FACULTY OF ARTS
Charles University

Analyse contrastive de la complexité syntaxique à l'aide de corpus parallèles annotés en Universal Dependencies :

Promesses et écueils

Olga Nádvorníková

Institut d'Etudes romanes, Faculté des Lettres de l'Université Charles à Prague

Conférence invitée dans le programme Translitteræ (PSL Paris)

Lattice CNRS – ENS UMR 8094

Paris 28-05-2024

PLAN

1. Introduction : motivation de la recherche et délimitation du champs d'étude
2. Les ingrédients :
   - 2.1 Mesures de la complexité syntaxique
   - 2.2 *Universal Dependencies* : principes de base
   - 2.3 Les données : Corpus parallèle (multilingue) InterCorp
3. Le résultat : corpus parallèle InterCorp v16ud (version pilote)
   - 3.1 Implémentation des mesures de la complexité syntaxique
   - 3.2 Exemples des requêtes sur corpus
4. Exemples d'application du corpus parallèle InterCorp v16ud
   - 4.1 Niveau de phrase (analyse contrastive et traductologique)
   - 4.2 Niveau de texte
   - 4.3 Niveau de genre textuel
   - 4.4 Niveau de langue (perspective typologique)
5. Conclusions : promesses et écueils

1. Introduction: Définition de la complexité syntaxique

- Définition générale de la complexité de systèmes :

„the number and variety of elements and the elaborateness of their interrelational structure“ (Rescher 1998:1, Hübler 2007:10)

Beaman (1984: 45; *Coordination and Subordination Revisited: Syntactic Complexity in Spoken and Written Narrative Discourse*):

***„syntactic complexity in language is related to the number, type, and depth of embedding in a text**. Syntactically simple authors use short, single clause sentences and rely more heavily on coordinated structures* [...]. *Syntactically complex authors* [...] *use longer sentences and more subordinate clauses that reveal more complex structural relationships*.“ (cf. définition similaire dans De Clercq 2016)

The syntactic complexity of a *sentence* can be defined in terms of the **number and the variability of clauses it contains**, and in terms of the **degree of their hierarchical relations**.

Délimitation du champs de recherche - 1

a) **diachronie** : An increase in complexity thus corresponds, at the most general level, to the increase in hierarchical organization, that is, an increase in the number of hierarchical levels within a system (Givón 2009:4) ;

d) **synchronie – analyses monolingues** (Mačutek, Čech & Milička 2019), propositions relatives (Hudelot 1980), Biber, Larsson & Hancock 2023 (English), etc.

c) **analyse contrastive** : clause-linking (Lehmann 1988), clause-combining (Cosme 2006, etc.), information packaging (Solfjeld 1996, Fabricius-Hansen 1999), *shared task UD* (Berdicevskis et al. 2018, etc.),

f) **traductologie** : Izquierdo & Marco 2000, corpus comparables de traductions (par exemple Canavese – Mori 2021; *readability of Italian legislative texts*) ou corpus parallèles (*universaux de la traduction* – simplification, normalisation, etc.)

b) *register variation* – oral/écrit (Beaman 1984 etc.), academic: Biber & Gray 2017, etc..

e) typologie (Levshina 2019, 2021) – Leipzig Corpora Collection (comparable, UD)

g) **readability** (Kincaid et al. 1975, Dell'Orletta et al. 2011, Gruszczyński & Ogrodniczuk 2015 *Jasnopis*).

h) ***language acquisition*** et ***language proficiency assessment*** (L1 et L2), Lu 2010, etc.

Délimitation du champs de recherche - 2

- **syntactic** complexity (e.g. Ferreira 1991; Givón 1991; Szmrecsanyi 2004, *complexité syntaxique* De Clercq 2016, Berdicevskis et al. 2018, Brunato & Venturi 2023, etc.)
- **cognitive** complexity (e.g. Mondorf 2003; Givón 1991; Rohdenburg 1996)
- **clause** complexity (e.g. Kuboň 2001)
- **linguistic** complexity (e.g. Schleppegrell 1992)
- **structural** complexity (e.g. Givón: 1991; Arnold et al. 2000)
- **grammatical** / **syntactic weight** (e.g. Wasow 1997; Wasow and Arnold 2003)
- **information density/packaging** (Fabricius-Hansen 1999, Solfjeld 1998, etc.)

Niveau d'analyse : phrase/texte/genre textuel/langue ?

Délimitation du champs de recherche - 3

- objectif : ***absolute complexity*** (***objective*** *complexity* = propriétés formelles, voir Brunato & Venturi 2023: 59) et non ***relative*** (*complexité subjective* – orientée vers les locuteurs et évaluant le degré de l'effort cognitif nécessaire pour le traitement de la phrase/du texte, „readability"), voir Szmrecsanyi and Kortmann 2012: 10.
- **complexité syntaxique de la *phrase*** (textes écrits) vs. *grammatical (linguistic) complexity* de **la langue** (Schleppegrell 1992, Koplening et al. 2017, etc.) ; niveaux d'analyse : phrase > texte > genre textuel > langue ?

Exemple (1)

1. Quelles sont les **causes** (déclencheurs) de ces changements en structure syntaxique ?
2. S'agit-il d'exceptions, ou de **tendances générales** (texte/genre/paire de langues) ?
3. Possibilités de **repérage automatique** dans les corpus et de **quantification** ?

FR (Camus-Peste)	en	cs
Au même moment, un coup de revolver partit du second et le chien se retourna comme une crêpe, agitant violemment ses pattes pour se renverser enfin sur le flanc, secoué par de longs soubresauts. (A. Camus, *La Peste*)	[…] when a revolver barked from the third-floor window. // The dog did a somersault like a tossed pancake, lashed the air with its legs, and floundered on to its side, its body writhing in long convulsions. (transl. S. Gilbert)	V té chvíli však vyšla z druhého patra rána a pes se otočil jako čamrda, prudce zatřepal packami, svalil se na zem a dodělal v škubavých křečích. (transl. M. Tomášková)

Tentative précédente 1

Formes verbales finies / non-finies comme déclencheurs des changements de la complexité syntaxique

- converbes (*gérondifs, gerudio, participial adjuncts, imiesłów przysłówkowy*, etc.) et participes > formes verbales finies

(Čermák, Nádvorníková et al. 2020 FR,IT,ES,PT > cs, Nádvorníková à par. 1 et 2 FR-CS-PL, etc.)

FR. *dit-il **en souriant*** > cs. *řekl **a usmál se***

*'(he) said **and (he) smiled**'*

Filiouchkina Krave 2012 (RU > no), Fabricius-Hansen 1999 (DE > no), etc.

MAIS : résultats partiels

Exemple (1)

FR (Camus-Peste)	en	cs
Au même moment, un coup de revolver partit du second et le chien se retourna comme une crêpe, agitant violemment ses pattes pour se renverser enfin sur le flanc, secoué par de longs soubresauts. (A. Camus, *La Peste*)	[…] when a revolver barked from the third-floor window. // The dog did a somersault like a tossed pancake, lashed the air with its legs, and floundered on to its side, its body writhing in long convulsions. (transl. S. Gilbert) **SPLITTING**	V té chvíli však vyšla z druhého patra rána a pes se otočil jako čamrda, prudce zatřepal packami, svalil se na zem a dodělal v škubavých křečích. (transl. M. Tomášková)

Tentative précédente 2 – Changements de la segmentation en phrases

Sentence splitting/joining pas une stratégie dominante : comment analyser les changements **à l'intérieur de la phrase?**

Direction of translation	Total N° of alignments	N° of split segments	%
EN>cs	1,168,881	59,639	5,10%
CS>en	212,373	16,824	**7,92%**
FR>cs	378,750	14,956	3,95%
CS>fr	242,579	15,116	**6,23%**
TOTAL	2,002,583	106,535	5,32%

Sources :

- NÁDVORNÍKOVÁ, Olga, 2017. Parallel Corpus in Translation Studies: Analysis of Shifts in the Segmentation of Sentences in the Czech-English-French Part of the InterCorp Parallel Corpus. In: *Language Use and Linguistic Structure*. Olomouc: Palacký University Olomouc, s. 445–461. https://linguisticapragensia.ff.cuni.cz/wp-content/uploads/sites/12/2017/10/LP_2017-2_Nadvornikova_35-57.pdf
- NÁDVORNÍKOVÁ, Olga, 2021. Contexts and Consequences of Sentence Splitting in Translation (English-French-Czech). *Research in Language*. *19(3),* pp. 229-250

Voir aussi : Ramm (2004), Serbina (2014), Frankenberg-Garcia (2019), etc.

Quantification des changements en structures syntaxiques : **mesures de la complexité syntaxique**

FR (Camus-Peste)	en	cs
Au même moment, un coup de revolver partit du second et le chien se retourna comme une crêpe, agitant violemment ses pattes pour se renverser enfin sur le flanc, secoué par de longs soubresauts. (A. Camus, *La Peste*)	[…] when a revolver barked from the third-floor window. // The dog did a somersault like a tossed pancake, lashed the air with its legs, and floundered on to its side, its body writhing in long convulsions. (transl. S. Gilbert)	V té chvíli však vyšla z druhého patra rána a pes se otočil jako čamrda, prudce zatřepal packami, svalil se na zem a dodělal v škubavých křečích. (transl. M. Tomášková)
Sub.ratio = 2,5 ((2+3)/2) Max.Tree.Depth = 3	Sub.ratio = 1,33 ((3+1)/3) Max.Tree.Depth = 1	Sub.ratio = 1 (5/5) Max.Tree.Depth = 0

Subordination ratio (Jagaiah et al. 2020: 2600) =

$$\frac{(\text{N° T-units} + \text{N° Sub})}{\text{N° T-units}}$$

- **T-unit** = *main clause and all the subordinate clauses attached to it* (Hunt 1965)
- **Sub = subordinate clauses (finite and non-finite)**

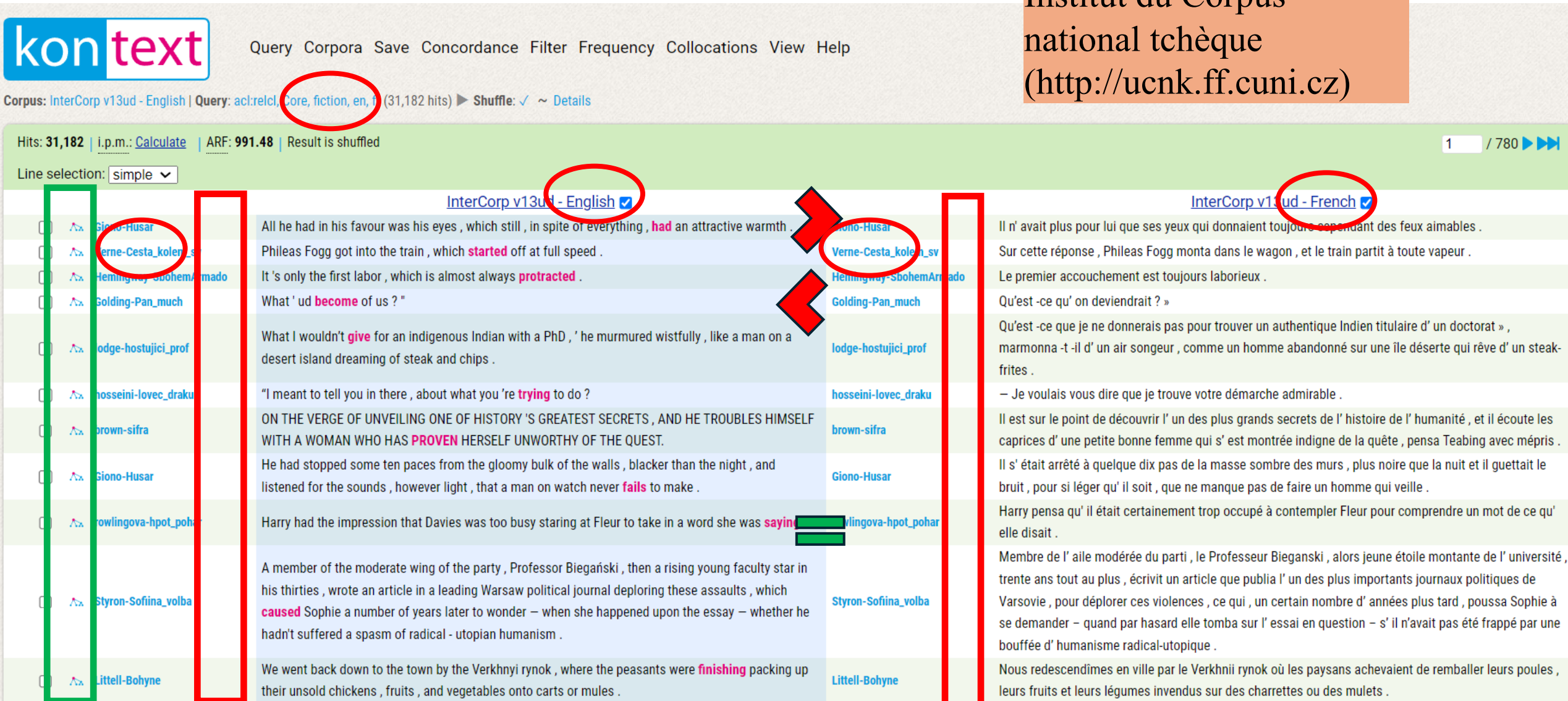
es :

cs

Au même moment, un coup de revolver partit du second et le chien se retourna comme une crêpe, agitant violemment ses pattes pour se renverser enfin sur le flanc, secoué par de longs soubresauts. (A. Camus, *La Peste*)	[…] when a revolver barked from the third-floor window. // The dog did a somersault like a tossed pancake, lashed the air with its legs, and floundered on to its side, its body writhing in long convulsions. (transl. S. Gilbert)	V té chvíli však vyšla z druhého patra rána a pes se otočil jako čamrda, prudce zatřepal packami, svalil se na zem a dodělal v škubavých křečích. (transl. M. Tomášková)
Sub.ratio = 2,5 ((2+3)/2) Max.Tree.Depth = 3	Sub.ratio = 1,33 ((3+1)/3) Max.Tree.Depth = 1	Sub.ratio = 1 (5/5) Max.Tree.Depth = 0

Corpus multilingue (parallèle = de traductions) & **annotation syntaxique** d'après le schéma commun

Institut du Corpus national tchèque (http://ucnk.ff.cuni.cz)



Ingrédients

- ***pommes :*** **mesures de la complexité syntaxique** (*sub.ratio, s_length, mdd*, etc.)
- ***oeufs :*** **annotation syntaxique** fiable pour l'analyse translinguistique (*Universal Dependencies*, www.universaldependencies.org)
- ***farine :*** **données** (corpus parallèle/ multilingue InterCorp)

Les cuisiniers

(Faculté des Lettres de l'Université Charles à Prague) :

Institut du Corpus national tchèque :

- informaticiens : Martin Vavřín (2021–2022) et Bohumil Šimčík (2023–2024)
- directeurs (Michal Křen a Michal Škrabal)
- Jiří Milička (mesures de *lexical diversity*)
- **Alexandr Rosen (chef de la section parallèle)**

Institut d'Etudes romanes : Olga Nádvorníková

PLAN

1. Introduction : motivation de la recherche du champs d'étude
2. **Les ingrédients :**
   - **2.1 Mesures de la complexité syntaxique**
   - 2.2 *Universal Dependencies* : principes de base
   - 2.3 Les données : Corpus parallèle (multilingue) InterCorp
3. Le résultat : corpus parallèle InterCorp v16ud (version pilote)
   - 3.1 Implémentation des mesures de la complexité syntaxique
   - 3.2 Exemples des requêtes sur corpus
4. Exemples d'application du corpus parallèle InterCorp v16ud
   - 4.1 Niveau de phrase (analyse contrastive et traductologique)
   - 4.2 Niveau de texte
   - 4.3 Niveau de genre textuel
   - 4.4 Niveau de langue (perspective typologique)
5. Conclusions : promesses et écueils

Attention : work in progress (28-05-2024)

- InterCorp v16ud – **version pilote limitée aux textes littéraires** (*fiction*) – version complète sera disponible très prochainement
- applications en recherche : ébauches (version pilote du corpus lancée en mars 2024)
- remarques, critiques, commentaires bienvenus

PLAN

1. Introduction : motivation de la recherche et délimitation du champs d'étude
2. **Les ingrédients :**
   - **2.1 Mesures de la complexité syntaxique**
   - 2.2 *Universal Dependencies* : principes de base
   - 2.3 Les données : Corpus parallèle (multilingue) InterCorp
3. Le résultat : corpus parallèle InterCorp v16ud (version pilote)
   - 3.1 Implémentation des mesures de la complexité syntaxique
   - 3.2 Exemples des requêtes sur corpus
4. Exemples d'application du corpus parallèle InterCorp v16ud
   - 4.1 Niveau de phrase (analyse contrastive et traductologique)
   - 4.2 Niveau de texte
   - 4.3 Niveau de genre textuel
   - 4.4 Niveau de langue (perspective typologique)
5. Conclusions : promesses et écueils

2.1 Mesures de la complexité syntaxique

= *simplification de la complexité*

- l'analyse de la complexité syntaxique doit prendre en considération **non seulement le nombre et la variabilité des entités**, mais également **le degré de leur organisation hiérarchique** (voir Beaman 1984: 45)
- complexité (syntaxique) est un phénomène **multidimensionnel** (Biber, Larsson & Hancock 2023), multifaceted (Brunato & Venturi 2023, UD, multilingual) :
  - une seule mesure de la complexité syntaxique produit toujours un biais > nécessaire de combiner plusieurs mesures (par exemple complexité de la phrase et du SN – Biber & Gray 2016).
  - De plus, la pertinence de la mesure peut varier en fonction du **genre textuel** (*register*) et en fonction des **propriétés structurelles de la langue**

