Evaluation d'outils d'annotation en entités nommées pour le français : que compare-t-on ?

Travaux avec Unitex-CasEN et SpaCy

Alexane Jouglar | sous la direction de Karën Fort, Alice Millour, Yoann Dupont

Le 25/06/2021

Sorbonne Université

Plan

Le domaine des entités nommées

Construction du travail

Les résultats obtenus

Lien vers le GitHub : https://github.com/AlexaneJ/memoireM1.

Le domaine des entités nommées

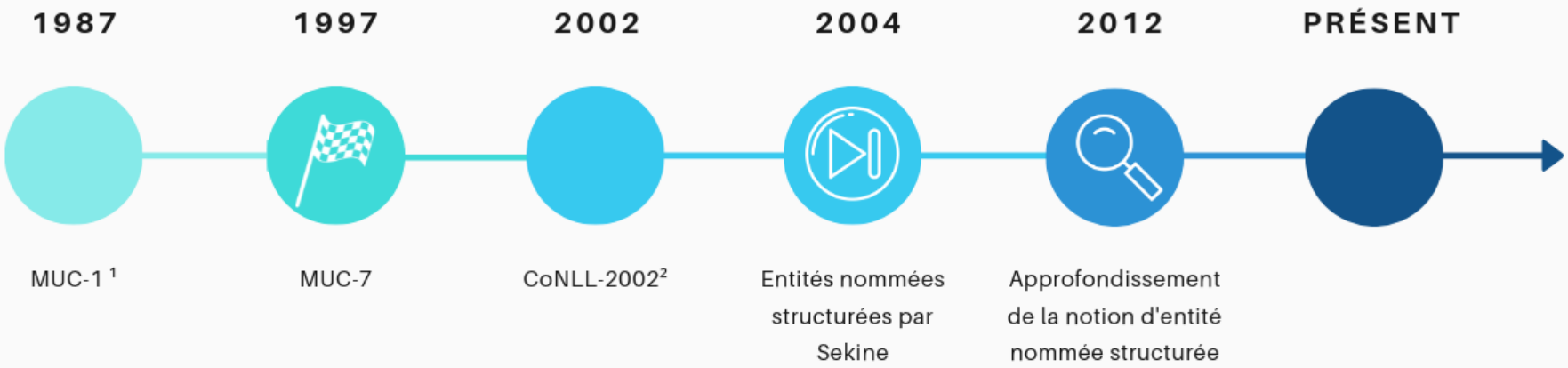
Définition

« Etant donnés un modèle applicatif et un corpus, on appelle entité nommée toute expression linguistique qui réfère à une entité unique du modèle de manière autonome dans le corpus. » - Ehrmann 2008

Le modèle applicatif

- le modèle applicatif repose sur un **choix**
- notion d'**essentialité**

Chronologie



1-MUC = Message Understanding Conference
2-CoNLL = Computational Natural Language Learning

