Faculté des langues
Université de Strasbourg

Master TdL

CORPUS PARALLÈLES ET ANNOTATION DES COLLOCATIONS

Soutenance de mémoire de master 2, Master *Technologies des langues*
Université de Strasbourg, lundi 19 septembre 2022

Soutenu par
Bastien Giordano

Co-dirigé par
Amalia Todirascu, professeure des universités, Université de Strasbourg
Frédéric Imbert, professeur des universités, Aix-Marseille Université

Plan de la présentation

Faculté des langues Université de Strasbourg

Plan de la présentation

1.1. Motivations (1/3)

Faculté des langues Université de Strasbourg

- Expressions polylexicales (EP) : omniprésentes dans la langue
- Posent de nombreux problèmes au traitement automatique des langues (TAL)
- Qualifiées de '*pain in the neck*' (Sag et al., 2002)
- Toujours le cas aujourd'hui : hétérogénéité importante
- Au centre de nombreuses recherches (Savary et al., 2015; 2017; 2018)

1.1. Motivations (2/3)

Faculté des langues Université de Strasbourg

- Plusieurs classes d'EP (Constant et al., 2017) :
  - Termes complexes : *canon à eau*
  - Mots composés : *porte-drapeau*
  - Expressions idiomatiques : *joindre les deux bouts*
  - Entités nommées : *Ministre de la culture*
  - Locutions : *peu à peu*
  - Collocations : *buveur invétéré, conclure un contrat*

1.1. Motivations (3/3)

- Relations syntaxiques / mesures statistiques insuffisantes
- De nombreux outils développés : `mwetoolkit` (Ramisch, 2012), Fips (Wehrli, 2007), Veyn (Zampieri et al., 2018)
- Ressources annotées en collocations : rares et disparates
- D'autant plus dans les corpus parallèles bi- et multilingues
- Manière d'annoter propre à chaque projet
- Recherches souvent concentrées sur un type de collocations

1.2. Objectifs

Faculté des langues Université de Strasbourg

- Construction d'un **corpus parallèle trilingue** (français, anglais, arabe) **multi-genre annoté en collocations verbales**
- Réalisation d'une **étude linguistique contrastive** sur les collocations et leur usage entre les langues

Plan de la présentation

2.1. Notions d'EP et de collocation (1/3)

Faculté des langues Université de Strasbourg

- Deux caractéristiques pour les EP :
  - Association d'au moins **2 unités lexicales**
  - **Idiomaticité prononcée** à différents niveaux (Baldwin et Kim, 2010) :
    - Lexical : *ex nihilo*, *ad hoc*
    - Syntaxique : *tout à coup* (syntaxe « déviante ») (Sag et al., 2002)
    - Sémantique : *passer au bleu* (forte opacité), *prêter l'oreille* (sens figuré)
    - Pragmatique : pragmatèmes (Tutin, 2015), clichés linguistiques (Mel'čuk, 2013)
    - Statistique : fréquence de cooccurrence élevée

2.1. Notions d'EP et de collocation (2/3)

Faculté des langues Université de Strasbourg

- **Approche statistique**
  - Combinaisons de mots récurrentes et arbitraires (Benson, 1990; Smadja, 1993)
  - Transparence sémantique (Cruse, 1996)
  - Baldwin et Kim (2010) : '[Collocation is] in our terms, a statistically idiomatic [multiword expression] (esp. of high frequency).'
- **Approche linguistique**
  - Patrons syntaxiques à l'œuvre (Tutin et Grossmann, 2002; Hausmann, 1989)
  - Relation syntaxique directe des constituants (Bartsch, 2004)

2.1. Notions d'EP et de collocation (3/3)

Faculté des langues Université de Strasbourg

- **Critères consensuels**
  - Arbitrarité de la combinaison
  - Récurrence de la cooccurrence
  - Transparence et absence de figement
  - Binarité (2 lexies ou syntagmes associés)
- Notre position : une collocation est une **EP statistiquement idiomatique avec relation de dépendance syntaxique** (sujet/verbe, verbe/objet pour ce projet)

2.2. Découverte et identification automatique (1/3)

- **Découverte** : inclut acquisition / extraction d'EP (Constant et al., 2017)
- **Identification** : ajout d'annotations aux EP découvertes
- Difficultés pour les 2 tâches :
  - Variabilité syntaxique : *il **a pris** des **mesures**, des **mesures furent prises***, etc.
  - Ambiguïté sémantique : *Kim **made** a **face** to the policeman, Kim **made** a **face** in pottery class* (Baldwin et Kim, 2010)
  - Éventuelle discontinuité : *elle **pose** réellement de drôles de **questions***
  - Imbrication : *elle **dirige**$_{1\ 2}$ le **débat**$_{1}$ et les **opérations**$_{2}$*

2.2. Découverte et identification automatique (2/3)

- Mesures d'association : **nécessaires** pour différencier les collocations des associations compositionnelles régulières
  - *Manger une salade* > association fréquente mais pas une collocation
  - وفي (*wa fī*, « et dans ») > association fréquente mais pas une collocation
  - Frontière parfois très fine
- Pas de mesure d'association plus performante qu'une autre (Constant et al., 2017)
  - MI, LLR, χ-squared, coefficient de Dice, etc.

2.2. Découverte et identification automatique (3/3)

- 3 types d'outils d'extraction de collocations (Todirascu et al., 2008) :
  - Approche statistique
    - Xtract (Smadja, 1993)
    - Inconvénient : trop grand nombre de candidats
  - Approche syntaxique / symbolique
    - Fips (Wehrli, 2007)
    - Résout les difficultés liées au passif, topicalisations, dislocations, clivées (Seretan, 2008)
  - Approche hybride
    - `mwetoolkit` (Ramisch, 2012)
    - Apprentissage automatique et plongements lexicaux (Garcia et al., 2017)
    - Réseaux de neurones récurrents : Veyn (Zampieri et al., 2018)