Exemples de mesures

sentence/clause length (No of tokens/words), *subordination ratio*, maximum tree depth (clausal), *mean number of subordinate clauses*, *number of adjective clauses per T-unit*, *mean number of morphemes per T-unit*, *number of gerunds, participles and absolutes* (***language proficiency assessment***, un des indices du niveau de la maîtrise de la langue – L1 ou L2 – 46 mesures de la complexité syntaxique différentes dans 36 études publiées entre 1970–2019 (méta-analyse dans Jagaiah, Olinghouse & Kearns 2020: 46)

noun phrase modification (Biber & Gray 2016)

sentence complexity ratio (clauses/sentences), *sentence coordination ratio* (T-units/sentences), *complex nominals per clause*, etc. (Xu & Li 2021) – 14 mesures, translational English

sentence length, *average length of dependency links*, etc. (Brunato & Venturi 2023) – ***multilingual treebanks***

2.1.1 *Sentence length* : mesure traditionnelle

Sentence length (nombre de tokens/words);

- préconisé par Szmreczányi (2004)
- fiable dans des recherches monolingues - *register variation* ou traductologie (Canavese – Mori 2021, Chlumská 2017)
- **not reliable cross-linguistically** (see Nádvorníková 2020) – référence au nombre de mots différents

fr. *Il est arrivé à la maison.(6 mots/7 tokens)*
cs: *Přišel domů (2 mots/3 tokens)*
come.PTCP.3.M.SG home
'(He) arrived home'

PONCTUATION en fr-en-cs

Conclusion : s_length est une mesure répandue, mais pas fiable dans des recherches multilingues

Analyse quantitative de la ponctuation (FR-EN-CS), Nádvorníková 2020 (doi:10.14712/18059635.2020.1.2)

langue synthétique (CS) vs. langue (plutôt) analytique (FR)

InterCorp	taille du corpus (tokens)	point final		deux-points		point d_interrog.		point d_exclam.	
		abs.fq.	ipm	abs.fq.	ipm	abs.fq.	ipm	abs.fq.	ipm
CS (orig.)	1 483 802	58 914	*39 705*	5 562	*3 748*	6 639	*4 474*	4 459	*3 005*
fr(CS)	1 735 949	62 114	*35 781*	6 452	*3 717*	6 969	*4 015*	5 036	*2 901*

Tailles de corpus différentes, même s'il s'agit de textes équivalents – les fréquences relatives (ipm – instances per million) peuvent s'en voir faussées

Mesures plus fiables pour la recherche multilingue

✓ **Subordination ratio** (Jagaiah et al. 2020: 2600) =

$$\frac{(\text{Nº T-units} + \text{Nº Sub})}{\text{Nº T-units}}$$

- **T-unit** = *main clause and all the subordinate clauses attached to it* (Hunt 1965)
- **Sub = subordinate clauses (finite and non-finite)**

✓ **Maximum Tree Depth** (in a sentence) = number of clauses sequentially nested inside one another

✓**mdd = mean dependency distance**

mdd:

Mean Dependency Distance (Yan & Li 2019, Yan 2021, Liu 2008, Mačutek et al. 2021, Brunato & Venturi 2023, etc.)

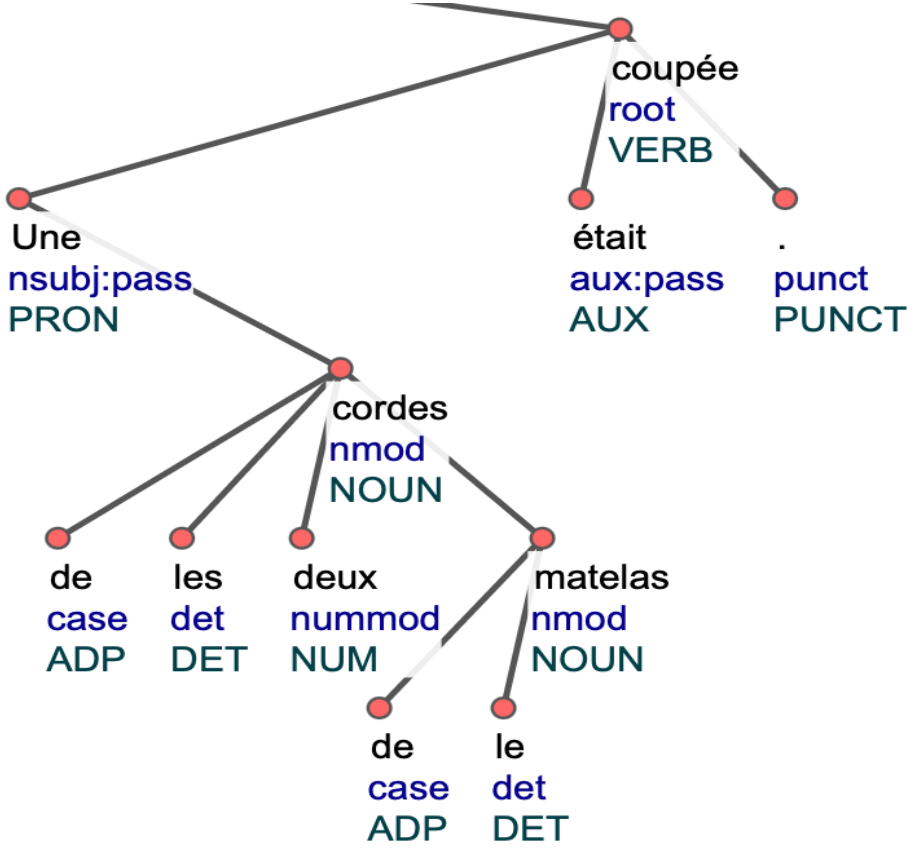
- *average number of words occurring between the syntactic head and the dependent in a text*
- **effort cognitif** > censée être fiable pour l'analyse translinguistique
- sans ponctuation
- calcul (*n* … nombre de mots dans la phrase)

  $DD_i = | ID_i - head_i |$

  $DD = \sum_{i = 0 \text{ až } n} DD_i$

  $mdd = DD / (n - 1)$
- DD = 26

  mdd = 26 / 9 ≅ 2,89



	Une	*des*		*deux*	*cordes*	*du*		*matelas*	*etait*	*coupée*
		de	*les*			*de*	*le*			
ID = *i*	1	2	3	4	5	6	7	8	9	10
$head_i$	10	5	5	5	1	8	8	5	10	0
DD_i	9	3	2	1	4	2	1	3	1	0

Critères du choix des mesures de la complexité syntaxique **pour l'analyse multilingue** basée sur un corpus parallèle annoté en **Universal Dependencies**

1. non seulement **le nombre** d'entités, mais aussi **le degré de leur organisation hiérarchique** (Max.NP.**Length** & Max.NP.**Depth**)
2. robuste **pour des langues typologiquement différentes** (cf. s_length)
3. ensemble couvrant des **aspects de la complexité syntaxique variés**, dans des **genres textuels variés** (clausal vs. phrasal complexity, e.g. *complex nominal phrases,* Xu & Li 2021, Biber & Gray 2016)
4. **comparabilité et replicabilité** (mesures déjà utilisées dans des recherches précédentes)
5. **„transparent"** pour l'utilisateur (pas l'aggrégation de mesures)

Critères techniques (implémentation dans l'interface du corpus) :

Compromis

6. en **nombre limité** (interface du corpus)
7. accessible pour le **traitement automatique** via *Universal Dependencies* (pas de distinction *finite* vs. *non-finite verb forms*) ou via d'autres données (token/word count)

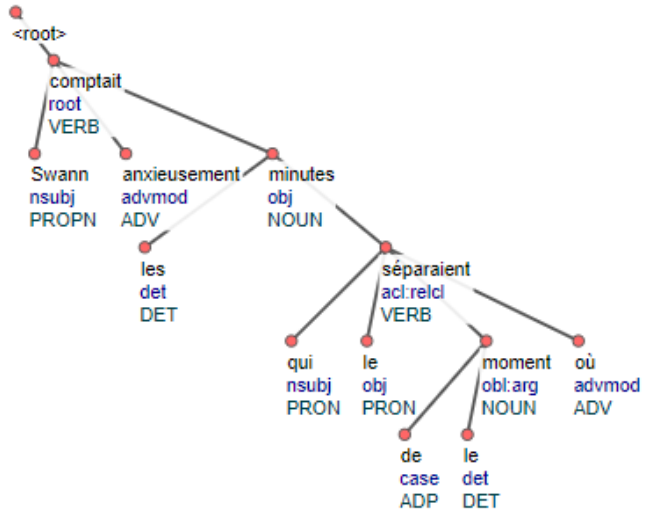
PLAN

1. Introduction : motivation de la recherche et délimitation du champs d'étude
2. **Les ingrédients :**
   - 2.1 Mesures de la complexité syntaxique
   - **2.2 *Universal Dependencies* : principes de base**
   - 2.3 Les données : Corpus parallèle (multilingue) InterCorp
3. Le résultat : corpus parallèle InterCorp v16ud (version pilote)
   - 3.1 Implémentation des mesures de la complexité syntaxique
   - 3.2 Exemples des requêtes sur corpus
4. Exemples d'application du corpus parallèle InterCorp v16ud
   - 4.1 Niveau de phrase (analyse contrastive et traductologique)
   - 4.2 Niveau de texte
   - 4.3 Niveau de genre textuel
   - 4.4 Niveau de langue (perspective typologique)
5. Conclusions : promesses et écueils

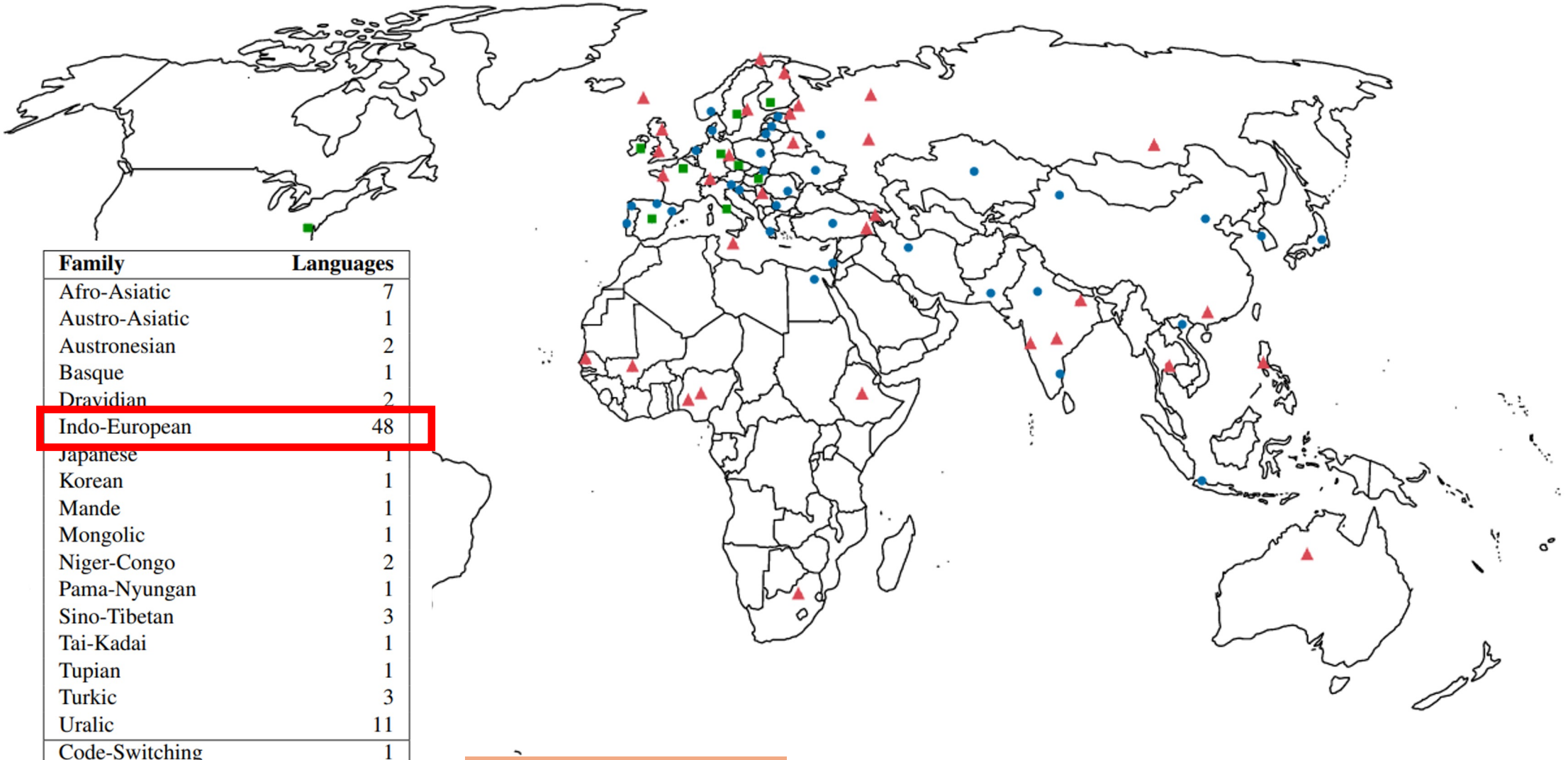
2.2 Universal Dependencies

- **projet collaboratif** – objectif : un **schéma d'annotation commun** („universel“) qui pourrait être utilisé pour un grand nombre de langues différentes (Nivre 2021, Nivre et al. 2020, Gerdes et al. 2018, Guillaume et al. 2019, de Marneffe et al. 2021, Zeman, 2018, etc.)
- **lancé** en 2014 (10 treebanks)
- **version la plus récente** (v2.13 – novembre 2023) contient 259 treebanks qui représentent 148 langues différentes de 31 familles linguistiques

(www.universaldependencies.org)



Annotation selon *Universal Depencencies*



Family	Languages
Afro-Asiatic	7
Austro-Asiatic	1
Austronesian	2
Basque	1
Dravidian	2
Indo-European	48
Japanese	1
Korean	1
Mande	1
Mongolic	1
Niger-Congo	2
Pama-Nyungan	1
Sino-Tibetan	3
Tai-Kadai	1
Tupian	1
Turkic	3
Uralic	11
Code-Switching	1
Creole	1
Sign Language	1

Table 4: Language families in UD v2.5.

Nivre et al. (2020)

Language	#	Sents	Words	Language	#	Sents	Words	Language	#	Sents	Words
Afrikaans	1	1,934	49,276	German	4	208,440	3,753,947	Old Russian	2	17,548	168,522
Akkadian	1	101	1,852	Gothic	1	5,401	55,336	Persian	1	5,997	152,920
Amharic	1	1,074	10,010	Greek	1	2,521	63,441	Polish	3	40,398	499,392
Ancient Greek	2	30,999	416,988	Hebrew	1	6,216	161,417	Portuguese	3	22,443	570,543
Arabic	3	28,402	1,042,024	Hindi	2	17,647	375,533	Romanian	3	25,858	551,932
Armenian	1	2502	52630	Hindi English	1	1,898	26,909	Russian	4	71,183	1,262,206
Assyrian	1	57	453	Hungarian	1	1,800	42,032	Sanskrit	1	230	1,843
Bambara	1	1,026	13,823	Indonesian	2	6,593	141,823	Scottish Gaelic	1	2,193	42,848
Basque	1	8,993	121,443	Irish	1	1,763	40,572	Serbian	1	4,384	97,673
Belarusian	1	637	13,325	Italian	6	35,481	811,522	Skolt Sámi	1	36	321
Bhojpuri	1	254	4,881	Japanese	4	67,117	1,498,560	Slovak	1	10,604	106,043
Breton	1	888	10,054	Karelian	1	228	3,094	Slovenian	2	11,188	170,158
Bulgarian	1	11,138	156,149	Kazakh	1	1,078	10,536	Spanish	3	34,693	1,004,443
Buryat	1	927	10,185	Komi Permyak	1	49	399	Swedish	3	12,269	206,855
Cantonese	1	1,004	13,918	Komi Zyrian	2	327	3,463	Swedish Sign Language	1	203	1,610
Catalan	1	16,678	531,971	Korean	3	34,702	446,996	Swiss German	1	100	1,444
Chinese	5	12,449	285,127	Kurmanji	1	754	1,0260	Tagalog	1	55	292
Classical Chinese	1	15,115	74,770	Latin	3	41,695	582,336	Tamil	1	600	9,581
Coptic	1	1,575	40,034	Latvian	1	13,643	219,955	Telugu	1	1,328	6,465
Croatian	1	9,010	199,409	Lithuanian	2	3,905	75,403	Thai	1	1,000	22,322
Czech	5	127,507	2,222,163	Livvi	1	125	1,632	Turkish	3	9,437	91,626
Danish	1	5,512	100,733	Maltese	1	2,074	44,162	Ukrainian	1	7,060	122,091
Dutch	2	20,916	306,503	Marathi	1	466	3,849	Upper Sorbian	1	646	11,196
English	7	35,791	620,509	Mbyá Guaraní	2	1,144	13,089	Urdu	1	5,130	138,077
Erzya	1	1,550	15,790	Moksha	1	65	561	Uyghur	1	3,456	40,236
Estonian	2	32,634	465,015	Naija	1	948	12,863	Vietnamese	1	3,000	43,754
Faroese	1	1,208	10,002	North Sámi	1	3,122	26,845	Warlpiri	1	55	314
Finnish	3	34,859	377,619	Norwegian	3	42,869	666,984	Welsh	1	956	16,989
French	7	45,074	1,157,171	Old Church Slavonic	1	6,338	57,563	Wolof	1	2,107	44,258
Galician	2	4,993	164,385	Old French	1	17,678	170,741	Yoruba	1	100	2,664

Table 3: Languages in UD v2.5 with number of treebanks (#), sentences (Sents) and words (Words).

