Transcultural Journal of Humanities & Social Sciences

Print ISSN 4239-2636 Online ISSN 4247-2636

An Online Academic Journal of
Interdisciplinary & transcultural topics in Humanities & social sciences

TJHSS

Designed by Abeer Azmy& Omnia Raafat

BUC Press House

Volume 6 Issue (4)

July 2025

Transcultural Journal for Humanities and Social Sciences (TJHSS) is a journal committed to disseminate a new range of interdisciplinary and transcultural topics in Humanities and social sciences. It is an open access, peer reviewed and refereed journal, published by Badr University in Cairo, BUC, to provide original and updated knowledge platform of international scholars interested in multi-inter disciplinary researches in all languages and from the widest range of world cultures. It's an online academic journal that offers print on demand services.

TJHSS Aims and Objectives:
To promote interdisciplinary studies in the fields of Languages, Humanities and Social Sciences and provide a reliable academically trusted and approved venue of publishing Language and culture research.

- **Print ISSN** **2636-4239**
- **Online ISSN** **2636-4247**

Transcultural Journal for Humanities & Social Sciences (TJHSS) Editorial Board

EDITORIAL BOARD

Ain Shams University, Egypt	
Prof. Rasha Kamal Professor of Chinese Language Vice- Dean of the School of Linguistics & Translation Badr University in Cairo & Faculty of Alsun, Ain Shams University, Egypt	Email: rasha.kamal@buc.edu.eg
Prof. Sun Yixue President of The International School of Tongji University	Email: 98078@tongji.edu.cn
Prof. Wang Genming President of the Institute of Arab Studies Xi'an International Studies University	Email: genmingwang@xisu.cn
Prof. Zhang hua Dean of post graduate institute Beijing language university	Email: zhanghua@bluc.edu.cn
Prof. Belal Abdelhadi Expert of Arabic Chinese studies Lebanon university	Email: Babulhadi59@yahoo.fr
GERMAN LANGUAGE & LITERATURE	
Prof. Baher El Gohary Professor of German Language and Literature Ain Shams University, Cairo, Egypt	Email: baher.elgohary@yahoo.com
Prof. El Sayed Madbouly Professor of German Language and Literature Badr University in Cairo & Ain Shams University, Cairo, Egypt	Email: elsayed.madbouly@buc.ed u.eg
Prof. George Guntermann Professor of German Language and Literature Universität Trier, Germany	Email: GuntermannBonn@t-online.de
Prof. Herbert Zeman Professor of German Language and Literature Neuere deutsche Literatur Institut für Germanistik Universitätsring 1 1010 Wien	Email: herbert.zeman@univie.ac.at
Prof. Lamyaa Ziko Professor of German Language and Literature Badr University in Cairo & Menoufia University, Egypt	Email: lamiaa.abdelmohsen@buc.edu.eg
Prof. p`hil. Elke Montanari Professor of German Language and Literature University of Hildesheim, Germany	Email: montanar@unihildesheim.de, elke.montanari@unihildesheim.de
Prof. Renate Freudenberg-Findeisen Professor of German Language and Literature Universität Trier, Germany	Email: freufin@uni-trier.de
ITALIAN LANGUAGE & LITERATURE	
Prof. Giuseppe Cecere Professore associato di Lingua e letteratura araba	Email: giuseppe.cecere3@unibo.it

Università di Bologna Alma Mater Studiorum, Italy	
Prof. Lamiaa El Sherif Professor of Italian Language & Literature BUC, Cairo, Egypt	Email: lamia.elsherif@buc.edu.eg
Prof. Shereef Aboulmakarem Professor of Italian Language & Literature Minia University, Egypt	Email: sherif_makarem@y ahoo.com
SPANISH LANGUAGE & LITERATURE	
Prof. Carmen Cazorla Professor of Spanish Language & Literature Universidad Complutense de Madrid, Spain	Email: mccazorl@filol.ucm.es
Prof. Elena Gómez Professor of Spanish Language & Literature Universidad Europea de Madrid, Spain	Email :elena.gomez@universidadeuropea.es Universidad de Alicante, Spain spc@ua.es
Prof. Isabel Hernández Professor of Spanish Language & Literature Universidad Complutense de Madrid, Spain	Email: isabelhg@ucm.es
Prof. Manar Abd El Moez Professor of Spanish Language & Literature Dean of the Faculty of Alsun, Fayoum University, Egypt	Email: manar.moez@buc.edu.eg
Prof. Mohamed El-Madkouri Maataoui Professor of Spanish Language & Literature Universidad Autónoma de Madrid, Spain	Email: elmadkouri@uam.es
Prof. Salwa Mahmoud Ahmed Professor of Spanish Language & Literature Department of Spanish Language and Literature Faculty of Arts Helwan University Cairo, Egypt	Email: Serket@yahoo.com
HUMANITIES AND SOCIAL SCIENCES	
Prof. Ahmad Zayed Professor of Sociology Faculty of Arts, Cairo University, Egypt Ex-Dean of the School of Humanities & Social Sciences Badr University in Cairo	Email: ahmedabdallah@buc.edu.eg
Prof. Amina Mohamed Baiomy Professor of Sociology Faculty of Arts Fayoum University, Egypt	Email: ama24@fayoum.edu.eg
Prof. Galal Abou Zeid Professor of Arabic Literature Faculty of Alsun, Ain Shams University	Email: gaalswn@gmail.com
Prof. M. **Safeieddeen Kharbosh** Professor of Political Science Dean of the School of Political Science and International Relations Badr University in Cairo, Egypt	Email: muhammad.safeieddeen@ buc.edu.eg

Prof. Sami Mohamed Nassar
Professor of Pedagogy Dean of the School of Humanities & Social Sciences Badr University in Cairo Faculty of Graduate Studies for Education, Cairo University

Email: sami.nassar@buc.edu.eg

خطاب رئيس مجلس الأمناء

BUC
BADR UNIVERSITY IN CAIRO
جامعة بـــدر بالقـــاهرة

أ. د. حسين محمود حسين حمودة
رئيس تحرير
(TJHSS) Transcultural Journal of Humanities & Social Sciences

تحية طيبة وبعد ،،،

تتقدم إليكم جامعة بدر بالقاهرة بالشكر على ما تبذلونه من جهد مادي ومعنوي لإصدار المجلة، فتميزكم المشهود خير قدوة، ممتنين لعملكم الدؤوب وتفوقكم الباهر، ونتمنى لكم المزيد من النجاحات المستقبلية.

تحريرًا في يوم الأربعاء الموافق 2024/08/07.

رئيس مجلس الأمناء
د/ حسن القلا

TABLE OF CONTENTS

Évaluation des schémas *Universal Dependencies* et *Surface Syntactic UD* pour l'annotation syntaxique de constructions complexes en arabe et en français

Mohamed Galal
Faculté des Lettres, Université de Sohag
MoDyCo (UMR 7114), Université Paris Nanterre & CNRS

mohamed_mostafa1@art.sohag.edu.eg
moh_galal4@yahoo.fr

Résumé. Ce travail de recherche se penche sur l'analyse contrastive de l'annotation syntaxique de certaines constructions complexes en arabe et en français, réalisée selon le schéma *Universal Dependencies* (UD) et sa variante *Surface-Syntactic Universal Dependencies* (SUD). L'analyse portera sur trois catégories de constructions : *les constructions relatives*, *les constructions à verbe support* et *les constructions copulatives et avec un auxiliaire*. La recherche a un double but : d'une part, elle vise à explorer comment les schémas UD et SUD représentent ces constructions syntaxiques en arabe et en français. D'autre part, il s'agira d'évaluer la pertinence et la capacité de ces schémas à rendre compte des particularités propres à chaque langue. Les résultats révèlent que les annotations syntaxiques, bien que cohérentes pour les cas standards, divergent dans les constructions non canoniques, une incohérence constatée tant entre les corpus arborés arabes et français qu'au sein de chacun d'eux. Des solutions s'imposent : harmoniser les *treebanks* arabes et français existants par la standardisation et la correction, et optimiser les guides d'annotation pour les langues concernées.

Mots-clés : *Universal Dependencies* ; *Surface-Syntactic Universal Dependencies* ; annotation syntaxique ; relative ; verbe support ; copule ; auxiliaire ; approche contrastive arabe-français

1. Introduction

Dans cet article, nous nous intéressons à l'annotation syntaxique de quelques constructions qui posent des défis d'analyse en arabe standard moderne et en français, en mettant en lumière certains cas idiosyncrasiques de l'arabe. Nous examinerons ces constructions problématiques en nous appuyant sur des données annotées selon le schéma du projet *Universal Dependencies* (UD) (Nivre *et al.*, 2016, 2020 ; de Marneffe *et al.*, 2021) et sa variante *Surface-Syntactic Universal Dependencies* (SUD) (Gerdes *et al*. 2018, 2019a, b ; 2021 ; 2024). Le projet UD est un travail collaboratif qui rassemble une centaine d'équipes de recherche à travers le monde ayant pour objectif de développer un schéma d'annotation applicable à toutes les langues dans le but de promouvoir l'étude typologique de différentes langues, faciliter l'apprentissage des langues et favoriser le développement d'outils de traitement automatique des langues (TAL) multilingues. Actuellement (version 2.16), le projet UD propose plus de 300 *treebanks*[i] (corpus arborés) dans plus de 170 langues, annotés avec le même schéma d'annotation. Les données UD et SUD sont largement disponibles et accessibles librement[ii] à la communauté scientifique via les adresses suivantes : UD (https://universaldependencies.org), SUD (https://surfacesyntacticud.github.io).

L'application des mêmes critères syntaxiques à des langues de types très différents, comme l'arabe et le français, permet souvent de dégager les constructions idiosyncrasiques, c'est-à-dire spécifiques à une langue. L'étude de ces constructions est utile pour la vérification de considérations théoriques en linguistique et en traductologie et pour l'enrichissement de l'enseignement multilingue, car elle met en lumière les points où les deux langues divergent dans leur fonctionnement.

Trois constructions seront examinées dans le cadre de cette étude : *les constructions relatives*, *les constructions à verbe support* et *les constructions copulatives et avec un auxiliaire*. L'objectif sera double : d'une part, il s'agit d'explorer comment ces deux schémas, UD et SUD, représentent ces constructions en arabe et en français. D'autre part, il s'agit d'évaluer leur pertinence et leur capacité à représenter les spécificités propres à chaque langue.

Cette étude aborde donc les questions suivantes :

- Quelles sont les points de convergence et de divergence entre les schémas d'annotation UD et SUD dans leur traitement des constructions étudiées, tant au niveau intra-langue (comparaison entre *treebanks* d'une même langue) qu'inter-langue (analyse contrastive arabe-français) ?
- Jusqu'à quel point les schémas d'annotation UD et SUD sont-ils aptes à rendre compte fidèlement des particularités linguistiques de l'arabe et du français, particulièrement dans le traitement des phénomènes morphosyntaxiques complexes et des constructions non canoniques ?

La structure de l'article est la suivante : la section 2 établit une comparaison détaillée entre les deux schémas d'annotation, UD et SUD. Notre objectif y est de souligner les particularités de SUD et ses distinctions par rapport à UD. Dans la section 3, nous présentons les *treebanks* arabes

et français qui sont accessibles via les projets UD et SUD, offrant ainsi un aperçu des données utilisées. La section 4 constitue le cœur de notre analyse empirique. Nous y examinerons les annotations syntaxiques des trois constructions ciblées par cette recherche, en tirant des conclusions basées sur l'observation des données.

2. UD vs SUD : une comparaison de schémas d'annotation

Deux perspectives théoriques dominent la représentation de la structure syntaxique d'une phrase. D'un côté, *la structure syntagmatique* (Chomsky, 1957, 1965), largement adoptée depuis le milieu du XX^{e} siècle, privilégie un découpage de la phrase en constituants (ou syntagmes) de plus en plus grands. De l'autre, *la structure de dépendance* (Tesnière, 1959 ; Mel'čuk, 1988 ; Kahane, 2001 ; Kahane et Gerdes, 2022), qui a connu un regain d'intérêt à partir des années 1980, met en évidence une représentation hiérarchique entre les mots.

Les deux schémas UD et SUD reposent sur le modèle de la syntaxe de dépendance. Bien qu'ils partagent l'objectif commun d'une annotation standardisée des langues, ils s'appuient sur des paradigmes théoriques distincts. Cette section se concentrera sur les convergences et les divergences des deux schémas.

2.1 Les principes et les critères

Depuis son lancement officiel en 2014, le projet UD a pris de l'ampleur pour devenir un projet d'annotation considérable. Comme nous l'avons déjà mentionné plus haut, le projet UD constitue indéniablement la plus vaste collaboration internationale pour annoter des corpus arborés en plusieurs langues. Il offre un système d'étiquetage universel et homogène pour les relations syntaxiques, les parties du discours et les traits morphologiques, tout en étant suffisamment flexible pour s'adapter aux spécificités de chaque langue.

Le schéma UD tire son inspiration de plusieurs sources : *Stanford Dependencies* (de Marneffe *et al*., 2014) pour les relations de dépendance, *Google's universal part-of-speech tags* (Petrov *et al*., 2012) pour l'annotation morphologique et *Interset interlingua* (Zeman, 2008) pour l'harmonisation des traits morphosyntaxiques. Ce schéma est basé sur la syntaxe profonde, suivant la distinction entre la syntaxe de surface et la syntaxe profonde, telle que proposée par la *Théorie Sens-Texte* (Mel'čuk, 1988), et privilégie les mots lexicaux par rapport aux mots fonctionnels en les choisissant comme têtes[iii] syntaxiques. Il marque ainsi une rupture significative avec la tradition des grammaires de dépendances, une décision qui a fait l'objet de controverses (cf. Osborne et Gerdes, 2019).

Quant au schéma SUD, il se présente comme une alternative à UD. Il fournit un ensemble d'annotations syntaxiques axé sur la syntaxe de surface plutôt que sur la syntaxe profonde. Comme l'expliquent ses concepteurs (Gerdes *et al.*, 2019a), les relations syntaxiques y sont définies sur des bases distributionnelles et fonctionnelles. Contrairement à UD, SUD privilégie l'identification des têtes fonctionnelles. Ainsi, il ne désigne pas systématiquement les mots

lexicaux comme têtes syntaxiques. Il attribue plutôt ce rôle aux éléments fonctionnels tels que les prépositions, les conjonctions de subordination, les auxiliaires et les copules, un point que nous aborderons plus en détail par la suite.

La figure 1 illustre la différence d'annotation entre UD (en haut) et SUD (en bas) pour la même phrase arabe : *lā yumkin li=l-ğamīʿ ʾan yataraffaʿū ʿan al-ʾamr* ('Tout le monde ne peut pas s'élever au-dessus de ça'). Elle révèle que SUD considère généralement les mots fonctionnels, comme la conjonction de subordination *ʾan* ('que'), les prépositions *lī* ('pour') et *ʿan* ('de/à') comme les têtes syntaxiques des éléments qu'ils gouvernent, au lieu des mots lexicaux *l-ğamīʿ* ('tout le monde'), *yataraffaʿū* ('ils s'élèvent') et *al-ʾamr* ('la chose').

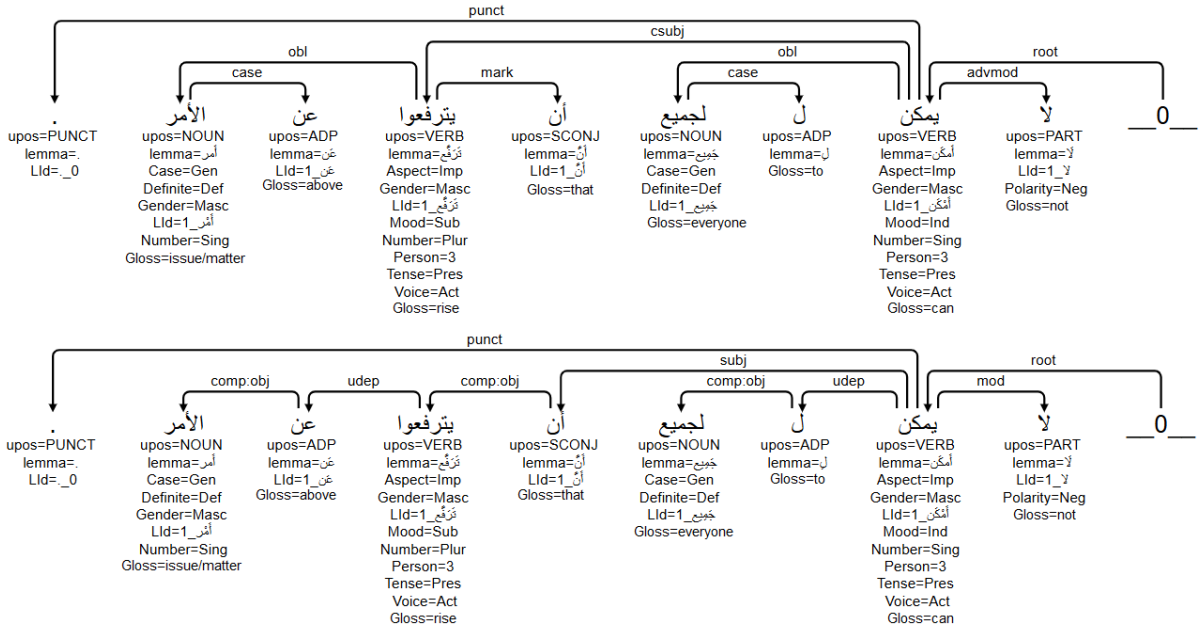


Figure 1 : l'annotation morphosyntaxique UD (en haut) et SUD (en bas) pour la même phrase de l'arabe ('*Tout le monde ne peut pas s'élever au-dessus de ça*'), tiré du *treebank* Arabic-PUD@2.15[iv]

2.2 Les relations syntaxiques

Comme mentionné précédemment, les relations syntaxiques utilisées dans UD sont une version révisée de celles initialement proposées pour la représentation des *Stanford Dependencies* (de Marneffe *et al*., 2014). La version 2 de UD compte 37 relations syntaxiques universelles. Dans le tableau ci-dessous issu du guide d'annotation UD, les rangées regroupent les relations selon leur rôle fonctionnel par rapport à leur tête et les colonnes classent les dépendants en fonction de leur catégorie structurelle. La partie inférieure de ce tableau inclut des relations qui ne sont pas considérées comme des dépendances au sens strict du terme.