Plan de la présentation

Faculté des langues Université de Strasbourg

3.1. Outils et ressources utilisés

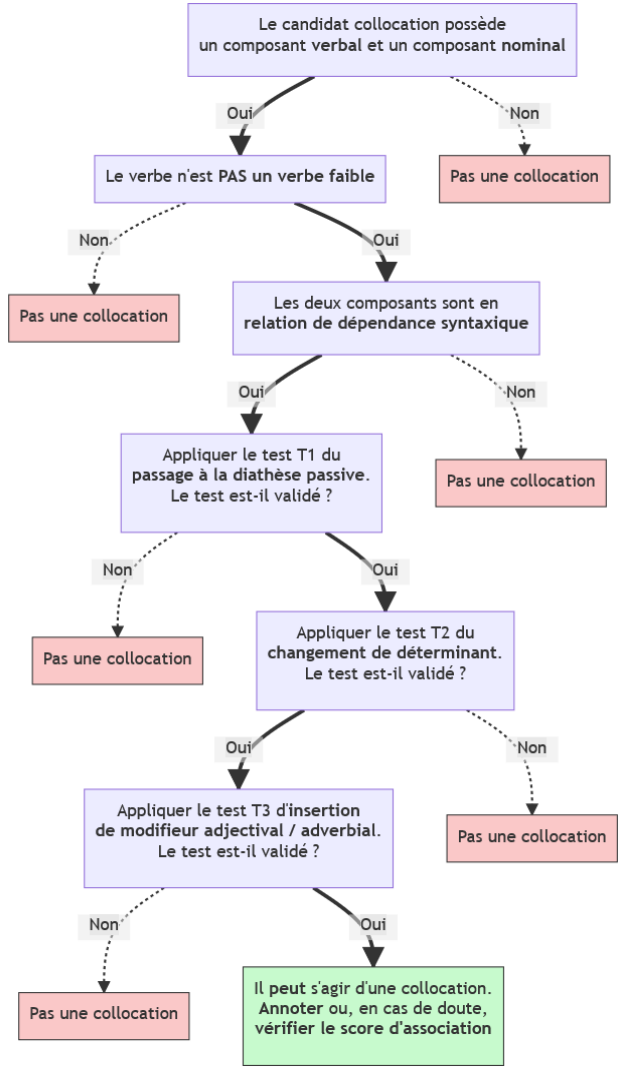
Faculté des langues Université de Strasbourg

- Corpus parallèles exploités :
  - **Global Voices** v2018q4 (textes journalistiques) (Tiedemann, 2012)
  - **TED2020** (transcriptions de conférences) (Reimers et Gurevych, 2020)
  - **Nations Unies v1.0** (textes juridiques) (Ziemski et al., 2016)
  - **WikiMatrix** (textes encyclopédiques) (Schwenk et al., 2019)
- Annotation automatique du corpus français avec **VarIDE** (Pasquer et al., 2018)
  - PARSEME Shared Task 1.1
  - **Hypothèse : identification des EP verbales plus efficace avec apprentissage des motifs de variabilité morphosyntaxique**

3.2. Guide d'annotation et corpus d'entraînement

Faculté des langues Université de Strasbourg

- Guide basé sur les travaux de PARSEME et SimpleApprenant (Todirascu et Cargill, 2019)
- Tests linguistiques (SimpleApprenant)
  - **Passage à la diathèse passive**
  - **Changement de déterminant**
  - **Ajout de modifieurs**
- Arbre de décision (PARSEME)



- Corpus d'entraînement PARSEME Shared Task 1.1 modifié
  - **Suppression** des :MVC, :VPC et :IRV
  - **Transformation** des :LVC et :VID en :COLL
  - **Vérification manuelle** des annotations :COLL

3.3. Annotation automatique et évaluation

Faculté des langues Université de Strasbourg

- Échantillonnage et création de **tritextes**
- Conversion des tritextes **au format .cupt** (CoNLL-U étendu)
- Entre 20 000 et 30 000 phrases selon le sous-corpus, soit un total légèrement supérieur à **100 000 triplets de phrases alignées**

- Évaluation standard réalisée sur un échantillon de 500 phrases extraites du corpus d'entraînement (avant entraînement)

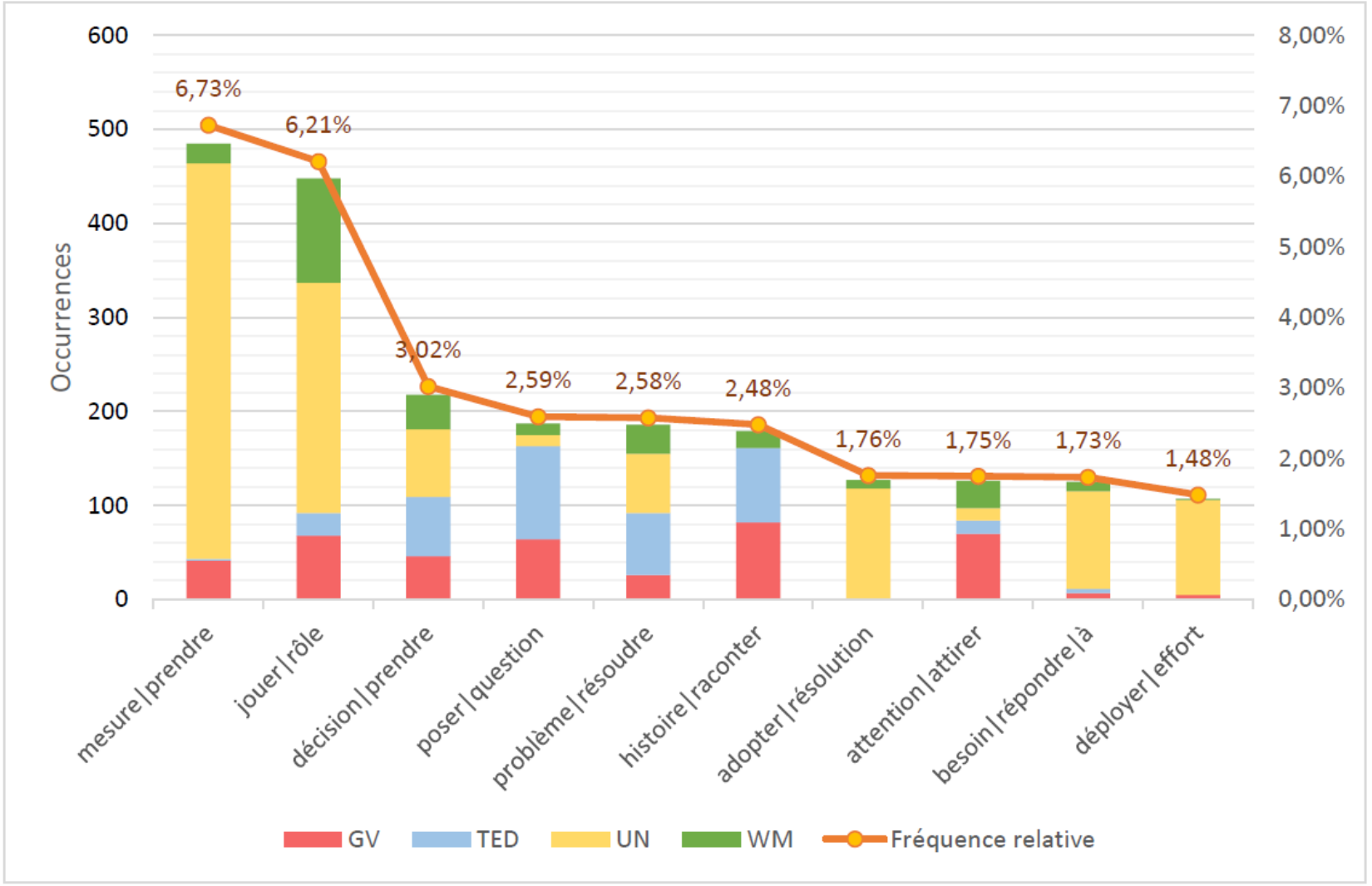
	Corpus d'entraînement	Corpus de test
Phrases	16725	500
Annotations	912	46

