innumeraßses: Viffes s citez: depuis vne mer/ iusques a
faufstre: par kinuincibse puissance de soy bias car il estoit muny
et garny de sayde et protectioy de nostre seigneur Besuchrist.
Et que fesdictes prouinces furent par suy ostees et debeffees/


tonique/ forraine/ bourgoingne/ ytalie/ bretaigne: ⁊ aultres re-
gions innumerables: villes ⁊ citez: depuis vne mer/ iusques a
faultre: par linuincible puissance de son bras. car il estoit muny
et garny de layde et protection de nostre seigneur Jesuchrist.
Et que lesdictes prouinces furent par luy ostees et debellees/

Anonyme, 1527. *Cronique et histoire faicte et composée par reverend pere en dieu Turpin*. Paris, P. Vidoue. f°ii.r.

tonique/ forraine/ bourgoingne/ ptalie/ bretaigne: s auftres te-
gions innumeraßses: Viffes s citez: depuis vne mer/ iusques a
fauftre: par kinuincibse puissance de soy bias car il estoit muny
et garny de sayde et protectioy de nostre seigneur Besuchrist.
Et que fesdictes prouinces furent par suy ostees et debeffees/


tonique/ forraine/ bourgoingne/ ytalie/ bretaigne: ⁊ aultres re-
gions innumerables: villes ⁊ citez: depuis vne mer/ iusques a
faultre: par linuincible puissance de son bras. car il estoit muny
et garny de layde et protection de nostre seigneur Jesuchrist.
Et que lesdictes prouinces furent par luy ostees et debellees/

Anonyme, 1527. *Cronique et histoire faicte et composée par reverend pere en dieu Turpin*. Paris, P. Vidoue. f°ii.r.

STAGIAIRES

- Stages d’excellence 2022
  - La recherche n’est pas toujours la gloire d’envoler en colloque
  - La réalité = temps devant l’écran, le texte
- Accents : pas toucher sauf pour désambigüiser (ou/où, a/à, des/dès…)
- Apostrophe : modernisation (Iay -> J’ay, quil -> qu’il…)
- Différenciation : j/i, u/v
- Guillemets : français, anglais doubles + simples pour 3 niveaux
- Pourquoi ? Faciliter la lemmatisation, POS-tagging et l’analyse des dépendances

STAGIAIRES

- Balisage XML des chapitres, paragraphes, titres de chapitre, pagination, romains/italiques, vers…
- Pourquoi ?
  - Préserver la structure du document
  - Permettre de remonter au document source :
    - Comparaison entre éditions d’un même ouvrage
    - Citations…

RELECTURES

- ≥ 2 relectures par texte
- Pourquoi : ?
  - L’erreur est humaine
  - Y a-t-il *toujours* des scories ?????
  - On vise la perfection, main on est humains

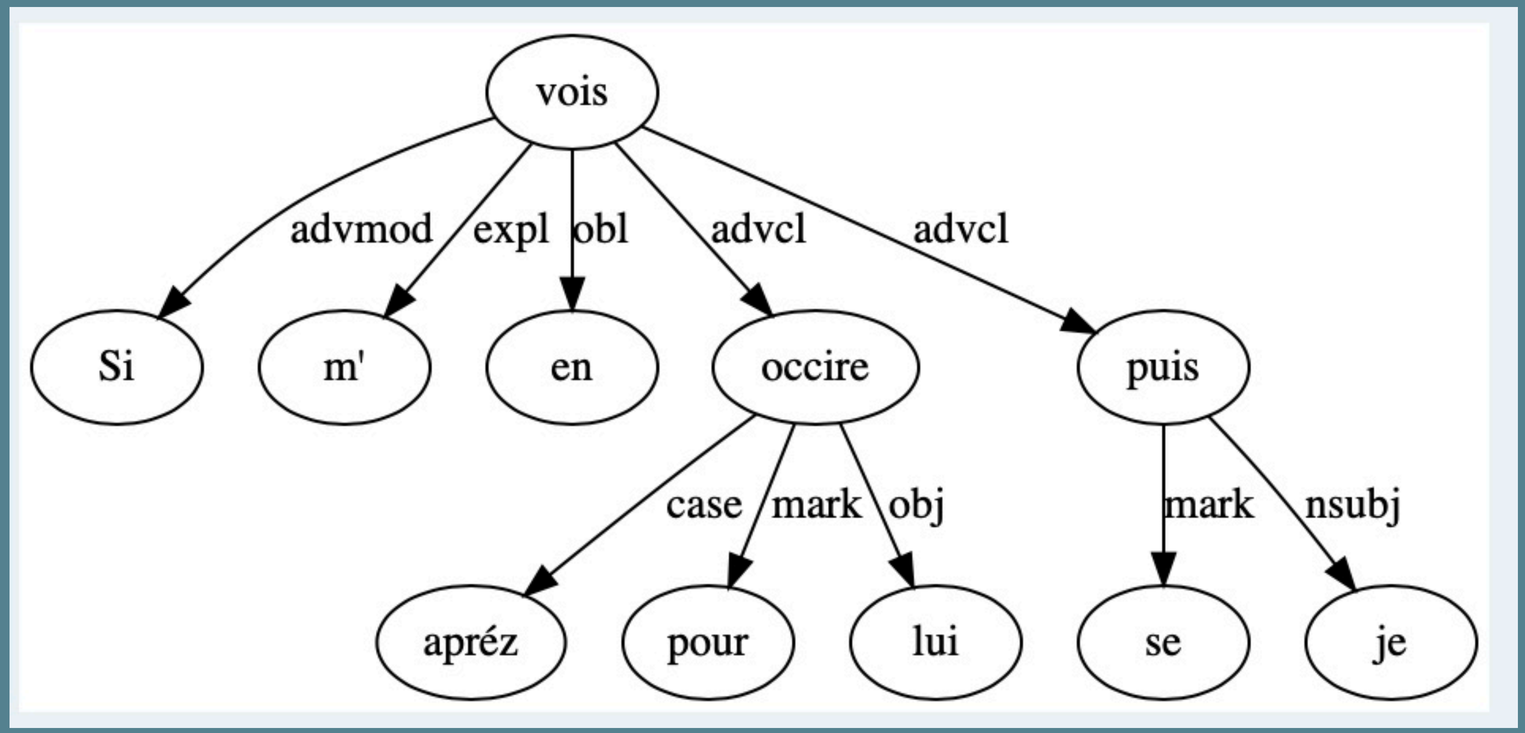
LGERM

- http://stella.atilf.fr/LGeRM/plateforme/ (Souvay & Pierrel 2009)
- Lemmatisation
- Étiquetage morphosyntaxique (POS tagging)
- Vérification manuelle des formes inconnues au lexique MF, signalement des NP, erreurs de segmentation (*l'açé* cf *laçé*), vérification des formes douteuses (*maintemant, avoieut)*
- Désambiguisation :
  - Orthographe inhabituelle d'un mot connu, ou erreur d'OCR/humaine par nous, ou compositeur ?
  - *que* PRON ou SCONJ?