Nivre et al. (2020)

578 collaborateurs à UD annotation dans le monde entier

Zeman, D., Nivre, J., Abrams, M., Aepli, N., Agić, Ž., Ahrenberg, L., Aleksandravičiūtė, G., Antonsen, L., Aplonova, K., Aranzabe, M. J., Arutie, G., Asahara, M., Ateyah, L., Attia, M., Atutxa, A., Augustinus, L., Badmaeva, E., Ballesteros, M., Banerjee, E., Bank, S., Barbu Mititelu, V., Basmov, V., Batchelor, C., Bauer, J., Bellato, S., Bengoetxea, K., Berzak, Y., Bhat, I. A., Bhat, R. A., Biagetti, E., Bick, E., Bielinskienė, A., Blokland, R., Bobicev, V., Boizou, L., Borges Völker, E., Börstell, C., Bosco, C., Bouma, G., Bowman, S., Boyd, A., Brokaitė, K., Burchardt, A., Candito, M., Caron, B., Caron, G., Cavalcanti, T., Cebiroğlu Eryiğit, G., Cecchini, F. M., Celano, G. G. A., Čéplö, S., Cetin, S., Chalub, F., Choi, J., Cho, Y., Chun, J., Cignarella, A. T., Cinková, S., Collomb, A., Çöltekin, Ç., Connor, M., Courtin, M., Davidson, E., de Marneffe, M.-C., de Paiva, V., de Souza, E., Diaz de Ilarraza, A., Dickerson, C., Dione, B., Dirix, P., Dobrovoljc, K., Dozat, T., Droganova, K., Dwivedi, P., Eckhoff, H., Eli, M., Elkahky, A., Ephrem, B., Erina, O., Erjavec, T., Etienne, A., Evelyn, W., Farkas, R., Fernandez Alcalde, H., Foster, J., Freitas, C., Fujita, K., Gajdošová, K., Galbraith, D., Garcia, M., Gärdenfors, M., Garza, S., Gerdes, K., Ginter, F., Goenaga, I., Gojenola, K., Gökırmak, M., Goldberg, Y., Gómez Guinovart, X., González Saavedra, B., Griciūtė, B., Grioni, M., Grūzītis, N., Guillaume, B., Guillot-Barbance, C., Habash, N., Hajič, J., Hajič jr., J., Hämäläinen, M., Hà Mỹ, L., Han, N.-R., Harris, K., Haug, D., Heinecke, J., Hennig, F., Hladká, B., Hlaváčová, J., Hociung, F., Hohle, P., Hwang, J., Ikeda, T., Ion, R., Irimia, E., Ishola, Ọ., Jelínek, T., Johannsen, A., Jørgensen, F., Juutinen, M., Kaşıkara, H., Kaasen, A., Kabaeva, N., Kahane, S., Kanayama, H., Kanerva, J., Katz, B., Kayadelen, T., Kenney, J., Kettnerová, V., Kirchner, J., Klementieva, E., Köhn, A., Kopacewicz, K., Kotsyba, N., Kovalevskaitė, J., Krek, S., Kwak, S., Laippala, V., Lambertino, L., Lam, L., Lando, T., Larasati, S. D., Lavrentiev, A., Lee, J., Lê H`ông, P., Lenci, A., Lertpradit, S., Leung, H., Li, C. Y., Li, J., Li, K., Lim, K., Liovina, M., Li, Y., Ljubešić, N., Loginova, O., Lyashevskaya, O., Lynn, T., Macketanz, V., Makazhanov, A., Mandl, M., Manning, C., Manurung, R., Mărănduc, C., Mareček, D., Marheinecke, K., Martínez Alonso, H., Martins, A., Mašek, J., Matsumoto, Y., McDonald, R., McGuinness, S., Mendonça, G., Miekka, N., Misirpashayeva, M., Missilä, A., Mititelu, C., Mitrofan, M., Miyao, Y., Montemagni, S., More, A., Moreno Romero, L., Mori, K. S., Morioka, T., Mori, S., Moro, S., Mortensen, B., Moskalevskyi, B., Muischnek, K., Munro, R., Murawaki, Y., Müürisep, K., Nainwani, P., Navarro Horñiacek, J. I., Nedoluzhko, A., Nešpore-Bērzkalne, G., Nguy˜ên Thị, L., Nguy˜ên Thị Minh, H., Nikaido, Y., Nikolaev, V., Nitisaroj, R., Nurmi, H., Ojala, S., Ojha, A. K., Olúòkun, A., Omura, M., Osenova, P., Östling, R., Øvrelid, L., Partanen, N., Pascual, E., Passarotti, M., Patejuk, A., Paulino-Passos, G., Peljak-Łapińska, A., Peng, S., Perez, C.-A., Perrier, G., Petrova, D., Petrov, S., Phelan, J., Piitulainen, J., Pirinen, T. A., Pitler, E., Plank, B., Poibeau, T., Ponomareva, L., Popel, M., Pretkalniņa, L., Prévost, S., Prokopidis, P., Przepiórkowski, A., Puolakainen, T., Pyysalo, S., Qi, P., Rääbis, A., Rademaker, A., Ramasamy, L., Rama, T., Ramisch, C., Ravishankar, V., Real, L., Reddy, S., Rehm, G., Riabov, I., Rießler, M., Rimkutė, E., Rinaldi, L., Rituma, L., Rocha, L., Romanenko, M., Rosa, R., Rovati, D., Roșca, V., Rudina, O., Rueter, J., Sadde, S., Sagot, B., Saleh, S., Salomoni, A., Samardžić, T., Samson, S., Sanguinetti, M., Särg, D., Saulīte, B., Sawanakunanon, Y., Schneider, N., Schuster, S., Seddah, D., Seeker, W., Seraji, M., Shen, M., Shimada, A., Shirasu, H., Shohibussirri, M., Sichinava, D., Silveira, A., Silveira, N., Simi, M., Simionescu, R., Simkó, K., Šimková, M., Simov, K., Smith, A., Soares-Bastos, I., Spadine, C., Stella, A., Straka, M., Strnadová, J., Suhr, A., Sulubacak, U., Suzuki, S., Szántó, Z., Taji, D., Takahashi, Y., Tamburini, F., Tanaka, T., Tellier, I., Thomas, G., Torga, L., Trosterud, T., Trukhina, A., Tsarfaty, R., Tyers, F., Uematsu, S., Urešová, Z., Uria, L., Uszkoreit, H., Utka, A., Vajjala, S., van Niekerk, D., van Noord, G., Varga, V., Villemonte de la Clergerie, E., Vincze, V., Wallin, L., Walsh, A., Wang, J. X., Washington, J. N., Wendt, M., Williams, S., Wirén, M., Wittern, C., Woldemariam, T., Wong, T.-s., Wróblewska, A., Yako, M., Yamazaki, N., Yan, C., Yasuoka, K., Yavrumyan, M. M., Yu, Z., Žabokrtský, Z., Zeldes, A., Zhang, M., and Zhu, H. (2019). Universal Dependencies 2.5. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. http://hdl.handle.net/11234/1-3105.

Origines de *Universal Dependencies*

Sylvain Kahane (2022)
séminaire ERTIM 20/02/2022
Joakim Nivre (2024)
https://ufal.mff.cuni.cz

- *Stanford Dependencies* 2005: **mots grammaticaux dépendent des mots lexicaux**
  https://nlp.stanford.edu/software/stanford-dependencies.html
- *Google Universal Tagset* 2007: 12 POS (parties du discours) (dans UD finalement 17 "UPOS") https://github.com/slavpetrov/universal-pos-tags
- *Interset* 2006: universal morphological categories for CONLL shared tasks: *Conference on Computational Natural Language Learning* https://github.com/dan-zeman/interset
- *CONLL-X* 2007: tabular format, also for encoding syntactic structure
  https://web.archive.org/web/20160814191537/http://ilk.uvt.nl/conll/#dataformat

Principes de l’annotation selon *Universal Dependencies*

- Compromis (tzv. *Manning’s law*):
  - UD needs to be satisfactory for **linguistic analysis of individual languages**
  - UD needs to be good for **linguistic typology**
  - UD must be suitable for rapid, consistent **annotation**
  - UD must be easily comprehended and used **by a non-linguist**
  - UD must be suitable **for computer parsing** with high accuracy
  - provide good **support for NLP** tasks

Il est facile d’améliorer le système dans un critère, mais difficile de garder le niveau élevé pour tous.

Principes de l'annotation I – 3 niveaux d'annotation (UD Guidelines version 2) (verze 1: 2014)

1) 17 parties du discours – `upos`

   https://universaldependencies.org/u/pos/index.html

2) 24 traits morphologiques – `feats`

   https://universaldependencies.org/u/feat/index.html

3) 37 fonctions syntaxiques – `deprel`

   https://universaldependencies.org/u/dep/index.html

Niveau 1 – **upos** (17 universal POS)

Open class words	Closed class words	Other
ADJ	ADP	PUNCT
ADV	AUX	SYM
INTJ	CCONJ	X
NOUN	DET	
PROPN	NUM	
VERB	PART	
	PRON	
	SCONJ	

```
[upos="ADJ"]
```

Niveau 2 : feats = 24 catégories morphologiques

Lexical features*	Inflectional features*	
	*Nominal**	*Verbal**
PronType	Gender	VerbForm
NumType	Animacy	Mood
Poss	NounClass	Tense
Reflex	Number	Aspect
Foreign	Case	Voice
Abbr	Definite	Evident
Typo	Degree	Polarity
		Person
		Polite
		Clusivity

```
[feats="Number=Sing"]
```

Index des `feats`:

Lexical features*	Inflectional features*	
	*Nominal**	*Verbal**
PronType	Gender	VerbForm
NumType	Animacy	Mood
Poss	NounClass	Tense
Reflex	Number	Aspect
Foreign	Case	Voice
Abbr	Definite	Evident
Typo	Degree	Polarity
		Person
		Polite
		Clusivity

Niveau 3. **deprel** : *universal dependency relations* `[deprel="acl:relcl"]`

selon les catégories morphosyntaxiques

Universels mais spécifiables : **`acl:relcl`**

selon les fonctions syntaxiques

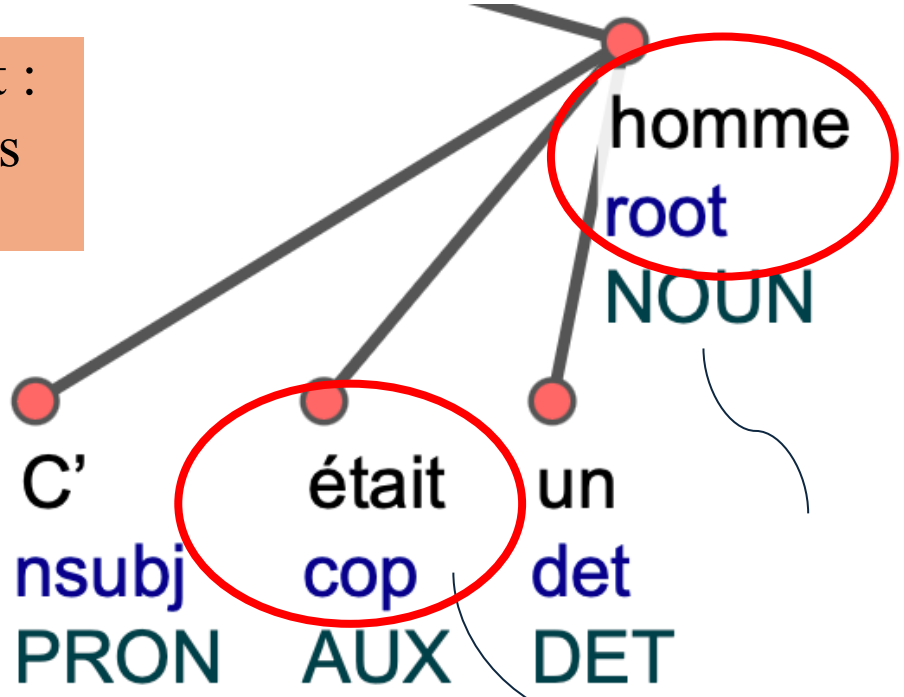
	Nominals	Clauses	Modifier words	Function words
Core arguments	**nsubj**	**csubj**		
	obj	**ccomp**		
	iobj	**xcomp**		
Non-core dependents	***obl***	***advcl***	***advmod***	***aux***
	vocative		***discourse***	***cop***
	expl			***mark***
	dislocated			
Nominal dependents	nmod	acl	amod	det
	appos			clf
	nummod			case

Quand il partait / En partant

Coordination	MWE	Loose	Special	Other
conj *conjunct*	**fixed** *multiword expression*	**list**	**orphan** (*when head is elided*)	**punct** *punctuation*
cc *coordinating conjunction*	**flat** *multiword expression*	**parataxis** (*direct speech*)	**goeswith** (*split words*)	**root**
	compound		**reparandum** *overridden disfluency*	**dep** *unspecified dependency*

Principe II : Priorité donnée aux mots lexicaux

Mots lexicaux mis en avant : **mots grammaticaux** traités comme des **dépendants**



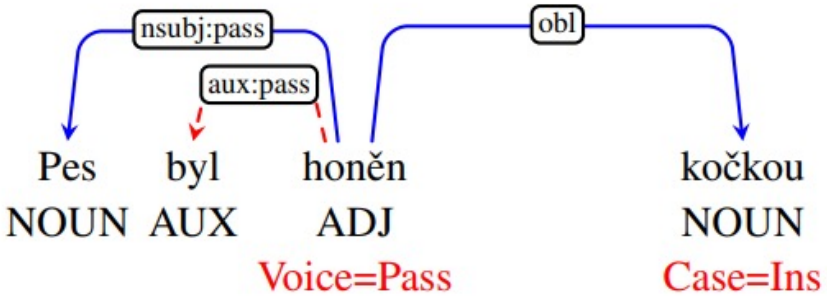
fonction syntaxique

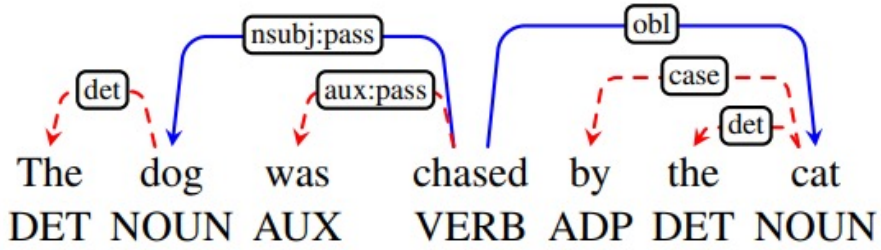
deprel: dependency relation

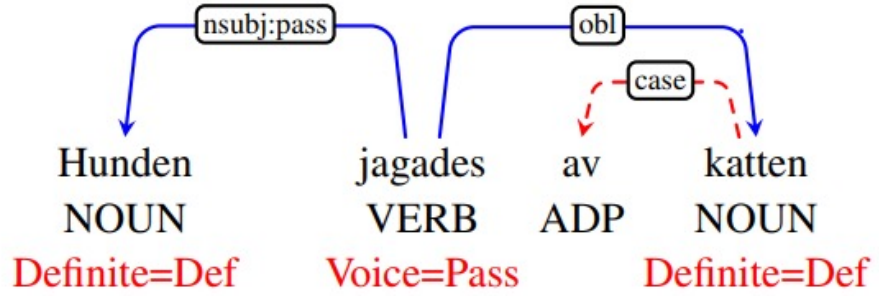
fonctio

deprel

SUD (Surface-Syntactic UD) : copule comme la tête (Gerdes et al. 2018)