	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
Coordination	**Headless**	**Loose**	**Special**	**Other**
conj	fixed	list	compound	punct
cc	flat	parataxis	orphan	root
			goeswith	dep
			reparandum	

Tableau 1 : les relations syntaxiques employées dans UD v.2, issu du guide d'annotation UD

Les langues peuvent, d'ailleurs, créer des relations plus spécifiques considérées comme des sous-types des relations universelles existantes. Ces sous-types sont identifiés par le type de base, suivi d'un deux-points et d'une description spécifique, par exemple, `nsubj:pass` pour un sujet passif.

Le schéma SUD adopte une approche plus concise de l'annotation syntaxique. Par exemple, pour traiter les compléments d'objet direct, SUD recourt à l'utilisation d'une relation unique : `comp:obj`, tandis que UD propose une classification plus fine avec les relations `obj` pour les objets nominaux, `ccomp` pour les objets propositionnels et `xcomp` pour les objet propositionnels sans sujet. SUD réduit ainsi les dix-sept relations de UD (`nsubj`, `csubj`, `obj`, `iobj`, `obl`, `xcomp`, `ccomp`, `amod`, `nmod`, `nummod`, `advmod`, `acl`, `advcl`, `aux`, `cop`, `case`, `mark`) à trois relations principales : `subj` pour le sujet, `comp` pour le complément, `mod` pour le modifieur. Ces relations peuvent ensuite être affinées avec des sous-relations spécifiques.

De plus, SUD introduit une relation générique, `udep` (underspecified dependency), qui peut englober à la fois `mod` et `comp`. Cette relation `udep` est utilisée pour rattacher les relations non spécifiées, c'est-à-dire lorsque les informations sont insuffisantes pour une classification précise. Cela peut arriver si les données proviennent de différents *treebanks* qui ne fournissent pas toutes les distinctions nécessaires, ou si la phrase elle-même est ambiguë, rendant une décision claire impossible (Gerdes *et al.*, 2019a).

Le schéma hiérarchique dans la figure 2 compare l'ensemble des relations utilisées dans UD et SUD : le cadre vert contient les relations qui sont communes aux deux schémas, c'est-à-dire celles que SUD utilise avec la même signification et le même objectif que UD et le cadre orange liste les relations spécifiques à UD qui n'ont pas été retenues dans le schéma SUD. Les aspects sémantiques de certaines constructions sont d'ailleurs traités comme des sous-spécifications des relations syntaxiques de base, afin de préserver la distinction entre les critères syntaxiques et sémantiques. SUD ajoute également des traits syntaxiques profonds sur les dépendances (boîtes bleues-claires dans le cadre bleu) : les informations sémantiques reliant deux unités lexicales sont indépendantes de la syntaxe et peuvent être ajoutées facultativement aux relations, séparées par le caractère arobase `@` (…`@x`, …`@agent`, …`@lvc`, …`@pass`, etc.).

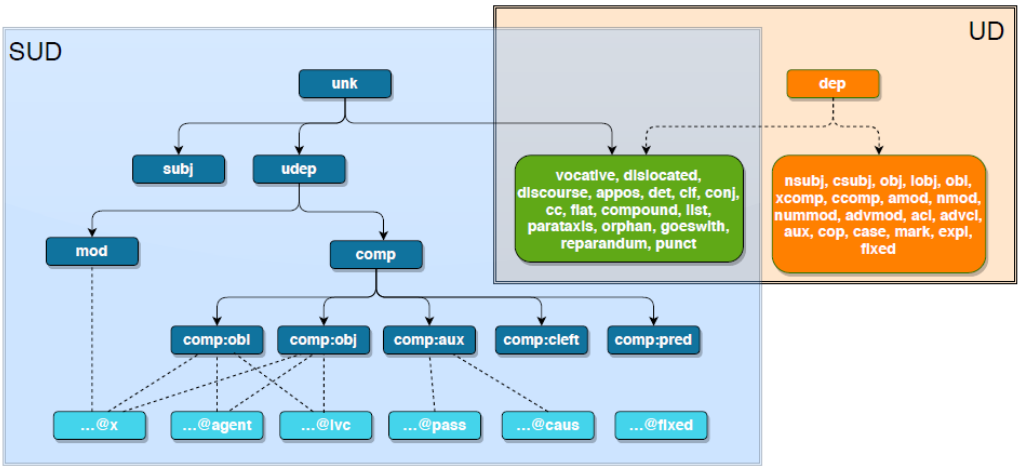


Figure 2 : le schéma des relations syntaxiques employées dans SUD issu de Gerdes *et al*. (2019a)

2.3 Les étiquettes POS et les traits morphosyntaxiques

Le schéma d'annotation UD propose un inventaire exhaustif de 17 catégories grammaticales (UPOS) et de 27 traits morphologiques, fournissant à la fois des informations lexicales (type d'adjectif, définitude) et morpho-syntaxiques (nombre, genre, temps, etc.). SUD se conforme aux normes de UD en adoptant les mêmes étiquettes de parties du discours et de traits morpho-syntaxiques universels. Par contre, SUD intègre un trait morphologique supplémentaire, `Shared`, afin de permettre une analyse plus fine des dépendants des conjonctions de coordination (cf. Gerdes *et al.*, 2024).

UPOS			
ADJ:	adjective	**PART:**	particle
ADP:	adposition	**PRON:**	pronoun
ADV:	adverb	**PROPN:**	proper noun
AUX:	auxiliary	**PUNCT:**	punctuation

CCONJ:	coordinating conjunction	**SCONJ:**	subordinating conjunction
DET:	determiner	**SYM:**	symbol
INTJ:	interjection	**VERB:**	verb
NOUN:	noun	**X:**	other
NUM:	numeral		

Traits morphologiques		
PronType	Gender	VerbForm
NumType	Animacy	Mood
Poss	NounClass	Tense
Reflex	Number	Aspect
Case	Voice	Abbr
Definite	Evident	Typo
Deixis	Polarity	Foreign
DeixisRef	Person	ExtPos
Degree	Polite	Clusivity
Shared (SUD)		

Tableau 2 : UPOS et traits morphologiques employés dans UD et SUD

Cette liste de traits morpho-syntaxiques qui paraît restreinte offre, néanmoins, une grande flexibilité en proposant plus de 180 valeurs potentielles. Par exemple, le trait `Gender` a pour valeurs : `Com` (common gender), `Fem` (feminine gender), `Masc` (masculine gender) et `Neut` (neuter gender). Toutes ces valeurs ne sont cependant pas utilisées dans toutes les langues.

2.4 Le format d'encodage

Le format d'encodage standard pour l'annotation des *treebanks* en UD et en SUD est le format CoNLL-U (.conllu) (https://universaldependencies.org/format.html). Ce format est une version révisée du CoNLL-X (Buchholz et Marsi, 2006). Dans ces fichiers tabulaires (figure 3), chaque *token*[v] d'une phrase est représenté sur une ligne distincte. Chaque ligne est divisée en dix colonnes, ce qui permet d'encoder diverses informations morphologiques et syntaxiques. Voici comment les informations sont réparties (voir aussi Kahane et Mazziotta, 2022) :

- Colonne 1 : identifiant du mot.
- Colonne 2 : le mot lui-même (le *token*).
- Colonne 3 : le lemme du mot (sa forme de base non fléchie).
- Colonne 4 : la partie du discours (UPOS).
- Colonne 6 : les traits morphosyntaxiques standard.

- Les colonnes 7 et 8 sont dédiées à l'arbre de dépendance :
  - Colonne 7 : l'identifiant du gouverneur du mot (le mot dont il dépend). Si le mot n'a pas de gouverneur (comme la racine de l'arbre), un zéro est indiqué.
  - Colonne 8 : la fonction syntaxique du mot par rapport à son gouverneur (par exemple, `det` pour un déterminant).
- Les colonnes 5 et 9 peuvent délibérément rester vides dans certains formats d'annotation, car elles sont spécifiquement réservées à l'utilisation par les analyseurs syntaxiques (*les parsers*) lors du traitement ultérieur des données.
- Colonne 10 : la colonne MISC, qui contient des informations additionnelles ou spécifiques non couvertes par les autres colonnes.

Les lignes vierges servent de séparateurs entre les phrases dans ces fichiers.

LId=1_لا	_	advmod	2	Polarity=Neg	RP	PART	لَا	لا	1
LId=1_أمكن	_	root	0	Aspect=Imp|Gender=Masc|Mood=Ind|Number=Sing|Person=3|Tense=Pres|Voice=Act	VBC	VERB	أمكن	يمكن	2
SpaceAfter=No	_	case	4	_	IN	ADP	لِ	ل	3
LId=1_جميع	_	obl	2	Case=Gen|Definite=Def	NN	NOUN	جَميع	لجميع	4
LId=1_أنْ	_	mark	6	_	IN	SCONJ	أنْ	أن	5
LId=1_ترفعوا	_	csubj	2	Aspect=Imp|Gender=Masc|Mood=Sub|Number=Plur|Person=3|Tense=Pres|Voice=Act	VBC	VERB	تَرفَع	يترفعوا	6
LId=1_عن	_	case	8	_	IN	ADP	عَن	عن	7
LId=1_أمر|SpaceAfter=No	_	obl	6	Case=Gen|Definite=Def|Gender=Masc|Number=Sing	NN	NOUN	أمر	الأمر	8
LId=،_0	_	Punct	2	_	,	PUNCT	,	,	9

Figure 3 : extrait de l'encodage au format CoNLL de la phrase dans la figure 1

2.5 Les outils de requête

Il existe de nombreux outils permettant d'interroger les *treebanks* UD (PML TQ, TEITOK, INESS). Dans le cadre de cette étude, nous aurons recours à l'outil Grew-match[vi] (Guillaume, 2021). La particularité de cette plateforme réside dans sa capacité à traiter et à visualiser simultanément les graphes des *treebanks* UD et SUD, ce qui le rend idéal pour mener des recherches comparatives/contrastives, comme celles menées dans la présente étude. La figure 4 illustre comment les couches d'annotations sont visualisées dans le cadre de Grew-match :

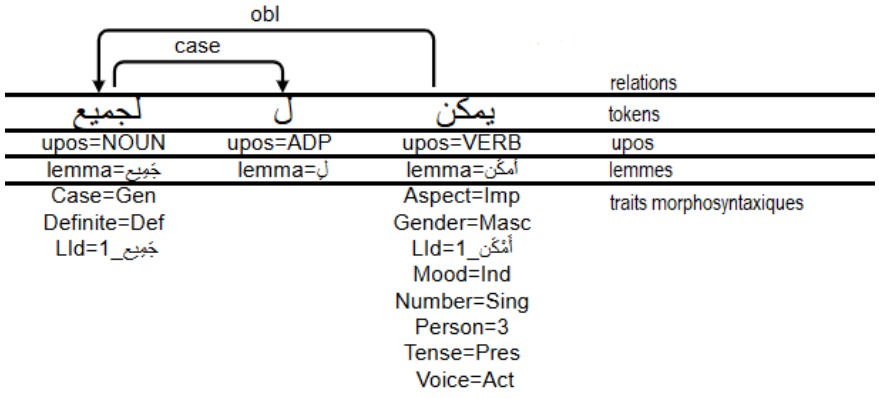


Figure 4 : la visualisation des couches d'annotations dans via l'outil Grew-match

Le système de requête Grew-match offre d'ailleurs des fonctionnalités avancées pour interroger les *treebanks*. Il permet notamment de regrouper les résultats selon différents critères. Par exemple, on peut lancer une requête qui nous permet de savoir la position du sujet par rapport au verbe en fonction de la fonction syntaxique du verbe. Ainsi à partir du langage de requête de Grew suivant:

```
pattern { e: G -> V; V -[subj]-> S; S[upos=NOUN] }
```

on peut déterminer si un sujet S précède ou non un verbe V (whether_1) en fonction de la fonction syntaxique de ce dernier (e.label). Les données de l'arabe[vii] (figure 5) révèlent, par exemple, que lorsque le verbe est la tête d'une subordonnée relative (`mod@relcl`), on observe un ordre inversé sujet-verbe dans un seul cas contre 62 cas où l'ordre est standard. En sélectionnant le chiffre 1, l'utilisateur peut accéder à l'exemple correspondant.

e.label / whether_1	510 root	159 comp:obj	63 mod@relcl	46 conj:coord	14 mod	8 parataxis	5 comp:obl	4 udep	1 subj
631 No	421	83	62	41	9	5	5	4	1
179 Yes	89	76	1	5	5	3			

Figure 5 : Requête Grew-match avec double clustering

2.6 La conversion

SUD et UD se présentent comme des schémas isomorphes (Gerdes *et al.* 2019b), permettant une conversion aisée des annotations d'un format à l'autre avec un minimum de pertes. Comme l'explique Gerdes *et al.* (2019b), ces pertes minimales sont généralement dues à deux facteurs principaux : soit des analyses non conformes aux les guides UD ou SUD, soit la nature plus plate de la structure UD qui ne préserve pas toutes les relations hiérarchiques entre les dépendants. En effet, la conversion des *treebanks* UD vers SUD est très bénéfique. Elle permet d'approfondir l'analyse comparative des structures syntaxiques et de favoriser ainsi des études typologiques plus rigoureuses.

2.7 Les *treebanks* disponible dans UD et SUD

Dans sa version 2.15 publiée le 15 novembre 2024, le projet UD recense 296 *treebanks* couvrant 168 langues. Il est déjà prévu d'ajouter 98 nouveaux *treebanks* couvrant 82 langues. Ce travail immense est le fruit d'une collaboration ouverte avec plus de 600 contributeurs. Quant au SUD, il compte 300 *treebanks* dans sa version 2.15 mise à disposition en novembre 2024 (d'après le site officiel : https://surfacesyntacticud.github.io/data). Ces *treebanks* sont classés comme suit : 5 *treebanks* développés au format mSUD[viii] (appelés « Native mSUD »)[ix], 9 *treebanks* au format SUD (appelés « Native SUD ») et 281 corpus sont automatiquement convertis en SUD à partir des données UD correspondantes (version 2.15)[x].

2.8 Autres caractéristiques

Bien que le schéma UD soit conçu pour être applicable à un large éventail de langues (notamment pour les textes standard), SUD a été mis au point dans le but d'analyser des textes non standards, notamment de l'oral (Gerdes *et al.* 2019b). Au niveau TAL, des études mettent en regard l'apprenabilité[xi] des deux schémas UD et SUD. Tuora, Przepiórkowski et Leczkowski (2021) ont testé l'hypothèse selon laquelle des critères syntaxiques améliorent la performance des analyseurs de dépendance. En comparant cinq analyseurs syntaxiques sur 21 *treebanks*, les résultats révèlent que SUD présente généralement une meilleure apprenabilité que UD, en fonction de l'analyseur syntaxique et le corpus considérés. Le choix de critères syntaxiques pour l'identification des têtes dans les arbres de dépendance, dans le cas de SUD, permet d'améliorer la performance des analyseurs syntaxiques de dépendance.

Le tableau ci-dessus récapitule les différentes caractéristiques de la comparaison des deux schémas d'annotation UD et SUD.

Caractéristique	**UD**	**SUD**
Principe	Syntaxe profonde	Syntaxe de surface
Critères	Sémantiques, fonctionnels	Distributionnels, fonctionnels
Relations	37 relations	17 relations
POS	14 UPOS	14 UPOS
Traits morpho-syntaxiques	27 (+ de 180 valeurs)	28 (+ de 180 valeurs)
Format	CoNLL-U	CoNLL-U
Conversion vers le format isomorphe	Possible	Possible
Outils de requête	Grew-match et autres	Grew-match
Treebanks	296 (novembre 2024)	300 (novembre 2024)
Type de texte ciblé	Textes standards	Textes non standards/oraux

Tableau 3 : récapitulatif de la comparaison des deux schémas d'annotation UD et SUD

3. Les *treebanks* arabes et français dans UD et SUD

Dans cette section, nous allons présenter les *treebanks* arabes et français qui seront étudiés dans le cadre de cette étude. Nous allons examiner en détail leurs origines ainsi que leurs caractéristiques spécifiques. Cette présentation des *treebanks* est fondamentale pour comprendre les données sur lesquelles notre recherche sera basée.