TokenID	Forme	etiq	lemme	code	Pp	TokenChap	sentID
0	Comment	ADVint	COMMENT2	adv. et conj.	f.1v	1	1
1	monsieur	NOMcom	MONSIEUR	subst. masc.	f.1v	2	1
2	sainct	ADJqua	SAINT	adj. et subst. masc.	f.1v	3	1
3	Jacques	NOMpro	JACQUES	nom propre	f.1v	4	1
4	l'	DETdef	LE	art. déf.	f.1v	5	1
5	apostre	NOMcom	APÔTRE	subst. masc.	f.1v	6	1
6	se	PROper	SE1	pron. pers.	f.1v	7	1
7	apparut	VERcjg	APPARAÎTRE	verbe	f.1v	8	1
8	au	PRE.DETdef	LE|À	art. déf.|prép.	f.1v	9	1
9	noble	ADJqua	NOBLE1	adj. et subst. masc.	f.1v	10	1
10	Roy	NOMcom	ROI1	subst. masc.	f.1v	11	1
11	Charlemaigne	NOMpro	CHARLEMAGNE	nom propre	f.1v	12	1
13	apres	ADVgen	APRÈS	prép. et adv.	f.1v	14	1
14	qu'	CONsub	QUE	conj., rel. interr.	f.1v	15	1
15	il	PROper	IL	pron. pers.	f.1v	16	1
16	eut	VERcjg	AVOIR1	verbe	f.1v	17	1
17	veu	VERppe	VOIR1	verbe	f.1v	18	1
18	le	DETdef	LE	art. déf.	f.1v	19	1
19	grand	ADJqua	GRAND	adj.	f.1v	20	1

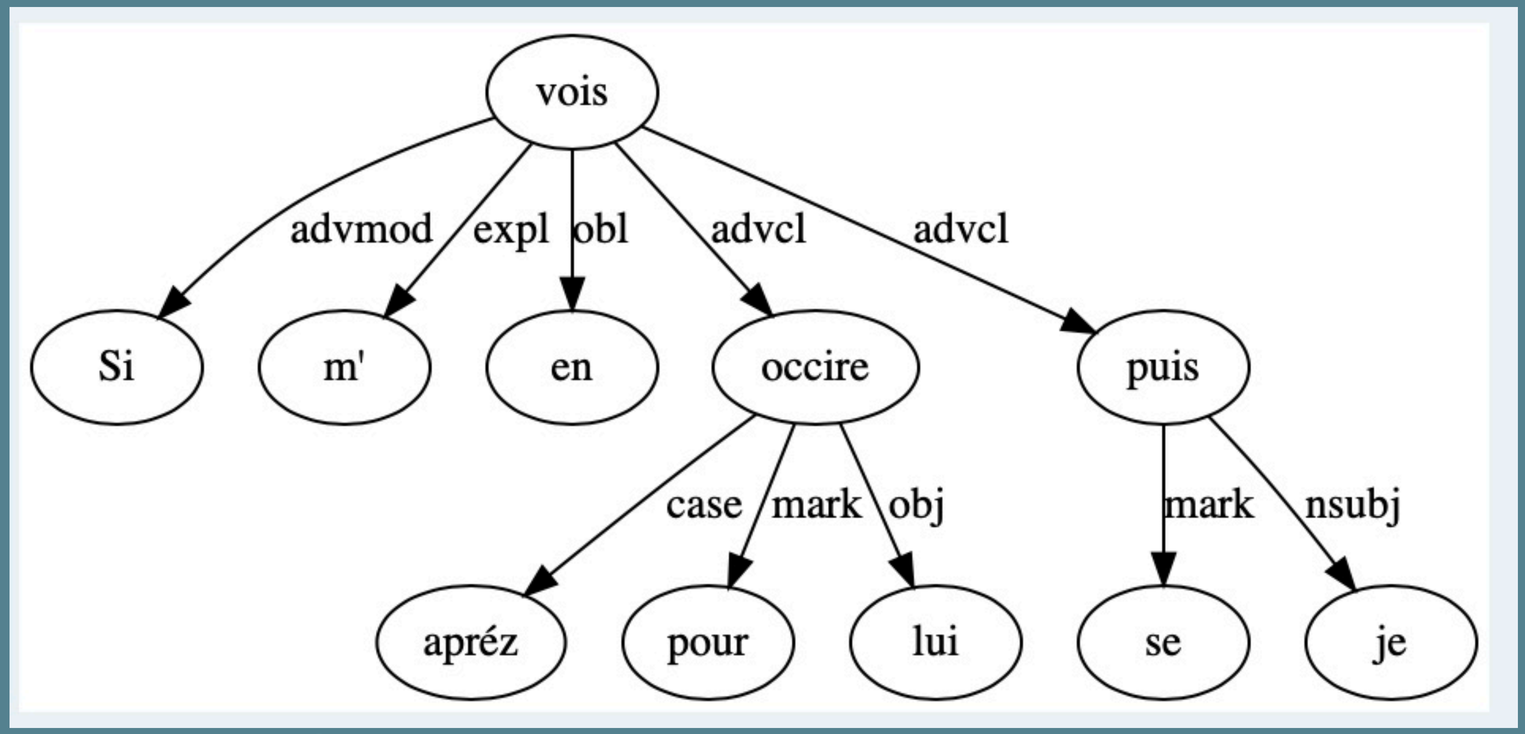
HOPS

- **HOPS « An honest parser of sentences » (Grobbol & Crabbé 2021)**
- **https://github.com/hopsparser/hopsparserLemmatisation**
- ***Si m'en vois apréz pour lui occire se je puis.***