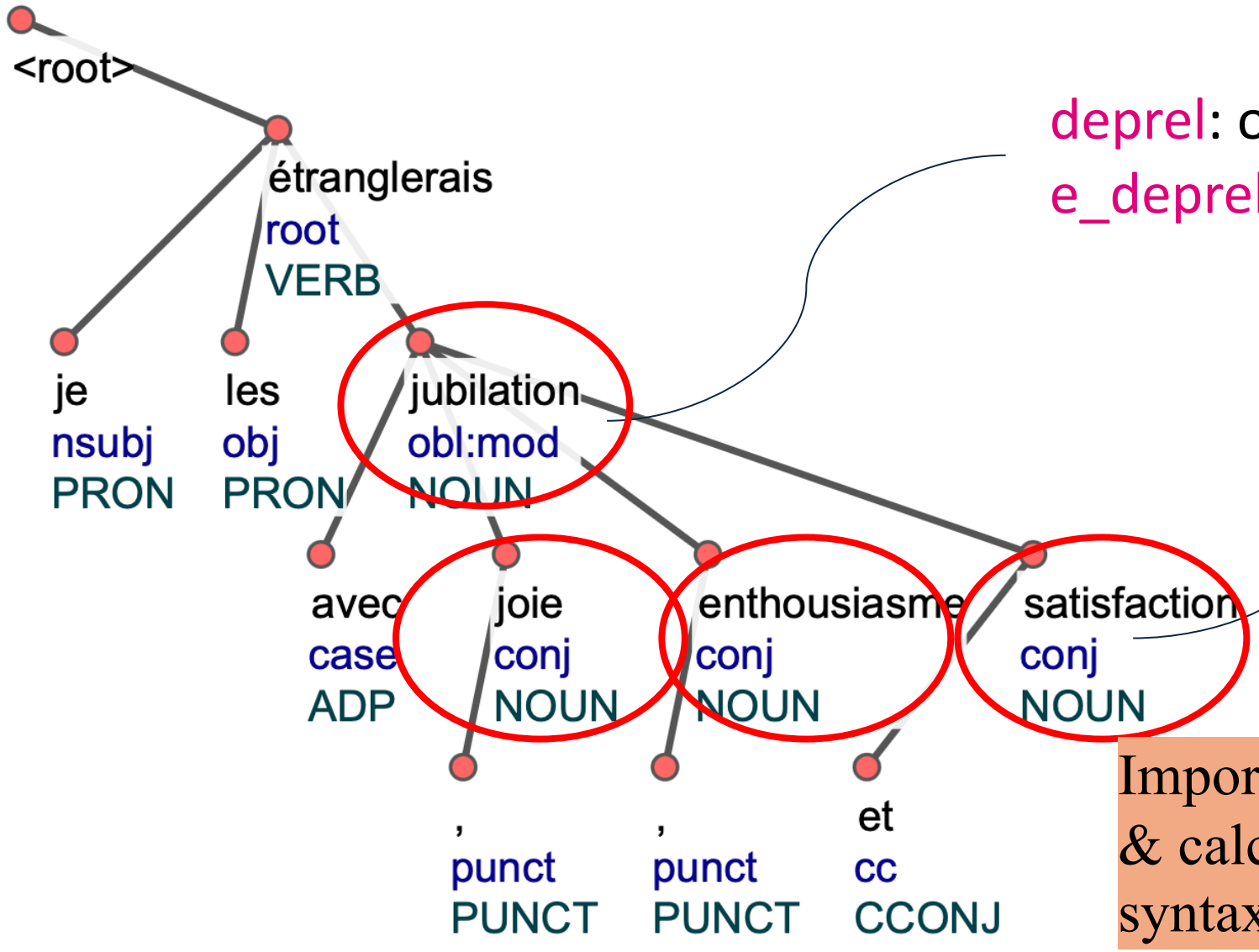






Principe III – coordinations en bouquet

je les étranglerais avec jubilation , joie , enthousiasme et satisfaction



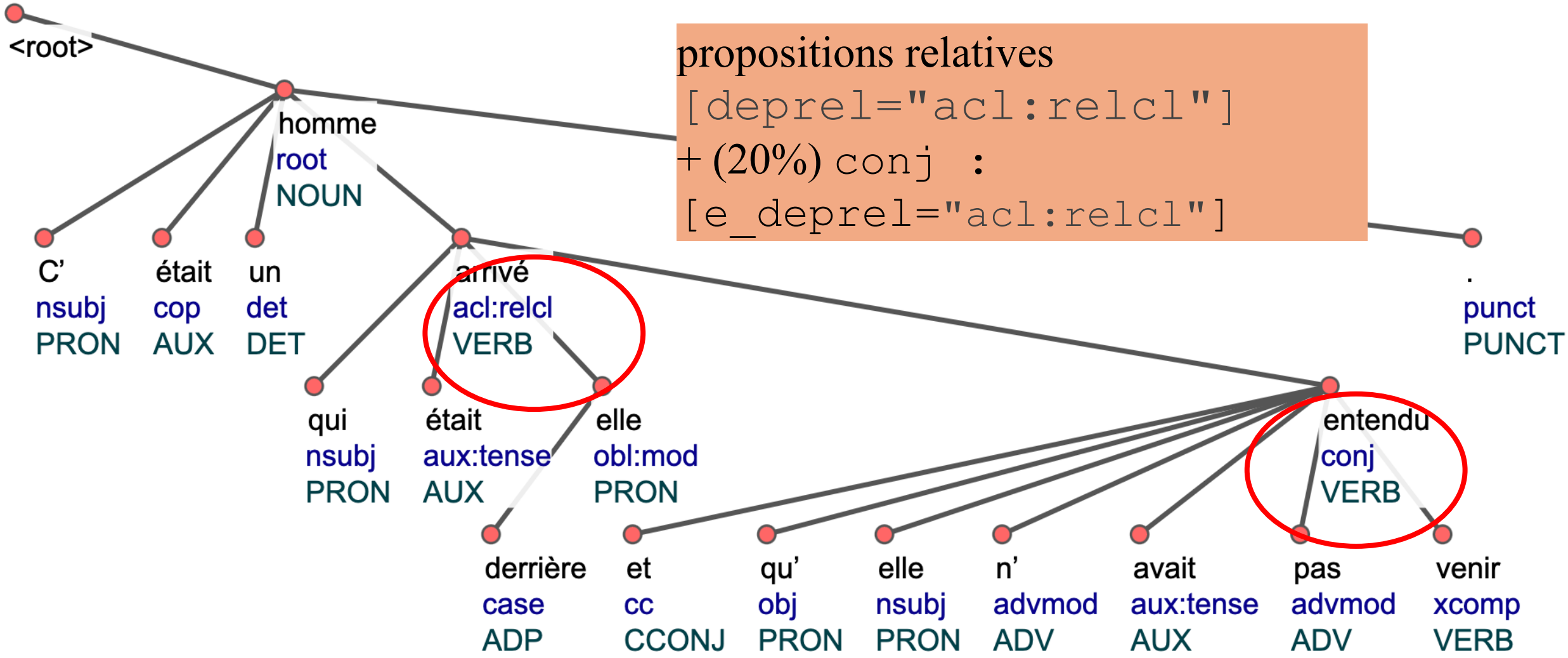
deprel: obl:mod
e_deprel: obl:mod

deprel: conj
e_deprel: obl:mod

Important pour la recherche sur corpus & calcul des mesures de la complexité syntaxique

ID	word	lemma	upos	feats	head	deprel	e_deprel
1	*C’*	c’	PRON	Gender=Masc|Number=Sing|Person=3|PronType=Dem	4	nsubj	nsubj
2	*était*	être	AUX	Mood=Ind|Number=Sing|Person=3|Tense=Imp|VerbForm=Fin	4	cop	cop
3	*un*	un	DET	Definite=Ind|Gender=Masc|Number=Sing|PronType=Art	4	det	det
4	*homme*	homme	NOUN	Gender=Masc|Number=Sing	0	root	root
5	*qui*	qui	PRON	PronType=Rel	7	nsubj	nsubj
6	*était*	être	AUX	Mood=Ind|Number=Sing|Person=3|Tense=Imp|VerbForm=Fin	7	aux:tense	aux:tense
7	*arrivé*	arriver	VERB	Gender=Masc|Number=Sing|Tense=Past|VerbForm=Part|Voice=Act	4	**acl:relcl**	**acl:relcl**
8	*derrière*	derrière	ADP	_	9	case	case
9	*elle*	lui	PRON	Emph=Yes|Gender=Fem|Number=Sing|Person=3|PronType=Prs	7	obl:mod	obl:mod
10	*et*	et	CCONJ	_	16	cc	cc
11	*qu’*	qu’	PRON	PronType=Rel	16	obj	obj
12	*elle*	lui	PRON	Emph=No|Gender=Fem|Number=Sing|Person=3|PronType=Prs	16	nsubj	nsubj
13	*n’*	n’	ADV	Polarity=Neg	16	advmod	advmod
14	*avait*	avoir	AUX	Mood=Ind|Number=Sing|Person=3|Tense=Imp|VerbForm=Fin	16	aux:tense	aux:tense
15	*pas*	pas	ADV	Polarity=Neg	16	advmod	advmod
16	*entendu*	entendre	VERB	Gender=Masc|Number=Sing|Tense=Past|VerbForm=Part|Voice=Act	**7**	**conj**	**acl:relcl**
17	*venir*	venir	VERB	VerbForm=Inf	16	xcomp	xcomp
18	.	.	PUNCT	_	4	punct	punct

C’ était un homme qui était arrivé derrière elle et qu’ elle n’ avait pas entendu venir .



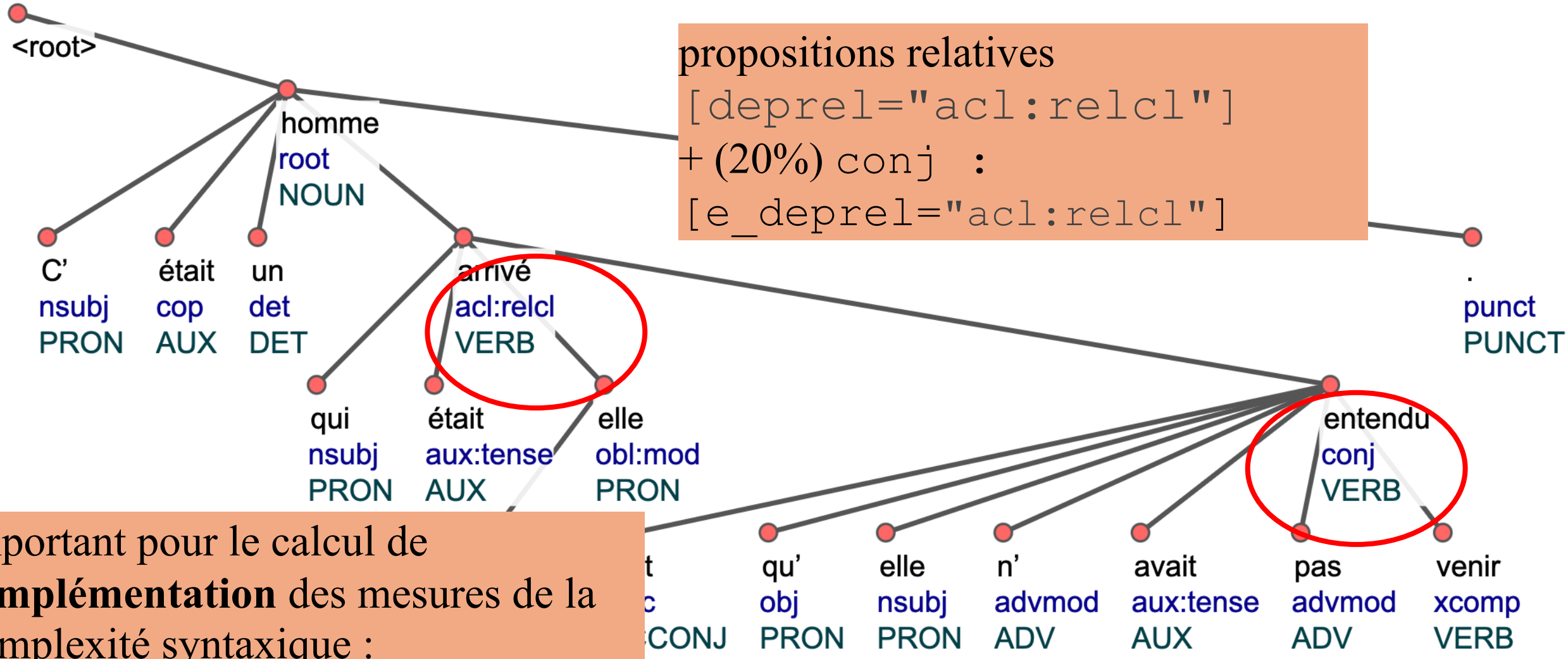
propositions relatives

`[deprel="acl:relcl"]`

+ (20%) `conj :`

`[e_deprel="acl:relcl"]`

C’ était un homme qui était arrivé derrière elle et qu’ elle n’ avait pas entendu venir .



propositions relatives
`[deprel="acl:relcl"]`
+ (20%) `conj :`
`[e_deprel="acl:relcl"]`

important pour le calcul de **l’implémentation** des mesures de la complexité syntaxique :
`conj` causerait de faux niveaux de dépendances

PLAN

1. Introduction : motivation de la recherche et délimitation du champs d'étude
2. Les ingrédients :
   - 2.1 Mesures de la complexité syntaxique
   - 2.2 *Universal Dependencies* : principes de base
   - **2.3 Les données : Corpus parallèle (multilingue) InterCorp**
3. Le résultat : corpus parallèle InterCorp v16ud (version pilote)
   - 3.1 Implémentation des mesures de la complexité syntaxique
   - 3.2 Exemples des requêtes sur corpus
4. Exemples d'application du corpus parallèle InterCorp v16ud
   - 4.1 Niveau de phrase (analyse contrastive et traductologique)
   - 4.2 Niveau de texte
   - 4.3 Niveau de genre textuel
   - 4.4 Niveau de langue (perspective typologique)
5. Conclusions : promesses et écueils

2.3 **Les données** : corpus multilingue InterCorp

- **projet lancé** en 2005 (Institut du Corpus national tchèque de la Faculté des Lettres de l'Université Charles à Prague, https://ucnk.ff.cuni.cz)
- **v1** disponible en ligne en 2008
- **2023 : v16** (Rosen, Šimčík, Vavřín & Zasina 2023)

Presentation du corpus (en anglais) :
https://wiki.korpus.cz/doku.php/en:cnk:intercorp

v13ud (et **v16ud** en préparation) :
- annotées selon **Universal Dependencies (UD)**
https://universaldependencies.org
- v16ud : implémentation de 6 mesures de la complexité syntaxique

Accès au corpus

Interface *KonText* :
https://kontext.korpus.cz

Accès gratuit et en ligne

a) **votre login institutionnel** (SSO Shibboleth)
b) inscription : https://www.korpus.cz/signup

`SYN2020 > all corpora >`

`InterCorp` **`v16`** – (version standard)

`InterCorp` **`v13ud`** – (version UD complète)

version annotée en UD et dotée de mesures de la complexité syntaxique :

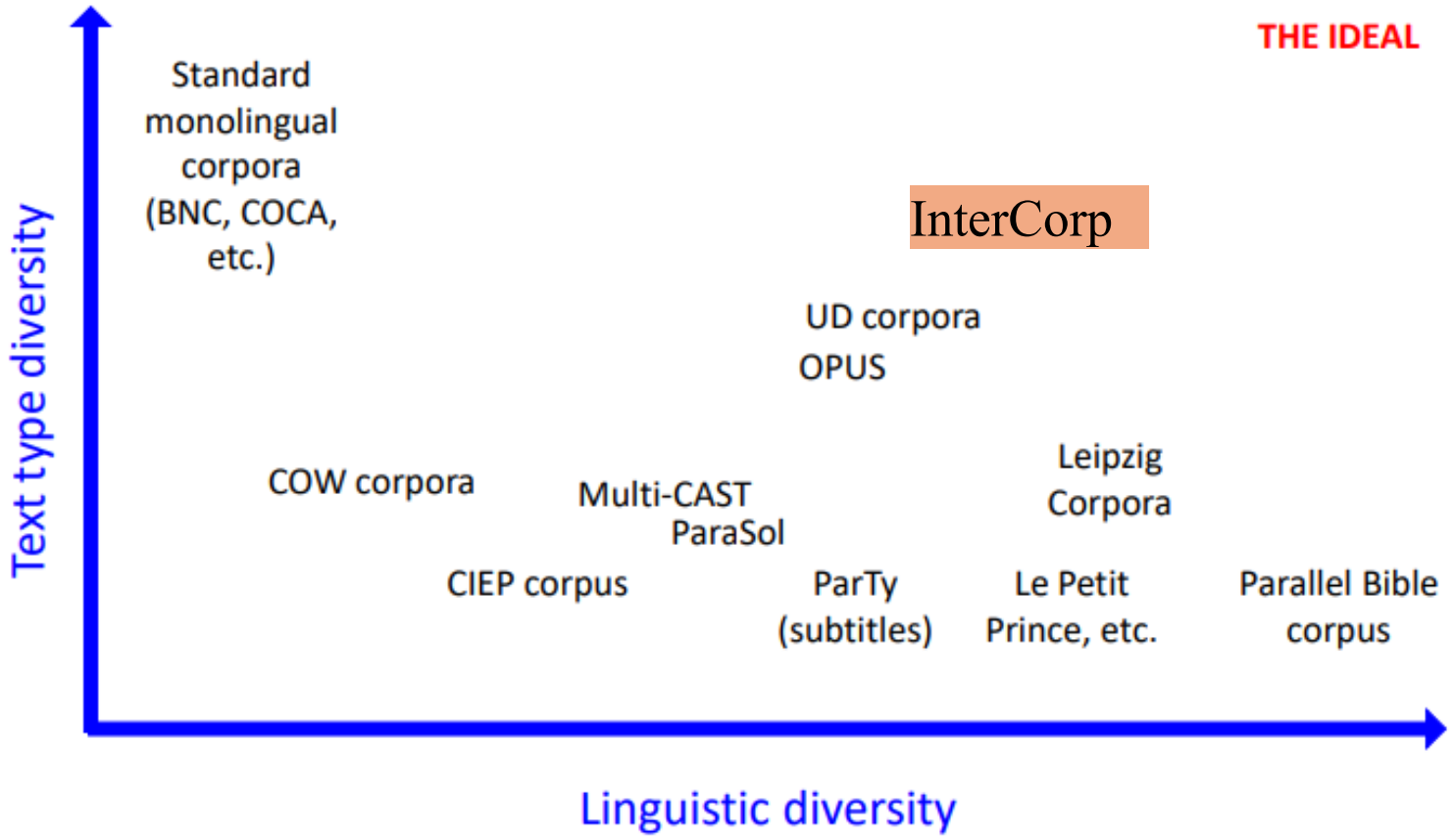
v16ud

pilote – *fiction* seulement

login provisoire :

login test16ud, mot de passe test16ud

Corpus parallèle (multilingue) InterCorp (composition et taille)

