La coordination dans les *treebanks* de dépendance : une approche contrastive arabe-français

العطف في بنوك المشجرات الاعتمادية: مقاربة تقابُليّة بين العربية والفرنسية

Dr. Mohamed Galal
Lecturer of linguistics, Department of French Language, Faculty of Arts, Sohag University
Associate Member, MoDyCo Laboratory (UMR 7114), Paris Nanterre University & CNRS
د. محمد جلال
مدرس اللغويات بقسم اللغة الفرنسية كلية الآداب، جامعة سوهاج
عضو مشارك بمعمل موديكو، جامعة باريس نانتير

Coordination in Dependency Treebanks: A Contrastive Arabic-French Approach

Abstract.

This study aims to evaluate the syntactic annotation of coordination in Arabic and French within the *Universal Dependencies* (UD) and *Surface Syntactic Universal Dependencies* (SUD) projects. It focuses on non-canonical cases (e.g., embedded coordination and non-constituents) and idiosyncratic Arabic constructions (initial *wa*, circumstantial *wa* and accompaniment *wa*). It evaluates the extent to which these two annotation schemes offer relevant solutions for these complex structures. The study reveals consistency for standard constructions but notable divergences for atypical and idiosyncratic cases, even within the treebanks of the same language. These inconsistencies pose a major challenge for both linguistics and Natural Language Processing (NLP). Furthermore, annotation corrections have been proposed for some idiosyncratic Arabic cases to increase coherence across different Arabic treebanks and with those of other languages.

Keywords: Syntactic annotation, Coordination, Universal Dependencies, Surface Syntactic Universal Dependencies, Arabic-French contrastive study

العطف في بنوك المشجرات الاعتمادية: مقاربة تقابُليّة بين العربية والفرنسية

الملخص.

يسعى هذا البحث إلى تقييم الوسم التركيبي لظاهرة العطف في اللغتين العربية والفرنسية، وذلك ضمن إطار مشروعي "الاعتمادات العالمية" (UD) و"الاعتمادات العالمية ذات البنية السطحية" (SUD). ويركز البحث على حالات غير نمطية من تراكيب العطف، مثل العطف المتداخل وعطف المكونات غير المتجانسة، كما يدرس بعض التراكيب العربية الفريدة، مثل واو الابتداء، وواو الحال، وواو المعية. ويهدف من ذلك إلى تقييم مدى فاعلية الوسم التركيبي لهذين المشروعين في تقديم حلول ملائمة لهذه التراكيب المعقدة. وتكشف الدراسة عن اتساق في التراكيب المعيارية للعطف، ولكنها تظهر اختلافات ملحوظة في الحالات غير النمطية والفريدة، وذلك حتى داخل بنوك المشجرات التابعة للغة نفسها. وتمثل هذه التناقضات تحديًا كبيرًا، سواء في مجال اللغويات أو في مجال المعالجات الآلية للغات، وقد اقترحت الدراسة بعض التصويبات في الوسم التركيبي لبعض تراكيب العطف غير النمطية في اللغة العربية، سعيًا لزيادة التجانس بين بنوك المشجرات العربية بعضها البعض ومع بنوك المشجرات في اللغات الأخرى.

الكلمات المفتاحية: الوسم التركيبي، العطف، مشروع الاعتمادات العالمية، مشروع الاعتمادات العالمية ذات البنية السطحية، دراسة تقابُليّة بين العربية والفرنسية

La coordination dans les *treebanks* de dépendance : une approche contrastive arabe-français

1. Introduction

L'importance des corpus arborés, appelés aussi *treebanks*[1], est aujourd'hui incontestable pour le domaine du traitement automatique des langues (TAL). Ils constituent le fondement des modèles d'apprentissage automatique de grammaires et de lexiques, essentiels pour le développement d'outils de fouille de données, traduction automatique, apprentissage assisté par ordinateur, etc. D'un point de vue linguistique, ces *treebanks* permettent la vérification de considérations théoriques en linguistique, et, quant aux *treebanks* parallèles, en traductologie et en typologie des langues.

Bien que la coordination ait fait l'objet de nombreuses recherches (voir entre autres Al Khalaf *et al.*, 2024 ; Biskri et Bensaber, 2008 ; Boukedi et Haddar, 2014 pour l'arabe ; Abeillé, 2005 ; Abeillé et Mouret, 2010 ; Gerdes et Kahane, 2015 ; Mouret, 2007 ; 2008 pour le français), ce phénomène continue de poser un défi majeur pour la validation de tout modèle de TAL. L'objectif de cet article sera donc d'examiner l'annotation syntaxique de la coordination dans une perspective contrastive arabe-français. Notre examen portera précisément sur l'annotation de la coordination telle qu'elle est représentée dans les *treebanks* du projet *Universal Dependencies*[2] (UD) (Nivre *et al.*, 2016, 2020 ; de Marneffe *et al.*, 2021) et ceux du projet parallèle *Surface-Syntactic Universal Dependencies*[3] (SUD) (Gerdes *et al*. 2018, 2019 ; 2021 ; 2024). Ces deux projets offrent un cadre d'annotation morpho-syntaxique basé sur l'approche de la syntaxe de dépendance (Tesnière, 1959 ; Mel'čuk, 1988 ; Kahane et Gerdes, 2022) et conçu pour être applicable à toutes les langues. La distinction majeure entre le schéma UD et celui de SUD est que le premier est basé sur la syntaxe profonde et favorise les mots lexicaux comme tête des dépendances, alors que le second est basé sur la syntaxe de surface et opte pour des têtes fonctionnelles. Dans ses dernières versions publiées en mai 2025, les projets UD et SUD mettent à disposition plus de 300 *treebanks*, pour plus de 170 langues, dont 3 pour l'arabe standard moderne[4] et 8 pour le

français[5]. Pour une présentation détaillée des principes et caractéristiques des projets UD et SUD, une comparaison de leurs schémas d'annotation, ainsi qu'un inventaire des *treebanks* disponibles pour l'arabe et le français, voir Galal (à paraître).