Les entités nommées structurées

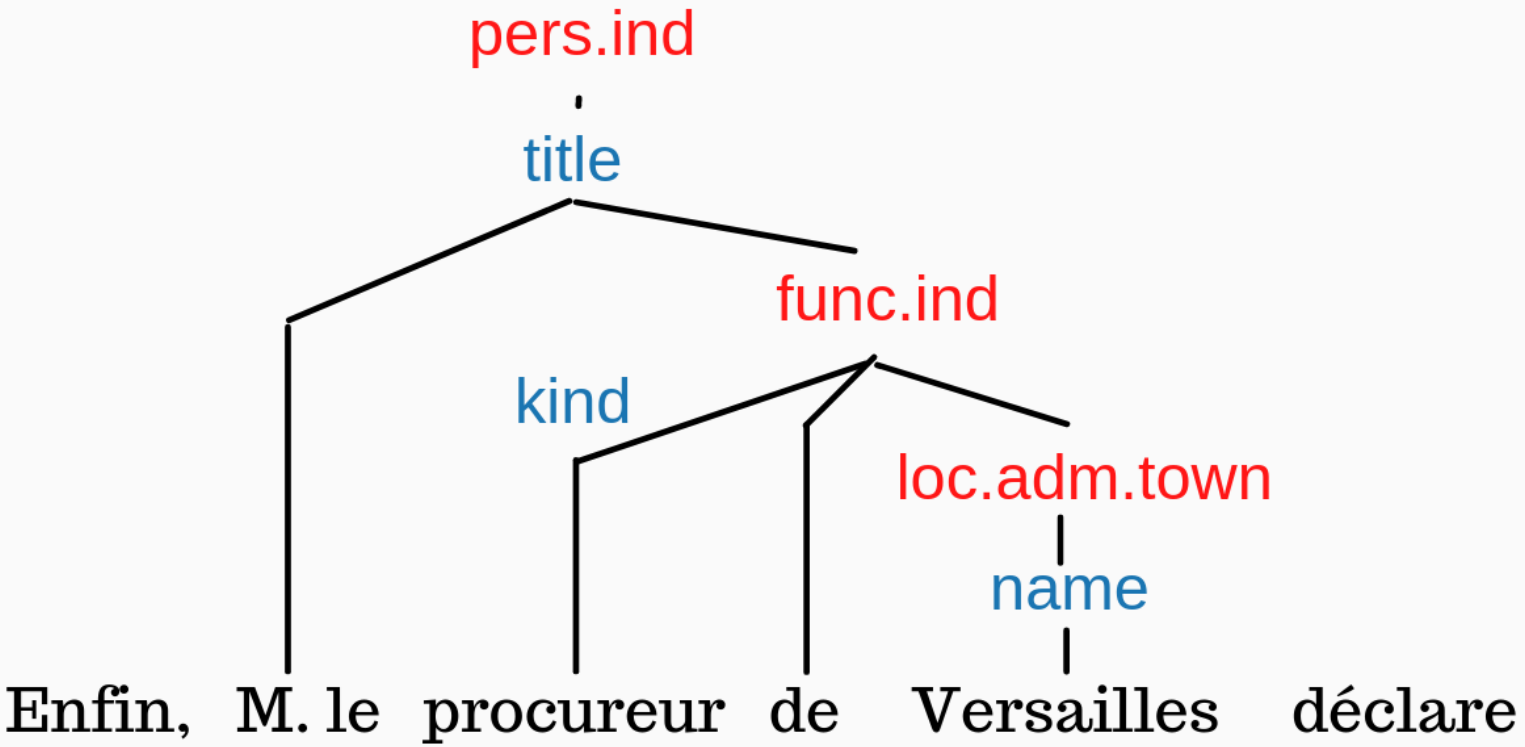
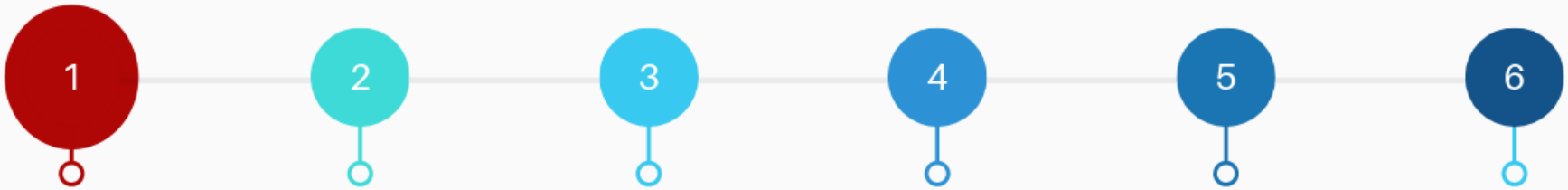


Schéma issu de Rosset et al. 2012.