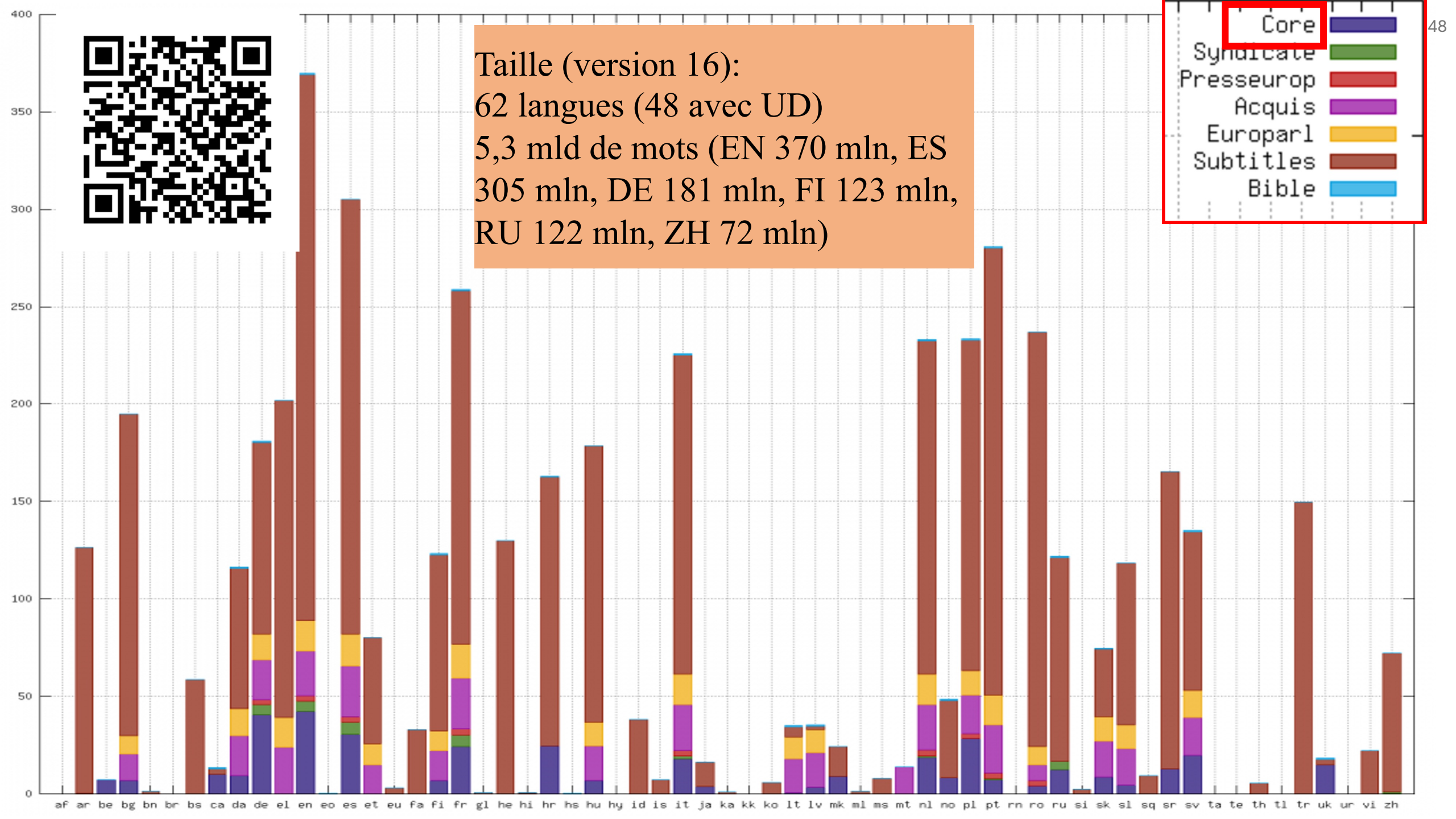


Crières de classification de corpus :

- diversité de genres textuels
- diversité de langues (corpus multilingues)

Natalia Levshina, UCCTS conference 2021

Taille (version 16):
62 langues (48 avec UD)
5,3 mld de mots (EN 370 mln, ES 305 mln, DE 181 mln, FI 123 mln, RU 122 mln, ZH 72 mln)



rouge : textes disponibles dans *core* du corpus

rouge en gras : *plus de 10 textes dans core*

InterCorp v16

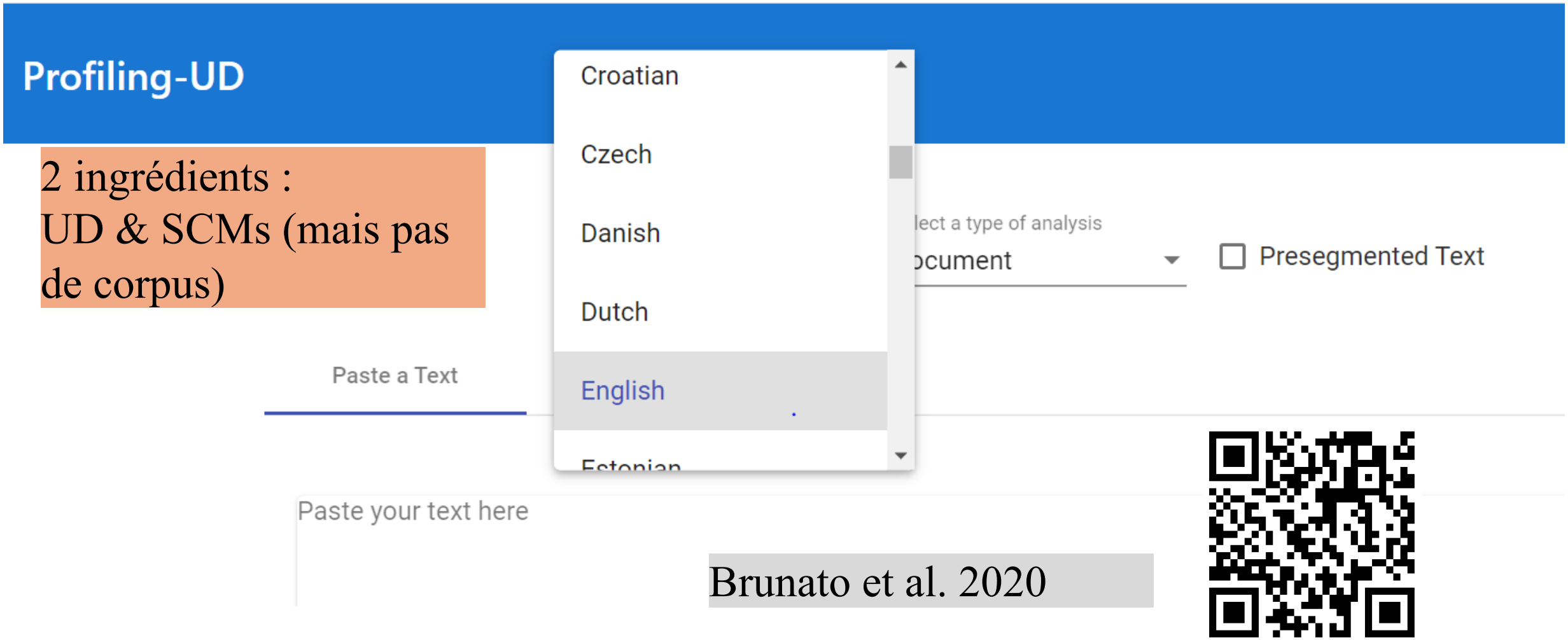
langues

Afrikaans Albanian Arabic Armenian Basque **Belarusian** Bengali Bosnian Breton **Bulgarian** Catalan Chinese **Croatian Czech** Danish **Dutch English** Esperanto Estonian **Finnish French** Galician Georgian **German** Greek Hebrew Hindi Hungarian Icelandic Indonesian **Italian Japanese** Kazakh Korean **Latvian** Lithuanian Macedonian Malay Malayalam Maltese **Norwegian** Persian **Polish Portuguese** Romani Romanian **Russian** Serbian Sinhala **Slovak Slovene Spanish Swedish** Tagalog Tamil Telugu Thai Turkish Ukrainian Upper Sorbian Urdu Vietnamese

PLAN

1. Introduction : motivation de la recherche et délimitation du champs d'étude
2. Les ingrédients :
   - 2.1 Mesures de la complexité syntaxique
   - 2.2 *Universal Dependencies* : principes de base
   - 2.3 Les données : Corpus parallèle (multilingue) InterCorp
3. **Le résultat : corpus parallèle InterCorp v16ud (version pilote)**
   - 3.1 Implémentation des mesures de la complexité syntaxique
   - 3.2 Exemples des requêtes sur corpus
4. Exemples d'application du corpus parallèle InterCorp v16ud
   - 4.1 Niveau de phrase (analyse contrastive et traductologique)
   - 4.2 Niveau de texte
   - 4.3 Niveau de genre textuel
   - 4.4 Niveau de langue (perspective typologique)
5. Conclusions : promesses et écueils

Linguistic Profiling Tool (UD) http://linguistic-profiling.italianlp.it/



2 ingrédients :
UD & SCMs (mais pas de corpus)

Brunato et al. 2020

Traits syntaxiques

Profiling UD output : Etranger

Etranger 1					
FR		CS		EN	
max_links_len	85	max_links_len	40	max_links_len	94
avg_prepositional_chain_len	1,080537	avg_prepositional_chai	1.0545454	avg_prepositional_ch	1.1008403
n_prepositional_chains	447	n_prepositional_chains	165	n_prepositional_chai	476
prep_dist_1	92,17002	prep_dist_1	94.545454	prep_dist_1	90.546218
prep_dist_2	7,606264	prep_dist_2	5.4545454	prep_dist_2	8.8235294
prep_dist_3	0,223714	obj_pre	49.113475	prep_dist_3	0.6302521
obj_pre	54,50763	obj_post	50.886524	obj_pre	2.4291497
obj_post	45,49237			obj_post	97.570850
subj_pre	99,60544	subj_pre	72.108108	subj_pre	97.840172
subj_post	0,394564	subj_post	27.891891	subj_post	2.1598272
dep_dist_acl	0,601527	dep_dist_acl	0.7182007	dep_dist_acl	0.5394502
dep_dist_acl:relcl	0,854269	dep_dist_advcl	1.2285012	dep_dist_acl:relcl	0.7364668
dep_dist_advcl	1,369863	dep_dist_advmod	7.0497070	dep_dist_advcl	2.1812552
dep_dist_advcl:cleft	0,080878	dep_dist_advmod:emp	1.4112014	dep_dist_advmod	7.0034712
dep_dist_advmod	6,475257	dep_dist_amod	3.3075033	dep_dist_amod	2.9599399

Profiling UD output : Etranger

Etranger 1					
FR		CS		EN	
dep_dist_root	6,121417	dep_dist_root	7.4781074	dep_dist_root	5.6055915
dep_dist_xcomp	1,713592	dep_dist_xcomp	1.5939015	dep_dist_vocative	0.0046908
principal_proposition_dist	49,91409	principal_proposition_(	55,85133	dep_dist_xcomp	1.7309316
subordinate_proposition_dis	50,08591	subordinate_propositic	44.148669	principal_proposition	45.098039
subordinate_post	88,25043	subordinate_post	89.306029	subordinate_proposit	54.901960
subordinate_pre	11,74957	subordinate_pre	10.693970	subordinate_post	82.798833
avg_subordinate_chain_len	1,244397	avg_subordinate_chain	1.2353760	subordinate_pre	17.201166
subordinate_dist_1	79,29562	subordinate_dist_1	80.083565	avg_subordinate_cha	1.2238010
subordinate_dist_2	17,50267	subordinate_dist_2	16.991643	subordinate_dist_1	80.373001
subordinate_dist_3	2,774813	subordinate_dist_3	2.2284122	subordinate_dist_2	17.229129
subordinate_dist_4	0,320171	subordinate_dist_4	0.6963788	subordinate_dist_3	2.0426287
subordinate_dist_5	0,106724			subordinate_dist_4	0.3552397

2 problèmes :

1. `conj` pas pris en compte
2. pas de corpus

3. Le résultat : InterCorp v16ud (SCMs)

implémentation	phrasal level	sentence/clausal level
nombre d'entités	maxNPLength *maximum NP length*	sLength *sentence length (words)*
		subRatio *subordination ratio*
organisation hiérarchique	maxNPDepth *maximum NP depth*	maxTreeDepth *maximum tree depth*
effort cognitif		mdd *mean dependency distance*

Implémentation en tant qu'attributs de :

View > Corpus-specific settings > Structures: **<text>, <s>**

<text>:

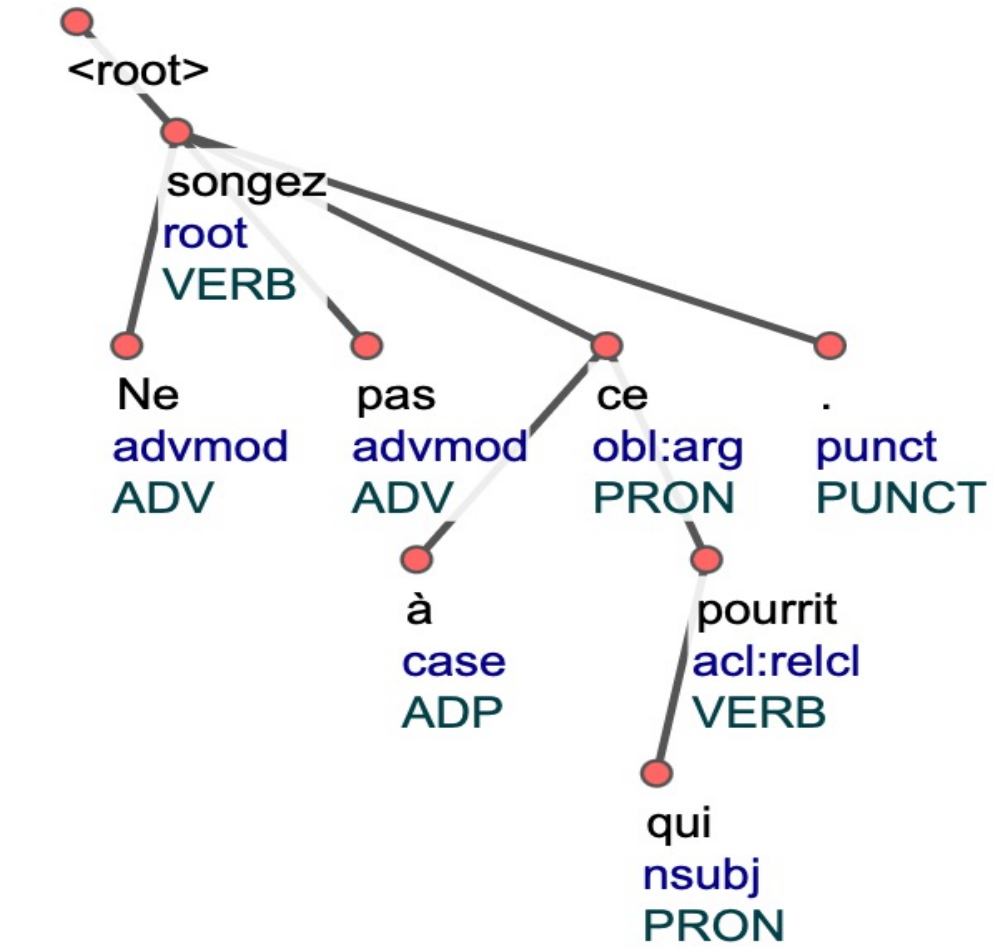
<text
author=Hugo, Victor
title=Les Misérables
lexDivWord=468.07
lexDivLemma=379.30
subRatioAvg=2.05
maxTreeDepthAvg=0.84
sLengthAvg=15.32
mdd=2.93
maxNPLengthAvg=6.49
maxNPDepthAvg=1.86
…
>

<s> (phrase) :

<s
id=fr:Hugo-Bidnici:0:130:9
maxNPDepth=2
subRatio=2.0
sLength=7
maxNPLength=4
mdd=1.5
maxTreeDepth=1
>

View > KWIC/**Sentence**

Ne songez pas à ce qui pourrit .