3.1 Les *treebanks* arabes

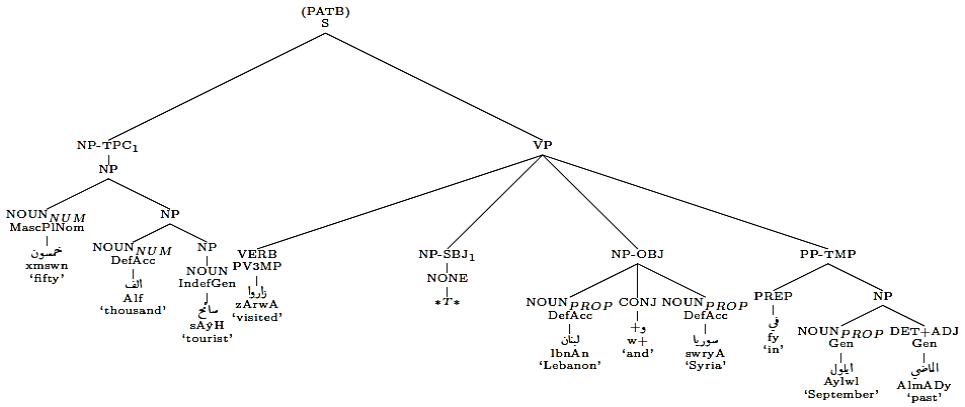
L'élaboration d'un corpus arboré pour l'arabe standard moderne n'est pas récente ; des initiatives existent depuis le début des années 2000. Parmi elles, *le projet Penn Arabic Treebank* (PATB) (Maamouri *et al.*, 2004), développé dès 2001 dans le cadre du projet *Linguistic Data Consortium* (LDC) à l'université de Pennsylvanie, est reconnu comme le premier *treebank* destiné à l'analyse syntaxique de l'arabe standard. Le PATB contient des textes de type majoritairement journalistique qui ont été publiés progressivement en quatre parties majeures. Ces ressources sont étiquetées en parties du discours (POS), glosées en anglais et annotées en arbres syntagmatiques, suivant le modèle prédicat-argument du *Penn Treebank* (cf. Marcus *et al.,* 1993). La figure 6, issue de Habash (2010, p.105-110), illustre l'arbre syntaxique du PATB pour la phrase donnée en (1) :

(1) خمسون ألف سائح زاروا لبنان وسوريا في أيلول الماضي

ẖamsūn	*alf*	*sāʾiḥ*	*zārū*	*lubnān*	*wa=sūriyā*	*fī*
cinquante	mille	touriste.	visiter.PST.3PL	Liban	COORD=Syrie	PREP
ʾaylūl	*al-māḍī*					
septembre	DEF-dernier					

'Cinquante mille touristes ont visité le Liban et la Syrie en septembre dernier'

La partie supérieure de la figure montre l'arbre tel qu'il était représenté initialement, tandis que la partie inférieure illustre le format sous lequel il est actuellement présenté. Une description détaillée du PATB est disponible dans le guide d'annotation (cf. Maamouri *et al.*, 2011).




```
(S (NP-TPC-1 (NOUN_NUM+NSUFF_MASC_PL_NOM xmswn)          fifty
             (NP (NOUN_NUM+CASE_DEF_ACC Alf)             thousand
                 (NP (NOUN+CASE_INDEF_GEN sA}H))))       tourist
   (VP (PV+PVSUFF_SUBJ:3MP zArwA)                        visited
       (NP-SBJ-1 (-NONE- *T*))
       (NP-OBJ (NOUN_PROP+CASE_DEF_ACC lbnAn)            Lebanon
               (CONJ w+)                                 and
               (NOUN_PROP+CASE_DEF_ACC swryA))           Syria
       (PP-TMP (PREP fy)                                 in
               (NP NOUN_PROP+CASE_DEF_GEN Aylwl          September
                   (DET+ADJ+CASE_DEF_GEN AlmADy)))))     past
```

Figure 6 : un exemple d'arbre du PATB issu de Habash (2010)

Si PATB est le premier *treebank* de l'arabe annoté en constituant, le premier *treebank* développé en dépendance pour l'arabe est celui du *Prague Arabic Dependency Treebank* (PADT) (Smrž *et al.*, 2002 ; Hajiˇc *et al.*, 2004 ; Smrž *et al.*, 2008). Ce projet a été initié à l'*Institut de Linguistique Formelle et Appliquée* de l'Université Charles de Prague. Les données du PADT proviennent en partie du PATB, qui ont été converties automatiquement vers une représentation basée sur les dépendances. Pour ses normes d'annotation, elles sont issues de celles du *Prague Dependency Treebank* (PDT) (Hajiˇc *et al.*, 2001), initialement conçu pour le tchèque, mais en les intégrant des modifications pour correspondre aux particularités de l'arabe. La figure 7, issue de Habash (2010, p.107), illustre la phrase en (1) représentée en PADT.

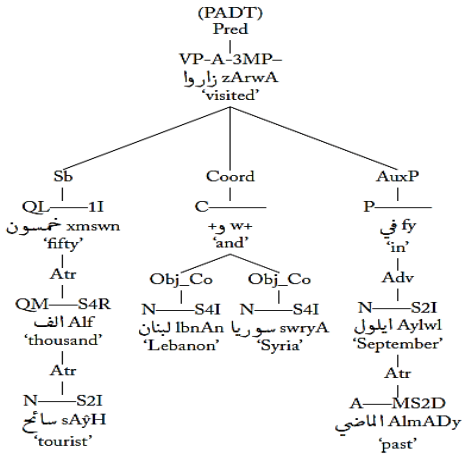
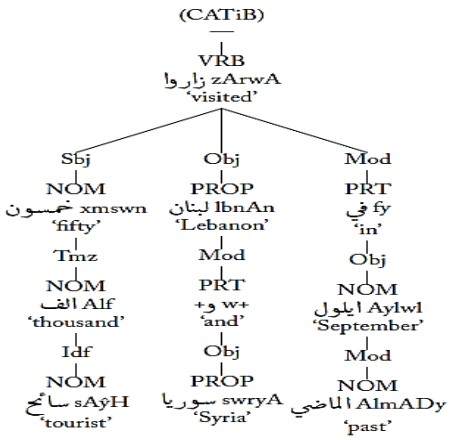


Figure 7 : un exemple d'arbre du PADT issu de Habash (2010)

Un autre *treebank* de dépendances pour l'arabe est le *Columbia Arabic Treebank* (CATiB) (Habash, Faraj et Roth, 2009). Ce corpus arboré s'inspire des principes de la grammaire traditionnelle arabe, ce qui permet une annotation plus aisée et rend cette ressource plus intuitive pour les arabophones (Habash, 2010, p. 108). Dans sa version initiale, le *treebank* CATiB emploie six étiquettes de parties du discours (POS) et huit relations syntaxiques. Il contient 273 000 *tokens* de textes journalistiques qui ont été annotés directement selon les normes CATiB, ainsi que l'ensemble des parties 1, 2 et 3 du PATB qui ont été automatiquement converties selon les normes CATiB (Taji, Habash, et Zeman, 2017, p. 167). La Figure 8, issue de Habash (2010, p.109-110), illustre l'arbre du CATiB (à gauche) de la phrase en (1) et son format actuel (à droite).



Columbia Arabic Treebank Example

1	xmswn	NOM	4	SBJ	fifty
2	Alf	NOM	1	IDF	thousand
3	sA}H	NOM	2	IDF	tourist
4	zArwA	VRB	0	---	visited
5	lbnAn	PROP	4	OBJ	Lebanon
6	w+	PRT	5	MOD	and
7	swryA	PROP	6	OBJ	Syria
8	fy	PRT	4	MOD	in
9	Aylwl	PROP	8	OBJ	September
10	AlmADy	NOM	9	MOD	past

Figure 8 : un exemple d'arbre du CATiB issu de Habash (2010)

Depuis la création de ces premiers *treebanks*, plusieurs autres ont vu le jour, tels que le *Quranic Arabic Corpus*[xii] (Dukes et Buckwalter, 2010), développé à l'Université de Leeds et dédié à l'analyse du texte arabe coranique. Il fournit des annotations en morphologie et en syntaxe de dépendance, se rapprochant le plus des descriptions de la grammaire traditionnelle arabe (figure 9a) ; *i3rab treebank* (Halabi *et al.*, 2021), qui adopte des structures de dépendances plus conformes à la théorie grammaticale traditionnelle arabe ; *Arabic Poetry Treebank* (ArPoT) (Al-Ghamdi *et al.*, 2021) spécifiquement conçu pour l'analyse de la poésie arabe classique ; et plus récemment le *Camel Treebank* (CamelTB)[xiii] (Habash *et al.*, 2022), qui comprend une riche diversité de textes arabes (poésie préislamique, romans, textes religieux et philosophiques, commentaires en ligne sur les réseaux sociaux, actualités, etc.). Ces textes ont été annotés automatiquement en morphologie et en syntaxe, puis corrigés manuellement. Les annotations suivent les normes de CATiB (figure 9b).

(a) (b)

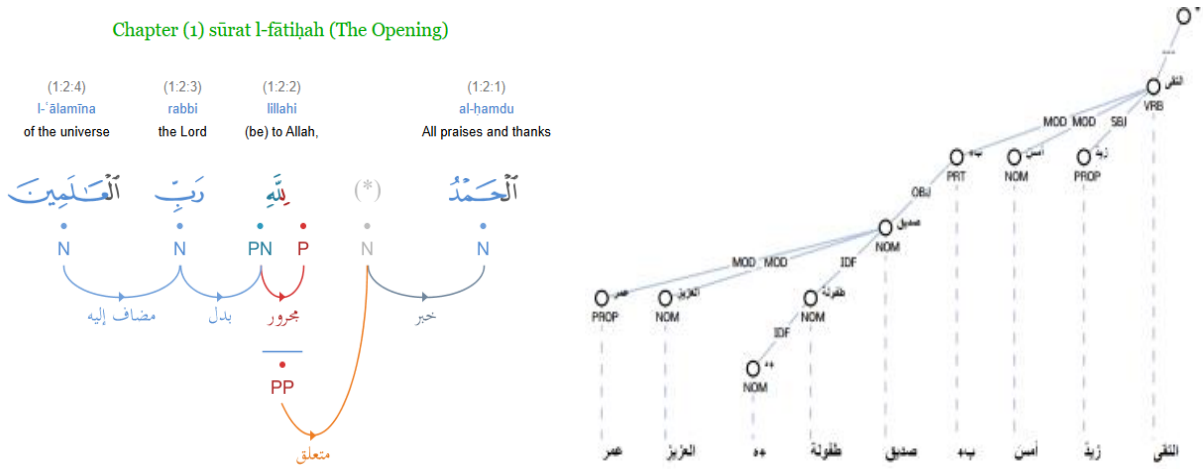


Figure 9 : un exemple de graphe de dépendance issue de Quranic Arabic Corpus et arbre de CamelTB

Avec l'élaboration du projet UD, dont l'objectif principal est de développer un système d'annotations uniformes, la voie a été ouverte pour l'élaboration des *treebanks* de diverses langues, tous construits sur le même schéma d'annotation. Cette uniformité facilite la comparaison de ces ressources entre elles et permet d'évaluer la pertinence et la cohérence des analyses syntaxiques adoptées. Dans ce contexte, un nombre de *treebanks* arabe ont été intégrés à UD.

Actuellement, la version UD 2.15 inclut trois *treebanks* de l'arabe standard moderne. Ces *treebanks* sont les suivants :

UD Arabic-PADT est le résultat de la conversion de *Prague Arabic dependency treebank* (PADT). Il fait partie de UD depuis sa version 1.2 et contient 7 664 phrases, 242056 *tokens* et 282 384 mots syntaxiques. Ses données proviennent de la collection *HamleDT* (HArmonized Multi-LanguagE Dependency Treebank), où 30 *treebanks* ont été harmonisées, d'abord selon les normes de Prague Dependencies, puis selon les normes de Stanford Dependencies (Rosa *et al.*, 2014), avant d'être finalement converties en UD (Taji, Habash, et Zeman, 2017). Les lemmes dans Arabic-PADT sont vocalisés avec les signes diacritiques arabes, tandis que les phrases normales sont écrites sans ces signes de vocalisation.

UD Arabic-NYUAD est construite à partir du *Penn Arabic Treebank* (PATB), parties 1, 2 et 3, grâce à une conversion des arbres en constituants du PATB vers des représentations de dépendances conformes au format CATiB. Ce processus est suivi de l'application de transformations morphologiques et syntaxiques (Taji, Habash et Zeman, 2017). Le *treebank* Arabic-NYUAD fait partie de UD depuis sa version 2.0 et contient 19 738 phrases, 629 295 *tokens* et 738 889 mots syntaxiques. La tokenisation suit le modèle de PATB : il segmente tous les clitiques sauf l'article défini *al-* ('le/la/les') (*ibid*.). Sur le plan syntaxique, CATiB, la base de Arabic-NYUAD, vise à reproduire une structure plus fidèle à l'analyse grammaticale arabe traditionnelle, en se concentrant sur la modélisation de l'attribution des cas grammaticaux. Il en résulte que les mots fonctionnels sont fréquemment les têtes de leurs structures de phrases. Comme mentionné précédemment (cf. §2.1), UD tend vers une représentation plus proche de la sémantique. Des ajustements significatifs ont été donc appliqués via des transformations syntaxiques lors de la conversion pour aligner Arabic-NYUAD avec les principes de UD (cf. Taji, Habash et Zeman, 2017). Le treebank UD Arabic-NYUAD, soumis à une licence LDC restrictive, ne fournit ni le texte des phrases ni les lemmes en accès libre. N'ayant pas eu accès à ces données, cette ressource ne pourra être exploitée dans la présente étude.

UD Arabic-PUD fait partie du projet Parallel Universal Dependencies (PUD), qui a pour but de développer des *treebanks* parallèles pour plusieurs langues, permettant la comparaison interlinguistique[xiv]. PUD offre 1 000 phrases alignées dans diverses langues (Zeman *et al.*, 2017). Ces phrases, issues de Wikipédia ou de textes journalistiques en cinq langues sources et traduites en 19 langues, ont été initialement annotées selon le schéma de McDonald *et al.* (2013) puis

converties au schéma UD par la communauté UD. UD Arabic-PUD contient 1000 phrases et 20 747 *tokens* et il fait partie de UD depuis sa version 2.1[xv].

***Treebank*s arabes**	**Sources**	**Version**	**Phrases**	**Tokens**	**Mots syntaxiques**
UD Arabic-PADT	Converti	v1.2	7 664	242 056	282 384
UD Arabic-NYUAD	Converti	v2.0	19 738	629 295	738 889
UD Arabic-PUD	Converti	v2.1	1 000	20 747	-

Tableau 4 : statistiques des treebanks arabes dans UD v2.15

Dans sa version actuelle (novembre 2024), le projet SUD comprend les trois *treebanks* de l'arabe standard moderne convertis à partir des *treebanks* UD (SUD_Arabic-NYUAD@2.15, SUD_Arabic-PADT@2.15 et SUD_Arabic-PUD@2.15). Il est prévu d'enrichir les données SUD avec trois nouveaux *treebanks* pour des dialectes arabes. Ces *treebanks* seront élaborés directement selon les normes SUD (native SUD) pour le darija marocain (cf. Dominique Caubet), l'arabe tunisien (cf. Aya Gherab *et al.*) et l'arabe égyptien (cf. Mohamed Galal). Ce travail s'inscrit dans le cadre du projet ANR Autogramm[xvi] (Kahane, 2022), dont l'objectif est d'élaborer à la fois des corpus arborisés et des grammaires descriptives pour les langues moins documentées. Une fois annotés en SUD, ces *treebanks* dialectaux seront ensuite convertis au format UD et intégrés au projet UD.

3.2 Les *treebanks* français

Les *treebanks* français dans UD présentent une certaine diversité en termes de taille et de niveau de langue (écrits vs. parlé), en comparaison avec ceux de l'arabe. Actuellement, le projet UD (v2.15) met à disposition huit *treebanks* pour le français :

(S)UD French-GSD fait partie du projet UD depuis la version UD v1. Il tire son origine de la version 2.0 du *Universal Dependency Treebank* de Google, publiée en 2013 (cf. Guillaume *et al.*, 2019) qui comprend des données provenant de diverses sources telles que la presse, la littérature, les textes officiels, les blogs et les pages Wikipédia. L'annotation initiale de ce corpus a été effectuée manuellement par deux groupes d'annotateurs différents dans le contexte d'un projet multilingue d'harmonisation (concernant l'anglais, l'allemand, le français, l'espagnol, le suédois et le coréen), ensuit, il a été converti au format UD en 2015. Depuis son intégration à UD, il est maintenu et enrichi séparément de sa version initiale du projet Google. Le corpus actuel contient 16 342 phrases, 389 362 *tokens* et 400 385 mots syntaxiques.

(S)UD French-Sequoia est le résultat d'une conversion automatique du treebank *SUD French-Sequoia*, lui-même issu de l'ancien corpus *French Sequoia* (Candito et Seddah, 2012) qui comporte des phrases et des textes provenant de quatre sources distinctes : l'Agence européenne des médicaments, Europarl, le journal régional *l'Est Républicain* et Wikipédia en français. Le

corpus *Sequoia* a été initialement annoté en structure de constituants, selon le schéma de *French Treebank* (FTB) (Abeillé et Barrier, 2004), puis converti automatiquement en dépendances (Candito et Seddah, 2012). UD French-Sequoia est intégré au projet UD depuis sa version 2.0 et contient 3 099 phrases, 685 93 *tokens* et 70 545 mots syntaxiques.

UD French-ParTUT est le résultat d'une conversion au format UD d'un corpus arboré parallèle multilingue (italien, français et anglais), *PARTUT* (Sanguinetti et Bosco, 2015). Ce dernier a été développé à l'Université de Turin et rassemble différents types de textes, notamment des discours, des documents légaux et des articles encyclopédiques de Wikipédia. Le corpus est disponible depuis la version UD 2.0 et contient 1 020 phrases, 27 638 *tokens* et 28 576 mots syntaxiques.

UD French-PUD est disponible depuis la version UD 2.1. Pareillement comme UD Arabic-PUD, le corpus a été intégré au projet *Parallel Universal Dependencies* (PUD). L'annotation du français dans ce corpus se distingue parfois des autres corpus français en restant plus proche de l'anglais ; par exemple, les possessifs y sont traités comme des pronoms, avec la relation `nmod:poss` au lieu de déterminants avec la relation `det` comme dans les autres corpus français (cf. Guillaume, de Marneffe et Perrier, 2019). Ce corpus contient 1 000 phrases, 24 131 *tokens* et 24 726 mots syntaxiques.