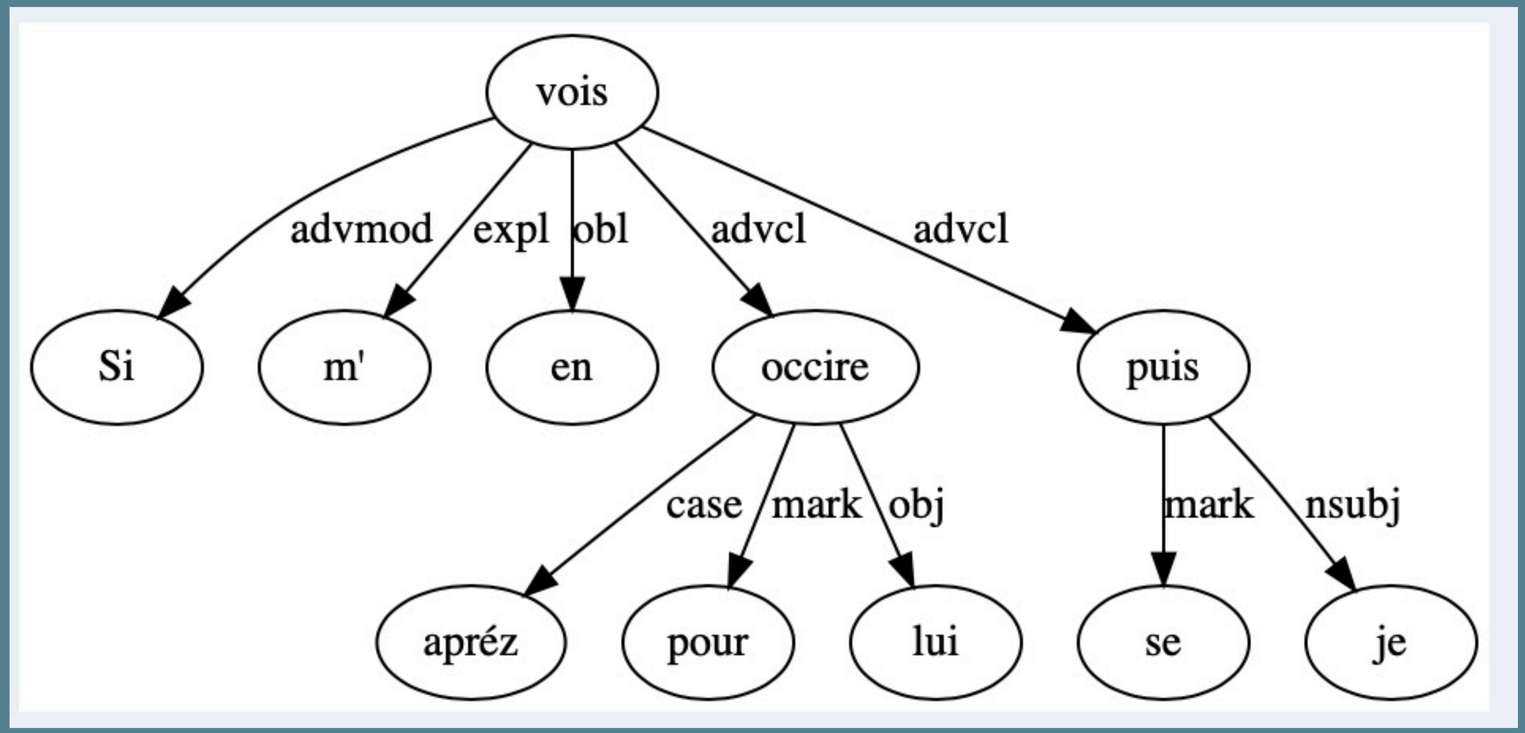
HOPS

- **HOPS « An honest parser of sentences » (Grobbol & Crabbé 2021)**
- **https://github.com/hopsparser/hopsparserLemmatisation**
- ***Si m'en vois apréz pour lui occire se je puis.***
- ***Alors je m’en vais après lui pour l’occire si je peux***



HOPS

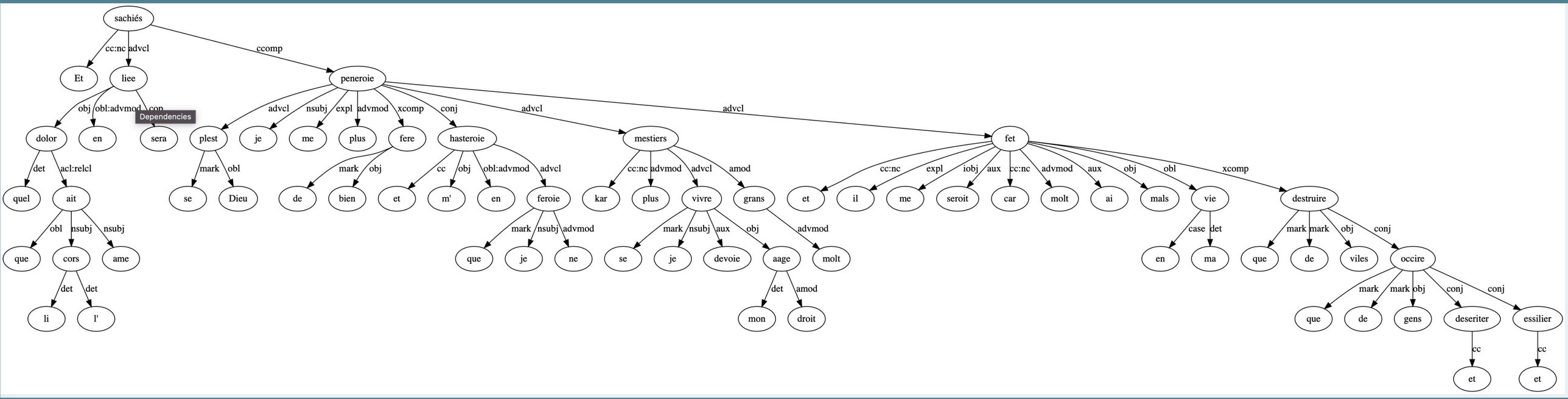
- **HOPS « An honest parser of sentences » (Grobbol & Crabbé 2021)**
- **https://github.com/hopsparser/hopsparserLemmatisation**
- ***Si m'en vois apréz pour lui occire se je puis.***
- ***Alors je m’en vais après lui pour l’occire si je peux***
- ***Alors, je pars le suivre, et le tuer si je peux***



QUESTIONS

- **2 outils c'est bien, mais difficile à différencier *automatiquement que_SCONJ/que_PRON, autre_PRON/autre_ADJ, le_det/le_PRON***
- **Les dépendances sont-elles correctes ?**
- **LGERM : lemmes + étiquettes morphysyntaxiques**
- **HOPS étiquettes morphysyntaxiques + dépendances**
- **Et s'ils sont en désaccord pour l'étiquette morphysyntaxique ?**
- **Peut-on recourir à d'autres outils similaires ?**