InterCorp v16ud (SCMs)

implémentation	phrasal level	sentence/clausal level
nombre d'entités	maxNPLength *maximum NP length*	sLength *sentence length (words)*
		subRatio *subordination ratio*
organisation hiérarchique	maxNPDepth *maximum NP depth*	maxTreeDepth *maximum tree depth*
effort cognitif		mdd *mean dependency distance*

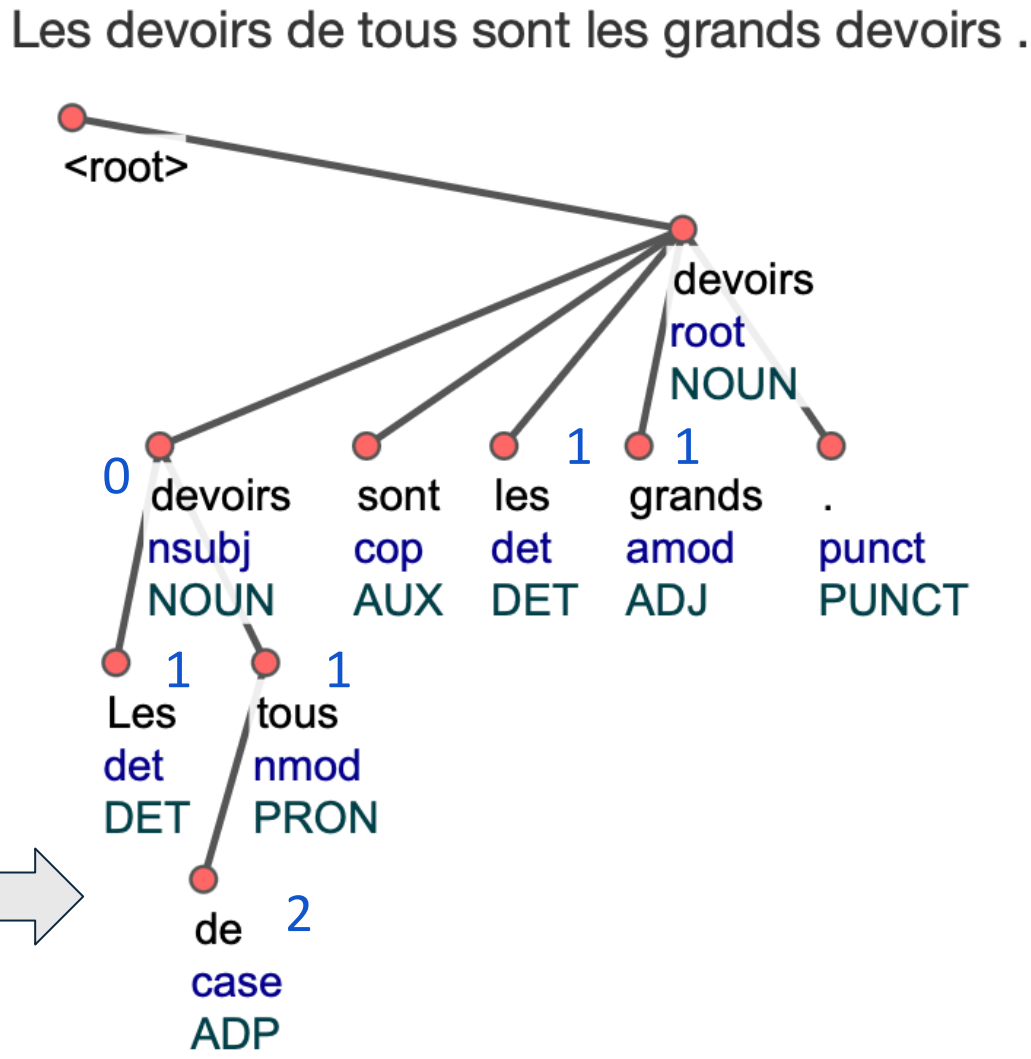
NP – mesures de la complexité syntaxique

MaxNPLength:

- nombre de mots dans NP le plus long
- *les devoirs de tous*
- = 4

MaxNPDepth:

- nombre maximal d'enchâssements dans NP
- *devoirs* … 0
- *les* … 1
- *tous* … 1
- *de* … 2
- = 2



InterCorp v16ud (SCMs)

implémentation	phrasal level	sentence/clausal level
nombre d'entités	maxNPLength *maximum NP length*	sLength *sentence length (words)*
organisation hiérarchique	maxNPDepth *maximum NP depth*	subRatio *subordination ratio*
		maxTreeDepth *maximum tree depth*
effort cognitif		mdd *mean dependency distance*

Phrase – mesures de la complexité syntaxique

sLength:

- nombre de mots dans la phrase
- sans ponctuation

MaxTreeDepth:

- nombre maximal d'enchâssements de propositions subordonnées (finies et non-finies)
- `conj` pas considéré comme niveau d'enchâssement

subRatio:

- subordination ratio
- $$\frac{(\text{N° T-units} + \text{N° Sub})}{\text{N° T-units}}$$

Phrase

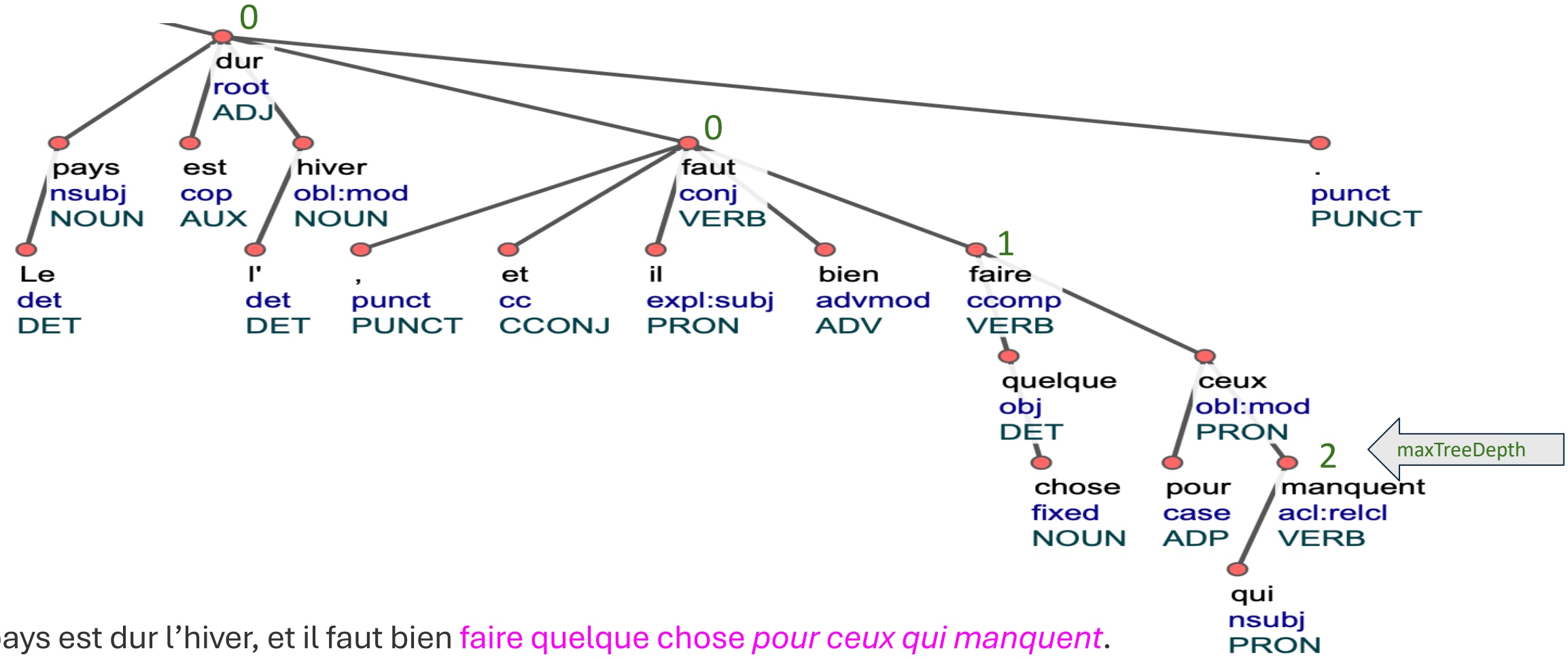
maxNPDepth=2 **subRatio=2.0** sLength=17

maxNPLength=4 mdd=1.75 **maxTreeDepth=2**

No T-units = 2

No sub. clauses = 2

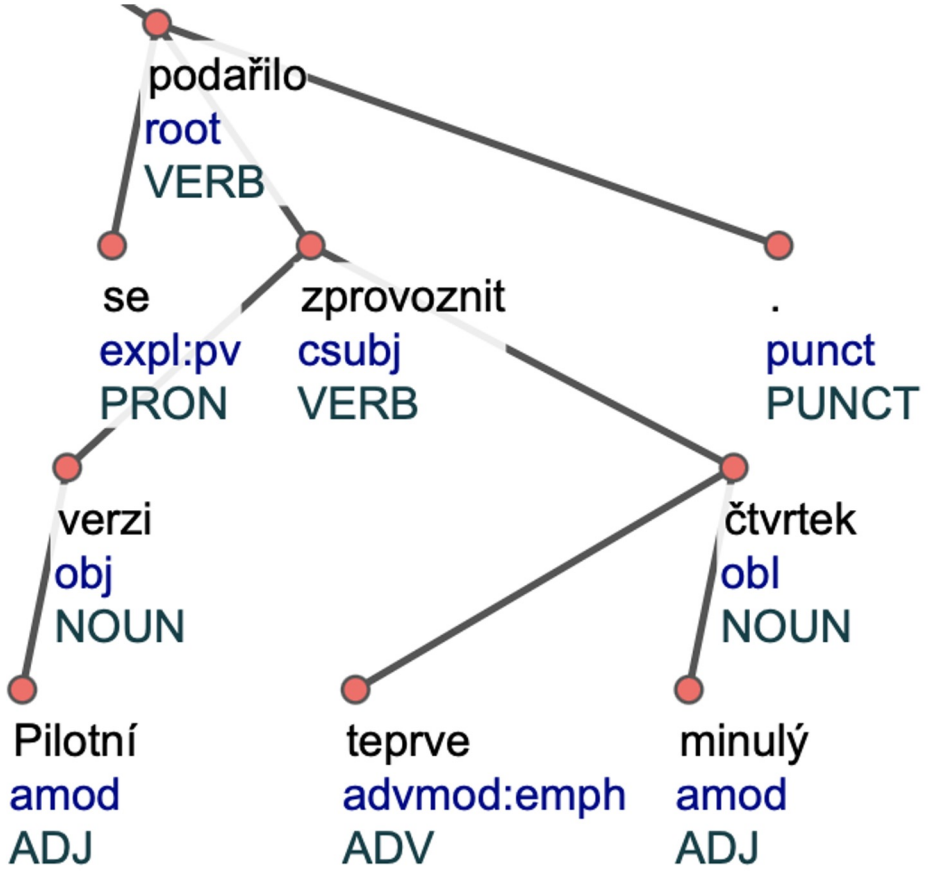
subRatio = (2 + 2) / 2 = **2**



Le pays est dur l'hiver, et il faut bien faire quelque chose *pour ceux qui manquent*.

mdd = Mean Dependency Distance

= average number of words occurring between the syntactic head and the dependent in a text (Yan & Li 2019, Yan 2021, Liu 2008, Mačutek et al. 2021, Brunato & Venturi 2023, etc.)



- censé refléter le degré de l'effort cognitif
- sans la ponctuation
- calcul (*n* … nombre de mots dans la phrase)

  $DD_i = | ID_i - head_i |$

  $DD = \sum_{i=0 \text{ à } n} DD_i$

  $mdd = DD / (n - 1)$
- DD = 12

  mdd = 12 / 7 ≅ 1,71

	Pilotní	verzi	se	podařilo	zprovoznit	teprve	minulý	čtvrtek
ID (= *i*)	1	2	3	4	5	6	7	8
$head_i$	2	5	4	0	4	8	8	5
DD_i	1	3	1	0	1	2	1	3

3.2 Exemples de requêtes sur corpus

a) informations sur la complexité syntaxique **incluses directement dans la requête**
b) informations sur la complexité syntaxique **visualisées dans la concordance** (monolingue ou parallèle)

3.2.1 SCMs directement dans la requête :

EXEMPLE 1 – quel sera le résultat de cette requête ?

```
<s maxTreeDepth="0" & sLength <= "10" />
```

Spécification : dans ce type de phrases, cherche les sujets coordonnés (*Marie et Pierre sont partis*)

```
[deprel="conj" & p_deprel="nsubj.*"] within
<s maxTreeDepth="0" & sLength <= "10" />
```

SCMs directement dans la requête : recherche contrastive

EXEMPLE 2 – propositions relatives coordonnées en contraste

French – aligned corpus English

[deprel="conj" & p_deprel="acl:relcl"] within <s maxTreeDepth >= "3" />

spécification dans le corpus aligné `- English ("advanced search", "Translations **contain** matching results")`

[deprel="conj" & p_deprel="acl:relcl"]


InterCorp v13ud - English
Advanced query
Insert tag
Insert within
Keyboard
Recent queries
[deprel="conj" & p_deprel="acl:relcl"]
TIP In CQL mode you can use Shift+ENTER to enter a new line (next tip)
Specify parameters
Query application:
Translations contain matching results
include non-aligned:

SCMs directement dans la requête–

maxNPDepth et maxNPLength

EXEMPLE 3 :

[deprel="nsubj:pass*"] within <s maxNPDepth >="10" &

maxNPLength >="40"/>

corpus alignés CS-EN :

https://www.korpus.cz/kontext/view?q=~GiOiMyS0uU6c (Concordance > Permanent link)

3.2.2 SCMs visualisées dans la concordance (monolingue ou parallèle)

InterCorp v16ud – French, aligned corpus English + advanced search

Participes (présents ou passés) :

```
[feats="VerbForm=Part" & upos="VERB"]
```

`View > Corpus specific settings > References (metadata) > <s> > sub.ratio + s.length + Max.tree.depth` (ou d'autres) > en bas : `APPLY VIEW OPTIONS`

(le cas échéant : `<doc> > doc.id`)

Corpus-specific settings for InterCorp v16ud - French

Positional attributes | Structures | References | Additional functions

<#>
- Token number

<doc>
- Document number
- doc.id
- doc.tag_model

<text>
- text.lang
- text.pubyear
- text.version
- text.pubmonth
- text.pubDateYear
- text.pubDateMonth
- text.id
- text.author
- text.title
- text.group
- text.publisher
- text.pubplace
- text.origyear
- text.isbn
- text.txtype
- text.comment
- text.original
- text.srclang
- text.translator
- text.transsex

<p>
- p.id

<s>
- s.id
- s.maxNPDepth
- s.subRatio
- s.sLength
- s.maxNPLength
- s.mdd
- s.maxTreeDepth

tree – doc_id – **sub.ratio – s_length – max.tree.depth**

CS

Vaculik-Cesky_snar ✦ 2.5 ✦ 18 ✦ 2

Dal jsem mu povinnou výstrahu , aby to nedělal , a
doporučil / mu dotáhnout pointu pro případ , že to udělá .

fr

Vaculik-Cesky_snar ✦ 3.5 ✦ 38 ✦ 2

Comme il se doit , je lui ai donné l' avertissement obligatoire en lui demandant de ne pas le faire et je lui ai recommandé de travailler la péroraison dans le cas où il le ferait tout de même .

Recommandation : télécharger la concordance (SCMs incluses)

Corpus multilingue (parallèle = de traductions) & **annotation syntaxique** d'après le schéma commun

kontext Query Corpora Save Concordance Filter Frequency Collocations View Help

Corpus: InterCorp v13ud - English | **Query:** acl:relcl, Core, fiction, en, f (31,182 hits) ▶ **Shuffle:** ✓ ~ Details

Hits: **31,182** | i.p.m.: Calculate | ARF: **991.48** | Result is shuffled 1 / 780

Line selection: simple

	InterCorp v13ud - English		InterCorp v13ud - French
Giono-Husar	All he had in his favour was his eyes , which still , in spite of everything , had an attractive warmth .	Giono-Husar	Il n' avait plus pour lui que ses yeux qui donnaient toujours cependant des feux aimables .
Verne-Cesta_kolem_s	Phileas Fogg got into the train , which started off at full speed .	Verne-Cesta_kolem_sv	Sur cette réponse , Phileas Fogg monta dans le wagon , et le train partit à toute vapeur .
Hemingway-SbohemArmado	It 's only the first labor , which is almost always protracted .	Hemingway-SbohemArmado	Le premier accouchement est toujours laborieux .
Golding-Pan_much	What ' ud become of us ? "	Golding-Pan_much	Qu'est -ce qu' on deviendrait ? »
lodge-hostujici_prof	What I wouldn't give for an indigenous Indian with a PhD , ' he murmured wistfully , like a man on a desert island dreaming of steak and chips .	lodge-hostujici_prof	Qu'est -ce que je ne donnerais pas pour trouver un authentique Indien titulaire d' un doctorat » , marmonna -t -il d' un air songeur , comme un homme abandonné sur une île déserte qui rêve d' un steak-frites .
hosseini-lovec_draku	"I meant to tell you in there , about what you 're trying to do ?	hosseini-lovec_draku	— Je voulais vous dire que je trouve votre démarche admirable .
brown-sifra	ON THE VERGE OF UNVEILING ONE OF HISTORY 'S GREATEST SECRETS , AND HE TROUBLES HIMSELF WITH A WOMAN WHO HAS PROVEN HERSELF UNWORTHY OF THE QUEST.	brown-sifra	Il est sur le point de découvrir l' un des plus grands secrets de l' histoire de l' humanité , et il écoute les caprices d' une petite bonne femme qui s' est montrée indigne de la quête , pensa Teabing avec mépris .
Giono-Husar	He had stopped some ten paces from the gloomy bulk of the walls , blacker than the night , and listened for the sounds , however light , that a man on watch never fails to make .	Giono-Husar	Il s' était arrêté à quelque dix pas de la masse sombre des murs , plus noire que la nuit et il guettait le bruit , pour si léger qu' il soit , que ne manque pas de faire un homme qui veille .
rowlingova-hpot_pohar	Harry had the impression that Davies was too busy staring at Fleur to take in a word she was saying	rowlingova-hpot_pohar	Harry pensa qu' il était certainement trop occupé à contempler Fleur pour comprendre un mot de ce qu' elle disait .
Styron-Sofiina_volba	A member of the moderate wing of the party , Professor Biegański , then a rising young faculty star in his thirties , wrote an article in a leading Warsaw political journal deploring these assaults , which caused Sophie a number of years later to wonder — when she happened upon the essay — whether he hadn't suffered a spasm of radical - utopian humanism .	Styron-Sofiina_volba	Membre de l' aile modérée du parti , le Professeur Bieganski , alors jeune étoile montante de l' université , trente ans tout au plus , écrivit un article que publia l' un des plus importants journaux politiques de Varsovie , pour déplorer ces violences , ce qui , un certain nombre d' années plus tard , poussa Sophie à se demander – quand par hasard elle tomba sur l' essai en question – s' il n'avait pas été frappé par une bouffée d' humanisme radical-utopique .
Littell-Bohyne	We went back down to the town by the Verkhnyi rynok , where the peasants were finishing packing up their unsold chickens , fruits , and vegetables onto carts or mules .	Littell-Bohyne	Nous redescendîmes en ville par le Verkhnii rynok où les paysans achevaient de remballer leurs poules , leurs fruits et leurs légumes invendus sur des charrettes ou des mulets .