(S)UD French-Rhapsodie (Gerdes et Kahane, 2017), nommé *UD French-Spoken* jusqu'à la version 2.8, résulte de la conversion automatique (avec des corrections manuelles) du *treebank* du projet Rhapsodie (Lacheret *et al.* 2014). La particularité de *Rhapsodie* est qu'il contient des annotations à la fois en prosodie et en syntaxe de transcriptions de données de langue orale. Le modèle d'annotation *Rhapsodie* s'inspire de la syntaxe de dépendance (Tesnière 1959 ; Mel'čuk 1988 ; Kahane, 2001) et des études sur la syntaxe de l'oral (Blanche-Benveniste 1990, 2010 ; Deulofeu 2003). Il est disponible dans UD depuis la version 2.2 et comprend 3 209 phrases, 43 699 *tokens* et 44 242 mots syntaxiques.

(S)UD French-FQB, (French Question Bank) a été intégré au projet UD depuis sa version 2.4. Il a été constitué à partir d'une conversion d'un corpus entièrement constitué de questions, *la French QuestionBank v1* (Seddah et Candito, 2016), issues de diverses sources (traductions d'ensembles de test, des questions fréquemment posées de sites web officiels, des questions tirées des forums de cuisine, etc.). Ce corpus est basé sur le schéma d'annotation du *French Treebank* (FTB) (Abeillé *et al.*, 2003) en adoptant modification pour tenir compte des caractéristiques spécifiques des syntagmes interrogatifs. Le corpus converti en UD comprend 22 89 phrases, 2 3347 *tokens* et 23 899 mots syntaxiques.

(S)UD French-ParisStories est un corpus de français parlé (Kahane *et al.*, 2021). Il contient des monologues et des dialogues de locuteurs vivant en région parisienne. Le corpus a été rassemblé et transcrit par des étudiants en linguistique de l'Université Sorbonne Nouvelle – Paris 3, puis révisé par des étudiants du *Master pluriTAL* (Inalco, Paris Nanterre, Sorbonne Nouvelle) entre

2017 et 2021. Les données ont été initialement annotées en morpho-syntaxique selon le schéma SUD, puis ont été automatiquement converties au format UD grâce au logiciel Grew. Le corpus fait partie de UD depuis la publication de la version UD v2.9 et contient 2 776 phrases, 42 257 *tokens* et 42 789 mots syntaxiques.

(S)UD French-ALTS (Automated Sixteenth-century corpus) est un *treebank* dédié au français juridique du XVI[e] siècle (Ziane et Romanova 2024). Il contient actuellement un unique texte, à savoir des procès-verbaux extraits du registre Crime I du Greffe de Guernesey. Ce document a été transcrit directement depuis le manuscrit original et annoté manuellement en POS, lemmes et fonctions syntaxiques, dans le cadre du projet franco-allemand *MICLE* (2021-2024)[xvii]. Le texte présente des traits et des formes dialectaux normands. Ce corpus arboré est intégré au projet UD depuis sa version 2.16 et contient 1 269 phrases, 43 088 *tokens* et 43 832 mots syntaxiques [xviii].

***Treebank*s arabes**	**Sources**	**Version**	**Phrases**	**Tokens**	**Mots syntaxiques**
UD French-GSD	SUD-Natif	v1	16 342	389 362	400 385
UD French-Sequoia	SUD-Natif	v2.0	3 099	68 593	70 545
UD French-ParTUT	Converti	v2.0	1 020	27 638	28 576
UD French-PUD	Converti	v2.1	1 000	24 131	24 726
UD French-Rhapsodie	SUD-Natif	v.2.2	3 209	43 699	44 242
UD French-FQB	SUD-Natif	v2.4	2 289	23 347	23 899
UD French-ParisStories	SUD-Natif	v2.9	2 776	42 257	42 789
UD French-ALTS	Converti	v2.16	1 269	43 088	43 832

Tableau 5 : statistiques des treebanks français dans UD v2.15

Ces *treebanks* français mentionnés, comme l'ensemble des *treebanks* UD, sont également disponibles au format SUD via une conversion, selon la dernière mise à jour en novembre 2024.

4. L'étude de l'annotation syntaxique des constructions en arabe et en français

Dans cette section, nous allons nous pencher sur l'annotation syntaxique des constructions complexes en arabe et en français. Trois constructions seront examinées : *les constructions relatives*, *les constructions à verbe support* et *les constructions copulatives et avec un auxiliaire*.

4.1 Les constructions relatives

Avant d'aborder les choix d'annotation des constructions relatives faits par UD et SUD, il convient, d'abord, de souligner quelques aspects comparatifs de ces constructions entre l'arabe et

le français. Bien que ces constructions partagent la fonction essentielle, celle de modifieurs adnominaux, leurs structures et leurs fonctionnements divergent nettement (cf. Lafhej, 2007 ; Youssef, 2012).

i) Les pronoms relatifs en arabe s'accordent en genre et en nombre avec leur antécédent[xix], contrairement au français où l'accord du pronom relatif en genre et en nombre n'est pas systématique :

(2) a. الولد الذي يبتسم لطيف

al-wald	***allḏī***	*yabtasim*	*laṭīf*
DEF-garçon.M.SG	**REL.M.SG**	sourir.PRS.M.3SG	gentil.M.SG

'Le garçon qui sourit est gentil'

b. البنت التي تبتسم لطيفة

al-bint	***allatī***	*tabtasim*	*laṭīfah*
DEF-fille.F.SG	**REL.F.SG**	sourir.PRS.F.3SG	gentil.F.SG

'La fille qui sourit est gentille'

ii) En arabe, une reprise pronominale se référant à l'antécédent est fréquente, même obligatoire dans certains contextes ; ce pronom est souvent cliticisé au verbe ou à une préposition à l'intérieur de la proposition relative. En français, cette reprise pronominale est inhabituelle et, généralement, évitée :

(3) الولد الذي رأيته كان مبتسمًا

al-walad	*allḏī*	*qarʾatu=**hu***	*kāna*	*mubtasiman*
DEF-garçon	REL	voir.PST.1SG=**PRO**	être.PST.3SG	souriant

'Le garçon que j'ai vu était souriant'

iii) L'arabe possède une structure relative particulière sans la réalisation explicite du pronom relatif, surtout avec des antécédents indéfinis, la relation étant alors souvent marquée par un pronom de reprise. En français, l'introduction d'une proposition relative par un pronom relatif est quasi systématique :

(4) أشاهد فيلمًا رأيته من قبل

ʾušāhidu	***film-a-n***	*raʾaytu=**hu***	*min*	*qabl*
regarder.PRES.1SG	**film.SG-ACC-INDEF**	voir.PST.1SG=**PRO**	PREP	avant

'Je regarde un film que j'ai déjà vu'

iv) L'arabe se distingue par la présence de pronoms relatifs dits intégratifs (cf. Le Goffic, 2002), sans antécédent nominal explicite : *mā* pour un inanimé et *man* pour un animé dont l'équivalent en français sont les constructions composées 'ce que/qui' et 'celui que/qui' respectivement. Notons que les pronoms relatifs prototypiques *allḏī*, *allatī*, etc. peuvent aussi introduire une relative sans antécédent (cf. Youssef, 2012). Ces éléments fonctionnent comme translatifs du

verbe en substantif. Une proposition introduite par un pronom intégratif ne requiert pas toujours un pronom de reprise ; dans certains contextes, la construction sans ce pronom est parfois la tournure privilégiée (El Kassas, 2005, p. 65) :

(5) a. أفهمُ جيدًا ما تقوله

ʾafhamu	*ǧayyidan*	***mā***	*taqūlu=**hu***
comprendre.PRS.1SG	bien	**REL**	dire.PRS.2SG=**PRO**

‘Je comprends bien ce que tu dis’

b. أفهمُ جيدًا ما تقول

ʾafhamu	*ǧayyidan*	***mā***	*taqūlu*
comprendre.PRS.1SG	bien	**REL**	dire.PRS.2SG

‘Je comprends bien ce que tu dis’

c. قرأت الذي كتبه

qarʾatu	***allḏī***	*kataba=hu*
lire.PST.1SG	**REL**	écrir.PST.3SG=PRO

‘J’ai lu ce qu’il a écrit’

De nombreux linguistes (voir pour le français Tesnière, 1959 ; Kahane 2002, pour l’anglais Sag, 1997) ont souligné que les pronoms relatifs fonctionnent également comme des subordonnants. Comme le montrent, Kahane et Gerdes (2020, p. 76), l’exemple de la relative *qui dort* dans *la fille qui dort est une amie* illustre le fait que *qui* ne se comporte pas comme un simple dépendant du verbe *dort*. En effet, on ne peut pas remplacer *qui dort* par *Marie dort* sans rendre la phrase agrammaticale, ce qui montre que *qui* modifie la distribution de la construction verbale.

Une solution envisagée pour rendre compte de ce phénomène est d’attribuer deux positions syntaxiques aux pronoms relatifs : une fonction translative, qui en fait la tête de la proposition relative, et une fonction pronominale, qui leur assure une place à l’intérieur de cette proposition. L’analyse en double position syntaxique est exemplifiée par la représentation SUD du segment de la phrase anglaise *She picked up the pans in which she’d made the potatoes and maple glaze* (figure 10), issue de Gerdes *et al*. (2024, p. 618). La position translative est désignée par REL.

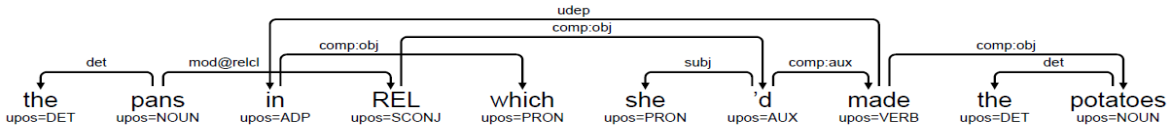
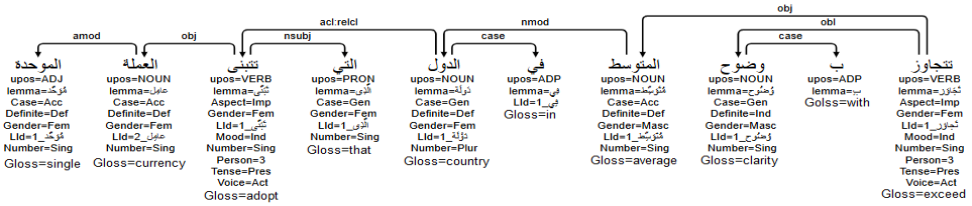


Figure 10 : analyse en double position de la construction relative en anglais, issue de Gerdes *et al*. (2024)

Cependant de nombreux *treebanks* (voir par exemple le projet Orféo[xx], Debaisieux, Benzitoun et Deulofeu, 2016) optent pour ne pas considérer les pronoms relatifs dans leur rôle de subordonnant et les traiter comme de simples pronoms afin de simplifier la structure syntaxique et de la maintenir sous forme d’arbre de dépendance. UD et SUD suivent également la même approche.

Dans le cadre de UD et de SUD[xxi], les relatives reçoivent un traitement uniforme soit en arabe, soit en français (figures 11 et 12), adoptant une fonction pronominale pour les pronoms relatifs qui leur confère une place au sein de la proposition relative, à l'instar de tout pronom ordinaire. La sous-relation UD `acl:relcl` (adnominal clause : relative clause)[xxii] rattache la tête de la proposition relative à son antécédent et le pronom relatif dépend du verbe de la proposition relative, souvent en tant que sujet `nsubj` ou objet `obj`.



'[…] dépasse nettement la moyenne des pays qui adoptent la monnaie unique'[xxiii]

Figure 11 : UD_Arabic-PUD@2.15[xxiv]

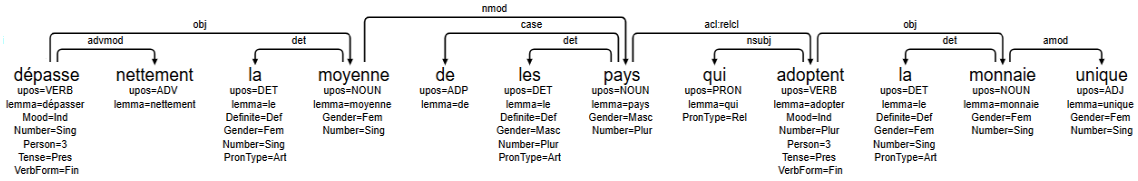
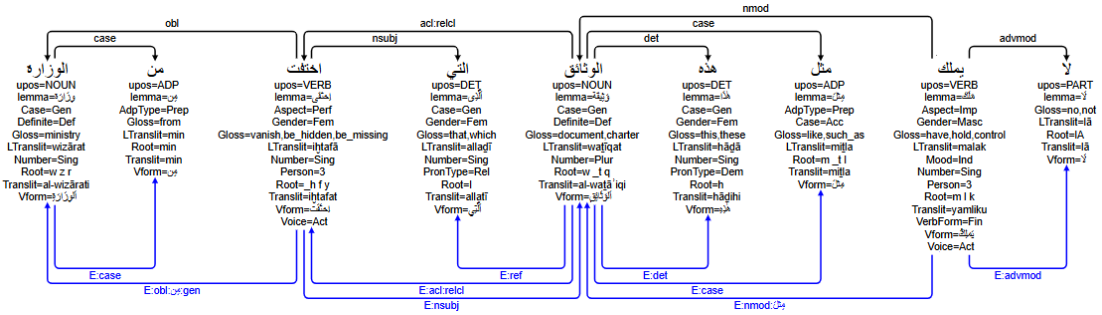


Figure 12 : UD_French-PUD@2.15

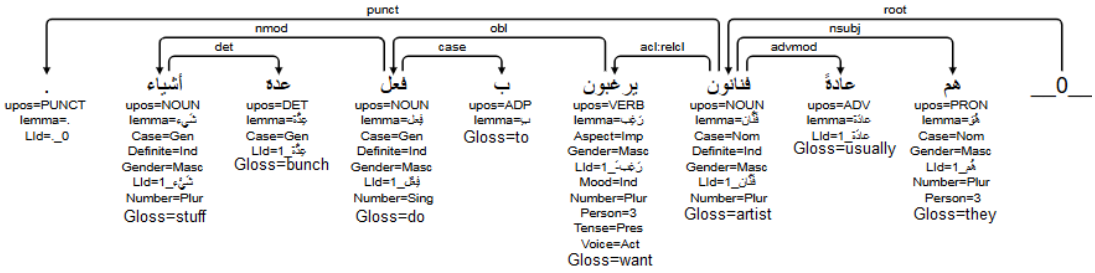
Certains *treebanks* UD, y compris Arabic-PADT, sont également fournis avec une strate d'annotation additionnelle (Enhanced Dependencies[xxv]) qui va au-delà de la structure de dépendances de base. Leur objectif est de clarifier les liens syntaxiques et sémantiques qui sont sous-entendus ou non spécifiés dans l'annotation UD de base. Dans les annotations enrichies de UD Arabic-PADT (en bleu) (figure 13), le pronom relatif[xxvi] se rattache à son antécédent avec la relation spéciale `ref`, utilisée exclusivement dans les annotations enrichies. L'antécédent, quant à lui, est annoté comme dépendant du prédicat principal de la proposition relative[xxvii], en lui attribuant la fonction syntaxique qu'il aurait eue dans sa position initiale. La sous-relation `acl:relcl` est toujours entre le prédicat de la relative et son antécédent.



'Il ne possède pas de tels documents qui ont disparu du ministère'

Figure 13 : UD_Arabic-PADT@2.15

En l'absence d'un pronom relatif explicite en arabe, la relation `acl:relcl` rattache toujours l'antécédent au verbe principal de la proposition relative, sans qu'aucune relation ne rattache ce verbe à un pronom (figure 14 vs 15).



'Ce sont généralement des artistes {qui} veulent faire un tas de trucs'

Figure 14 : UD_Arabic-PUD@2.15

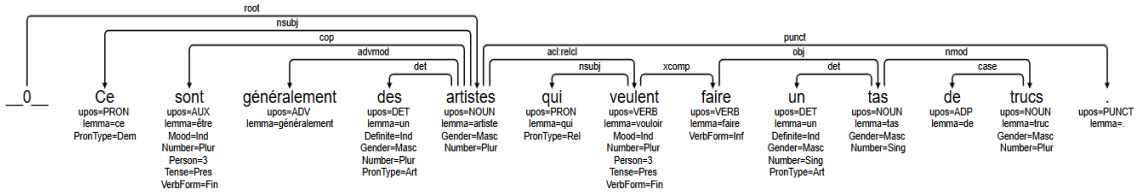
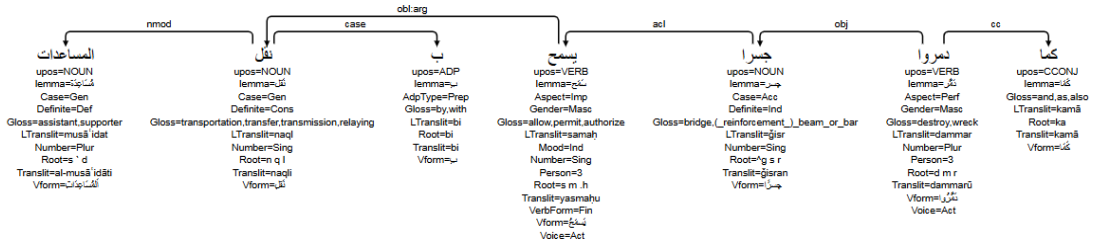


Figure 15 : UD_French-PUD@2.15

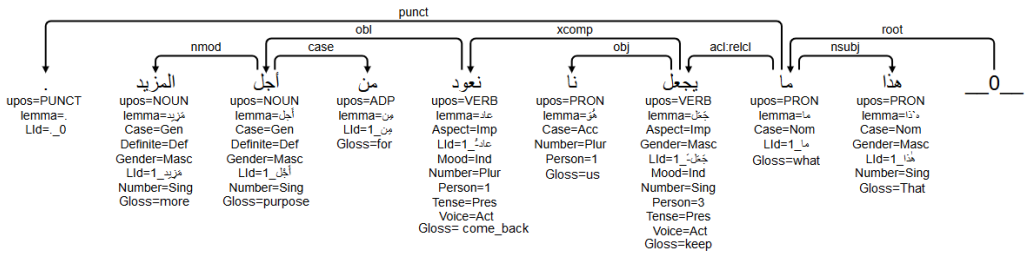
Le fait que la deuxième proposition modifie le nom, mais n'est pas une proposition relative introduite par un pronom relatif prototypique, UD Arabic-PADT choisit d'étiqueter le prédicat de la deuxième proposition avec la relation plus générale `acl` (figure 16).