- Précision : 0.89, Rappel : 0.74, **F-mesure : 0.81**

3.3. Annotation automatique et évaluation

Faculté des langues Université de Strasbourg

- Après annotation automatique, **étape très chronophage de correction / augmentation manuelle**
- Annotations avant / après correction : **6968 / 7211**
- **F-mesure : 0.84**



Plan de la présentation

Faculté des langues Université de Strasbourg

4.1. Outils et ressources utilisés (1/2)

- **GIZA++** (Och et Ney, 2003)
  - Création de tables de traduction bilingues
- **ZAP** (Akbik et Vollgraff, 2018)
  - Projection des annotations canoniques (parties du discours, relations de dépendances, entités nommées, cadres sémantiques)
  - Depuis l'anglais vers le français, l'espagnol et l'allemand
  - **S'appuie sur des tables de traduction bilingues**
  - **Limites** : impossibilité de projeter depuis le français, impossibilité de projeter des annotations personnalisées (COLL), pas d'arabe

4.2. Outils et ressources utilisés (2/2)

Faculté des langues Université de Strasbourg

- **Génération de fichiers d'alignements** phrase à phrase

```
{18 Dubai={20 Dubaï=1.0}, 7 blogger={6 blogueur=1.0}, 20 reputation={18
réputation=1.0}, 13 n't={16 pas=1.0}, 10 the={17 la=1.0}, 2 the={5 le=1.0},
8 Ammaro={9 Ammaro=1.0}, 17 tarnish={15 ternira=1.0}, 11 case={13
affaire=1.0}, 4 hand={3 côté=1.0}}
```

- **Transformation** de la sortie

```
[identifiant_phrase, identifiant_token_anglais, forme_token_anglais,
identifiant_token_français, forme_token_français, score]
```

- **Enrichissement** avec annotations COLL

```
[['195', '18', 'Dubai', '20', 'Dubaï', '1.0', '*'], ['195', '7', 'blogger',
'6', 'blogueur', '1.0', '*'], ['195', '20', 'reputation', '18',
'réputation', '1.0', '1'], ['195', '13', "n't", '16', 'pas', '1.0', '*'],
['195', '10', 'the', '17', 'la', '1.0', '*'], ['195', '2', 'the', '5', 'le',
'1.0', '*'], ['195', '8', 'Ammaro', '9', 'Ammaro', '1.0', '*'], ['195',
'17', 'tarnish', '15', 'ternira', '1.0', '1:COLL'], ['195', '11', 'case',
'13', 'affaire', '1.0', '*'], ['195', '4', 'hand', '3', 'côté', '1.0', '*']]
```

- **Projection** vers le fichier .cupt anglais

4.2. Projection FR > EN et évaluation (1/2)

Faculté des langues Université de Strasbourg

- Résultats bruts de la projection

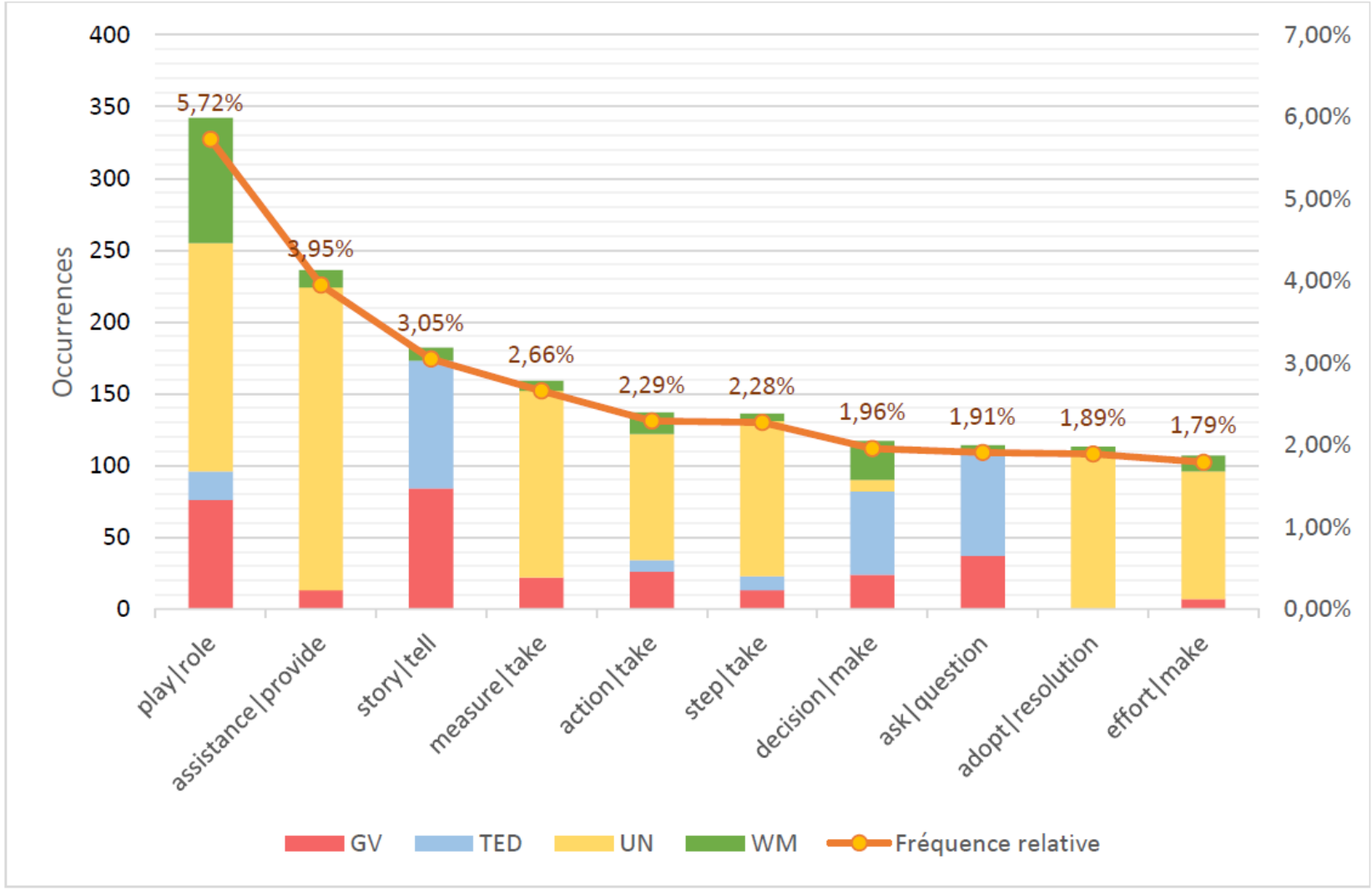
Corpus	Tokens annotés (FR)	Tokens annotés (EN)	Différence
GV	3522	2224	-36,85%
TED	1565	1205	**-23,00%**
UN	5867	3596	-38,70%
WM	3735	2532	-32,21%
Tous	14 669	9557	**-34,85%**