Construction du travail

Première étape - création du corpus



CRÉATION DU CORPUS

7 textes de 1000 tokens environ

CAMPAGNE D'ANNOTATION

WebAnno

ANNOTATION AUTOMATIQUE

Unitex-CasEN et SpaCy

EVALUATION DES SORTIES

COMMENTAIRE DES RÉSULTATS

CONCLUSION

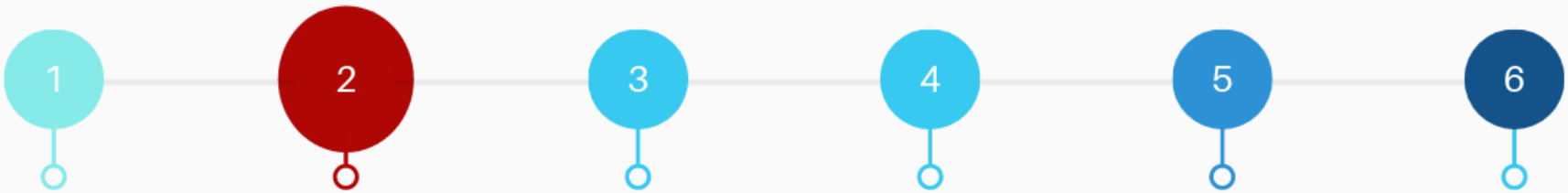
Données sur le corpus

Document	**Année**	**Source**	**Licence**	**# tokens[1]**
Aij-wikiner	2017	UD	CC-BY-4.0	859
GSD	2015	UD	CC-BY-SA 4.0	839
pg6470	1873	Gutenberg	Domaine public	821
Sequoia	2017	UD	CC-BY-NC-SA 4.0	818
Spoken	2018	UD	CC-BY-SA 4.0	1003
Wikinews (extraits)	2018	Wikinews	CC-BY 2.5	854
APIL	2011	GAL[2] Othe-Armance	Licence LGPLLR	620

[1] Nombre total de tokens : 69 246.

[2] GAL = Groupe d'Action Locale.