'Ils ont également détruit un pont qui permettait le passage de l'aide

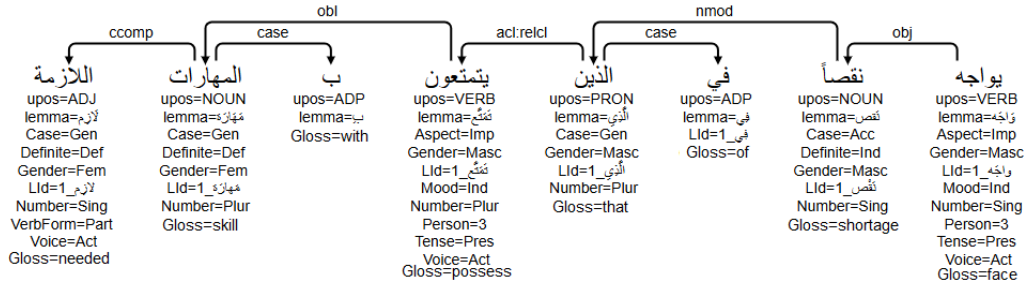
Figure 16 : UD_Arabic-PADT@2.15

Les *treebanks* UD arabes choisissent d'annoter les pronoms relatifs sans antécédent *mā*, *man*, *allḏī*, *allḏīna* etc. comme la tête de la subordonnée relative (figures 17, 18 et 19). Notons que ces éléments ont été catégorisés comme DET dans UD Arabic-PADT et gouvernent le prédicat de la subordonnée par la relation `acl`.



‘C’est ce qui nous incite à revenir pour davantage […]’

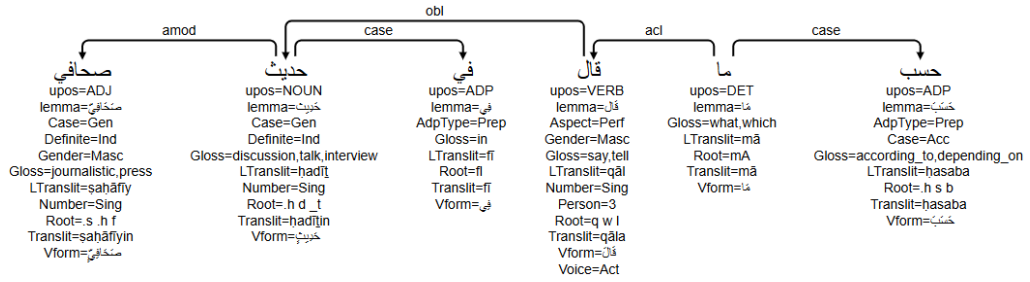
Figure 17 : UD_Arabic-PUD@2.15



‘Il manque de personnes ayant les compétences nécessaires’

Litt. Il fait face à une pénurie de {ceux} qui possèdent les compétences nécessaires.

Figure 18 : UD_Arabic-PUD@2.15



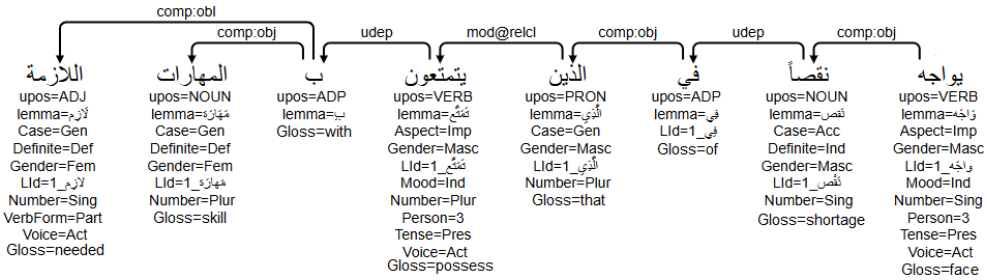
‘selon ce qu’il a dit dans une interview de presse […]’

Figure 19 : UD_Arabic-PADT@2.15

Comme le soulignent Gerdes *et al.* (2024) concernant les constructions anglaises introduites par un mot *wh*-word, l’analyse qui considère les pronoms relatifs intégratifs comme la tête de la subordonnée relative semble problématique. Ceci est d’autant plus vrai pour les pronoms relatifs prototypiques arabes tels que *allḏī*, *allatī*, etc., et ce, pour plusieurs raisons : i) cette analyse s’éloigne de l’approche pronominale qui est habituellement retenue pour les propositions relatives standard, ii) elle rend la fonction pronominale de ces éléments intégratifs difficilement perceptible, voire complètement masquée, et iii) l’extension `:relcl` ne correspond plus à la relation entre la proposition relative et son gouverneur (car les pronoms relatifs font partie de la proposition relative) ni à la fonction translative.

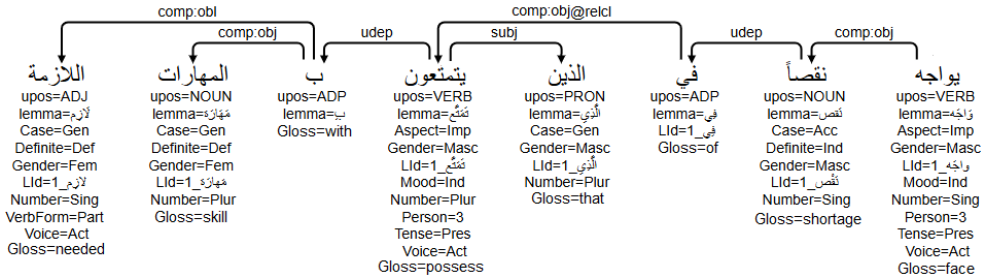
L’analyse proposée sera donc de retenir l’analyse pronominale pour ces constructions. En comparant l’analyse actuelle de SUD (figure 20), obtenue par conversion depuis la représentation UD, avec l’analyse proposée (figure 21), on constate que le verbe principal

devient la tête de la proposition relative dans l'analyse proposée, gouvernant le pronom *allḏīna* via la relation `subj`. De plus, la fonction pronominale de ces éléments est constamment signalée par l'ajout du trait `@relcl` sur la relation liant la proposition relative à son gouverneur. Ce trait additionnel `@relcl`, qui accompagne les relations, ne se limite donc pas à indiquer les cas où la proposition relative modifie directement un nom (son antécédent), comme c'est le cas avec les propositions relatives classiques. L'un des avantages de cette analyse, comme le montrent Gerdes *et al.* 2024, réside dans sa facilité de conversion vers une analyse considérant ces éléments comme des subordonnants, L'inverse, en revanche, n'est pas possible.



'Il manque de personnes ayant les compétences nécessaires'

Figure 20 : SUD_Arabic-PUD@2.15



'Il manque de personnes ayant les compétences nécessaires'

Figure 21 : l'analyse proposée dans SUD

4.2 Les constructions à verbe support

Les constructions à verbe support sont des combinaisons de deux constituants : un verbe, souvent sémantiquement pauvre, et un nom prédicatif (ou éventuellement un adjectif, un verbe ou un syntagme prépositionnel) (6a, b). Ces deux constituants forment une unité prédicative complexe. Différentes langues semblent partager le phénomène des verbes supports (voir Ahnaiba, 2006 ; Ibrahim, 2008 ; Madkhali, 2024 pour l'arabe ; Gross, 1998 ; Mel'čuk, 2004 ; Danlos, 2010 pour le français) :

(6) a. اتخذ قرارًا مفاجئ [V + N]

ittaẖaḏa	*qarār-a-n*	*mufāǧiʾ*
prendre.PST.3SG	décision-ACC-INDF	soudaine

'Il a pris une décision soudaine'

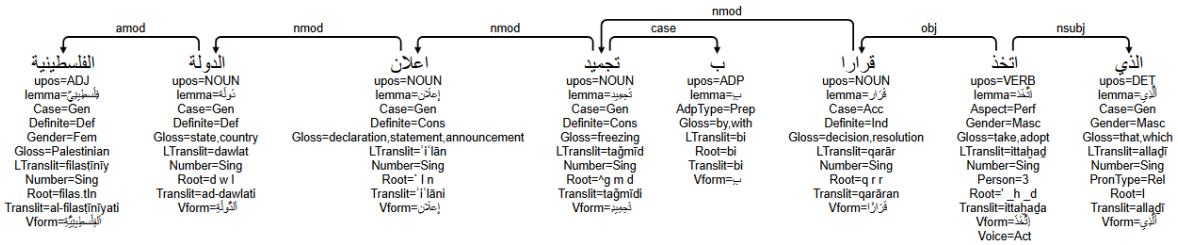
b. قام بتجهيز القاعة [V + SP]

qāma	*bi=taġhīz*	*al-qāʿah*
effectuer.PST.3SG	PREP=préparation	DEF-sall

‘Il a préparé la salle’

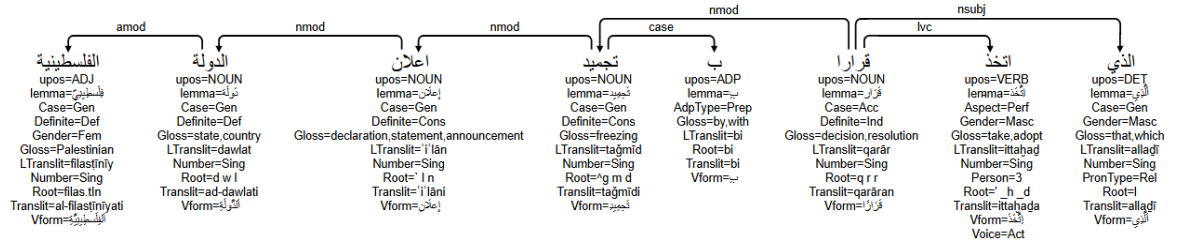
Comme nous l’avons mentionné plus haut (§2.1), les relations de dépendance dans UD s’établissent principalement entre les mots lexicaux, alors que les mots fonctionnels sont perçus comme des porteurs de traits morphosyntaxiques, qui se rattachent normalement à un mot lexical. Le choix de la tête syntaxique devient notamment complexe dans le cas des verbe supports qui représentent un cas d’expressions polylexicales (cf. Stephen et Zeman, 2024 ; voir aussi Kahane, Courtin et Gerdes, 2018 pour le traitement des expressions polylexicales dans le cadre de UD) où deux ou plusieurs mots se combinent en une seule unité lexicale.

Les *treebanks* UD, soit en arabe, soit en français, adoptent une approche particulière dans leur analyse des constructions à verbe support, s’écartant ainsi des principes de base du schéma UD qui favorise une analyse sémantique des mots fonctionnels. Ils traitent le verbe comme la tête du nom prédicatif qu’il gouverne par la relation `obj`[xxviii]. Cette approche ne correspond pas, par exemple, à l’analyse standard de la copule (voir *infra*), qui est considérée comme dépendante du prédicat nominal ou adjectival. La figure 22 illustre l’analyse actuelle de la construction à verbe support en arabe, alors que la figure 23 présente une analyse plus cohérente avec les principes UD (inspirée de Gerdes et Kahane, 2016).



‘[…] qui a pris la décision de geler la déclaration de l’État palestinien’

Figure 22 : UD_Arabic-PADT@2.15



‘[…] qui a pris la décision de geler la déclaration de l’État palestinien’

Figure 23 : l’analyse cohérente avec les principes UD

Un autre point d’incohérence avec les principes se manifeste dans certains *treebanks* français[xxix]. Dans UD French-GSD, les compléments de la construction à verbe support sont rattachés au verbe plutôt qu’au nom prédicatif (fait→obl:arg→problèmes) (figure 24). Les principes stipulent que les compléments devraient être liés au nom prédicatif, qui est le porteur du sens

(face→obl:arg→problèmes). Sur le plan syntaxique, rattacher le complément de cette construction au nom est privilégié (cf. l'analyse adoptée par UD French-PUD, figure 25), car celui-ci peut former une unité significative avec son complément. On peut reprendre le nom prédicatif et son complément de l'exemple (*Face à de sérieux problèmes*) dans une nouvelle phrase (*Face à de sérieux problèmes environnementaux, des mesures urgentes doivent être prises*), ce qui suggère une forte cohésion syntaxique entre le nom et son complément. Le nom en tant que prédicat contrôle la valence, le verbe sert davantage à introduire ou à grammaticaliser ce nom.

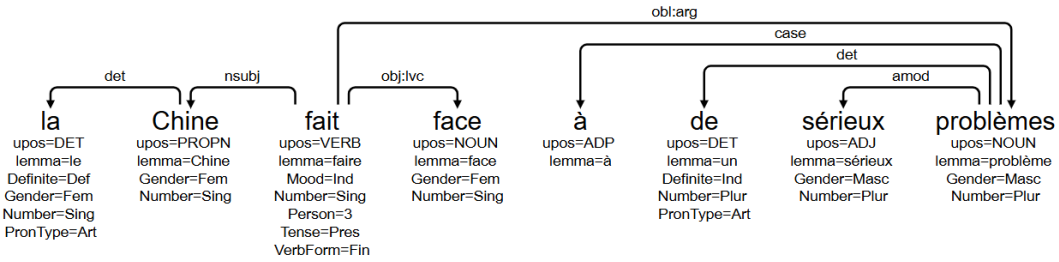


Figure 24 : UD_French-GSD@2.15

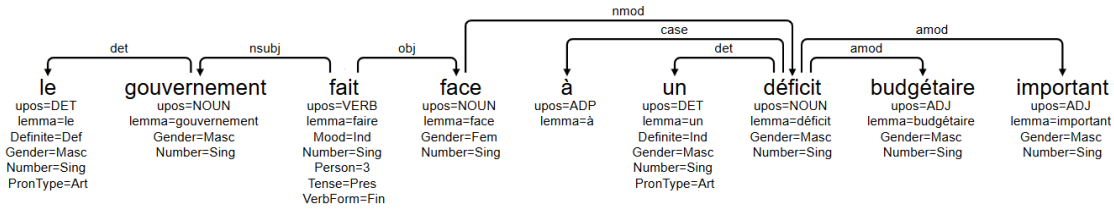


Figure 25 : UD_French-PUD@2.15

Toutefois, dans certains cas, le rattachement des compléments au verbe plutôt qu'au nom est privilégié. Ceci est notamment le cas lorsque la combinaison du nom et de son complément ne constitue pas une unité sémantique cohérente (cf. le guide d'annotation de SUD[xxx]). Prenons l'exemple du français *les enfants doivent prendre garde aux voitures en traversant la rue*. On ne dirait pas naturellement **garde aux voitures en traversant la rue, […]*. Dans ce contexte, le verbe *prendre* est considéré comme la tête du complément *aux voitures*.

Si l'ambiguïté persiste quant à la détermination de la tête, la pronominalisation peut servir de test pertinent. Bien que ce test puisse être difficile à appliquer dans un exemple du type *la carte donne accès à plusieurs centres* (il est peu naturel de dire *la carte le donne à plusieurs centres* en remplaçant *accès*), il s'avère efficace pour d'autres constructions. Par exemple, pour *on a fait une promenade dans la ville de Liège*, la pronominalisation fonctionne bien : *cette promenade, on l'a faite dans la ville de Liège*. Dans ce dernier cas, la pronominalisation de *une promenade* en *l'* indique clairement que *promenade* est la tête du syntagme et peut être séparée du verbe support *fait*.

Dans le cadre de SUD, les constructions à verbe support sont marquées avec le trait syntaxique profond `@lvc` (figure 26).

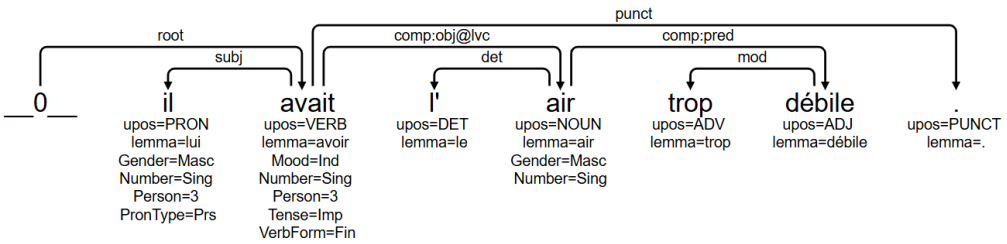


Figure 26 : SUD_French-ParisStories@2.15

4.3 Les copules et les auxiliaires

Les structures copulatives et les constructions verbales complexes incluant des auxiliaires soulèvent de nombreuses questions linguistiques dans diverses langues, y compris l'arabe et le français. Ces structures constituent d'ailleurs un point de désaccord fondamental entre les schémas UD et SUD. Avant d'aller plus loin, il est nécessaire de présenter les aspects suivants concernant les copules et les auxiliaires en arabe.