- Évaluation standard réalisée sur 500 phrases extraites aléatoirement, comparée à une projection manuelle

Base	Précision	Rappel	F-mesure
Collocation	0.49	0.83	**0.62**
Token	0.64	**0.88**	**0.74**

4.2. Projection FR > EN et évaluation (2/2)

- Erreurs liées à :
  - Encapsulation
  - Ellipse
  - Choix lexical différent
  - Transfert des parties du discours
  - Reformulation
  - Verbes à particules
- **F-mesure EP : 0.41**
- **F-mesure token : 0.51**



4.2. Projection EN > AR et évaluation (1/3)

- **Quelques différences méthodologiques :**
  - Création d'une **table de traduction bilingue EN-AR avec GIZA++**
  - **Modifications dans le code source de ZAP** pour exploiter cette ressource EN-AR
- Même méthodologie pour le reste
- **Hypothèse : avec une double projection, on pourrait augmenter le rappel de la projection :**
  - Création d'une **table de traduction bilingue FR-AR avec l'anglais comme langue pivot**
  - Après projection EN > AR, ajout d'une projection FR > AR pour **couvrir les tokens qui n'auraient pas été annotés avec la première**

4.2. Projection EN > AR et évaluation (2/3)

Faculté des langues Université de Strasbourg

- **Résultats bruts** de la projection simple EN > AR

Corpus	Tokens annotés (EN)	Tokens annotés (AR)	Différence
Complet	14 088	7261	-39,44%

- **Comparaison avec projection double** EN+FR > AR

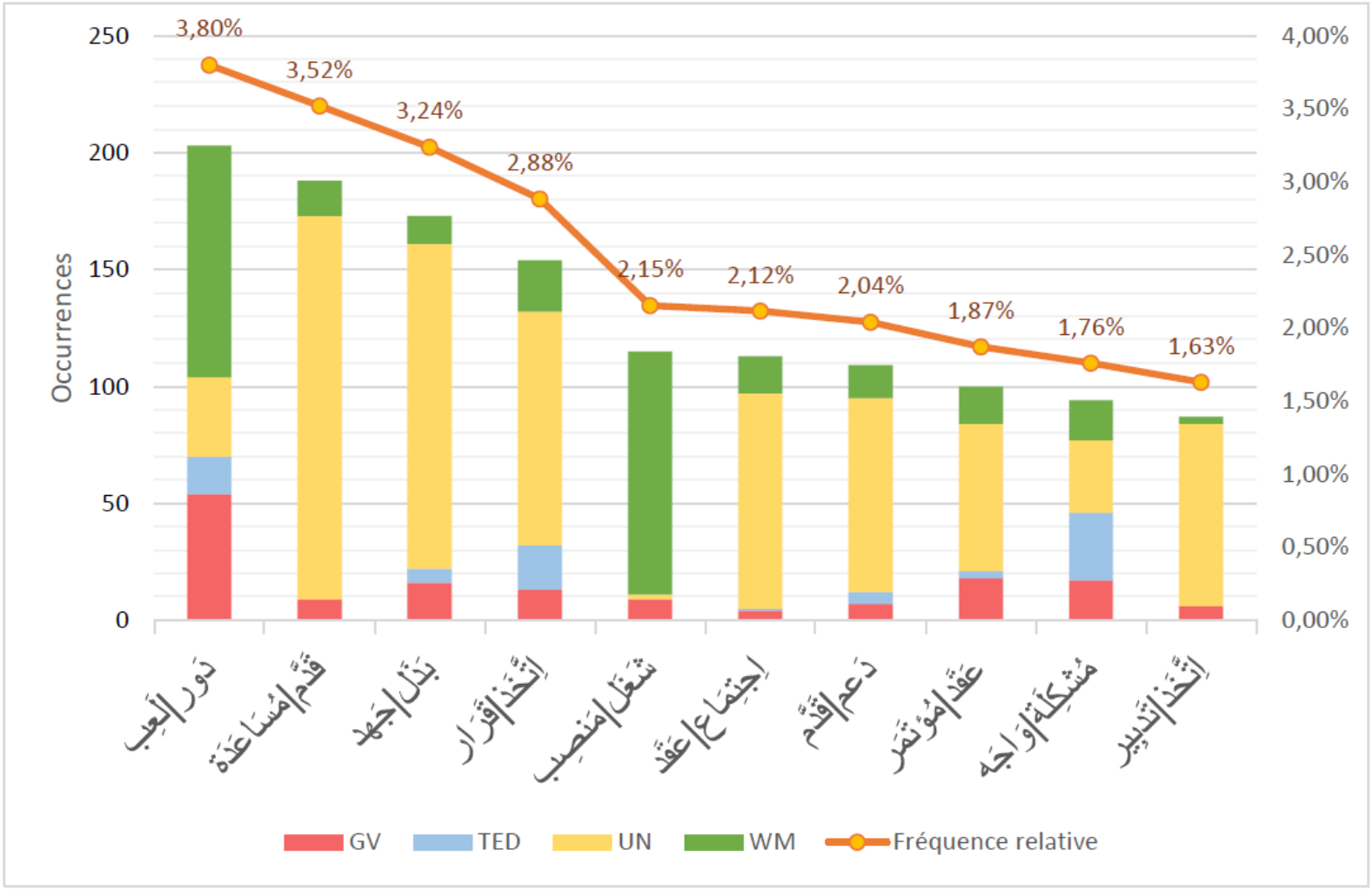
Corpus	Tokens annotés (SP)	Tokens annotés (DP)	Différence
Complet	7261	9882	+26,52%