Deuxième étape - campagne d'annotation



CRÉATION DU CORPUS

7 textes de 1000 tokens environ

CAMPAGNE D'ANNOTATION

WebAnno

ANNOTATION AUTOMATIQUE

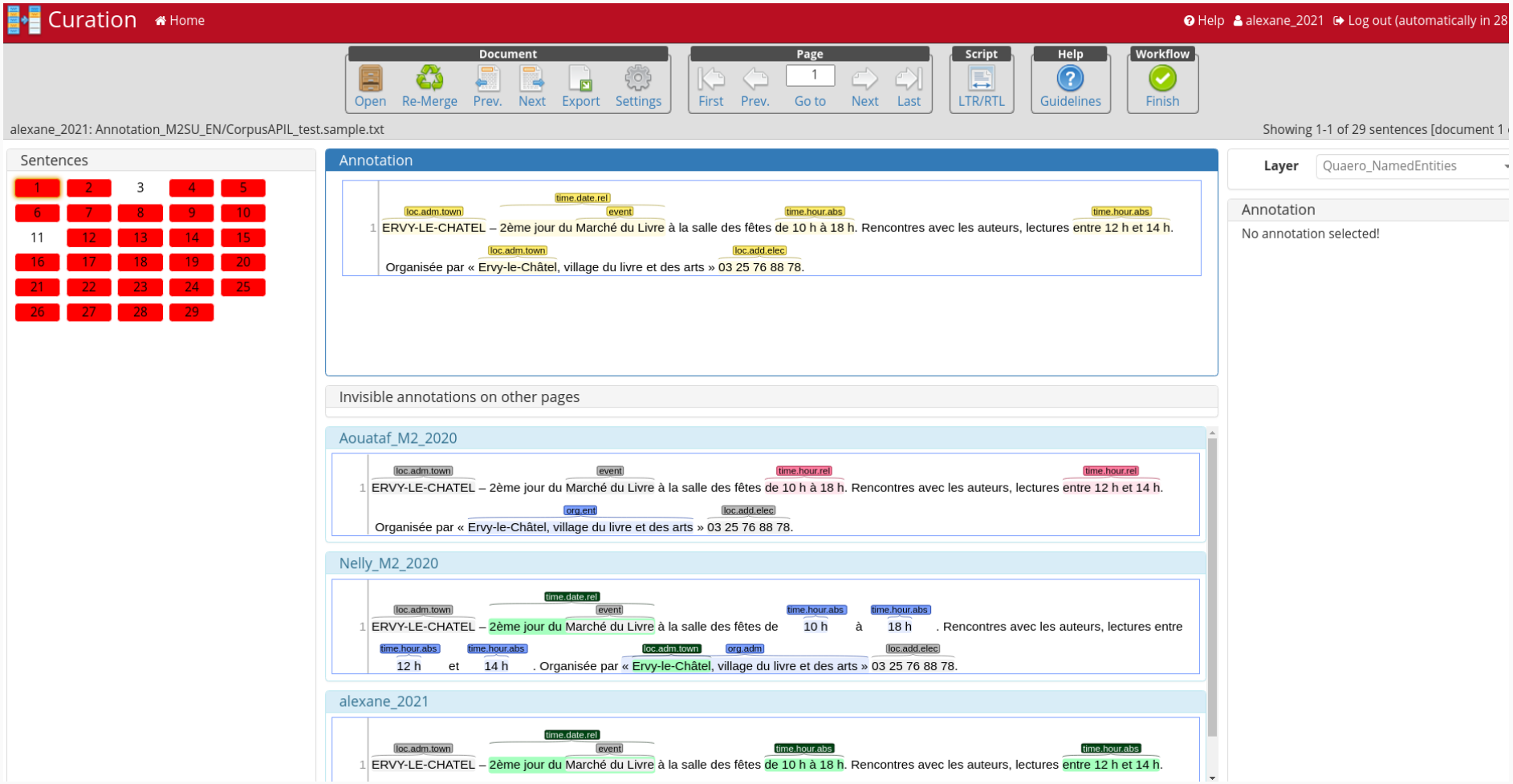
Unitex-CasEN et SpaCy

EVALUATION DES SORTIES

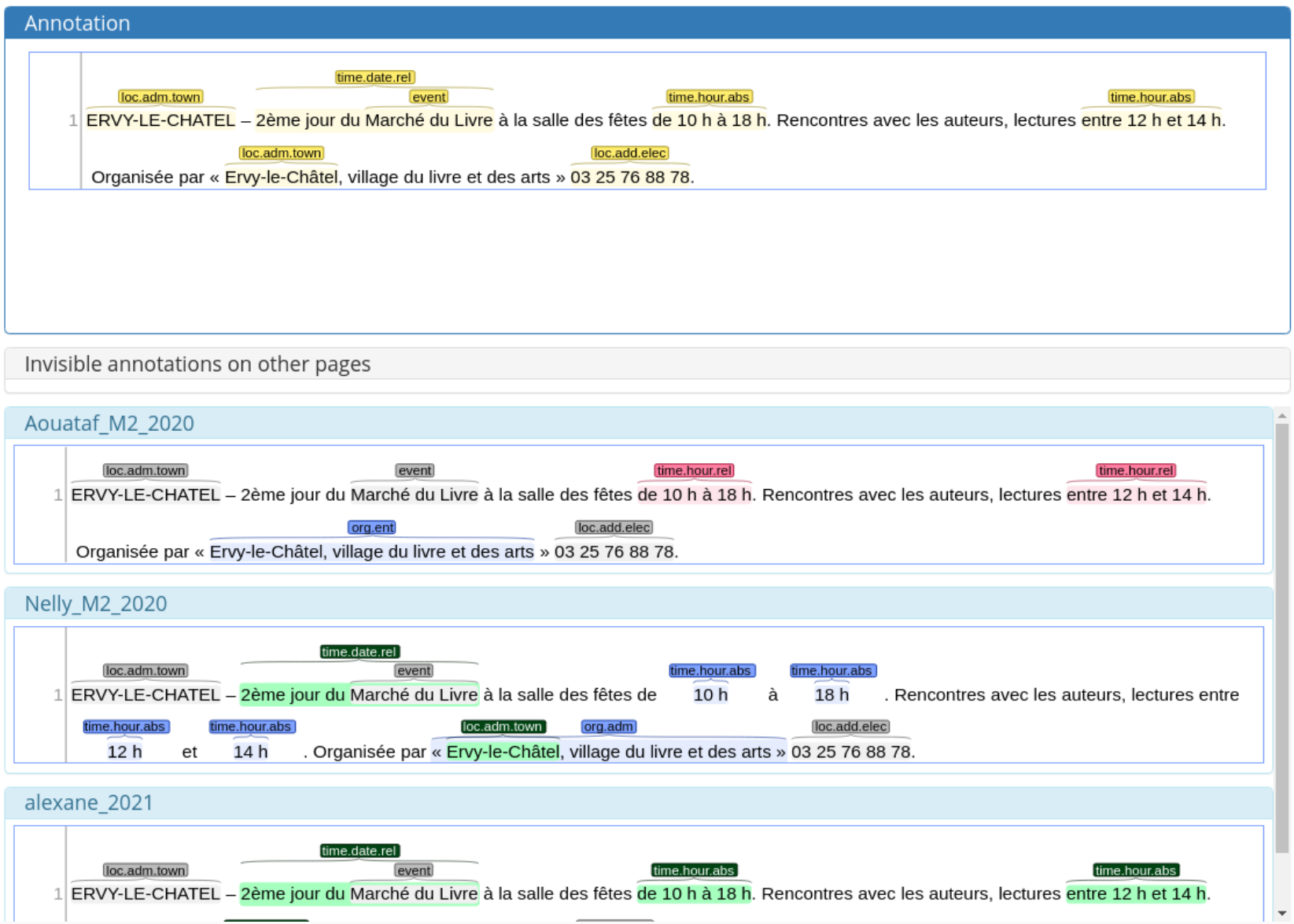
COMMENTAIRE DES RÉSULTATS

CONCLUSION

Visualisation de la plateforme WebAnno



Visualisation de la plateforme WebAnno - adjudication



Accords inter-annotateurs (alpha de Krippendorff)

	K	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	Ale
K							0,79							0,81
A1								0,91						0,90
A2				0,64		0,72							0,69	0,77
A3														0,91
A4														0,84
A5													0,77	0,86
A6														0,82
A7														0,92
A8											0,80			0,88
A9												0,38		0,52
A10														0,83
A11														0,93
A12														0,83