Le verbe *kāna*, dont l'équivalent français est le verbe *être*, occupe une place fondamentale dans la langue arabe. Il assume deux fonctions principales : celle de copule et celle de verbe auxiliaire. Lorsqu'il fonctionne comme un verbe copule, il régit un sujet, appelé *ʾism kāna* ('le nom de *kāna*'), qui se met au cas nominatif et un prédicat, appelé *ḫabar kāna* ('l'attribut de *kāna*'), qui se met au cas accusatif (7a). Quand il est employé comme auxiliaire, le verbe, sémantiquement vide, constitue avec un autre verbe une forme verbale composée, indiquant principalement un aspect temporel (cf. El Kassas, 2005 ; Pinon, 2013) (7b).

(7) a. كان الرجل بشوشًا

kāna	*ar-raǧul-u*	*bašūš-a-n*
être.PST.3SG	DEF-homme-NOM	souriant-ACC-INDF

'L'homme était souriant'

b. كان الرجل قد غادر

kāna	*ar-rǧl-u*	*qad*	***ġādara***
être.PST.3SG	DEF-homme-NOM	déjà	**partir.PST.3SG**

'L'homme était déjà parti'

L'une des spécificités de l'arabe par rapport au français est le fait qu'au présent de l'indicatif, la copule *kāna* n'est généralement pas réalisée. Dans ce cas, il s'agit d'une construction prédicative qui comprend un *mubtadaʾ* ('sujet de la proposition à N-initial') et un le *ḫabar* ('un attribut'), les deux se mettent au cas nominatif (8a). Pour nier un état au présent, l'arabe recourt au verbe *laysa* ('ne pas être') qui reste souvent invariable et gouverne un nom à l'accusatif (8b).

(8) a. الرجلُ بشوشٌ

*ar-raǧul-**u***	*bašūš-**u-n***
DEF-homme-**NOM**	souriant-**NOM-INDF**

'L'homme {est} souriant'

b. الرجل ليس بشوشًا

ar-rağul-u	***laysa***	*bašūš-**a-n***
DEF-homme-NOM	**ne_pas_être**	souriant-**ACC-INDF**

'L'homme n'est pas souriant'

D'autres verbes arabes agissent comme des copules, connus dans la grammaire traditionnelle arabe sous l'appellation *ʾaḫawāt kāna* ('les sœurs de *kāna*'), comme *ʾaṣbaḥa*, *ʾaḍḥā*, *ʾamsā*, *bāta*, *ẓalla*, etc. ('devenir', 'rester'). Ils introduisent donc dans une proposition nominale avec un *ʾism* au nominatif et un *ḫabar* au cas accusatif (9a). L'arabe possède également un groupe de verbes auxiliaires aspectuels, communément appelés dans la grammaire arabe classique *ʾafʿāl al-mqārabah wal-šurūʿ* ('les verbes d'imminence et de commencement') qui signalent soit le rapprochement de la réalisation d'un événement, soit son commencement, comme *kāda*, *ʾawšaka*, etc. ('être sur le point de') et *šaraʿa*, *badaʾa*, etc. ('commencer à', 'se mettre à') (9b). Ces verbes se comportent de manière similaire au verbe *kāna*. Ils gouvernent un *ʾism* au nominatif et un *ḫabar* propositionnel au cas accusatif dont la tête est un verbe au présent de l'indicatif. Ce dernier régit un pronom sujet coréférant avec le *ʾism*.

(9) a. ظل الرجلُ صامتًا

ẓalla	*ar-rağul-**u***	*ṣāmtit-**a-n***
rester.PST.3SG	DEF-homme-**NOM**	silencieux-**ACC-INDEF**

'L'homme est resté silencieux'

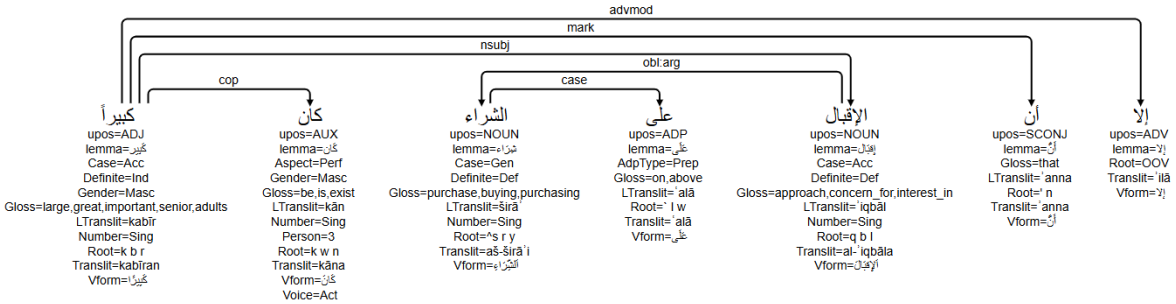
b. بدأ الولدُ يلعب

badaʾa	*al-walad-**u***	***yalʿab***	***{huwa}***
commencer.PST.3SG	DEF-garçon-**NOM**	**jouer.PRS.3SG**	**{il}**

'Le garçon a commencé à jouer'

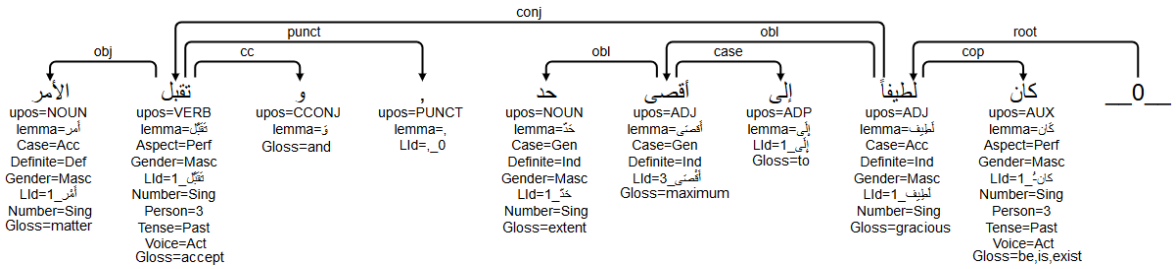
Dans le cadre des projets UD et SUD, seulement *kāna* et *laysa* en arabe et *être* en français sont considérés comme copules. Ceci s'inscrit dans l'approche du projet qui limite la relation `cop` (copula) seulement aux copules pures n'ajoutant que des marques de temps, d'aspect ou de modalité au prédicat. Cela implique que la majorité des langues possède au plus un seul verbe copule. Tous les autres verbes que la grammaire arabe traditionnelle qualifie de copules sont donc analysés comme des verbes lexicaux ordinaires.

Les *treebanks* UD pour l'arabe présentent une annotation standardisée concernant *kāna*/*laysa* en tant que copules. Dans UD Arabic-PADT (figure 27) et Arabic-PUD (figure 28), les deux copules prennent la catégorie grammaticale AUX, conformément au guide UD. L'approche du schéma UD consiste à traiter le prédicat nominal/adjectival comme la tête de la construction copulative, qui gouverne la copule via la relation `cop`. Cette approche trouve sa légitimité, comme l'explique le guide d'annotation UD, dans le constat que de nombreuses langues, à l'instar du russe (et de l'arabe, comme on vient de le voir), omettent souvent, voire toujours, une copule explicite dans de telles constructions.



‘[…] cependant, l’affluence pour l’achat était grande’

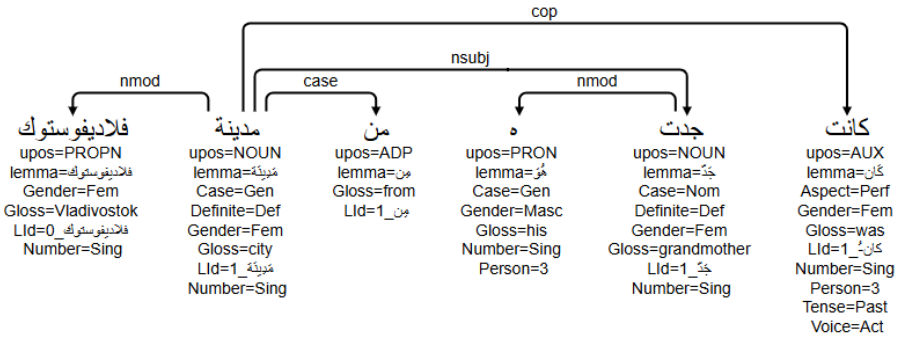
Figure 27 : UD_Arabic-PADT@2.15



‘Il était extrêmement gentil et a accepté la situation’

Figure 28 : UD_Arabic-PUD@2.15

Cette analyse est également retenue dans le cadre du projet UD lorsque le prédicat est un syntagme prépositionnel. Dans ce cas, le noyau nominal est identifié comme la tête de la proposition (figure 29).



‘[…] sa grand-mère était de la ville de Vladivostok’

Figure 29 : UD_Arabic-PUD@2.15

La même analyse concernant la copule *être* est également adoptée dans tous les *treebanks* UD français (figure 30).

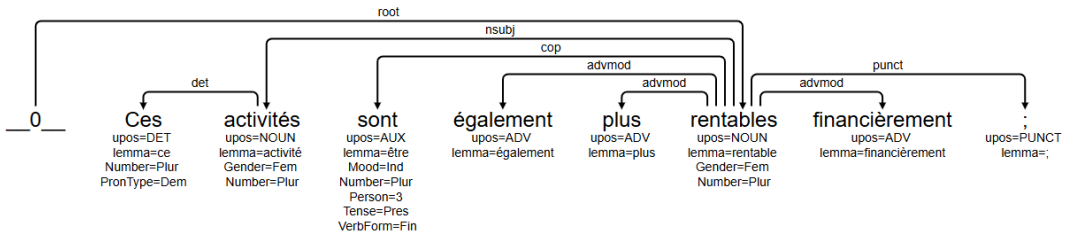
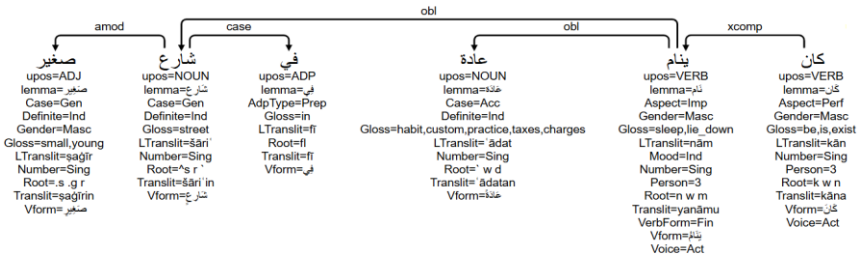


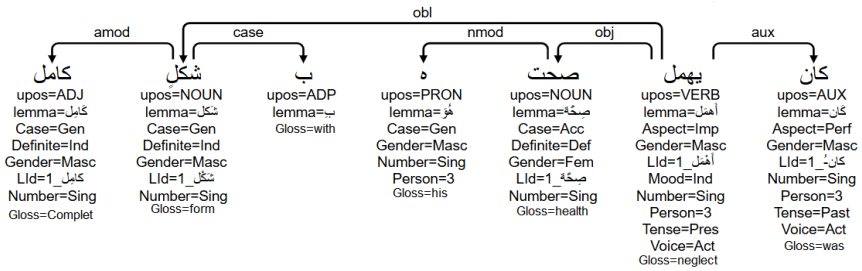
Figure 30 : UD_French-ParTUT@2.15

Les *treebanks* arabes présentent, cependant, des incohérences dans le traitement de *kāna* en tant qu'auxiliaire temporel (*kāna* + $V_{inaccompli}$), reflétant des choix d'annotation différents entre UD Arabic-PADT et UD Arabic-PUD. Ainsi, UD Arabic-PADT (figure 31) opte pour une analyse de *kāna* comme verbe lexical[xxxi] (étiqueté par la catégorie grammaticale VERB). Il régit un prédicat verbal via la relation `xcomp`, tout comme les verbes d'imminence comme *kāda* ('être sur le point de') ou de commencement comme *šaraʿa* ('se mettre à'). Le principal inconvénient de cette analyse est que l'utilisation de la relation `xcomp` occulte la fonction aspectuelle/temporelle de *kāna*, qui est pourtant fondamentale. De plus, cette analyse ne s'aligne pas sur le traitement standard des auxiliaires aspectuels/temporels dans les autres langues, comme le démontrent les auxiliaires *avoir* et *être* en français (Voir *infra*). Par contre, UD Arabic-PUD (figure 32) adopte une analyse uniforme de *kāna*. Il est systématiquement catégorisé comme AUX et dépend du verbe principal via la relation `aux` (auxiliary), qui est consacrée aux mots fonctionnels associés à un prédicat verbal et exprimant des catégories telles que le temps, le mode, l'aspect, la voix ou l'évidentialité.



'Il avait l'habitude de dormir habituellement dans une petite rue […]'

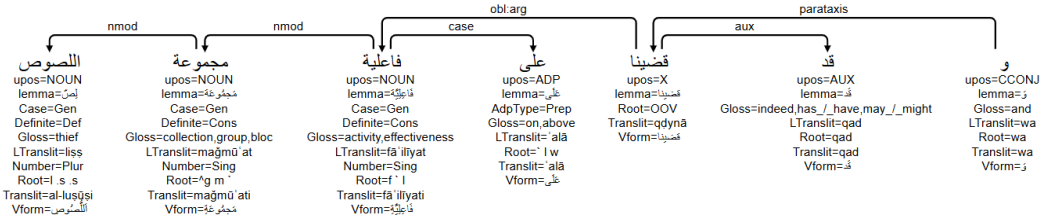
Figure 31 : UD_Arabic-PADT@2.15



'Il négligeait complètement sa santé'

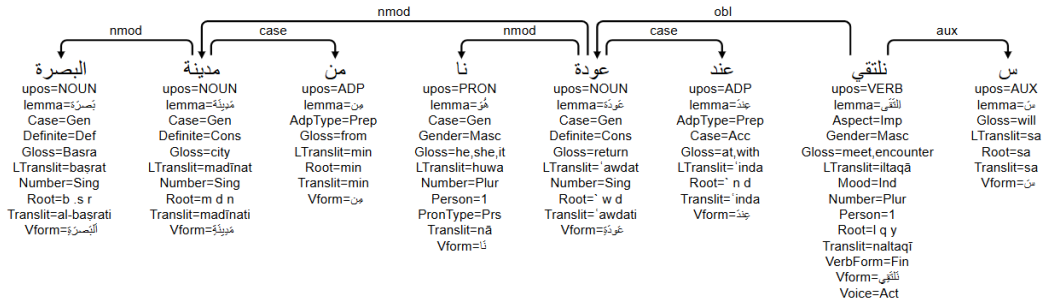
Figure 32 : UD_Arabic-PUD@2.15

Notons que, d'après le guide de UD, la relation `aux` n'est pas toujours restreinte aux verbes. Des marqueurs TAME (Temps-Aspect-Mode-Évidentialité) peuvent aussi se rattacher par la relation `aux`. Bien que UD Arabic-PUD réserve exclusivement la relation `aux` à l'auxiliaire *kāna*, UD Arabic-PADT attribue cette relation à des marqueurs tels que *qad* ('déjà') et *sawfa*, ainsi qu'au préverbe *sa-* exprimant le futur (figures 33 et 34).



‘Et nous avons mis fin à l’efficacité d’un groupe de voleurs […]’

Figure 33 : UD_Arabic-PADT@2.15



‘Nous nous rencontrerons à notre retour de la ville de Bassorah’

Figure 34 : UD_Arabic-PADT@2.15

Dans les *treebanks* UD du français, seuls trois verbes sont considérés comme auxiliaires, en plus de la copule *être* : les auxiliaires temporels *avoir* et *être*, l’auxiliaire passif *être* et l’auxiliaire causatif *faire*. La justification principale de cette sélection restrictive est basée sur le fait que ces trois verbes partagent exclusivement la propriété syntaxique définitoire des auxiliaires, celle de *la montée des clitiques* (cf. Gerdes *et al.*, 2024). Cela signifie que le pronom clitique peut se déplacer et se rattacher à un verbe qui n’est pas celui dont il dépend syntaxiquement (10a vs, b).

(10) a. Un message **lui** a été envoyé.
b. Je vais **le** voir.

Ces auxiliaires sont donc rattachés au verbe principal par la relation `aux`, avec parfois une extension précisant leur fonction : `aux:tense`, `aux:pass`, `aux:caus` (figure 35).