- **Certains phrases sont longues !**

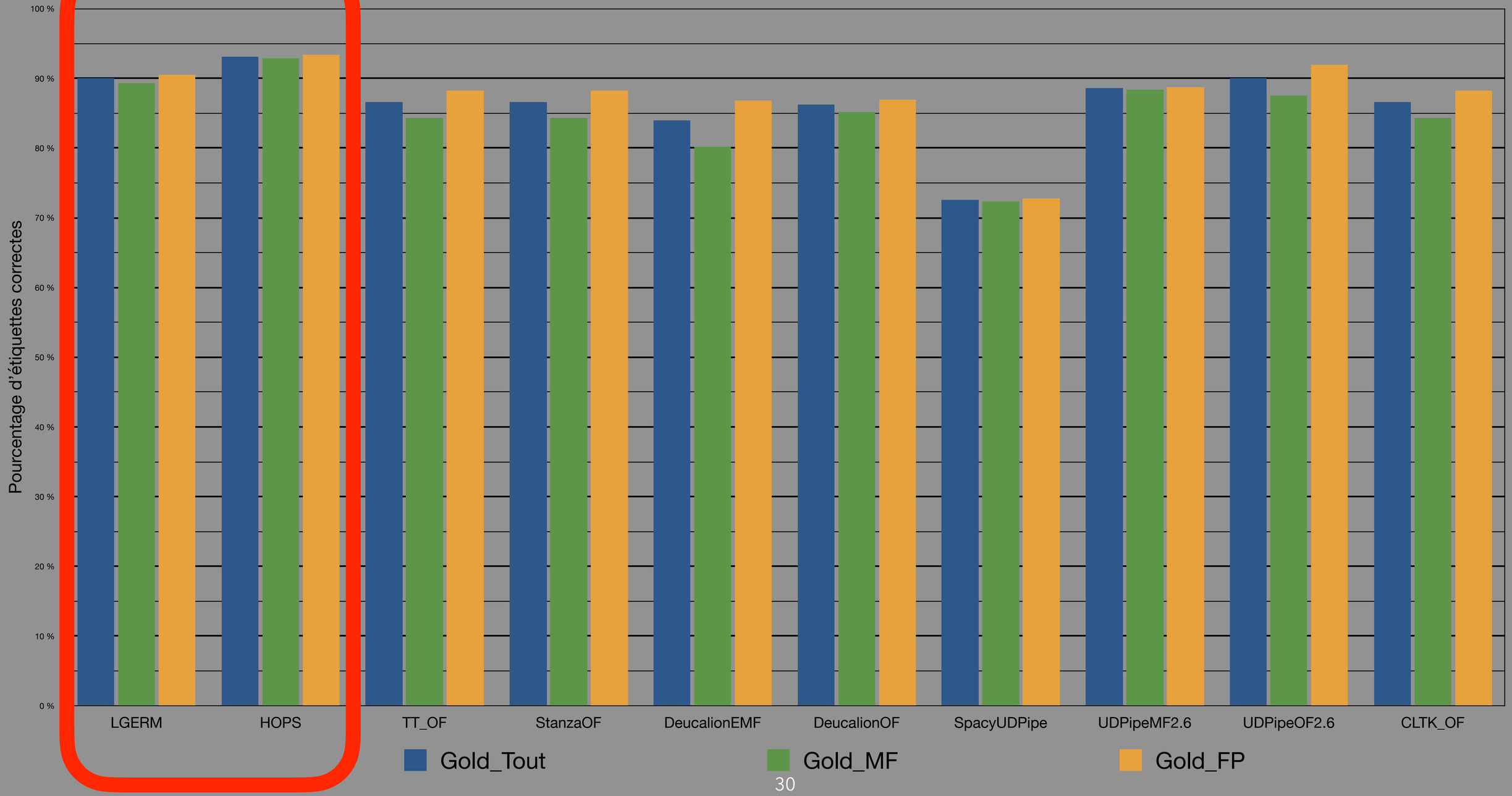


	Modèle	Interface	API REST	Adresse	Lemmes	ÉtiquettesXPOS	EtiquettesUPOS	Rôle=fonction	Dépendance=Num	Feats
						CONcoo	*CCONJ*	*amod*	*17*	*Gender=FemlNumber=Sing*
LGERM	**AF, MF**	**URL/G Souvay**	Non	Non	Oui	Oui	Non	Non	Non	Non
HOPS	**AF**	**Python, Shell**	Non	Non	Non	Non	Oui	Oui	Oui	Non
TT_OF	**AF**	**Python, AppIndép.**	Non	Non	Oui	Non	Oui	Non	Non	Non
TT_MF	**FMod**	**Python, AppIndép.**	Non	Non	Oui	Non	Oui	Non	Non	Non
StanzaOF	**AF**	**Python**	Non	Non	Non	Oui	Oui	Non	Oui	Oui
StanzaMF	**MF**	**Python**	Non	Non	Non	Oui	Oui	Non	Oui	Oui
DeucalionEMF	**EMF=FPC**	**URL, Python**	Non	https://dh.chartes.psl.eu/deucalion/freem	Oui	Oui	Non	Non	Non	Oui
DeucalionFR	**FMod**	**URL, Python**	Non	https://dh.chartes.psl.eu/deucalion/fr	Oui	Oui	Non	Non	Non	Oui
DeucalionOF	**AF**	**URL, Python**	Non	https://dh.chartes.psl.eu/deucalion/fro	Oui	Oui	Non	Non	Non	Oui
SpacyUDPipe	**AF**	**Python**	Non	Non	Non	Oui	Oui	Oui	Oui	Non
UDPipeMF2.6	**FMod**	**URL/API**	Oui	https://lindat.mff.cuni.cz/services/udpipe/	Oui	Oui	Oui	Oui	Oui	Oui
UDPipeOF2.6	**AF**	**URL/API**	Oui	https://lindat.mff.cuni.cz/services/udpipe/	Non	Oui	Oui	Oui	Oui	Oui
CLTK_OF	**AF**	**Python**	Non	Non	Non	Oui	Oui	Oui	Oui	Non