- **Évaluation standard** réalisée sur 500 phrases extraites aléatoirement, comparée à une projection manuelle

Projection	Base	Précision	Rappel	F-mesure
Simple	Collocation	0.21	0.40	**0.27**
	Token	0.46	0.64	**0.53**
Double	Collocation	0.21	0.40	**0.27**
	Token	0.46	**0.64**	**0.53**

4.2. Projection EN > AR et évaluation (3/3)

- Erreurs liées à :
  - *Maṣdar* + nom
  - Tournures passives avec تَمَّ *(tamma)*
  - Participes passés étiquetés ADJ
  - Mêmes phénomènes que précédemment
- **F-mesure EP : 0.17**
- **F-mesure token : 0.32**



Plan de la présentation

5.1. Approche quantitative (1/2)

- **Français (7211 annotations), anglais (5976)** et **arabe (5342)**
- **Corrélation entre la fréquence relative des collocations verbales et la dimension diamésique du discours et le degré de normalisation de la langue**
- Pourcentage de phrases contenant au moins une collocation :

Corpus	FR	EN	AR	TRI
GV	0.40%	0,41%	0,36%	**0,39%**
TED	0.25%	0,26%	0,20%	**0,24%**
UN	0.64%	0,61%	0,53%	**0,60%**
WM	0,38%	0,34%	0,39%	**0,37%**

5.1. Approche quantitative (2/2)

- **Même corrélation est observable au niveau de la distance** entre les composants des collocations (plus élevée pour les textes juridiques, plus basse pour les transcriptions de conférences)
- Distance plus élevée en moyenne **en français (3,08 tokens)**, devant l'**arabe (2,92 tokens)** puis l'**anglais (2,80 tokens)**
- Dans le corpus complet, **distance maximum = 42 tokens (UN@EN)**
- **Proportion collocations continues / discontinues** intéressante au niveau des langues

Corpus	Type	FR	EN	AR	TRI
Complet	Continues	737	1309	1835	1293,67
	Discontinues	6474	4667	3507	4882,67
	Proportion	**10,22%**	**21,90%**	**34,35%**	**20,95%**

5.2. Approche qualitative (1/2)

- **Caractère nominal de l'arabe prédominant**
  - ***Maṣdar* + nom** (formes non finies remplacées par un nom d'action)
    - تشويه سمعة دبي - *tašwīh sumʿat dubayy*
    - (litt. « *le ternissement de la réputation de Dubai* »)
  - **Participes passés** fondamentalement adjectifs (étiquetés ADJ)
    - القرارات المتّخذة - *al-qarārāt al-muttaḫiḏa*
    - (« *les décisions prises* » -> NOUN+ADJ)
  - **Tournure passive avec تَمَّ *(tamma)*** accompagnée d'une annexion
    - تمّ اتّخاذ الإجراء - *tamma ittiḫāḏ al-ʾiǧrāʾ*
    - (litt. « *a été la prise de la mesure* »)

5.2. Approche qualitative (2/2)

- **Phénomènes liés à une traduction de qualité**
  - **Encapsulation** : *'to ask' /* سأل (saʾala) pour « *poser / question* »
  - **Ellipse** : élision du verbe car non nécessaire à la compréhension
  - **Transposition** : *'transmitted diseases'* -> « *maladies transmissibles* »
  - **Reformulation** : lié au style et / ou à la fluence (p. ex. « *se noie dans le sang* » pour *'reached its bloodiest peak'*)

- **Phénomènes liés à une traduction de qualité moindre**
  - **Appauvrissement lexical** : « *faire un film* » <- *'to shoot a film*
  - **Calque** : « *servir sa peine* » pour *'serve his sentence'*
  - **Omission** : segment tout simplement omis par la traduction

Plan de la présentation

6.1. Apports

- **Corpus parallèle trilingue entièrement annoté en collocations verbales (3 * 100 000 phrases)**
- Pourrait servir à **l'évaluation d'outils d'annotation automatique**
- Évaluation de VarIDE avec nos données annotées (répartition 80/20) :

Langue	Précision	Rappel	F-mesure
FR	0.82	0.94	**0.88**
EN	0.79	0.73	**0.76**
AR	0.58	0.06	**0.12**

- **Méthode de projection originale**, valable mais perfectible

6.2. Limites

- **Travail d'annotation solitaire** => difficilement acceptable
- **Méthode de projection largement perfectible** avec plus de temps
  - Prise en considération des étiquettes grammaticales
  - Génération d'un rapport de projection
  - Enrichissement des tables de traduction
- **Travail de correction / augmentation de corpus après projection beaucoup trop important**

6.3. Perspectives

Faculté des langues Université de Strasbourg

- **Facilitation de l'étape de nettoyage de corpus** : semi-automatisation
- **ZAP mériterait d'être amélioré**
  - Projection depuis **d'autres langues** que l'anglais
  - **Ajout de langues-cibles** (tables de traduction)
  - Possibilité de **projeter les annotations personnelles**
- **Progrès significatifs** effectués dans l'identification des EP verbales depuis le début de ce projet
  - PARSEME ST 1.1 : meilleur outil **F-mesure@19 langues à 0.54**
  - PARSEME ST 1.2 : meilleur outil **F-mesure@14 langues à 0.70**
- **Réemploi de la méthodologie pour d'autres types d'EP**

Merci de votre attention !

Thank you for your attention!

أشكركم لإصغائكم شكراً جزيل !

Références (1/2)

Akbik, A., & Vollgraf, R. (2018). ZAP : An Open-Source Multilingual Annotation Projection Framework. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Baldwin, T., & Kim, S. N. (2010). Multiword Expressions. In N. Indurkhya & F. J. Damerau (Éds.), *Handbook of Natural Language Processing* (Second Edition, p. 267-292). CRC Press.

Bartsch, S. (2004). *Structural and functional properties of collocations in English : A corpus study of lexical and pragmatic constraints on lexical co-occurrence*. Gunter Narr Verlag.

Benson, M. (1990). Collocations and general-purpose dictionaries. *International Journal of Lexicography*, *3*(1), 23-34.