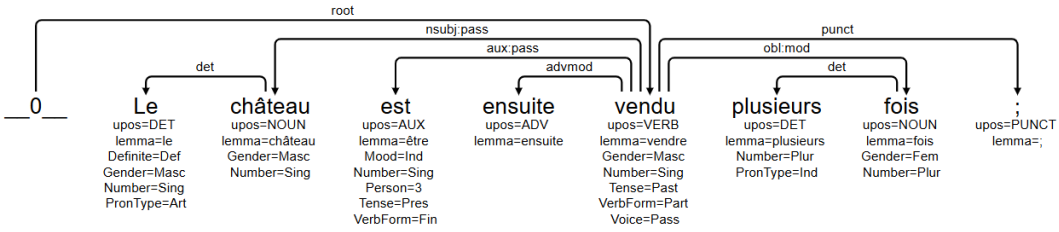


Figure 35 : UD_French-GSD@2.15

Le projet SUD rejette l’analyse UD qui considère que le prédicat non verbal d’une construction copulative ou le verbe principal d’une construction avec auxiliaire comme la tête syntaxique, pour adopter une analyse alternative fondée sur des critères purement distributionnels. SUD

propose d'identifier la copule et l'auxiliaire comme têtes syntaxiques, et ce pour les quatre raisons suivantes (cf. Gerdes *et al*., 2024) :

i) les auxiliaires et les copules portent souvent les marques de la force illocutoire de la phrase, c'est-à-dire sa fonction communicative : assertive, interrogative, etc. : ***As-tu*** *mangé ?*

ii) les auxiliaires et les copules sont le locus morphosyntaxique, c'est-à-dire les éléments qui portent les marques de flexion (mode, temps, personne, nombre) : *Je ne crois pas qu'il* ***fasse*** *venir ses enfants*.

iii) les auxiliaires et les copules imposent des contraintes distributionnelles sur les autres éléments, par exemple en fixant la position des clitiques, comme mentionné précédemment : *Un message* ***lui*** *a été envoyé*.

iv) les auxiliaires et les copules sont des composants obligatoires dans diverses langues (l'arabe autorise l'ellipse de la copule au présent de l'indicatif, voir *supra*), dont la présence conditionne l'organisation de la phrase et ne peuvent être omis sans altérer sa grammaticalité : *Il* ***fait*** *rire* → **il rire*.

Dans les figures 36 et 37, l'analyse SUD optera pour les auxiliaires *est*/*fait* comme têtes syntaxiques. Les verbes qui portent le sens lexical *vendu*/*rire* seront alors leurs dépendants. Pour identifier ces dépendants, la relation `comp:aux` sera employée. Cette relation correspond à la relation `aux` telle que définie par UD. De plus, la relation `comp:aux` pourra être enrichie par trois traits de la syntaxe profonde : `@tense` (pour le temps), `@pass` (pour le passif) et `@caus` (pour le causatif).

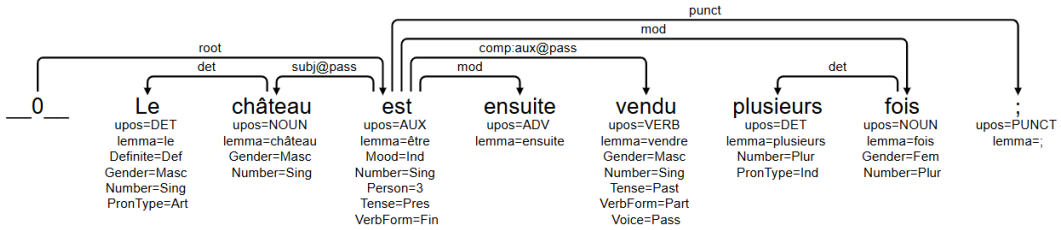


Figure 36 : SUD_French-GSD@2.15

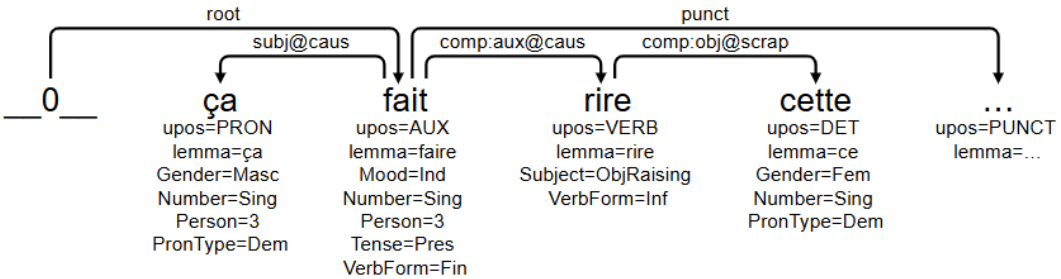
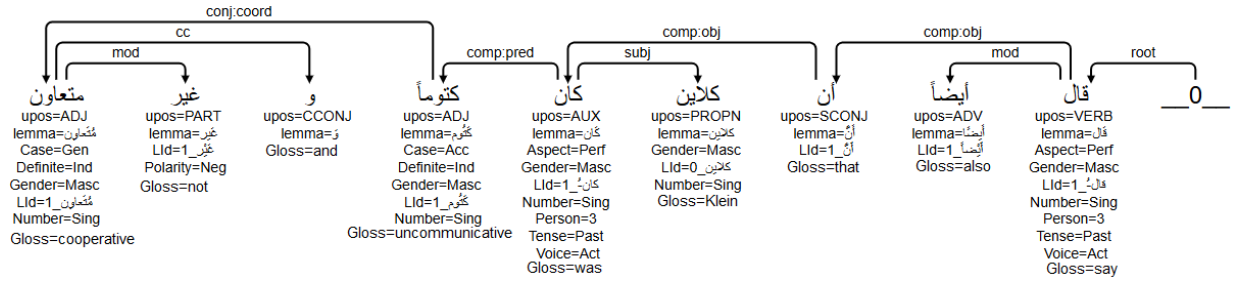


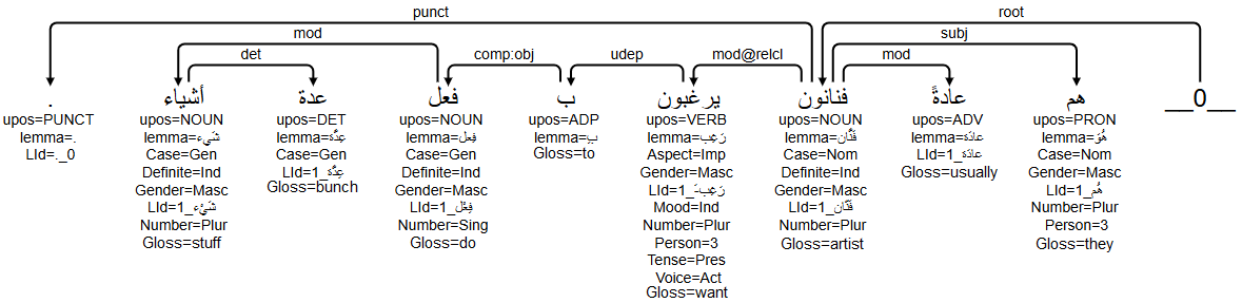
Figure 37 : SUD_French-Rhapsodie@2.15

Contrairement à UD qui vise une analyse uniforme des constructions prédicatives, qu'elles contiennent ou non une copule (le sujet étant toujours dépendant du prédicat), SUD établit une distinction entre ces deux types de constructions en arabe. Selon SUD, le sujet dépend de la copule lorsque celle-ci est présente (figure 38), tandis qu'en son absence, il dépend directement du prédicat (figure 39).



'Il a également dit que Klein était secret et peu coopératif'

Figure 38 : SUD_Arabic-PUD@2.15



'Ils sont habituellement des artistes qui désirent faire plusieurs choses'

Figure 39 : SUD_Arabic-PUD@2.15

En adoptant une analyse distinguant systématiquement les constructions avec et sans copule, SUD respecte les critères syntaxiques distributionnels. En présence de la copule, SUD se conforme à la grammaire arabe traditionnelle, qui voit la copule *kāna* comme un verbe bivalent régissant son sujet et son attribut. Lorsque la copule est absente, c'est le nom prédicatif qui est la tête syntaxique (voir Gerdes *et al.* 2024 concernant le cas des constructions prédicatives sans copule en russe).

En effet, dans la phrase :

(11) أخي طبيبُ أسنانٍ

ʾaḫ=ī	*ṭabīb-u*	*ʾasnān-i-n*
frère=POSS	médecin-NOM	dent.PL-GEN-INDEF

'Mon frère {est} dentiste'

on peut substituer le sujet *ʾaḫī* ('mon frère') par un autre nom comme *ʾabī* ('mon père') sans altérer la structure de la phrase. Néanmoins, si l'on tente de commuter *ʾakī* par un adjectif superlatif, comme *ʾamharu* ('le plus compétent'), cela donne lieu à un groupe nominal *ʾamharu*

ṭabību ʾasnānin (‘le dentiste le plus compétent’) dont la fonction et la distribution diffèrent de la phrase initiale *ʾaḫī ṭabību ʾasnānin* (‘mon frère {est} dentiste’).

Selon le *critère distributionnel sans effacement* (Gerdes *et al*. 2024, p. 598) :

> Si U = AB, A peut commuter avec un A’ et U et U’ = A’B n’ont pas la même distribution, alors A est la tête de U.

on pourrait suggérer que *ʾaḫī* (‘mon frère’) est la tête de la construction. Toutefois, cette conclusion est rejetée, car le sujet *ʾaḫī* (‘mon frère’) ne commute pas avec l’adjectif superlatif *ʾamharu* (‘le plus compétent’), étant donné que la commutation requiert que les éléments substituables soient mutuellement exclusifs. Or, l’adjectif superlatif *ʾamharu* (‘le plus compétent’) peut coexister dans la même phrase *ʾaḫī ʾamharu ṭabībi ʾasnānin* (‘mon frère {est} le dentiste le plus compétent’), où *ʾamharu* modifie le prédicat *ṭabībi ʾasnānin* (‘dentiste’). Si *ʾaḫī* était réellement la tête de la phrase, la structure n’autoriserait pas sa coexistence avec un élément (l’adjectif superlatif) qui pourrait le remplacer s’il était la tête. Cette cooccurrence viole le principe fondamental d’exclusion mutuelle qui est essentiel pour identifier la tête d’une construction syntaxique. En d’autres termes, l’absence d’exclusion entre *ʾaḫī* et *ʾamharu* invalide l’hypothèse que *ʾaḫī* est la tête.

5. Conclusion

Dans le cadre de cette recherche, nous avons examiné l’annotation syntaxique de certaines constructions complexes dans une approche contrastive arabe-français. Pour ce faire, nous nous sommes appuyés sur les corpus arborés (*treebanks*) en dépendances, annotés selon les schémas d’annotation *Universal Dependencies* (UD) et *Surface Syntactic Universal Dependencies* (SUD). Le premier privilégie les mots lexicaux comme têtes des relations de dépendance et le second favorise les têtes fonctionnelles. L’objectif était de comprendre les choix faits par ces deux schémas pour modéliser des phénomènes linguistiques complexes et d’évaluer leur pertinence pour capturer les spécificités linguistiques propres à chaque langue.

Après avoir exposé en détail les caractéristiques des deux schémas d’annotation UD et SUD, ainsi que les *treebanks* arabes et français accessibles via ces deux projets, nous avons procédé à l’étude de trois constructions complexes : *les constructions relatives*, *les constructions à verbe support* et *les constructions copulatives et avec un auxiliaire.* Les résultats de notre étude révèlent que, bien que les annotations syntaxiques de ces constructions montrent une uniformité dans les cas standards, elles affichent des différences marquées dans des situations plus idiosyncrasiques. Ces divergences ne se limitent pas uniquement aux comparaisons entre les *treebanks* de l’arabe et du français, mais apparaissent également au sein des *treebanks* d’une même langue. Ces différences intra-langues peuvent naturellement être expliquées par le fait que les cadres à travers lesquels ces données ont été initialement annotées sont différents, tout comme les équipes d’annotateurs qui ont réalisé ces travaux. Ces incohérences ont pour conséquence de rendre difficile la comparabilité entre les différents *treebanks*, ce qui éloigne le

projet de son objectif général, à savoir la typologie des langues. De plus, il devient complexe d'agréger les données, de les interroger uniformément ou d'y appliquer des traitements automatiques standardisés.

Ce constat met en évidence une double nécessité pour l'amélioration des annotations syntaxiques: il est devenu essentiel d'harmoniser les *treebanks* arabes déjà existants. Cela implique un travail de standardisation et de correction pour assurer une plus grande cohérence entre les différents corpus. Cette démarche d'harmonisation n'est pas sans précédent ; l'équipe française a déjà réalisé des efforts notables en ce sens sur les deux *treebanks* du français : UD French-GSD et UD French-Sequoia (cf. Guillaume *et al.*, 2019). Cependant, malgré les progrès déjà réalisés en français, il reste un travail considérable à accomplir pour améliorer les données en vue des prochaines versions du projet.

Au-delà de l'harmonisation des données existantes, ce travail doit nécessairement être complété par un enrichissement du guide d'annotation UD pour l'arabe standard[xxxii] qui présente des limitations par rapport à la version française[xxxiii]. En effet, le guide général d'annotation UD ne prend pas suffisamment en compte certaines constructions syntaxiques spécifiques, ni les particularités uniques de chaque langue. C'est pourquoi il est nécessaire que les guides d'annotation spécifiques à chaque langue comblent ces lacunes. Ces documents dédiés doivent inclure les phénomènes et les spécificités de la langue que le guide général ne couvre pas suffisamment.

En comparant les deux schémas d'annotation, SUD présente une capacité à offrir des solutions plus claires et robustes pour des constructions où les principes de UD deviennent complexes. D'ailleurs, pour l'arabe, SUD semble offrir une approche plus intuitive pour l'annotation syntaxique. SUD se focalise sur la syntaxe de surface et l'utilisation de critères distributionnels, c'est-à-dire que la tête d'une unité syntaxique est l'élément qui contrôle son comportement et sa distribution dans la phrase. Pour l'arabe, cela est particulièrement avantageux, car des mots fonctionnels, comme les prépositions ou les particules, peuvent jouer un rôle central dans la structure syntaxique. Cette approche, comme celle adoptée par le *Columbia Arabic Treebank* (CATiB) (Habash, Faraj et Roth, 2009)[xxxiv], est souvent plus conforme aux descriptions grammaticales traditionnelles de l'arabe, et, par conséquent, plus intuitive pour les annotateurs arabophones. Toutefois, malgré ces avantages théoriques, il est important de tester concrètement le schéma SUD sur un corpus natif de l'arabe, c'est-à-dire élaboré directement selon les normes SUD. Cette application pratique sur des données réelles est essentielle pour en évaluer pleinement la pertinence.

Notes

[i] *Treebank* est un terme technique de l'anglais couramment employé dans le domaine de la linguistique informatique et du TAL pour désigner à un ensemble des phrases extraites de textes authentiques et accompagnées d'une analyse syntaxique détaillée, souvent représentée sous la forme d'un arbre de dépendances ou de constituants (Kahane et Mazziotta, 2022, p. 64).

[ii] Certains *treebanks* du projet UD ne mettent pas à disposition le texte des phrases ni les lemmes en raison de restrictions de licence.

[iii] La tête d'une unité syntaxique U se définit comme « *toute sous-unité de U qui n'est gouvernée par aucune autre sous-unité de U* » (Kahane et Gerdes, 2022, p. 303, les concepts de *tête* et de *dépendance* sont expliqués en profondeur dans le chapitre 10).

[iv] Les caractéristiques des *treebank* arabes et français et la représentation des données sous forme de graphes via l'outil Grew-match serons présentés plus loin.

[v] Le *token* représente l'unité minimale de découpage d'un texte, servant de base au traitement linguistique. Il correspond à la plus petite entité reconnue comme élément autonome dans l'analyse (voir la page UD concernant la tokenisation et la segmentation des mots : https://universaldependencies.org/u/overview/tokenization.html).

[vi] Disponible à l'adresse : https://match.grew.fr/.

[vii] https://universal.grew.fr/?corpus=SUD_Arabic-PUD@2.16.

[viii] mSUD (et mUD) sont de nouveaux formats d'annotation conçus pour être compatibles avec le schéma UD et qui ont pour but de prendre des unités infra-mot comme *tokens* dans les corpus arborés (cf. Guillaume *et al*., 2024).

[ix] Les cinq *treebanks* sont également disponibles au format SUD.

[x] Certains *treebanks* UD n'ont pas été convertis au format SUD en raison de leurs licences interdisant la création d'œuvres dérivées.

[xi] L'apprenabilité signifie que les annotations sont plus facilement reproductibles par un analyseur syntaxique automatique.

[xii] https://corpus.quran.com/.

[xiii] https://sites.google.com/nyu.edu/camel-treebank/home.

[xiv] https://github.com/UniversalDependencies/UD_Arabic-PUD.

[xv] En plus de ces trois *treebanks* de l'arabe standard moderne, UD contient un certain nombre de *treebanks* pour un certain nombre de dialectes arabes, comme l'arabe levantin méridional (UD South Levantine Arabic MADAR, Zahra, 2020 ; Bouamor *et al*., 2018) et l'arabe maghrébin-français (UD Maghrebi Arabic French Arabizi, Arij, Menel et Djamé, 2023). UD contient également des *treebanks* pour des langues anciennes parlées dans la région arabe, comme l'akkadien (UD Akkadian RIAO ; UD Akkadian PISANDUB, Luukko *et al*. 2020), l'égyptien pré-copte (UD Egyptian-UJaen, Hernández et Passarotti, 2024) et le copte sahidique (UD Coptic Scriptorium, Zeldes et Abrams, 2018).

[xvi] https://autogramm.github.io/.

[xvii] https://anr.fr/fr/projets-finances-et-impact/projets-finances/projet/funded/project/anr-20-fral-0001/

[xviii] Le français bénéficie, d'ailleurs, de *treebanks* diachroniques au sein du projet *PROFITEROLE*. Le UD Middle_French-PROFITEROLE (cf. Prévost *et al.*, 2024) cible le moyen français (XIV^{e} - XV^{e} siècles), tandis que le UD Old_French-PROFITEROLE (cf. Romanova, Ziane, Daoudi) est dédié à l'ancien français. Il est, d'ailleurs, à noter que le *treebank* UD French-FTB, issu de la conversion au format UD du *French Treebank* (Abeillé *et al*., 2003), a été retiré de la dernière version UD.

[xix] La marque du cas (nominatif, accusatif, génitif) sur les pronoms relatifs est visible uniquement pour le duel.