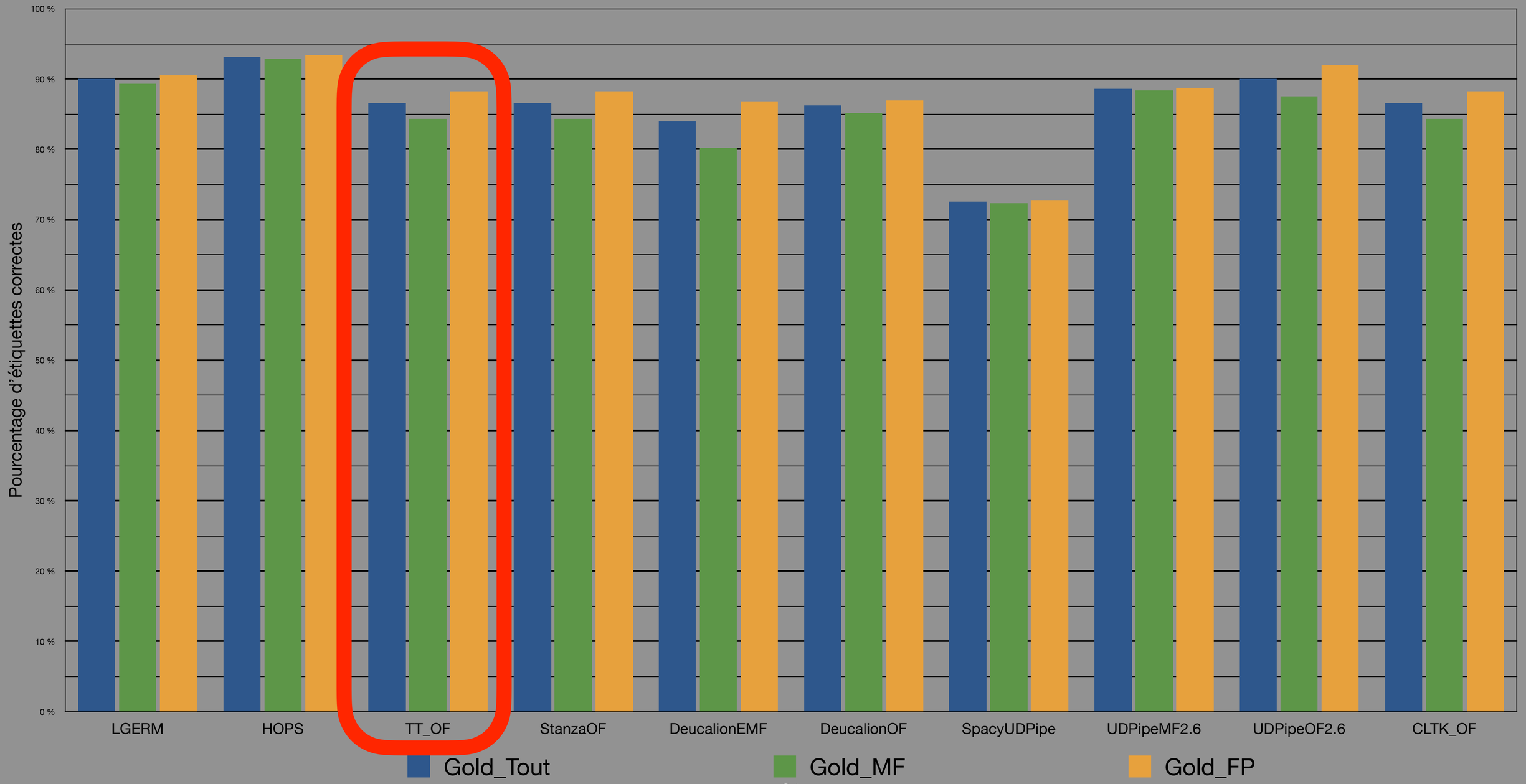
CORPUS GOLD

- 23 phrases représentatives, étiquetées manuellement (POS, dépendances, lemmes), 1100 mots
- Analysées avec chaque outil
  - Scripts Python pour prendre le texte d'origine, préparer les données pour chaque outil (certains gèrent la ponctuation, d'autres ne peuvent gérer lignes vides et d'autres en ont besoin, enfin d'autres ne peuvent gérer la numérotation des lignes…)
  - Scripts Python pour prendre la sortie de chaque outil et les réintégrer dans un fichier qui regroupe toutes les sorties, traduire entre jeux d'étiquettes, aligner les données, vérifier la correspondance entre forme_sortie et forme_rajoutée
- Analyses des dépendances actuellement en cours
- Analyses des lemmes + POS : finies

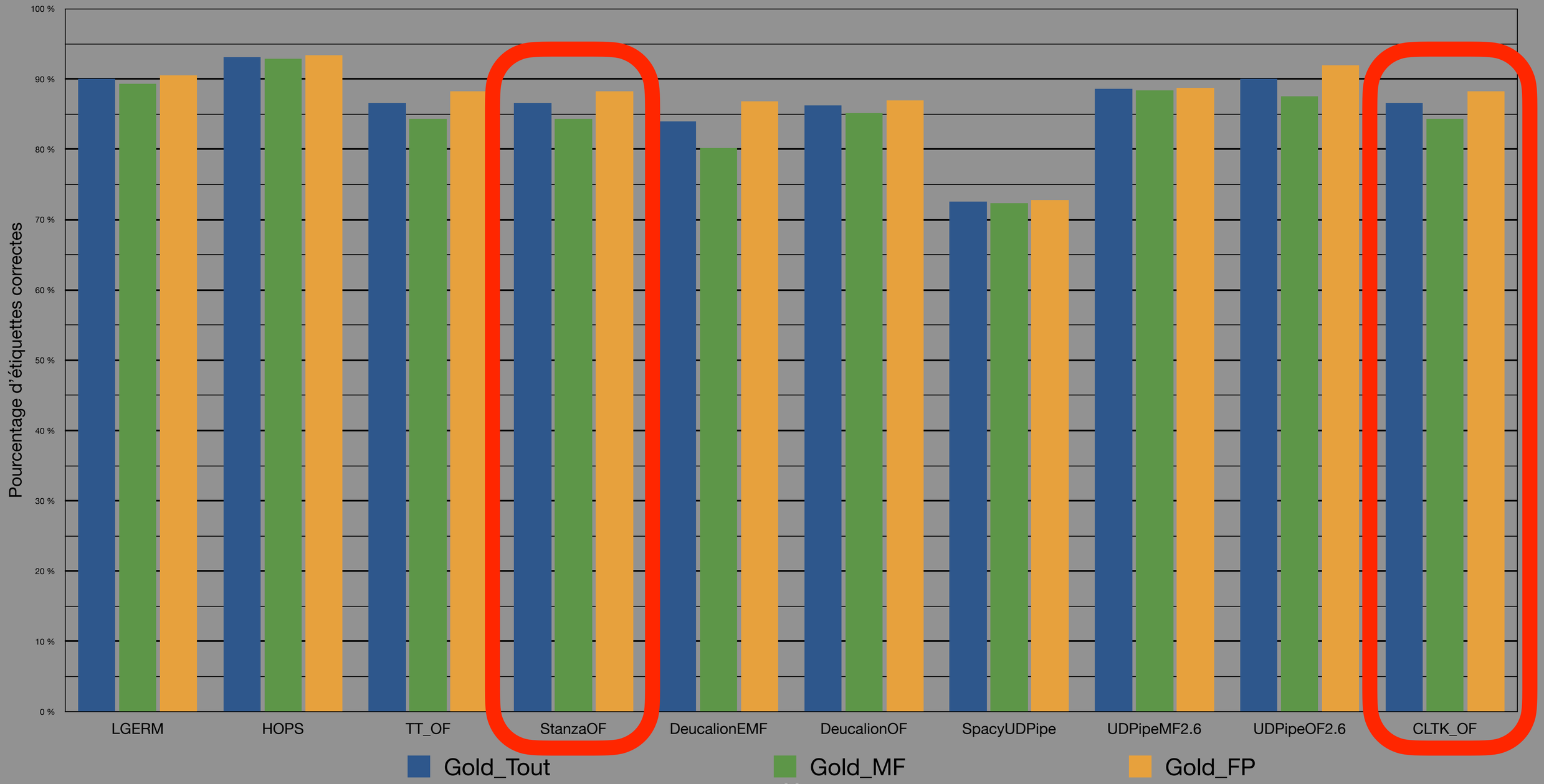
Étiquettes morpho. correctes en fonction des étiqueteurs



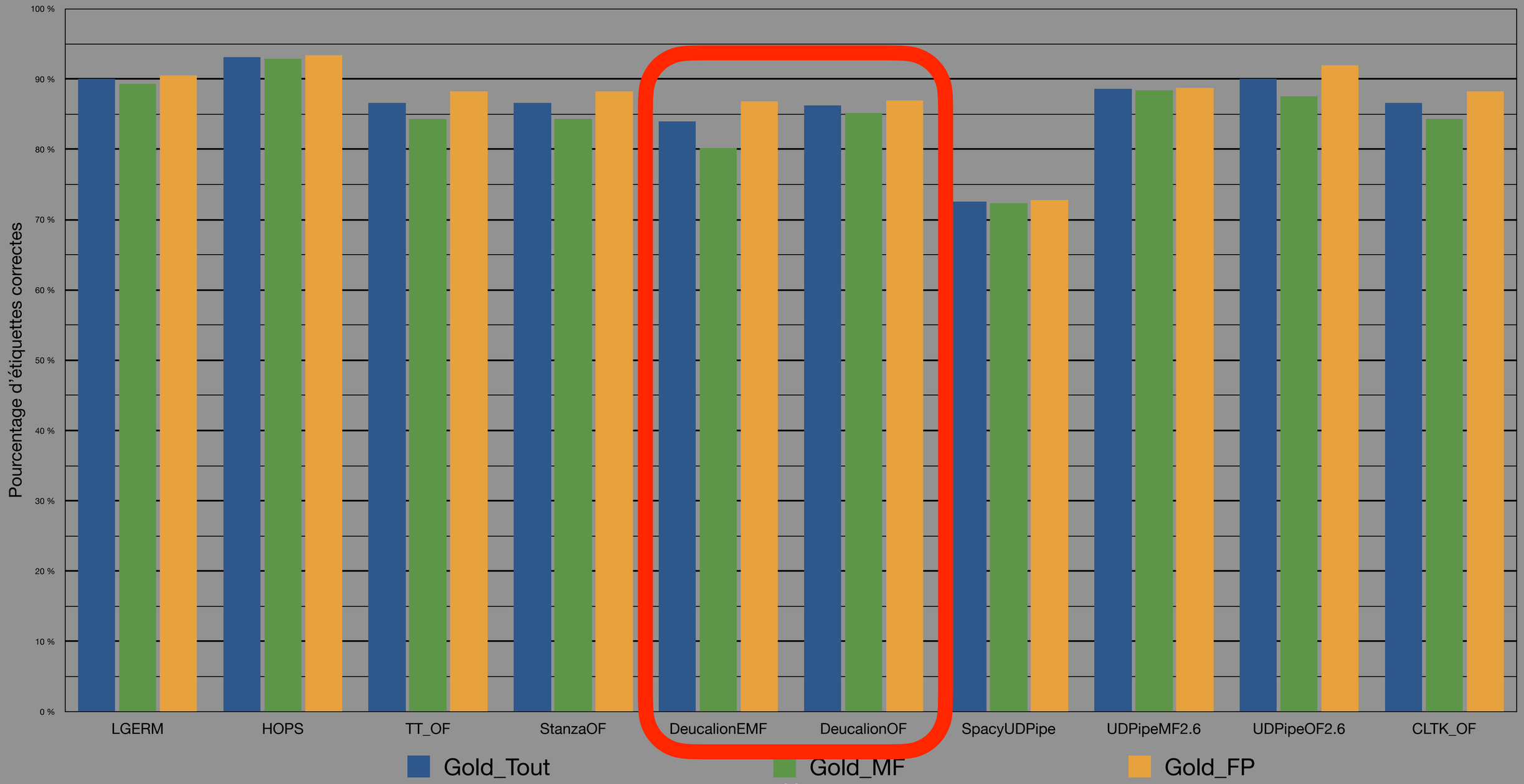
Étiquettes morpho. correctes en fonction des étiqueteurs