Brashi, A. S. (2005). *Arabic collocations : Implications for translation* [PhD Thesis]. University of Western Sydney.

Constant, M., Eryiğit, G., Monti, J., van der Plas, L., Ramisch, C., Rosner, M., & Todirascu, A. (2017). Multiword expression processing : A survey. *Computational Linguistics*, *43*(4), 837-892. https://doi.org/10.1162/COLI_a_00302

Cruse, D. A. (1986). *Lexical semantics*. Cambridge university press.

Dagan, I., Church, K., & Gale, W. (1999). Robust bilingual word alignment for machine aided translation. In *Natural Language Processing Using Very Large Corpora* (p. 209-224). Springer.

Erjavec, T. (2004). MULTEXT-East Version 3 : Multilingual Morphosyntactic Specifications, Lexicons and Corpora. *LREC*.

Erjavec, T. (2012). MULTEXT-East : Morphosyntactic resources for Central and Eastern European languages. *Language resources and evaluation*, *46*(1), 131-142.

Gale, W. A., & Church, K. (1993). A program for aligning sentences in bilingual corpora. *Computational linguistics*, *19*(1), 75-102.

Garcia, M., García-Salido, M., & Alonso-Ramos, M. (2017). Using bilingual word-embeddings for multilingual collocation extraction. *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, 21-30. https://doi.org/10.18653/v1/W17-1703

Habash, N. Y. (2010). Introduction to Arabic natural language processing. *Synthesis Lectures on Human Language Technologies*, *3*(1), 1-187.

Hausmann, F. J. (1989). *Wörterbücher : Ein internationales Handbuch zur Lexikographie*. W. de Gruyter.

Ide, N., & Pustejovsky, J. (2017). *Handbook of linguistic annotation*. Springer.

Kraif, O. (2001). Exploitation des cognats pour l'alignement : Architecture et évaluation. *Traitement automatique des langues*, *42*(3), 833-867.

Kraif, O. (2015). Multialignement vs bialignement : À plusieurs, c'est mieux! *TALN 2015, 22e conférence sur le Traitement automatique des langues naturelles*.

Mel'čuk, I. (1995). Phrasemes in language and phraseology in linguistics. *Idioms: Structural and psychological perspectives*, 167-232.

Mel'čuk, I. (2013). Tout ce que nous voulions savoir sur les phrasèmes, mais. *Cahiers de lexicologie*, *102*(1), 129-149.

Och, F. J., & Ney, H. (2003). A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, *29*(1), 19-51. https://doi.org/10.1162/089120103321337421

Pasquer, C., Ramisch, C., Savary, A., & Antoine, J.-Y. (2018). VarIDE at PARSEME Shared Task 2018 : Are Variants Really as Alike as Two Peas in a Pod? *COLING Workshop on Linguistic Annotation, Multiword Expressions and Constructions*.

Ramisch, C. (2012). *A generic and open framework for multiword expressions treatment : From acquisition to applications* [PhD Thesis]. Université de Grenoble.

Reimers, N., & Gurevych, I. (2020, octobre 5). Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. http://arxiv.org/abs/2004.09813

Références (2/2)

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., & Flickinger, D. (2002). Multiword Expressions : A Pain in the Neck for NLP. In A. Gelbukh (Éd.), *Computational Linguistics and Intelligent Text Processing* (Vol. 2276, p. 1-15). Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-45715-1_1

Savary, A., Candito, M., Mititelu, V. B., Bejček, E., Cap, F., Čéplö, S., Cordeiro, S. R., Gülşen Eryiğit, Giouli, V., Gompel, M. V., HaCohen-Kerner, Y., Kovalevskaitė, J., Krek, S., Liebeskind, C., Monti, J., Escartín, C. P., Plas, L. V. D., Behrang QasemiZadeh, Ramisch, C., … Vincze, V. (2018). PARSEME multilingual corpus of verbal multiword expressions. In *Multiword expressions at length and in depth : Extended papers from the MWE 2017 workshop*. Language Science Press. https://doi.org/10.5281/ZENODO.1471591

Savary, A., Ramisch, C., Cordeiro, S. R., Sangati, F., Vincze, V., Qasemi Zadeh, B., Candito, M., Cap, F., Giouli, V., & Stoyanova, I. (2017). The PARSEME shared task on automatic identification of verbal multiword expressions. *Proceedings of the 13th Workshop on Multiword Expression (MWE 2017)*, 31-47.

Savary, A., Sailer, M., Parmentier, Y., Rosner, M., Rosén, V., Przepiórkowski, A., Krstev, C., Vincze, V., Wójtowicz, B., & Losnegaard, G. S. (2015). PARSEME–PARSing and Multiword Expressions within a European multilingual network. *7th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC 2015)*.

Schwenk, H., Chaudhary, V., Sun, S., Gong, H., & Guzmán, F. (2019, juillet 15). WikiMatrix : Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia. *arXiv:1907.05791 [cs]*. http://arxiv.org/abs/1907.05791

Seretan, V. (2008). *Collocation extraction based on syntactic parsing* [PhD Thesis, University of Geneva]. http://archive-ouverte.unige.ch/unige:78

Smadja, F. (1993). Retrieving collocations from text : Xtract. *Computational linguistics*, *19*(1), 143-178.

Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. In N. Calzolari, K. Choukri, T. Declerck, M. Ugur Dogan, B. Maegaard, J. Mariani, J. Odijk, & S. Piperidis (Éds.), *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)* (p. 2214-2218). European Language Resources Association (ELRA).

Todirascu, A., & Cargill, M. (2019). SimpleApprenant : A platform to improve French L2 learners' knowledge of multiword expressions. *CALL and complexity*, 356.

Todiraşcu, A., Heid, U., Ştefănescu, D., Tufiş, D., Gledhill, C., Weller, M., & Rousselot, F. (2008). Vers un dictionnaire de collocations multilingue. *Cahiers de linguistique*, *33*(1), 161-186.

Tutin, A., Esperança-Rodier, E., Iborra, M., & Reverdy, J. (2015). Annotation of multiword expressions in French. *European Society of Phraseology Conference (EUROPHRAS 2015)*, 60-67.

Tutin, A., & Grossmann, F. (2002). Collocations régulières et irrégulières : Esquisse de typologie du phénomène collocatif. *Revue francaise de linguistique appliquée*, *Vol. VII*(1), 7-25.

Varga, D., Halácsy, P., Kornai, A., Nagy, V., Németh, L., & Trón, V. (2007). Parallel corpora for medium density languages. *Amsterdam Studies In The Theory And History Of Linguistic Science Series 4*, *292*, 247.