[xx] https://orfeo.grew.fr/?corpus=cefc-gold.

[xxi] Rappelons que faute d'accès au texte des phrases et des lemmes du *treebank* UD Arabic-NYUAD, nous n'avons pas pu exploiter cette ressource pour cette étude.

[xxii] `mod@relcl` (modifier relative clause) dans SUD.

[xxiii] Pour mieux cibler l'analyse du phénomène linguistique en question, des segments de la phrase ont été volontaires omis, ce qui a conduit à la suppression de certaines relations syntaxiques présentes dans la structure initiale.

[xxiv] Étant donné que la plupart des phrases dans le *treebank* Arabic-PUD ne sont pas glosées, nous avons procédé à l'ajout manuel des gloses en dessous des traits morphosyntaxiques (Gloss=). Afin de garantir leur uniformité avec l'ensemble des autres treebanks UD, nous les avons insérées en anglais.
[xxv] https://universaldependencies.org/u/overview/enhanced-syntax.html#ellipsis
[xxvi] Les pronoms relatifs prennent la catégorie DET avec le trait PronType=Rel.
[xxvii] Compte tenu du fait que ce n'est pas toujours le cas qu'il s'agisse du verbe.
[xxviii] Ou `obj:lvc` dans certains *treebank* français (UD French-GSD ; UD French-Sequoia ; UD French-Rhapsodie ; UD French-ParisStories). D'autres *treebanks*, notamment des langues asiatiques, ont employé la relation `compound` ou la sous-relation `Compound:lvc` pour marquer les constructions à verbes supports.
[xxix] Dans les *treebanks* arabes UD Arabic-PADT et UD Arabic-PUD, le complément est toujours rattaché au nom prédicatif.
[xxx] https://surfacesyntacticud.github.io/guidelines/u/particular_phenomena/lvc/.
[xxxi] Dans des contextes relativement peu fréquents, *kāna* conserve aussi un sens lexical plein, exprimant principalement l'existence (= *exister*, *avoir lieu*) (El Kassas, 2005, pp. 174-175) :

(i) كان هناك حريق في المدينة
kāna *hunāka* *ḥarīq* *fī* *al-madīnah*
être DEM incendie PREP DEF-ville
'Il y avait un incendie dans la ville'

[xxxii] https://universaldependencies.org/ar/.
[xxxiii] http://universaldependencies.org/fr.
[xxxiv] Je tiens à remercier Nizar Habash pour avoir attiré mon attention sur le rapprochement entre les schémas SUD et CATiB.

Remerciements

Je remercie Sylvain Kahane pour ses nombreuses remarques sur la précédente version de cet article, ainsi que les deux relecteurs de la revue pour leurs corrections.

Références bibliographiques

Abeillé, A., Barrier, N. (2004). Enriching a French Treebank. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal.

Abeillé, A., Clément, L., et Toussenel, F. (2003). Building a Treebank for French. In Abeillé, A. (Ed.), *Treebanks: Building and using parsed corpora*. Kluwer. 165–187.

Ahnaiba, A. (2006). *Les verbes supports en arabe classique et en arabe moderne. Le cas de'Akhadha/Ittakhadha l'équivalent du verbe support français prendre*. (Thèse de doctorat, Université Paris 4).

Al-Ghamdi, S., Al-Khalifa, H., et Al-Salman, A. (2021). A dependency treebank for classical Arabic poetry. In *Proceedings of the Sixth International Conference on Dependency Linguistics (Depling, SyntaxFest 2021)*. 1-9.

Arij, R., Menel, M., et Djamé, S. (2023). Enriching the NArabizi treebank: A multifaceted approach to supporting an under-resourced language. In *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*. 266-278.

Blanche-Benveniste, C., Chervel, A., et Le Franc, P. (1990). *Le français parlé : études grammaticales*. Éditions du CNRS.

Bouamor, H., Habash, N., Salameh, M., Zaghouani, W., Rambow, O., Abdulrahim, D., ... et Oflazer, K. (2018). The MADAR Arabic dialect corpus and lexicon. In *Proceedings of the Eleventh*

International Conference on Language Resources and Evaluation (LREC 2018). 3387-3396

Buchholz, S., et Marsi, E. (2006). CoNLL-X shared task on multilingual dependency parsing. In *Proceedings of the tenth conference on computational natural language learning (CoNLL-X)*. 149-164.

Candito, M., et Seddah, D. (2012). Le corpus Sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical. In *Proceedings of TALN'2012*. 49–60.

Chomsky, N. (1957). *Syntactic structures*. The MIT Press.

Chomsky, N. (1965). *Aspects of the theory of syntax*. The MIT Press.

Danlos, L. (2010). Extension de la notion de verbe support. In T. Nakamura, É. Laporte, A. Dister, et C. Fairon (Eds.), *Les Tables, La grammaire par le menu, Volume d'hommage à Christian Leclère*. Presses Universitaires de Louvain. 81-90.

De Marneffe, M. C., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J., et Manning, C. D. (2014). Universal Stanford dependencies: A cross-linguistic typology. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)* (Vol. 14). 4585-4592.

De Marneffe, M. C., Manning, C., Nivre, J., et Zeman, D. (2021). Universal Dependencies. *Computational Linguistics*, *47*(2). 255–308.

Debaisieux, J. M., Benzitoun, C., et Deulofeu, H. J. (2016). Le projet ORFEO : Un corpus d'études pour le français contemporain. *Revue Corpus*, *15*, 91-114.

Deulofeu, J. (2003). L'approche macrosyntaxique en syntaxe : un nouveau modèle de rasoir d'Occam contre les notions inutiles. *Scolia*, *16*, 77-95.

Dukes, K., et Buckwalter, T. (2010). A dependency treebank of the Quran using traditional Arabic grammar. In *2010 the 7th International Conference on Informatics and Systems (INFOS)*. IEEE. 1-7

El Kassas, D. (2005). *Une étude contrastive de l'arabe et du français dans une perspective de génération multilingue*. (Doctoral dissertation, Université Paris 7).

Gerdes, K., et Kahane, S. (2016). Dependency annotation choices: Assessing theoretical and practical issues of universal dependencies. In *Proceedings of the 10th Linguistic Annotation Workshop (LAW X)*. 131–142.

Gerdes, K., et Kahane, S. (2017). Trois schémas d'annotation syntaxique en dépendance pour un même corpus de français oral : le cas de la macrosyntaxe. In *Actes de la 24e conférence sur le traitement automatique des langues (TALN)*.1-9.

Gerdes, K., Guillaume, B., Kahane, S., et Perrier, G. (2018). SUD or Surface-Syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD. In *Universal Dependencies Workshop 2018*.

Gerdes, K., Guillaume, B., Kahane, S., et Perrier, G. (2019a). Improving Surface-syntactic Universal Dependencies (SUD): surface-syntactic relations and deep syntactic features. In *TLT 2019-18th International Workshop on Treebanks and Linguistic Theories*. 126-132.

Gerdes, K., Guillaume, B., Kahane, S., et Perrier, G. (2019b). Pourquoi se tourner vers le SUD : L'importance de choisir un schéma d'annotation en dépendance surface-syntaxique. In *Actes des Journées scientifiques « Linguistique informatique, formelle et de terrain »*, Orléans, France.

Gerdes, K., Guillaume, B., Kahane, S., et Perrier, G. (2021). Starting a new treebank? Go SUD! Theoretical and practical benefits of the Surface-Syntactic distributional approach. In *SyntaxFest Depling 2021-6th International Conference on Dependency Linguistics*. 35-46.

Gerdes, K., Guillaume, B., Kahane, S., Perrier G. (2024), Function words in Surface-Syntactic Universal Dependencies, in T. Osborne (Ed.), *The status of function words in dependency grammar, Linguistic Analysis*, *43(3-4)*. 589-628.

Gross, G. (1998). La fonction sémantique des verbes supports. *Travaux de linguistique*, *37*(1), 25-46.

Guillaume, B. (2021). Graph Matching and Graph Rewriting: GREW tools for corpus exploration, maintenance and conversion. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*. 168-175.

Guillaume, B., de Marneffe, M. C., et Perrier, G. (2019). Conversion et améliorations de corpus du français annotés en Universal Dependencies. *Traitement automatique des langues*, *60*(2), 71-95.

Guillaume, B., Gerdes, K., Guiller, K., Kahane, S., et Li, Y. (2024). Joint Annotation of Morphology and Syntax in Dependency Treebanks. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. 9568-9577.

Habash, N. Y. (2010). *Introduction to Arabic natural language processing*. Morgan et Claypool Publishers.

Habash, N., AbuOdeh, M., Taji, D., Faraj, R., El Gizuli, J., et Kallas, O. (2022). Camel treebank: An open multi-genre Arabic dependency treebank. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC 2022)*. 2672-2681.

Habash, N., Faraj, R., et Roth, R. (2009). Syntactic Annotation in the Columbia Arabic Treebank. In *Proceedings of MEDAR International Conference on Arabic Language Resources and Tools*. 125- 132

Hajič, J., Hajiˇcová, E., Pajas, P., Panevová, J., Sgall, P., et Vidová-Hladká, B. (2001). *Prague Dependency Treebank 1.0*. Linguistic Data Consortium. LDC2001T10.

Hajič, J., Smrž, O., Zemánek, P., Pajas, P., Šnaidauf, J., Beška, E., Kráˇcmar, J., et Hassanová, K. (2004). *Prague Arabic dependency treebank 1.0*. Linguistic Data Consortium. LDC2004T23.

Halabi, D., Fayyoumi, E., et Awajan, A. (2021). I3rab: A new Arabic dependency treebank based on Arabic grammatical theory. *Transactions on Asian and Low-Resource Language Information Processing*, *21*(2). 1-32.

Hernández, R. A. D., et Passarotti, M. C. (2024). Developing the Egyptian-UJaen Treebank. In *Proceedings of the 22nd Workshop on Treebanks and Linguistic Theories (TLT 2024)*. 1-10.

Ibrahim, A. H. (2005). Light verbs in standard and Egyptian Arabic. In M. T. Alhawary et E. Benmamoun (Eds.), *Perspectives on Arabic linguistics : Papers from the annual symposium on Arabic linguistics. Volume XVII–XVIII : Alexandria, 2003 and Norman, Oklahoma 2004*. John Benjamins Publishing Company. 117–131.

Kahane, S. (2001). Grammaires de dépendance formelles et théorie Sens-Texte. In *Actes de la conférence sur le Traitement Automatique des Langues Naturelles (TALN)* (Vol. 2). 1-63.

Kahane, S. (2002). À propos de la position syntaxique des mots qu-. *Verbum*, *24*(4). 399-435.

Kahane, S., Caron, B., Strickland, E., et Gerdes, K. (2021). Annotation guidelines of UD and SUD treebanks for spoken corpora. In *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021)*.

Kahane, S., et Gerdes, K. (2020). Annotation syntaxique du français parlé : Les choix d'ORFÉO. *Langages*, *219*(3), 69-86.

Kahane, S., et Gerdes, K. (2022). *Syntaxe théorique et formelle : Volume 1 : Modélisation, unités, structures*. Language Science Press.

Kahane, S., et Mazziotta, N. (2022). Les corpus arborés avant et après le numérique. *Revue TAL : traitement automatique des langues*, *63*(3), 63-88.

Kahane, S., Gerdes, K., et Courtin, M. (2018). Multi-word annotation in syntactic treebanks: Propositions for Universal Dependencies. In *16th International Conference on Treebanks and Linguistic Theories (TLT)*. 181-189.

Kahane, S., Vanhove, M., Ziane, R., et Guillaume, B. (2022). A morph-based and a word-based treebank for Beja. In *TLT 2021-20th International Workshop on Treebanks and Linguistic Theories*. 48-60.

Lacheret, A., Kahane, S., Beliao, J., Dister, A., Gerdes, K., Goldman, J.-P., Obin, N., Pietrandrea, P. et Tchobanov, A. (2014). Rhapsodie : un Treebank annoté pour l'étude de l'interface syntaxe-prosodie en français parlé. In *Actes du 4e congrès mondial de linguistique française (CMLF)*. 2675–2689

Lafhej, I. (2007). Aspects de variations entre le français et l'arabe : Le cas des subordonnées relatives. *Revue Maghrébine des Langues*, *5*(1). 238-255.

Le Goffic, P. (2002). Marqueurs d'interrogation/indéfinition/subordination : essai de vue d'ensemble. *Verbum*, *24*(4), 315-340.

Luukko, M., Sahala, A., Hardwick, S., et Lindén, K. (2020). Akkadian treebank for early neo-assyrian royal inscriptions. In *International Workshop on Treebanks and Linguistic Theories*. 124-134.

Maamouri, M., Bies, A., Buckwalter, T., et Mekki, W. (2004). The Penn Arabic Treebank: Building a Large-Scale Annotated Arabic Corpus. In *Proceedings of the International Conference on Arabic Language Resources and Tools*. 102–109.

Maamouri, M., Bies, A., Krouna, S., Gaddeche, F., et Bouziri, B. (2011). *Penn Arabic Treebank Guidelines*. Linguistic Data Consortium.

Madkhali, S. A. (2024). Light Verb Constructions in MSA. *Critical Studies in Languages and Literature*, *3*(1). 57-82.

Marcus, M. P., Santorini, B., et Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn treebank. *Computational Linguistics*, *19*(2), 313–330.

McDonald, R. T., Nivre, J., Quirmbach-Brundage, Y., Goldberg, Y., Das, D., Ganchev, K., ... Lee, J. (2013). Universal Dependency Annotation for Multilingual Parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 92–97.

Mel'čuk, I. A. (1988). *Dependency syntax: Theory and practice*. State University of New York Press.

Mel'čuk, I. A. (2004). Verbes supports sans peine. *Lingvisticæ investigationes*, *27*(2), 203-217.

Nivre, J., De Marneffe, M. C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., ... et Zeman, D. (2016). Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. 1659-1666.

Nivre, J., de Marneffe, M.-C., Ginter, F., Hajič, J., Manning, C., Pyysalo, S., Schuster, S., Tyers, F., et Zeman, D. (2020). Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020)*. 4034-4043.

Osborne, T., et Gerdes, K. (2019). The status of function words in dependency grammar: A critique of Universal Dependencies (UD). *Glossa: a journal of general linguistics. 4(1).* 17.1–28.

Petrov, S., Das, D., et McDonald, R. (2012). A universal part-of-speech tagset. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*. 2083-2090.

Pinon, C. (2013). Les valeurs de kâna en arabe contemporain. *Romano-Arabica, 13*. 161–183.

Prévost, S., Grobol, L., Dehouck, M., Lavrentiev, A., et Heiden, S. (2024). Profiterole : un corpus morpho-syntaxique et syntaxique de français médiéval. *Corpus*, *(25)*. https://doi.org/10.4000/corpus.8538

Rosa, R., Masek, J., Marecek, D., Popel, M., Zeman, D., et Zabokrtský, Z. (2014). HamleDT 2.0: Thirty Dependency Treebanks Stanfordized. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*. 2334-2341.

Roubaud, M. N. (2000). *Les constructions pseudo-clivées en français contemporain*. Honoré Champion.

Sag, I. A. (1997). English relative clause constructions. *Journal of Linguistics*, *33*(02). 431-483.

Sanguinetti, M., et Bosco, C. (2015). PartTUT: The Turin University Parallel Treebank. In *Language Resources and Evaluation*. 51-69.

Seddah, D., et Candito, M. (2016). Hard time parsing questions: Building a questionbank for french. In *Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. 376-382.

Smrž, O., Bielický, V., Kouˇrilová, I., Kráˇcmar, J., Hajiˇc, J., et Zemánek, P. (2008). Prague Arabic dependency treebank: A word on the million words. In *Proceedings of the Workshop on Arabic and Local Languages (LREC 2008)*. 16–23.

Smrž, O., Šnaidauf, J., et Zemánek, P. (2002). Prague Dependency Treebank for Arabic: Multi-Level Annotation of Arabic Corpus. In *Proceedings of the International Symposium on Processing of Arabic*. 147-155).

Stephen, A., et Zeman, D. (2024). Light Verb Constructions in Universal Dependencies for South Asian Languages. In *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies*. 163-177.

Taghizadeh, N., Faili, H., et Maleki, J. (2018). Cross-language learning for Arabic relation extraction. *Procedia Computer Science*, *142*. 190-197.

Taji, D., Habash, N., et Zeman, D. (2017). Universal dependencies for Arabic. In *Proceedings of the Third Arabic Natural Language Processing Workshop*. 166-176.

Tesnière, L. (1959). *Éléments de syntaxe structurale*. Klincksieck.

Tuora, R., Przepiórkowski, A., et Leczkowski, A. (2021, November). Comparing learnability of two dependency schemes: 'semantic'(UD) and 'syntactic'(SUD). In *Findings of the Association for Computational Linguistics (EMNLP 2021)*. 2987-2996.

Youssef, S. (2012). *Les relatives : comparaison entre le français, l'arabe classique, l'arabe moderne et l'arabe égyptien*. (Thèse de doctorat, Université de Franche-Comté).

Zahra, S. (2020). Parsing low-resource Levantine Arabic: Annotation projection versus small-sized annotated data.

Zeldes, A., et Abrams, M. (2018). The coptic universal dependency treebank. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*. 192-201.

Zeman, D. (2008). *Reusable tagset conversion using tagset drivers*. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*. 213–218.

Zeman, D., Popel, M., Straka, M., Hajic, J., Nivre, J., Ginter, F., ... Yli-Jyrä, A. (2017). CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. 1–19.

Ziane, R., et Romanova, N. (2024). Pistes pour l'optimisation de modèles de parsing syntaxique. In *LIFT 2-2024: Journées de lancement*, Orléans.