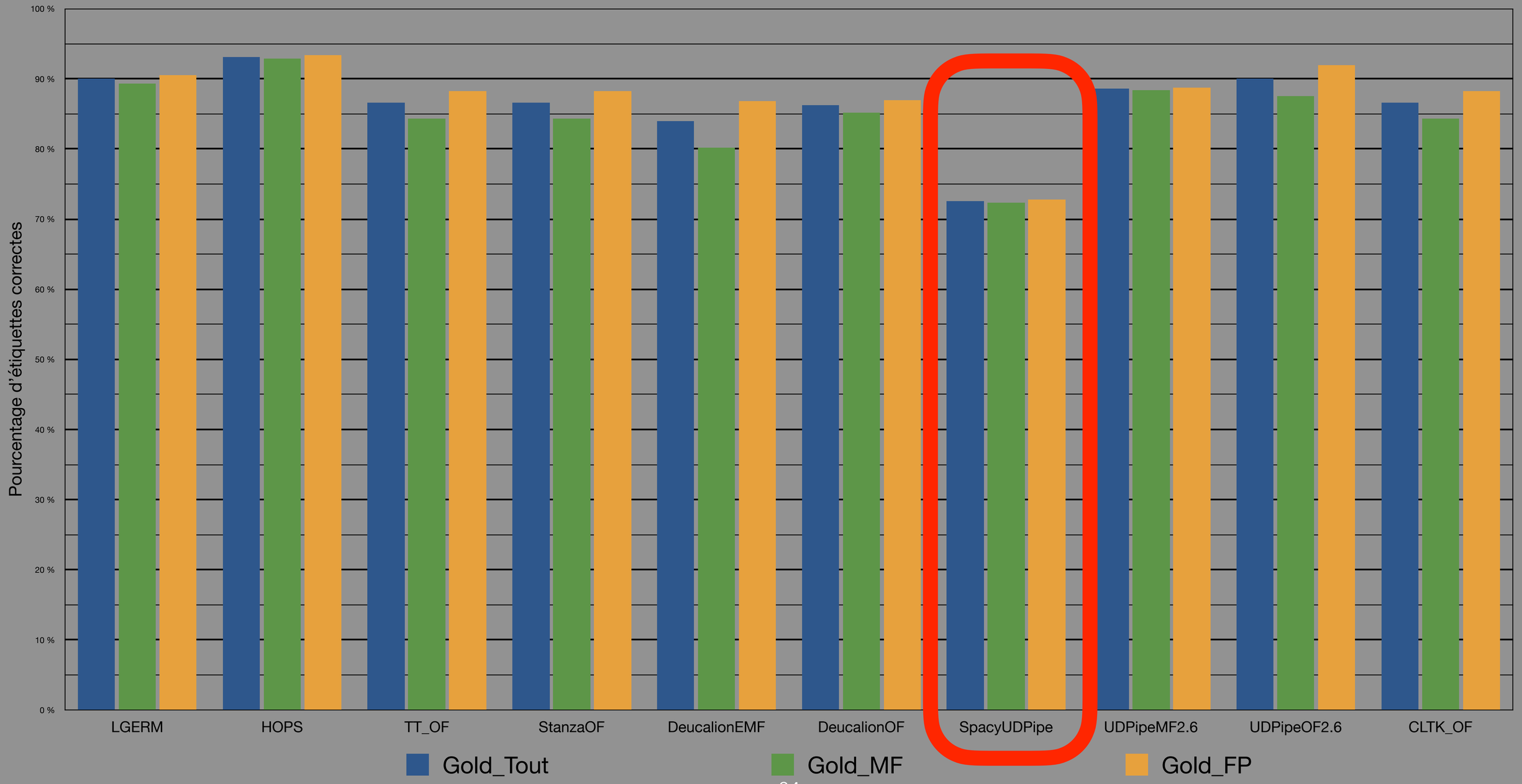
Étiquettes morpho. correctes en fonction des étiqueteurs