Wehrli, E. (2007). Fips, a « deep » linguistic multilingual parser. *Proceedings of the Workshop on Deep Linguistic Processing - DeepLP '07*, 120. https://doi.org/10.3115/1608912.1608931

Zampieri, N., Scholivet, M., Ramisch, C., & Favre, B. (2018). Veyn at parseme shared task 2018 : Recurrent neural networks for vmwe identification. *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, 290-296.

Ziemski, M., Junczys-Dowmunt, M., & Pouliquen, B. (2016). The United Nations Parallel Corpus v1.0. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 3530-3534. https://www.aclweb.org/anthology/L16-1561

Spécificités de l'arabe

- **Collocations issues du Coran**
  - عاهد عهداً *(ʿāhada ʿahdan, « conclure un pacte »)* / قصّ قصّة *(qaṣṣa qiṣṣatan, « raconter une histoire »)*
- **Emprunts calqués**
  - أبدى اهتمام *(ʾabdā ihtimām, « montrer de l'intérêt »)*
- **Morphologie agglutinante**
  - ضربكم مثلها *(ḍarabakum maṯalahā, « il vous a donné son exemple (à elle) »)*
- **Patrons de collocations uniques**
  - Verbe + *mafʿūl mutlaq* (complément absolu) : فرح فرحاً شديداً *(faraḥa fariḥan šadīdan, « il s'est réjoui d'une forte joie (litt.) »)*
  - Verbe + *tamyīz* (complément spécificatif) : نظر إليه شزراً *(naẓara ʾilayhi šazran, « il l'a regardé de travers »)*
  - Verbe + *ḥāl* (complément de manière) : ولّى هارباً *(wallā hāriban, « tourner en fuyant (litt.) »)*

Collocations et équivalents

- Exemples tirés de (Brashi, 2005) et augmentés avec le français

	Traduction transparente	Arbitrarité du verbe	Pas d'équivalence
FR	gagner la confiance	donner une leçon	se suicider (v. pron.)
EN	to win confidence	to teach a lesson	to commit suicide (collocation)
AR	كسب ثقة (kasaba ṯiqatan)	لقّن درسًا (laqqana darsan)	انتحر (intaḥara)

Exemple de triplet de phrases

Faculté des langues Université de Strasbourg

Source : corpus parallèle TED2020

FR : *Le premier est le simple pouvoir de bons outils de visualisation pour aider à démêler la complexité et vous encourager à* ***poser*** *les* ***questions*** *auxquelles vous n'avez pas pensé avant.*

EN : *First is the simple power of good visualization tools to help untangle complexity and just encourage you to* ***ask questions*** *you didn't think of before.*

AR : الاولى هي .. قوة العرض البصري للمشكلة .. وهي طريقة ناجعة لفك التشعب
فالتمثيل البصري يشجعك على **طرح اسئلة** لم تكن تخطر ببالك من ذي قبل

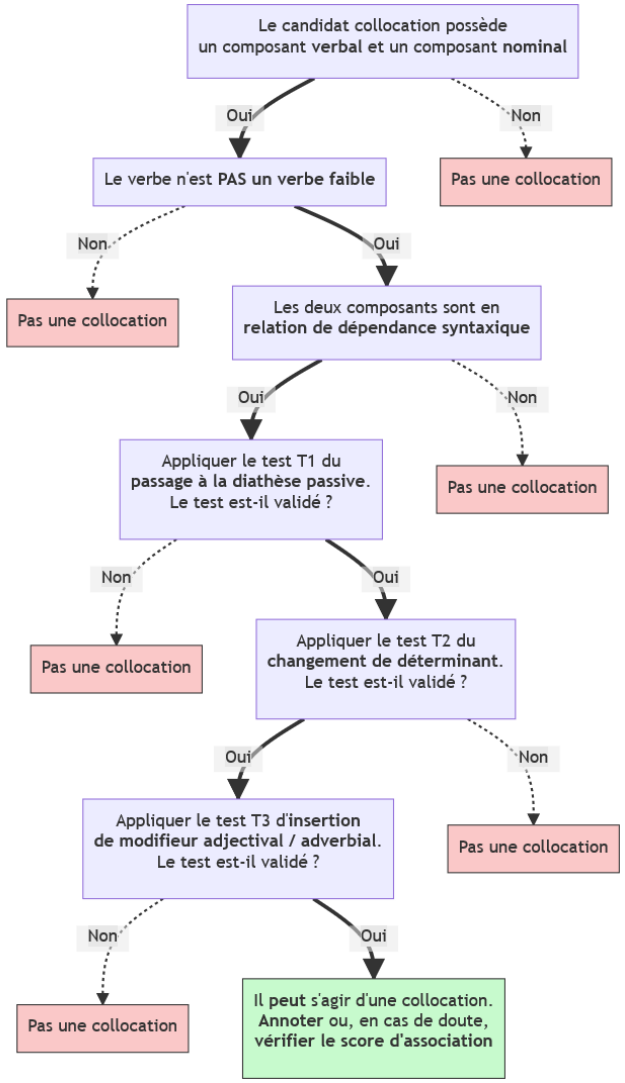
Distance maximum entre composants

Faculté des langues Université de Strasbourg

Source : corpus parallèle United Nations

Although no specific ***[action]_1:COLL*** *to review the fundamental factors that negatively affect the observance of the principles of national sovereignty and non-interference in the internal affairs of States in their electoral processes, as requested by the General Assembly in its resolution 47/130, was* ***[taken]_1*** *by the Commission at that session, references were made, in a number of resolutions, to the issue of elections in the context of guaranteeing the free expression of the will of peoples and ensuring respect for national sovereignty and non-interference in the internal affairs of the States concerned.*

Arbre de décision guide d'annotation (exemple)



- *« Laissez-moi vous **raconter** une **histoire**… »*
- Composant **verbal** et **nominal** ? Oui
- Verbe **faible** ? Non
- Relation de **dépendance syntaxique** ? Oui
- Passage à la **diathèse passive** possible ? Oui
  - *une **histoire** est **racontée***
- **Changement de déterminant** possible ? Oui
  - *raconter **mon / cette / l'**histoire*
- **Insertion d'un modifieur adjectival** possible ? Oui
  - *raconter une histoire **rocambolesque***
- **Information mutuelle** sur VW : >6 (1730 occ.)