Troisième étape - annotation automatique

CRÉATION DU CORPUS

7 textes de 1000 tokens environ

CAMPAGNE D'ANNOTATION

WebAnno

ANNOTATION AUTOMATIQUE

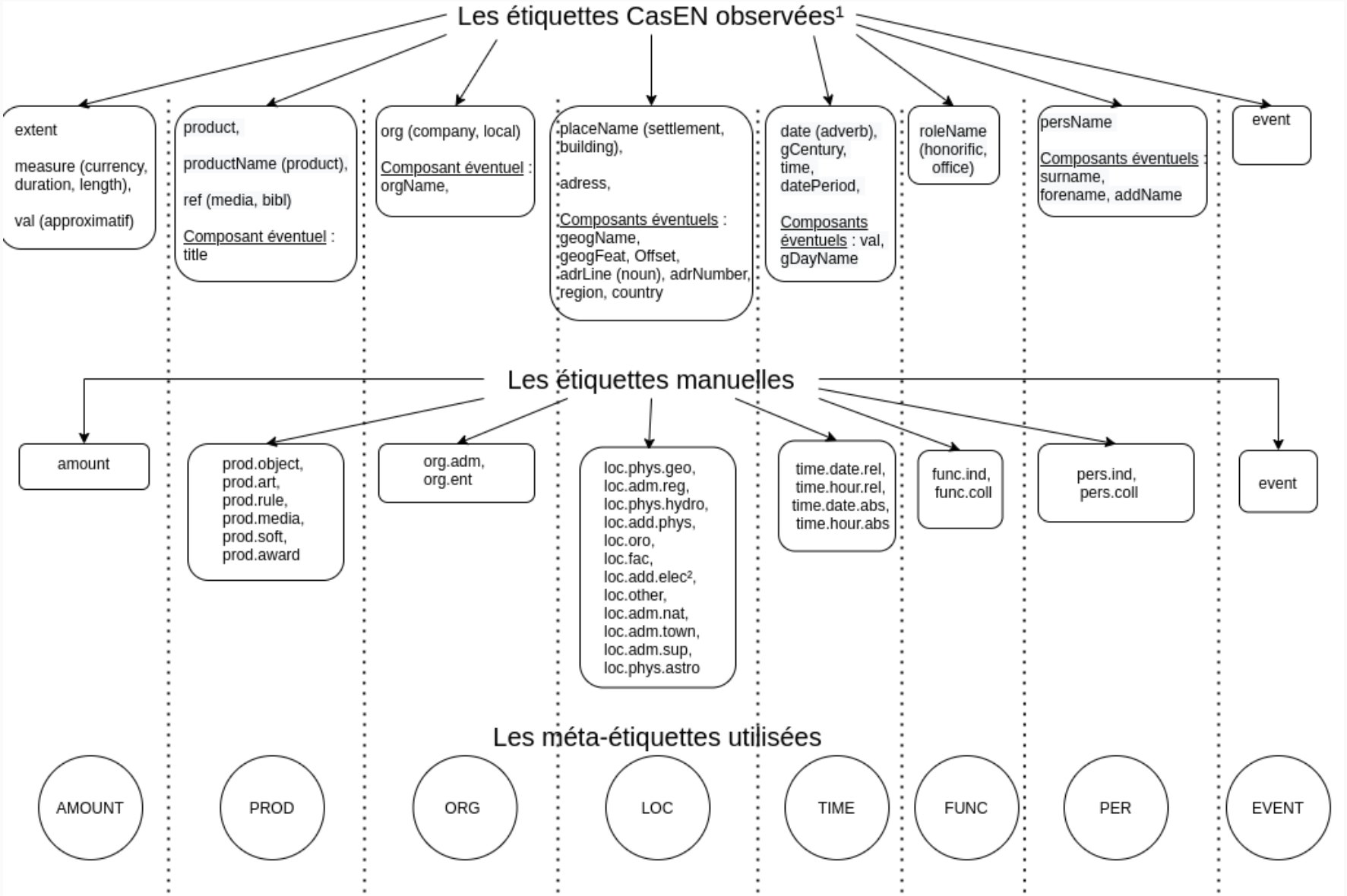
Unitex-CasEN et SpaCy

EVALUATION DES SORTIES

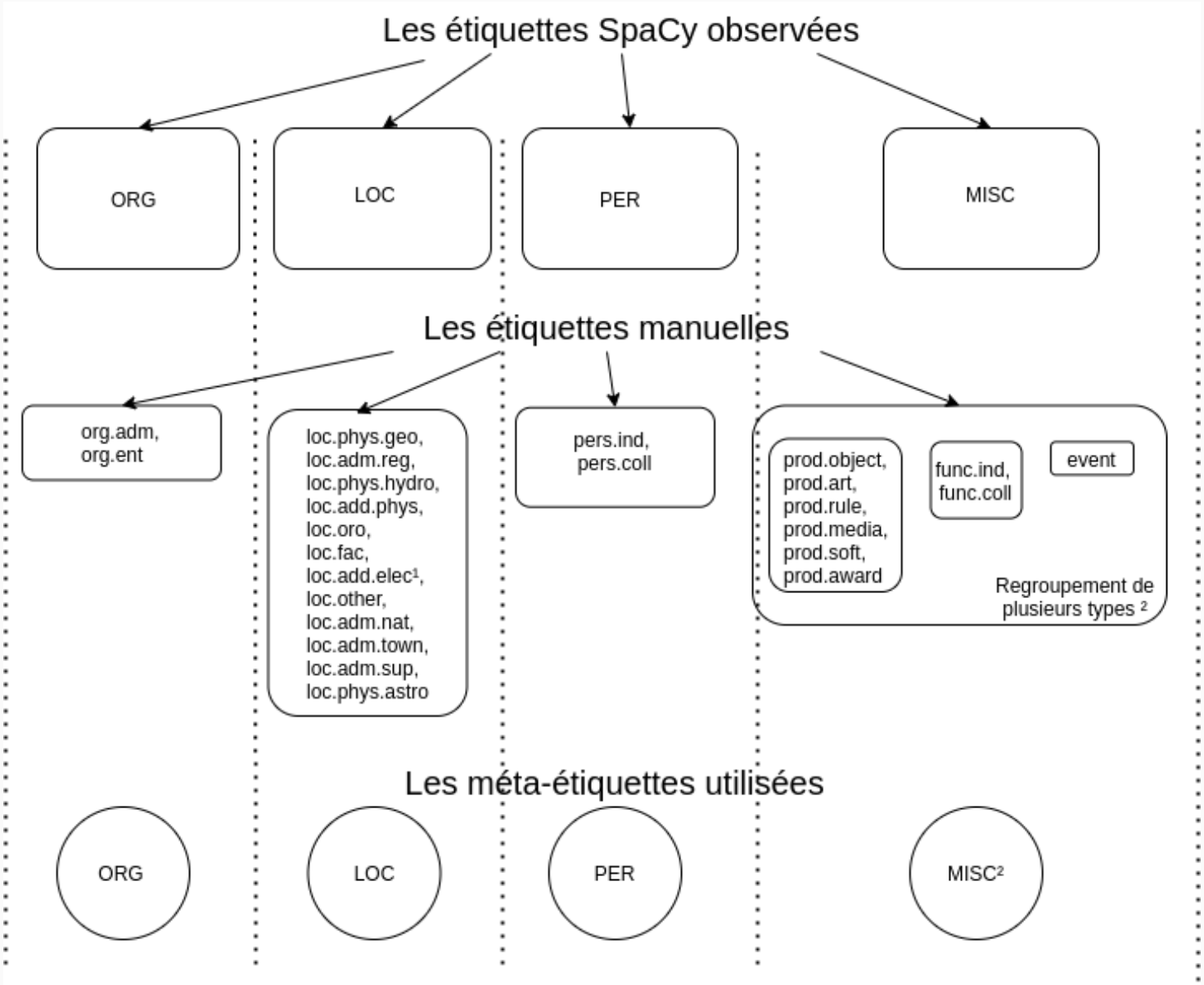
COMMENTAIRE DES RÉSULTATS

CONCLUSION

Réorganisation des étiquettes manuelles et Unitex-CasEN

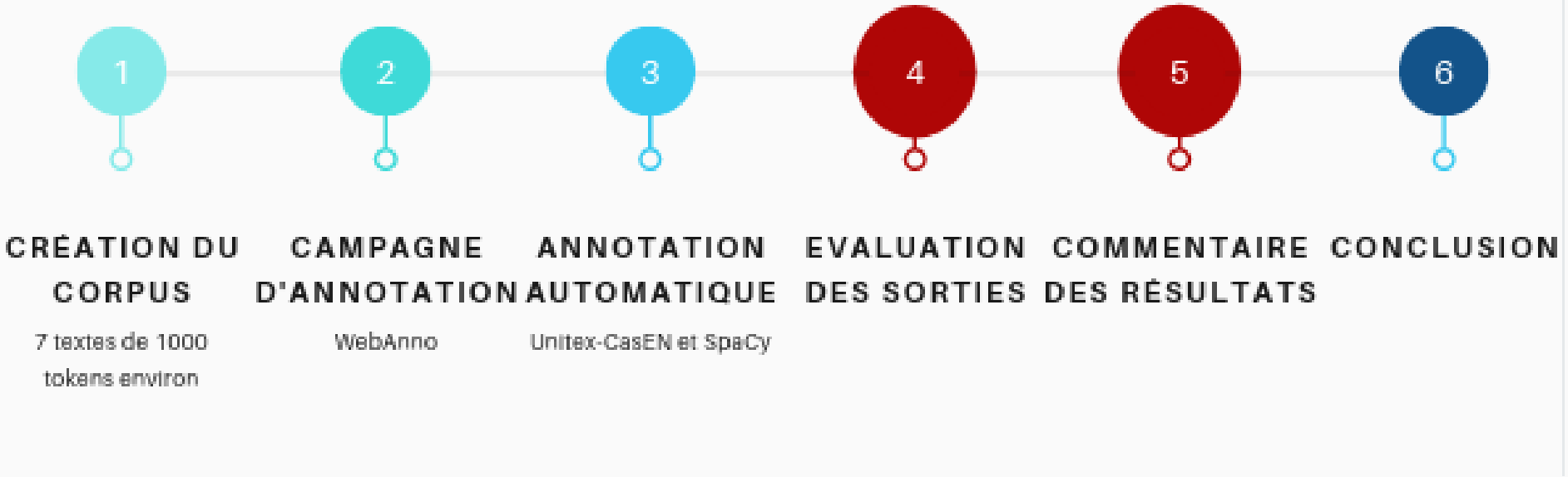


Réorganisation des étiquettes manuelles pour SpaCy



Les résultats obtenus

Evaluation et interprétation



Silence et bruit avec CasEN et SpaCy

Exactitude	Stricte			Partielle		
	VP	FN (silence)	FP (bruit)	VP	FN	FP
Unitex-CasEN	281	359	50	299	341	50
SpaCy	245	145	105	284	106	67

Résultats obtenus avec Unitex-CasEN

Exactitude	Stricte			Partielle		
	P	R	F1	P	R	F1
APIL	0,84	0,33	0,47	0,89	0,35	0,51
Wikinews	0,81	0,48	0,60	0,87	0,52	0,65
Sequoia	0,91	0,51	0,66	0,94	0,53	0,68
Aijwikiner	0,81	0,49	0,61	0,90	0,53	0,67