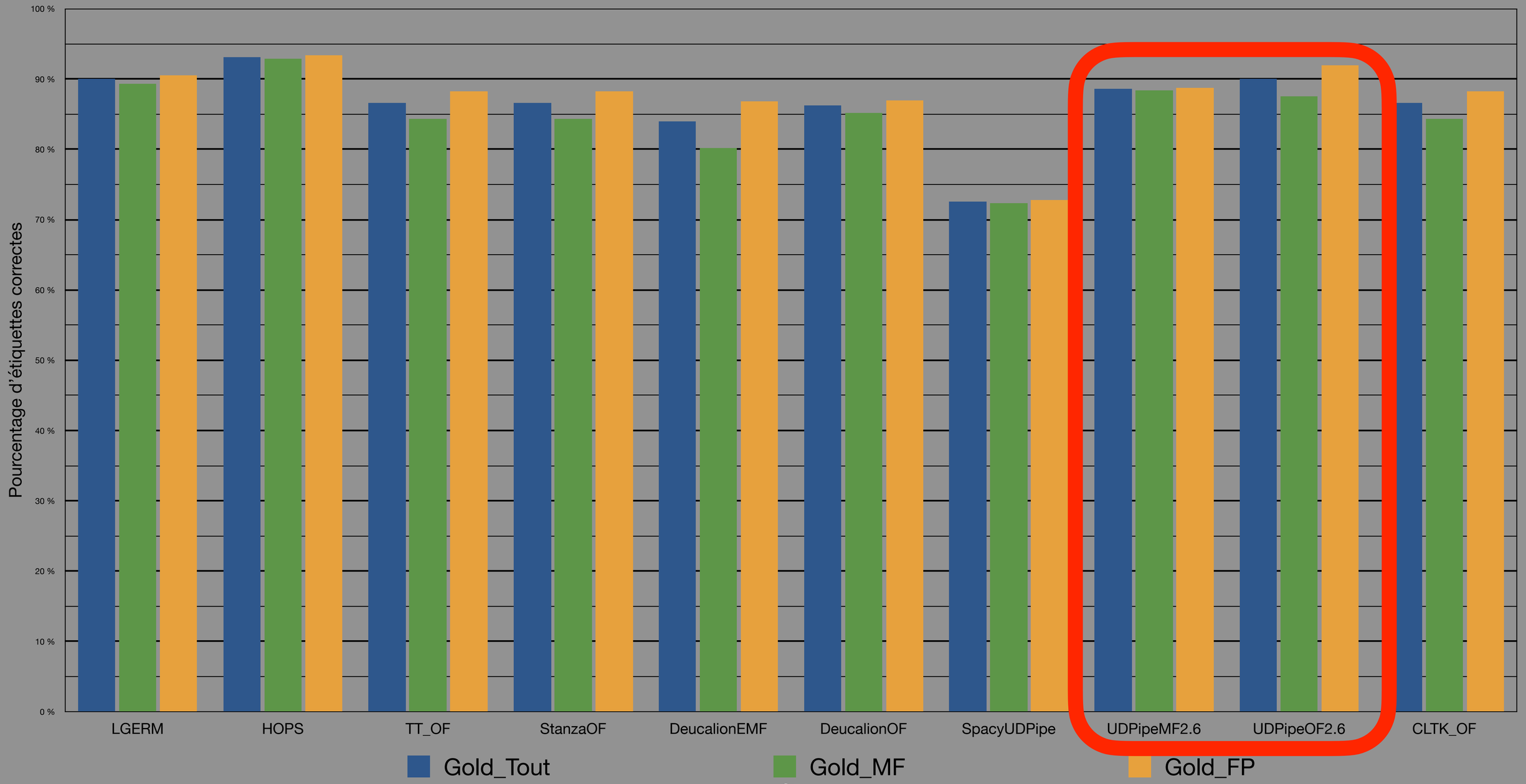
Étiquettes morpho. correctes en fonction des étiqueteurs