PLAN

1. Introduction : motivation de la recherche et délimitation du champs d'étude
2. Les ingrédients :
   - 2.1 Mesures de la complexité syntaxique
   - 2.2 *Universal Dependencies* : principes de base
   - 2.3 Les données : Corpus parallèle (multilingue) InterCorp
3. Le résultat : corpus parallèle InterCorp v16ud (version pilote)
   - 3.1 Implémentation des mesures de la complexité syntaxique
   - 3.2 Exemples des requêtes sur corpus
4. **Exemples d'application du corpus parallèle InterCorp v16ud**
   - 4.1 Niveau de phrase (analyse contrastive et traductologique)
   - 4.2 Niveau de texte
   - 4.3 Niveau de genre textuel
   - 4.4 Niveau de langue (perspective typologique)
5. Conclusions : promesses et écueils

4. Exemples d’application du corpus parallèle InterCorp v16ud

4.1 Niveau de phrase (analyse contrastive, traductologique) CS-FR-EN

4.2 Niveau de texte

4.3 Niveau de genres textuels

4.4 Niveau de langue (perspective typologique)

Type 1 : différences structurelles (non-finite subordinate vs finite coordinate verb forms)

FR - INFINITIVE sub. - xcomp

Elinor ne peut [root-LEVEL 0] s'empêcher **de lui signaler [xcomp-LEVEL 1]** qu'elle exagère [ccomp-LEVEL 2]. (non-fiction)

*'Elinor cannot refrain **from pointing out** that she exaggerates.*

sub.ratio=3 ((1+2)/1)

max.tree.depth = 2

EN similar to FR:

Elinor cannot refrain from pointing out to her sister that she is being extravagant.

cs - finite verb – root (conj – coordination)

Elinor se nedokáže [root-LEVEL 0] přemoct **a naznačí jí [conj-LEVEL 0]**, že přehání [ccomp-LEVEL 1]. (non-fiction)

*'Elinor cannot refrain **and points out** that she exaggerates.'*

sub.ratio=1.5 ((2 + 1)/2)

max.tree.depth = 1

Différences en préférences stylistiques

FR – étoffement

les disparités

opposant [deprel=acl]

les classes populaires et les classes moyennes.

*‚differences **opposing** the lower and middle classes‘*

cs, en - dépouillement (Vinay & Darbelnet 1971)

cs

rozdílů **mezi [between, ADP]** lidovou a střední vrstvou ~~[deprel=acl]~~

‚differences between the lower and middle class‘

Anglais similaire au tchèque :

disparity **between** the lower and middle classes.

Type 2 : *Specific features of translation* (*normalization*), Camus – *L'Etranger (>cs, en)*

FR – phrases courtes comme choix de l'auteur

J'ai bu.

‚I drank.'

J'ai eu alors envie de fumer.

‚I thus wanted to smoke.'

cs + en: fusion de phrases courtes

cs traduction :

Vypil jsem ji **a dostal** jsem chuť si zakouřit.

*‚I drank it **and I wanted** to smoke.'*

en traduction :

I drank the coffee, **and** then **I wanted** a cigarette.

Type 3 – différences d'annotation

FR – participles [deprel=acl] (clausal)

fr - […] *des* ***formes dérivées*** **[deprel=acl]** ***des idées suprêmes*** *du Bien*. (Sedláček *Ekonomie*)

en similar to fr:

[…] **forms derived [deprel=acl]** from the utmost ideas of Good .

acl deprel = 5% of clausal deprels

CS – participles [deprel=amod] (non-clausal)

CS – **formy odvozené [deprel=amod]** od nejzazší ideje Dobra.

‚forms derived from the utmost idea of Good.'

4.2 Application au niveau de **textes**

	phrasal	**clause/sentence**
nombre d'entités	maxNPLength**Avg** ***average*** *NP length*	sLength**Avg** ***average*** *sentence length (words)*
		subRatio**Avg** ***average*** *subordination ratio*
organisation hiérarchique	maxNPDepth**Avg** ***average*** *maximum NP depth*	maxTreeDepth**Avg** ***average*** *maximum tree depth*
effort cognitif		mdd *mean dependency distance*

mdd - mean dependency distance (MAX)

En tchèque (fiction et non-fiction, CS et cs) : styles spécifiques d'auteurs, par exemple Bohumil Hrabal

author	title	srclang	wordcount	subRatioAvg	maxTreeDepth	sLengthAvg	mdd	maxNPLengt	maxNPDepthAvg
Hrabal, Bohumil	Taneční hodiny pro s	cs	17460	2,37	2,70	873,05	10,37	72,20	5,20
Céline, Louis Ferdinan	Od zámku k zámku	fr	104807	1,76	0,86	41,30	8,33	9,10	1,66
García Márquez, Gabri	Podzim patriarchy	es	70478	4,32	4,29	310,53	7,36	68,23	7,72
Gersaová, Telinda	Mlčení	pt	22581	2,11	1,11	38,21	6,94	8,92	2,25
Zabužko, Oksana	Polní výzkum ukrajir	uk	35534	2,36	1,04	48,87	6,72	13,75	2,21
Hrabal, Bohumil	Obsluhoval jsem an	cs	67992	2,45	2,00	100,89	6,69	15,51	3,10
Céline, Louis Ferdinan	Sever	fr	132383	1,61	0,66	28,56	6,39	5,25	1,24
Hrabal, Bohumil	Kouzelná flétna	cs	3586	2,73	1,78	65,22	5,80	11,58	2,85
Macourek, Miloš	Mach a Šebestová	cs	14202	2,37	1,82	72,88	5,37	8,93	2,23
Delibes, Miguel	Pět hodin s Mariem	es	71494	2,43	1,41	39,22	5,37	4,96	1,58
Melchor, Fernanda	Období hurikánů	es	60085	4,53	2,18	63,19	5,26	12,94	2,71
Saramago, José	Baltasar a Blimunda	pt	111378	2,59	1,78	56,59	5,06	11,47	2,96
Pánek, Josef	Láska v době globál	cs	45703	2,14	0,95	27,26	5,01	4,82	1,40
Hrabal, Bohumil	Příliš hlučná samot	cs	25615	2,40	1,94	71,38	4,94	12,82	3,28
Macourek, Miloš	Pohádky	cs	44608	2,00	1,10	27,62	4,89	3,94	1,39
Hrabal, Bohumil	Postřižiny	cs	29216	1,76	1,02	36,05	4,87	6,30	1,91

lexical diversity : lexDivLemma (MIN)

CS, cs

lexDivLemma

(minima)

author	title	srclang	wordcount	lexDivWord	lexDivLemma
Tále, Samko	Kniha o hřbitově	sk	40193	421,49	278,60
Havel, Václav	Hry - Audience	cs	5175	452,43	301,26
Čapek, Josef	Povídání o pejskov	cs	11559	463,83	304,68
Karafiát, Jan	Broučci	cs	23590	463,81	309,10
Wittgenstein, Ludwig	Tractatus logico-ph	de	15375	506,44	318,21
Jarunková, Klára	Můj tajný zápisník	sk	14129	488,85	331,87
Čapek, Karel	Matka	cs	18062	518,42	348,42
Milne, Alan Alexander	Púovo zátiší	en	20795	496,31	349,31
Macourek, Miloš	Pohádky	cs	44608	496,65	350,16
Milne, Alan Alexander	Medvídek Pú	en	16967	496,93	350,29
Lindgrenová, Astrid	Děti z Bullerbynu	sv	50404	503,80	351,75
Havel, Václav	Hry - Vernisáž	cs	5170	503,98	352,22
Pánek, Josef	Láska v době globál	cs	45703	482,42	359,23
Havel, Václav	Largo desolato	cs	13378	498,65	361,03
Fuks, Ladislav	Myši Natálie Moosh	cs	97372	502,23	365,08

4.3 Niveau de **genres textuels** : non-fiction vs. fiction ?

H: higher SCMs in non-fiction than in fiction (Cvrček et al. 2020, etc.)

Nádvorníková (2023), UCCTS conference, Poznań

Comparison of EN-CS-FR non-fiction:

Results:

(Non-translated) Fiction vs non-fiction				sub.ratio	
	size (tokens)	root	sub. clauses	non-fiction	fiction
CS	887,089	61,471	45,647	1.74	1.63
EN	2,224,265	111,013	165,639	2.49	1.96
FR	2,383,843	97,312	178,042	2.83	2.11

Stat. significance (chi-squared test):

Significance (FDR correction)		
	CS	EN
EN	<2e-16	-
FR	<2e-16	<2e-16

Acquis communautaire

sub.ratio	Cohen's h	
1,56	CS-EN	0,345077
2,13	CS-FR	0,404877
2,29	EN-FR	0,064273

Effect size (Cohen's h):

Comparison of sub.clause and T-unit:		
First	Second	h
CS	EN	0,312638
CS	FR	0,405693
EN	FR	0,101912

4.4 Niveau de langues (perspective typologique)

Analyse pilote :

- 6 textes (fiction) dans 12 langues différentes (noyau commun – extrême comparabilité du sous-corpus), 524 240 tokens

Languages (4 Slavic, 3 Germanic, 3 Romance, Finnish and Japanese)

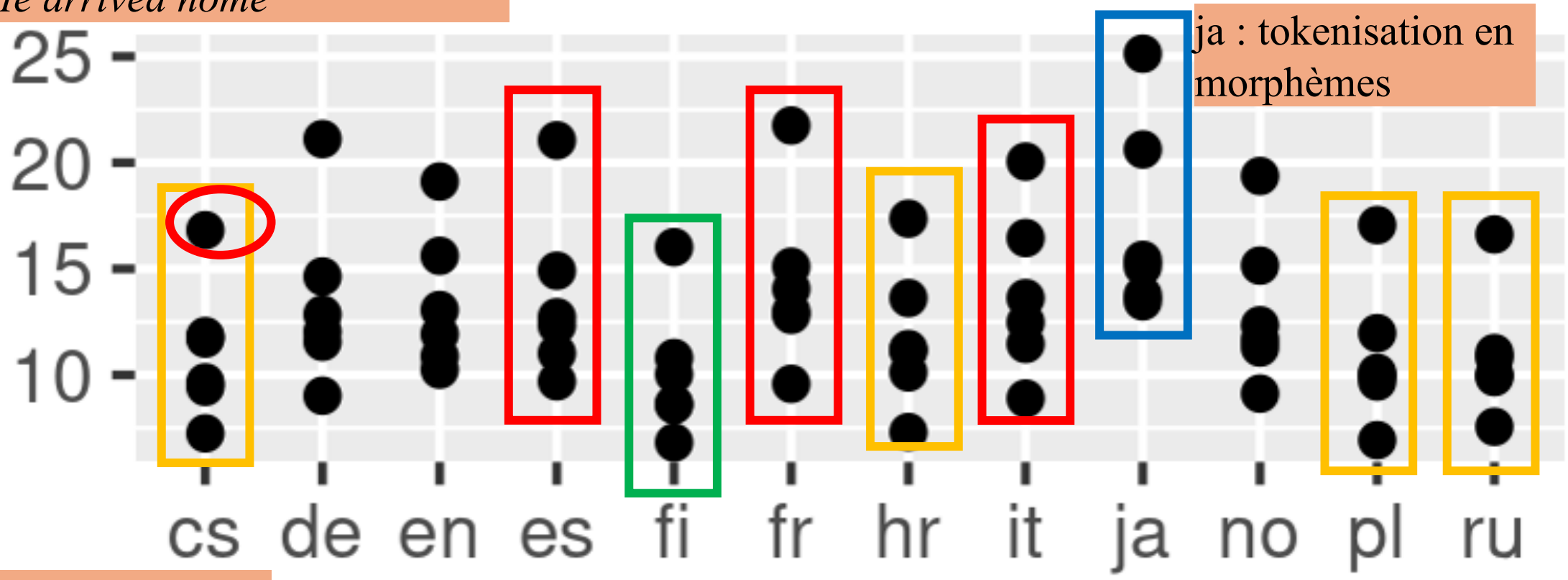
- **cs, ru, hr, pl**
- de, en, no
- **es, it, fr**
- **fi**, **ja**

4.4.1 *s_length*

fr *Il est arrivé à la maison.(6 w.)*
cs: *Přišel domů (2 words)*
‚He arrived home‘

Scatterplot – average sentence length (6 textes / 12 langues)

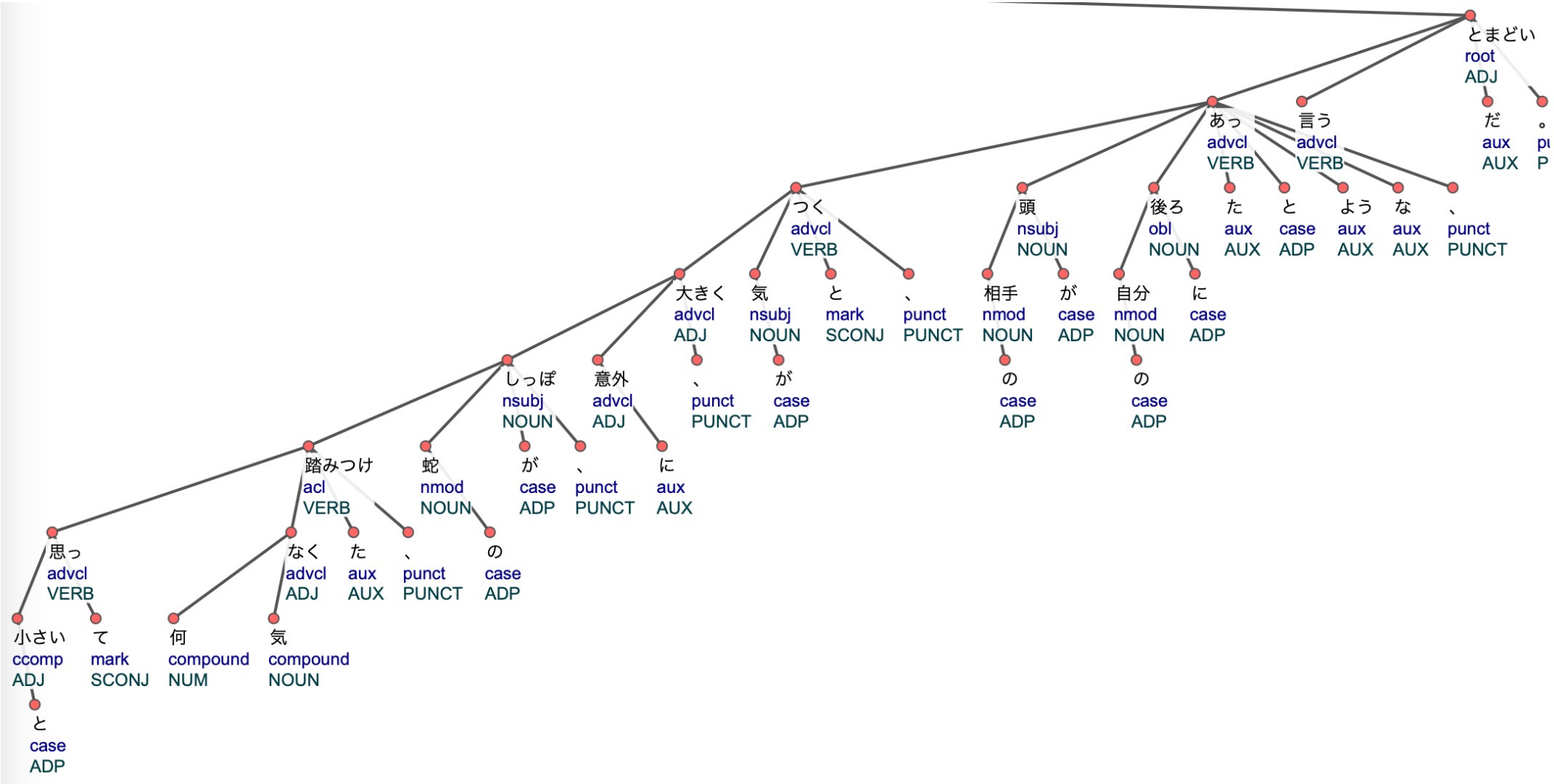
mesure dépendant des propriotés typologiques des langues et/ou des spécificités de l’annotation en UD