GSD	0,74	0,42	0,53	0,80	0,45	0,58
Spoken	0,61	0,46	0,52	0,63	0,48	0,55
pg6470	0,72	0,44	0,55	0,72	0,44	0,55
tous	**0,81**	**0,44**	**0,57**	**0,83**	**0,45**	**0,58**

Résultats obtenus avec SpaCy

Exactitude	Stricte			Partielle		
	P	R	F1	P	R	F1
APIL	0,45	0,46	0,46	0,61	0,61	0,61
Wikinews	0,78	0,65	0,71	0,86	0,71	0,78
Sequoia	0,73	0,62	0,67	0,82	0,70	0,77
Aijwikiner	0,80	0,76	0,78	0,91	0,86	0,88
GSD	0,73	0,69	0,71	0,84	0,79	0,81
Spoken	0,78	0,55	0,64	0,87	0,61	0,71
pg6470	0,86	0,76	0,80	0,93	0,81	0,87
tous	**0,70**	**0,63**	**0,66**	**0,81**	**0,73**	**0,77**

Comparaison des exactitudes strictes sur le corpus entier

Exactitude	Stricte		
	P	R	F1
Unitex-CasEN	0,81	0,44	0,57
SpaCy	0,70	0,63	0,66

Conclusion

Comment comparer deux outils qui ne réalisent pas exactement la même tâche ?

- choix
- prise en compte de la particularité des sorties

Pour aller plus loin :

- combinaison d'outils
- annotation automatique + correction manuelle

Merci pour votre attention.

References

Richard Eckart De Castilho et al. "A web-based tool for the integrated annotation of semantic and syntactic structures". In: *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*. 2016, pp. 76–84.

Maud Ehrmann. "Les Entitées Nommées, de la linguistique au TAL: Statut théorique et méthodes de désambiguïsation". PhD thesis. Paris Diderot University, 2008.

Matthew Honnibal et al. *spaCy: Industrial-strength Natural Language Processing in Python*. 2020. DOI: 10.5281/zenodo.1212303. URL: https://doi.org/10.5281/zenodo.1212303.

Denis Maurel et al. "Cascades de transducteurs autour de la reconnaissance des entités nommées". In: *Traitement automatique des langues* 52.1 (2011), pp. 69–96.

Sophie Rosset et al. "Structured named entities in two distinct press corpora: Contemporary broadcast news and old newspapers". In: *6th*