Corpus d'entraînement original vs. adapté (1/2)

Faculté des langues Université de Strasbourg

```
# global_columns = ID FORM LEMMA UPOS XPOS FEATS HEAD DEPREL DEPS MISC PARSEME:MWE
# source_sent_id = http://handle.net/11234/1-2515 UD_French/fr-ud-train.conllu fr-ud-train_07811
# text = Cela a pu jouer un rôle, mais avouez que si tel est le cas cela est regrettable.
1   Cela        cela        PRON  _  Number=Sing|PronType=Dem                                  3   nsubj  _  _  *
2   a           avoir       AUX   _  Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin   3   aux:tense  _  _  *
3   pu          pouvoir     VERB  _  Tense=Past|VerbForm=Part                                 0   root   _  _  *
4   jouer       jouer       VERB  _  VerbForm=Inf                                            3   xcomp  _  _  2:LVC.full
5   un          un          DET   _  Definite=Ind|Gender=Masc|Number=Sing|PronType=Art       6   det    _  _  *
6   rôle        rôle        NOUN  _  Gender=Masc|Number=Sing                                 4   obj    _  _  2
7   ,           ,           PUNCT _  _                                                       3   punct  _  _  *
8   mais        mais        CCONJ _  _                                                       9   cc     _  _  *
9   avouez      avouer      VERB  _  Mood=Imp|Number=Plur|Person=2|Tense=Pres|VerbForm=Fin  3   conj   _  _  *
10  que         que         SCONJ _  _                                                       18  mark   _  _  *
11  si          si          SCONJ _  _                                                       15  mark   _  _  *
12  tel         tel         ADJ   _  Gender=Masc|Number=Sing                                 15  nsubj  _  _  1:VID
13  est         être        AUX   _  Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin   15  cop    _  _  1
14  le          le          DET   _  Definite=Def|Gender=Masc|Number=Sing|PronType=Art       15  det    _  _  1
15  cas         cas         NOUN  _  Gender=Masc|Number=Sing                                 18  obl:mod  _  _  1
16  cela        cela        PRON  _  Number=Sing|PronType=Dem                                18  nsubj  _  _  *
17  est         être        AUX   _  Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin   18  cop    _  _  *
18  regrettable regrettable ADJ   _  Number=Sing                                            9   ccomp  _  _  *
19  .           .           PUNCT _  _                                                       3   punct  _  _  *
```

Corpus d'entraînement original vs. adapté (2/2)

Faculté des langues Université de Strasbourg

```
# global_columns = ID FORM LEMMA UPOS XPOS FEATS HEAD DEPREL DEPS MISC PARSEME:MWE
# source_sent_id = http://handle.net/11234/1-2515 UD_French/fr-ud-train.conllu fr-ud-train_07811
# text = Cela a pu jouer un rôle, mais avouez que si tel est le cas cela est regrettable.
1   Cela        cela        PRON  _  Number=Sing|PronType=Dem                                  3   nsubj  _  _  *
2   a           avoir       AUX   _  Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin   3   aux:tense  _  _  *
3   pu          pouvoir     VERB  _  Tense=Past|VerbForm=Part                                 0   root   _  _  *
4   jouer       jouer       VERB  _  VerbForm=Inf                                            3   xcomp  _  _  1:COLL
5   un          un          DET   _  Definite=Ind|Gender=Masc|Number=Sing|PronType=Art       6   det    _  _  *
6   rôle        rôle        NOUN  _  Gender=Masc|Number=Sing                                 4   obj    _  _  1
7   ,           ,           PUNCT _  _                                                       3   punct  _  _  *
8   mais        mais        CCONJ _  _                                                       9   cc     _  _  *
9   avouez      avouer      VERB  _  Mood=Imp|Number=Plur|Person=2|Tense=Pres|VerbForm=Fin  3   conj   _  _  *
10  que         que         SCONJ _  _                                                       18  mark   _  _  *
11  si          si          SCONJ _  _                                                       15  mark   _  _  *
12  tel         tel         ADJ   _  Gender=Masc|Number=Sing                                 15  nsubj  _  _  *
13  est         être        AUX   _  Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin   15  cop    _  _  *
14  le          le          DET   _  Definite=Def|Gender=Masc|Number=Sing|PronType=Art       15  det    _  _  *
15  cas         cas         NOUN  _  Gender=Masc|Number=Sing                                 18  obl:mod  _  _  *
16  cela        cela        PRON  _  Number=Sing|PronType=Dem                                18  nsubj  _  _  *
17  est         être        AUX   _  Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin   18  cop    _  _  *
18  regrettable regrettable ADJ   _  Number=Sing                                            9   ccomp  _  _  *
19  .           .           PUNCT _  _                                                       3   punct  _  _  *
```

Annotations PARSEME

- **Annotations impossibles à transformer en COLL**
  - **Constructions multi-verbes (MVC)** : *faire savoir*, *laisser tomber*
  - **Constructions verbes à particules (VPC)** : *give up (abandonner), sleep in (faire la grasse matinée)*
  - **Verbes intrinsèquement réflexifs (IRV)** : *s'incliner, se rapprocher*
- **Annotations potentiellement transformables en COLL**
  - **Constructions à verbe faible (LVC)** : avec « vrais » verbes faibles (*faire l'historique, avoir conscience*), avec verbes pleins (*dresser un bilan, apporter un témoignage*)
  - **Idiomes verbaux (VID)** : expressions idiomatiques (*jeter l'éponge, couper l'herbe sous le pied*), tournures idiomatiques (*il y a, il est question de*), associations compositionnelles (*poser une question, attirer l'attention*)

Fonctionnement de VarIDE

Faculté des langues Université de Strasbourg

- **Etape 1 : extraction de candidats EP (entraînement)**
  - Normalisation de tuples avec POS pour chaque EP annotée, p. ex. (NOUN,VERB)
  - Normalisation de tuples de lemmes, p. ex. (hommage,rendre)
  - Génération de toutes les formes fléchies du tuple de lemmes
- **Etape 2 : extraction des caractéristiques morphosyntaxiques (entraînement)**
  - Caractéristiques absolues : obtenues localement (*« il a rendu un hommage poignant »*, `ABS_morph_NOUN_Number=singular`)
  - Caractéristiques relatives : obtenues par comparaison avec les autres tuples de lemmes normalisés correspondant au candidat (`false`, `true` ou `-1`)
- **Etape 3 : prédiction et annotation**
  - Tous les candidats du corpus de test sont extraits suivant la même méthodologie
  - Caractéristiques relatives obtenues en comparant les tuples de lemmes normalisés du corpus d'entraînement (valeur booléenne)
  - Classifieur bayésien naïf de `nltk`

Architecture complète