ja : tokenisation en morphèmes

facteur dans *fiction* : style du texte

Tokenisation du japonais

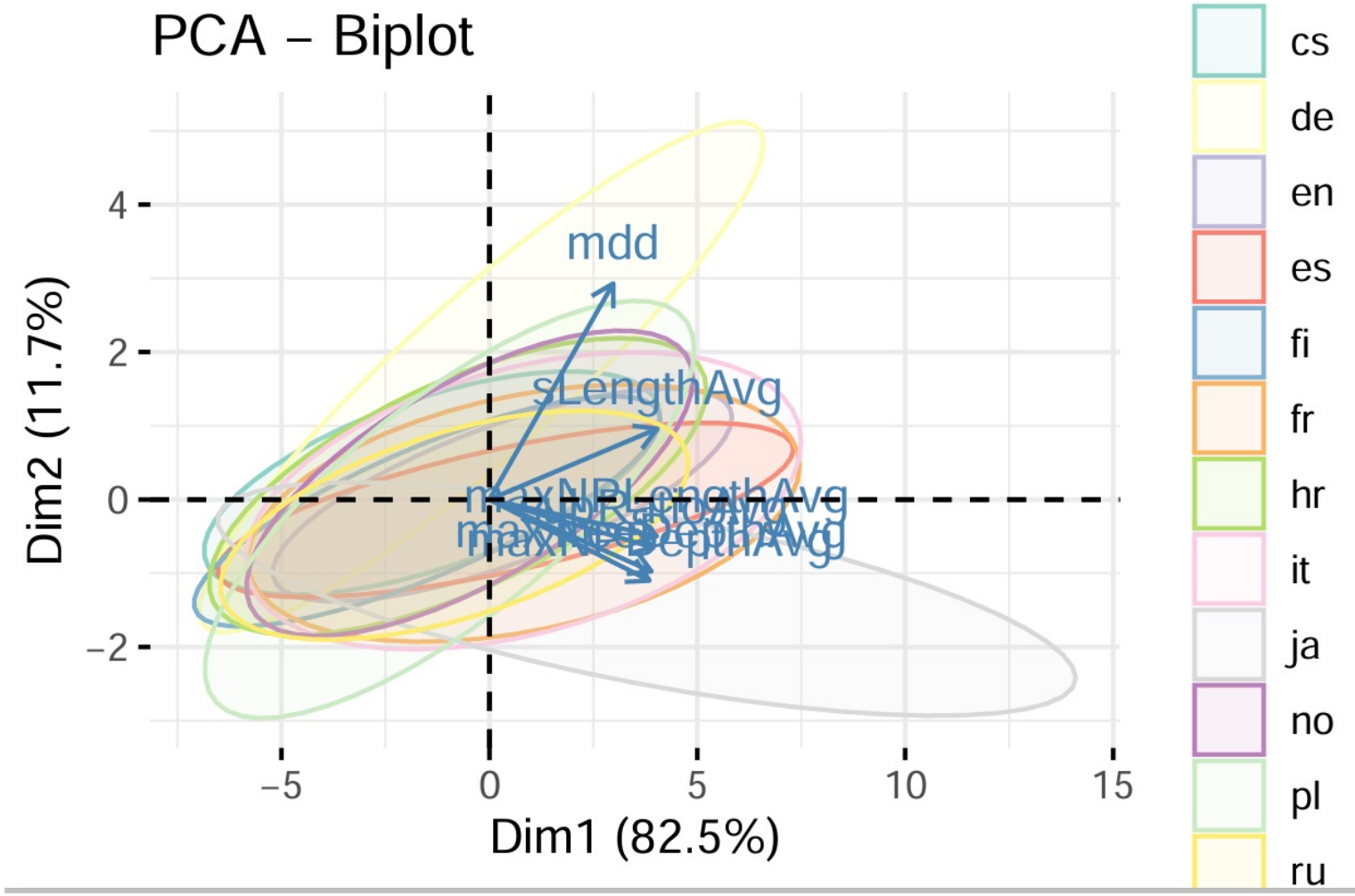


4.4.2 PCA – *principle component analysis* (échantillon 6 textes / 12 langues)

- sub.ratio, max.tree.depth, max.NPdepth et length **vs** s_length et mdd
- **ja** : spécificités de l'annotation
- **de** : mdd (mean dependency distance) élevé (SOV ?)

questions ouvertes :

- corpus large de *fiction* ?
- comparaison avec d'autres genres textuels (*non-fiction, Bible, subtitles*, etc.)
- large échantillon de langues, y compris non-indoeuropéennes ?



5. Conclusions – promesses et écueils

UD & **SCMs** & **corpus multilingue** = élargissement des possibilités de recherche :

- contrastive & traductologique
- typologique
- variationiste (*genres textuels*)
- applications pédagogiques
- analyse de style
- littérature de jeunesse (public cible)
- langues non-indoeuropéennes, *low-resourced languages*

Ecueils, questions ouvertes, tâches à terminer

- **finalisation de la version 16ud** (tous les genres textuels et mise en ligne) – EuroParl, Acquis; Bible; textes journalistiques, Subtitles
- **documentation** – corpus et SCMs (en anglais)
- **comparabilité** des (sous-)corpus
- « **noyau commun** » de textes limité
- **fiabilité de l'annotation** pour différentes langues (taille de treebanks)
- **corrélations** entres les différentes mesures de la complexité syntaxique


remercie
root
VERB
je
nsubj
PRON
vous
iobj
PRON
de
case
ADP
votre
det
DET
attention
oblig:arg
NOUN
.
punct
PUNCT

Bibliographie

Álvarez González, A., Zarina Estrada Fernández and a Claudine Chamoreau (2019). *Diverse scenarios of syntactic complexity*. Amsterdam: John Benjamins Publishing Company.

Arnold J., Wasow T., Losongco A. and Ginstrom R. (2000). Heaviness vs. Newness: The Effects of Structural Complexity and Discourse Status on Constituent Ordering. Language, vol. (17/1): 28-55.

Beaman K. (1984). Coordination and Subordination Revisited: Syntactic Complexity in Spoken and Written Narrative Discourse. In Tannen D. and Freedle R. (Eds), Coherence in Spoken and Written Discourse: 45-80.

Aleksandrs Berdicevskis, Çağrı Çöltekin, Katharina Ehret, Kilu von Prince, Daniel Ross, Bill Thompson, Chunxiao Yan, Vera Demberg, Gary Lupyan, Taraka Rama, and Christian Bentz (2018). Using Universal Dependencies in cross-linguistic complexity research. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*. Bruxelles: Association for Computational Linguistics, pp. 8–17.

Biber, Douglas & Gray, Bethany (2016). Grammatical complexity in academic English: Linguistic change in writing. Applied Linguistics, 37(6), 887–890.

Biber, Douglas, Larsson, Tove & Hancock, Gregory R. (2023). The linguistic organization of grammatical text complexity: comparing the empirical adequacy of theory-based models. Corpus Linguistics and Linguistic Theory. DOI: 10.1515/cllt-2023-0016

Brunato, Dominique, Andrea Cimino, Felice Dell'Orletta, Giulia Venturi, Simonetta Montemagni (2020). Profiling-UD: a Tool for Linguistic Profiling of Texts. in: Proceedings of the Twelfth Language Resources and Evaluation Conference, 2020, pp. 7145–7151.

Brunato, Dominique & Venturi, Guilia (202¨3). Why is this language complex? Cherry-pick the optimal set of features in multilingual treebanks. *Linguistics Vanguard* (1)9, pp. 59-72. https://doi.org/10.1515/lingvan-2021-0017

Canavese, P. and L. Mori (2021). Testing the hypothesis of "translation as a catalyst for plain legislation" on the syntactic level: A comparison of different varieties of legislative Italian. In: Castagnoli, S., S. Bernardini, A. Ferraresi, M. Miličević Petrović (eds) 2021. Using Corpora in Contrastive and Translation Studies Conference (6th Edition). Bertinoro (Italy), 9-11 September 2021.

Čermák, Petr et al. (2020). Complex Words, Causatives, Verbal Periphrases and the Gerund: Romance Languages Versus Czech (A Parallel Corpus-Based Study). Praha: Karolinum. http://hdl.handle.net/20.500.11956/117388
Chlumská, L. (2017). *Překladová čeština a její charakteristiky*. Praha: Nakladatelství Lidové noviny.
Cosme, Ch. (2006). Clause combining across languages. A corpus-based study of English-French translation shifts. *Languages in Contrast* 6(1), 71-108.
Croft, W., Nordquist, D., Looney, K., and Regan, M. (2017). Linguistic typology meets Universal Dependencies. In Dickinson, M., Hajič, J., Kübler, S., and Przepiórkowski, A., editors, *Proceedings of the 15th International Workshop on Treebanks and Linguistic Theories (TLT15)*, pages 63–75. Indiana University, Bloomington, Bloomington, IN, USA.
Cvrček, V. et al. (2020). *Registry v češtině*. Praha: NLN, 2020.
De Clercq, B. (2016) Le développement de la complexité syntaxique en français langue seconde : complexité structurelle et diversité. SHS Web of Conferences (27) 07006 (2016). DOI: 10.1051/shsconf/20162707006
Dell'Orletta F., Montemagni S., Venturi G. (2011). "*READ-IT: assessing readability of Italian texts with a view to text simplification".* In: SLPAT '11 – SLPAT '11 Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies (Edimburgo, UK, 30 Luglio 2011). Proceedings, pp. 73 – 83. Association for Computational Linguistics Stroudsburg, PA, USA, 2011.
Ebeling Oksefjell, S., Ebeling, J. (2020). Dialogue vs. narrative in fiction: A cross-linguistic comparison. *Languages in Contrast* 20(2), 2020, pp. 288-313.
Fabricius-Hansen, C. (1999). Information packaging and translation: aspects of translational sentence splitting (German–English/Norwegian). In Monika Doherty (ed.), *Sprachspezifische Aspekte der Informationsverteilung*. 175–214. Berlin: Akademie Verlag.
Ferreira F. (1991). Effects of Length and Syntactic Complexity on Initiation Times for Prepared Utterances. Journal of Memory and Language, vol. (30/2): 2110-2233.

Gerdes, Kim, Bruno Guillaume, Sylvain Kahane, and Guy Perrier. (2018). SUD or Surface-Syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 66–74, Brussels, Belgium. Association for Computational Linguistics.

Givón T. (1991). Markedness in grammar: distributional, communicative and cognitive correlates of syntactic structure. Studies in Language, vol. (15/2): 335-370.

Gruszczyński, W. & Ogrodniczuk, M. (2015). *Jasnopis czyli mierzenie zrozumiałości polskich tekstów uzitkowych.* Warszawa: SWPS.

Guillaume, Bruno, de Marneffe, Marie-Catherine, Perrier, Guy (2019). Conversion et améliorations de corpus du français annotés en Universal Dependencies. Revue TAL, ATALA, 2019, (2)60, pp. 71–95.

Himmelmann, N. P. & Eva Schultze-Berndt. (2006). *Secondary Predication and Adverbial Modification (The Typology of Depictives).* Oxford: OUP.

Hübler, Alfred W. (2007). Understanding complex systems. Complexity 12(5): 9–11.https://doi.org/10.1002/cplx.20178

Johansson, S. (2007). Seeing through Multilingual Corpora. On the Use of Corpora in Contrastive Studies. Amsterdam: John Benjamins.

Fabricius-Hansen, Cathrine (1998). Informational density and translation, with special reference to German – Norwegian – English. In Johansson, S. and Oksefjell, S., editors, Corpora and Cross-linguistic Research, page 197–234. Rodopi. DOI: 10.1163/9789004653665_012

Jagaiah, Thilagha, Olinghouse, Natalie G. & Kearns, Devin M. (2020). Syntactic complexity measures: variation by genre, grade-level, students' writing abilities, and writing quality. Read Writ 33, 2577–2638. DOI: 10.1007/s11145-020-10057-x

Kuboň, V. (2001). A Method for Analyzing Clause Complexity. Prague Bulletin of Mathematical Linguistics, vol. (75): 5-28

Levshina, N. (2019). "Token-based typology and word order entropy: A study based on Universal Dependencies " *Linguistic Typology*, vol. 23, no. 3, 2019, pp. 533-572. https://doi.org/10.1515/lingty-2019-0025

Levshina N. (2022). Corpus-based typology: applications, challenges and some solutions. *Linguistic Typology*, 25;26(1), pp. 129–160. doi: 10.1515/lingty-2020-0118.

Liu, Haitao. (2008). Dependency Distance as a Metric of Language Comprehension Difficulty. Journal of Cognitive Science. 9(2), 159–191. DOI:10.17791/jcs.2008.9.2.159.
Mačutek, J., Čech, R., Milička, J. (2019). Length of non-projective sentences: A pilot study using a Czech UD treebank. Proceedings of the First Workshop on Quantitative Syntax (Quasy, SyntaxFest 2019), Association for Computational Linguistics, Paris, 110-117.
Marneffe, M.-C. de ; Christopher Manning, Joakim Nivre, Daniel Zeman (2021). Universal Dependencies. In: *Computational Linguistics*, ISSN 1530-9312, vol. 47, no. 2, pp. 255-308.
Nádvorníková, O. (2013). Les gérondifs antéposés : quelles relations avec les contextes de gauche et de droite ? Verbum. 35(1–2), 161–174.
Nádvorníková, O. (2017a). « Pièges méthodologiques des corpus parallèles et comment les éviter », Corela [Online], HS-21 | 2017, Online since 20 February 2017, connection on 24 November 2020. URL: http://journals.openedition.org/corela/4810 ; DOI : https://doi.org/10.4000/corela.4810
Nádvorníková, O. (2017b). Le corpus multilingue InterCorp : nouveaux paradigmes de recherche en linguistique contrastive et en traductologie. Studii de Lingvistica. 7(déc.), 67–88. http://studiidelingvistica.uoradea.ro/docs/7-2017/pdf_uri/Nadvornikova.pdf
Nádvorníková, O. (2017c). Parallel Corpus in Translation Studies: Analysis of Shifts in the Segmentation of Sentences in the Czech-English-French Part of the InterCorp Parallel Corpus. In: *Language Use and Linguistic Structure*. Olomouc: Palacký University Olomouc, s. 445–461. http://olinco.upol.cz/wp-content/uploads/2017/06/olinco-2016-proceedings.pdf
Nádvorníková, O. (2020). The use of English, Czech and French punctuation marks in reference, parallel and comparable web corpora: a question of methodology. Linguistica Pragensia. 30(2), 30-50. ISSN 1805-9635, https://doi:10.14712/18059635.2020.1.2
Nádvorníková, O. (2021a). Le gérondif et le participe présent en français contemporain : Différence revisitée a la lumière de leur compatibilité avec les verbes de perception. In: Christelle Lacassain-Lagoin, Fabrice Marsac, Witold Ucherek, Katarína Chovancova & Monika Zázrivcová (éds). Sens (inter)dits. 2. Verbes et architectures syntatico-discursives. Paris: L'Harmattan. tome 2, p. 67–84.

Nádvorníková, O. (2021b). Contexts and Consequences of Sentence Splitting in Translation (English-French-Czech). *Research in Language 19(3),* pp. 229–250. https://czasopisma.uni.lodz.pl/research/issue/view/1045

Nádvorníková, O. (2020). Differences in the lexical variation of reporting verbs in French, English and Czech fiction and their impact on translation". Languages in Contrast 20:2, pp. 209-234. https://doi.org/10.1075/lic.00016.nad .

Nádvorníková, Olga (2021). Contexts and Consequences of Sentence Splitting in Translation (English-French-Czech). Research in Language, 19(3), pp. 229-250. https://doi.org/10.18778/1731-7533.19.3.01

Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. (2020). Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Rescher, N. (1998). Complexity: A Philosophical Overview. New Brunswick NJ: Transaction Rohdenburg G. (1996). Cognitive complexity and increased grammatical explicitness in English. Cognitive Linguistics, vol. 7, 149–182.

Rohdenburg G. (1996). Cognitive complexity and increased grammatical explicitness in English. Cognitive Linguistics, vol. (7): 149-182.

Schleppegrell M. (1992). Subordination and Linguistic Complexity. Discourse Processes: A Multidisciplinary Journal, vol. (15/1): 117-131.

Solfjeld, Kåre (1996). Sententiality and translation strategies German-Norwegian. Linguistics, 34, 567–590.

Szmrecsanyi, Benedikt (2004). On operationalizing syntactic complexity. In Le poids des mots. Proceedings of the 7th International Conference on Textual Data Statistical Analysis Louvain-la-Neuve, March 10–12, 2004, Vol. 2, Gérard Purnelle, Cédrick Fairon & Anne Dister (eds.), 1032–1039. Louvain-la-Neuve: Presses Universitaires de Louvain.

Xu, Jiajin & Jialei Li. (2021). A syntactic complexity analysis of translational English across genres. *Across Languages and Cultures* [online]. 2021-11-16, **22**(2), 214-232 [cit. 2023-09-01]. doi:10.1556/084.2021.00015

Yan, Chunxiao (2021). Complexité syntaxique et flux de dépendance : études quantitatives dans les treebanks universal dependencies. Linguistique. Université de Nanterre - Paris X, 2021. Français. ffNNT : 2021PA100127ff. fftel-03649621f
Zeman, Daniel (2018): The World of Tokens, Tags and Trees. Praha: ÚFAL. ISBN 978-80-88132-09-7.
Zeman, Daniel, Joakim Nivre, Mitchell Abrams, et al. (2020). Universal Dependencies 2.6, LINDAT/ CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. Available at: http://hdl.handle.net/11234/1- 3226. See also http://universaldependencies.org.

Internet :
https://universaldependencies.org/guidelines.html
Lindat UD Corpora (online search): https://lindat.mff.cuni.cz/services/kontext/corpora/corplist
Lindat UDPipe: https://lindat.mff.cuni.cz/services/udpipe/
Daniel Zeman: Universal Dependencies and the Slavic Languages. Warszawa, 19.11.2018.