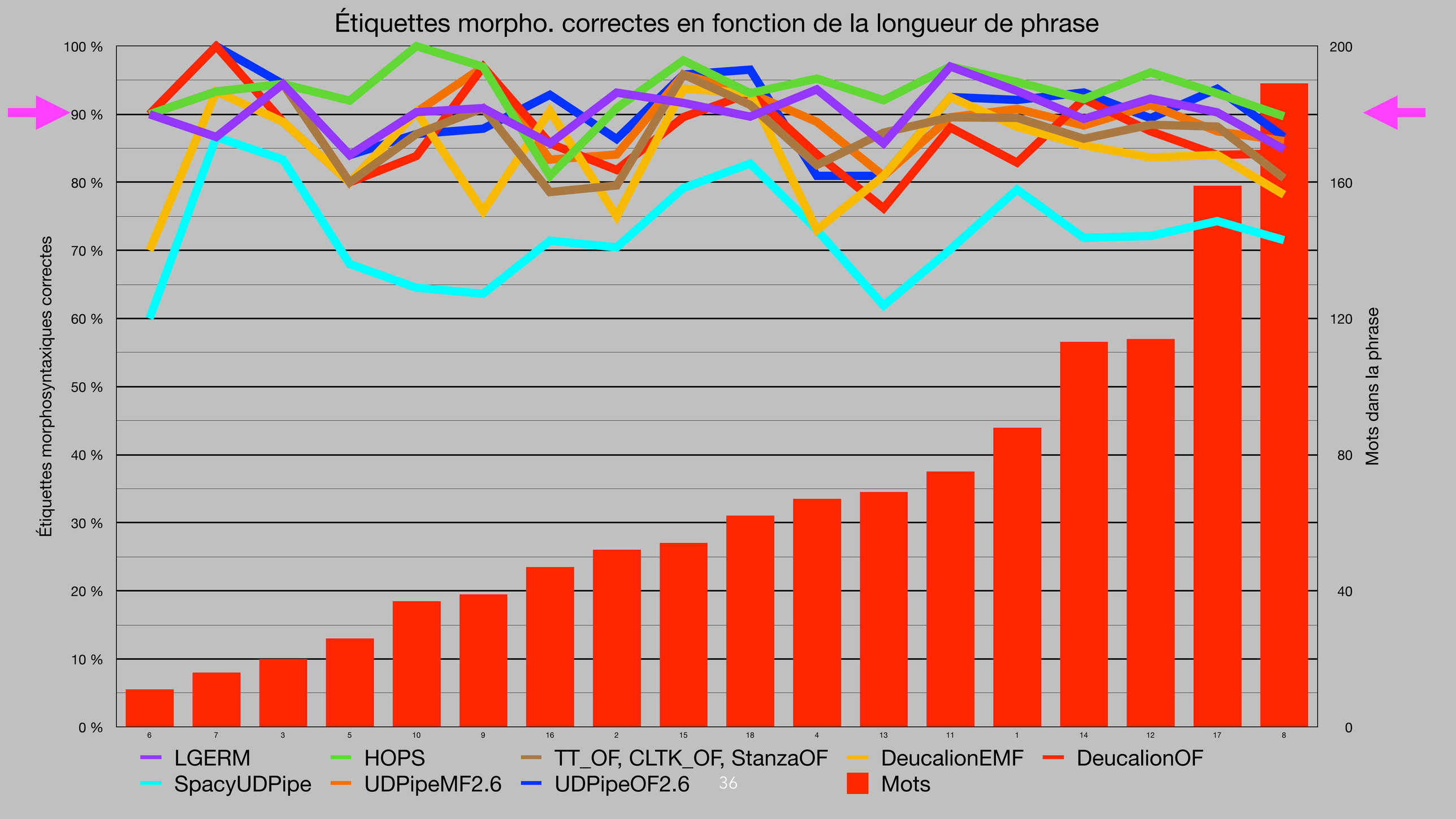


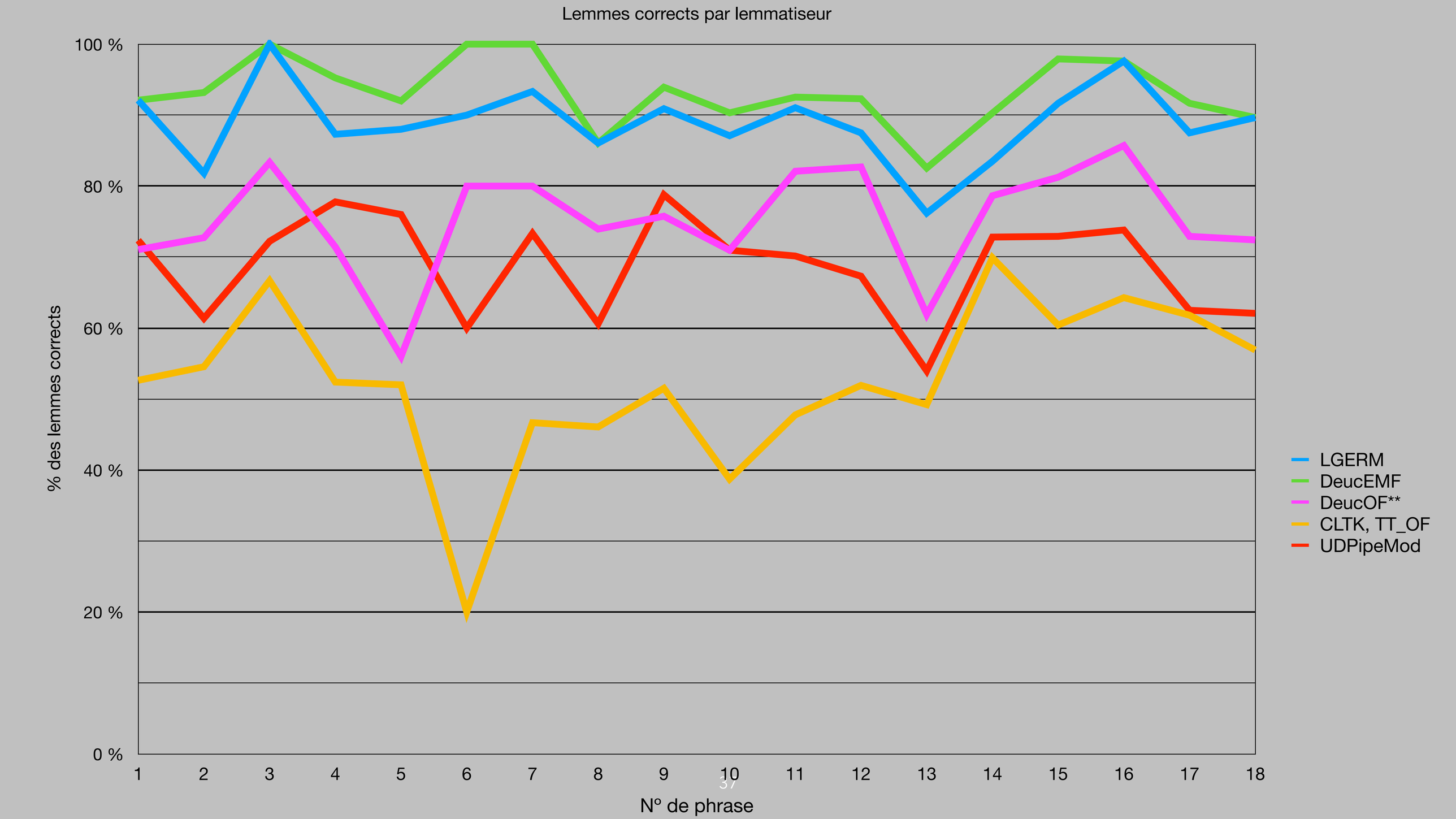
Étiquettes morpho. correctes en fonction des étiqueteurs



Étiquettes morpho. correctes en fonction des étiqueteurs




Étiquettes morpho. correctes en fonction de la longueur de phrase
Étiquettes morphosyntaxiques correctes
Mots dans la phrase
100 %
90 %
80 %
70 %
60 %
50 %
40 %
30 %
20 %
10 %
0 %
200
160
120
80
40
0
6 7 3 5 10 9 16 2 15 18 4 13 11 1 14 12 17 8
LGERM
HOPS
TT_OF, CLTK_OF, StanzaOF
DeucalionEMF
DeucalionOF
SpacyUDPipe
UDPipeMF2.6
UDPipeOF2.6
Mots


Lemmes corrects par lemmatiseur
100 %
80 %
60 %
40 %
20 %
0 %
% des lemmes corrects
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
N° de phrase
LGERM
DeucEMF
DeucOF**
CLTK, TT_OF
UDPipeMod

LES OUTILS

	OCR	Lemmatisation	POS tags	Dépendances
Transkribus (ReadCoOp 2022)	Oui			
LGERM (Souvay + Pierrel 2009)		Oui	Oui	
HOPS (Grobbol & Crabbé 2021)			Oui	Oui
UDPipe_OF (Straka 2018)			Oui	Oui
Stanza_OF (Qi et al. 2020)		Oui	Oui	Oui
Deucalion_EMF (Camps et al. 2020)		???		
Deuc_OF (Clérice, Camps et Pinche 2019)			Oui	
TT_AF (Schmid 1994)			Oui	

AMÉLIORATIONS…

- **Pour le corpus AF : identification d'une série de règles des sorties de LGERM, TT, HOPS, Stanza en confrontant leurs discordances**
- **Augmentation de taux d'étiquettes correctes passe de 90 à 95 %**
- ***que* : *Si TT dit* SCONJ *et que HOPS, LGERM et Stanza disent* PRON*, ignorer TT:* que**
- **Si la forme est dans Liste1, l'étiquette correcte est NOUN (limitation des NP)**
- **Liste1 = {damoiselle, madame, dame, moine, monseigneur, seigneur, maître, maitre, sieur, toussaint, paradis, enfer, ynfer, pâques, noël, pasque, pasques, penthecouste}**

AMÉLIORATIONS...

- **Si 5 étiqueteurs disent X et 1 dit Y, choisissons X**

- **Si 6 étiquettes sont différentes, solliciter l'intervention humaine...**

RÉFÉRENCES

- Camps, J-B *et al*. 2020. Corpus and Models for Lemmatisation and POS-tagging of Classical French Theatre. https://arxiv.org/abs/2005.07505.
- Clérice, T, Camps, J-B et Pinche, A. 2019. Deucalion, Modèle Ancien Francais (0.2.0). doi: 10.5281/zenodo.3237455
- Grobol, L and Crabbé, B. 2021. « Analyse en dépendances du français avec des plongements contextualisés » Dans *Actes de la 28ème Conférence sur le Traitement Automatique des Langues Naturelles, Lille, France*. URL https://hal.archives-ouvertes.fr/hal-03223424.
- Kraif O & Diwersy S. 2012. « Le Lexicoscope : un outil pour l'étude de profils combinatoires et l'extraction de constructions lexico-syntaxiques » dans *Actes de la conférence TALN 2012*, Grenoble, pp. 399-406.
- Qi, P *et al*. 2020. « Stanza: A Python Natural Language Processing Toolkit for Many Human Languages » dans *Association for Computational Linguistics (ACL) System Demonstrations*. 2020.
- ReadCoOp 2022. Transkribus. URL <https://readcoop.eu/transkribus>
- Schmid, H. 1994. « Probabilistic Part-of-Speech Tagging Using Decision Trees » dans *Proceedings of International Conference on New Methods in Language Processing*. Manchester, UK.
- Straka, M. 2018. « UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task» dans *Proceedings of CoNLL 2018: The SIGNLL Conference on Computational Natural Language Learning*, pp. 197-207, Association for Computational Linguistics, Stroudsburg, PA, USA.
- Vielliard, F. 2007. « Qu'est-ce que le "roman de chevalerie" ? Préhistoire et histoire d'une formule » dans Diu, I., Parinet, É., & Vielliard, F. (Eds.), *Mémoire des chevaliers : Édition, diffusion et réception des romans de chevalerie du XVII*e *au XX*e *siècle*. Publications de l'École nationale des chartes. doi :10.4000/books.enc.793