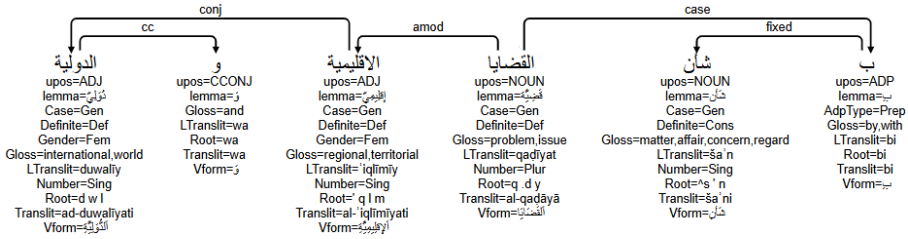
En plus des cas standards de la coordination, cette recherche met l'accent sur certains cas non canoniques des structures coordonnées, comme la coordination enchâssée et la coordination de non-constituants. Elle examine également certaines constructions idiosyncrasiques de l'arabe où le *wa* ('et') ne sert plus de conjonction de coordination, mais joue un rôle syntaxique différent au sein de la phrase, comme le *wa* initial, le *wa* circonstanciel et le *wa* d'accompagnement. Le but est d'évaluer si les deux schémas d'annotation UD et SUD ont fourni des solutions pertinentes et cohérentes pour la représentation de ces structures syntaxiques complexes. Afin de répondre à cette problématique, notre investigation se penchera sur les questions suivantes : i) Comment les deux schémas isomorphes se différencient-ils et se rejoignent-ils dans leur traitement de la coordination ? ii) Dans quelle mesure les *treebanks* arabes et français offrent-ils une analyse cohérente de la coordination ? iii) Les deux schémas d'annotation UD et SUD peuvent-ils saisir avec précision les phénomènes des structures coordonnées atypiques et idiosyncrasiques de l'arabe ?

Le plan de l'article est le suivant : nous aborderons successivement l'analyse de l'annotation de la coordination standard à deux ou à plusieurs conjoints (§2), la coordination enchâssée (§3), la coordination de non-constituants (§4). Ensuite, nous aborderons certains cas idiosyncrasiques de l'arabe (§5) où le *wa* ne fonctionne pas comme un simple coordonnant : le cas *wa* initial (§5.1), le cas de *wa* circonstanciel (§5.2) et le cas de *wa* d'accompagnement (5.3).

2. La coordination standard

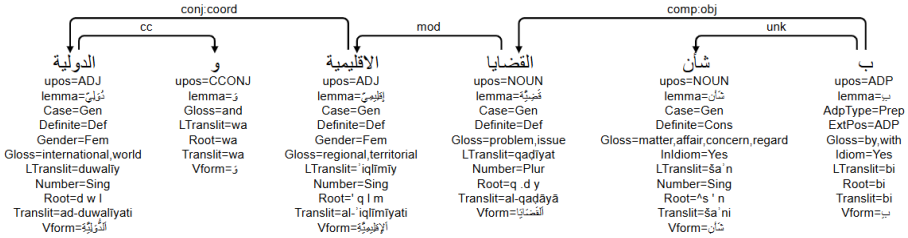
Que ce soit pour l'arabe[6] ou pour le français, la coordination à deux conjoints est traitée de manière identique dans l'ensemble des *treebanks*[7] UD et SUD (figures 1 et 3)[8]. La seule différence est le nom de la relation : `conj` (conjonction) dans UD et `conj` avec l'extension : `coord` dans SUD (figure 2). Comme le montre l'annotation

de l'exemple en arabe, la relation `conj` rattache la tête du premier conjoint *al-ʾiqlīmiyyah* ('régionales') à la tête du deuxième conjoint *ad-dawliyyah* ('internationales') et la relation `cc` (conjonction de coordination) est tiré de la tête du deuxième conjoint vers la conjonction de coordination *wa* ('et').



'[…] concernant les questions régionales et internationales'[9]

Figure 1 : UD_Arabic-PADT@2.15



'[…] concernant les questions régionales et internationales'

Figure 2 : SUD_Arabic-PADT@2.15

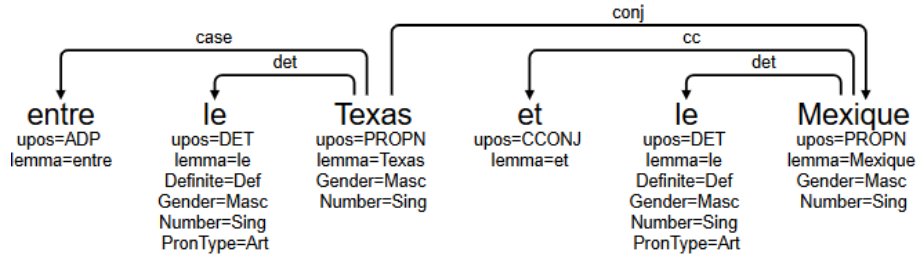


Figure 3 : UD_French-FQB@2.15

Cette analyse est motivée par un certain nombre de critères (cf. Gerdes *et al*. 2024 ; Gerdes et Kahane, 2015) :

Le critère i) est l'identification de *l'unité syntaxique*[10]. Prenons cet exemple de la coordination en français : « *entre le Texas et le Mexique* ». Le segment « *et le Mexique* » peut constituer à lui seul un énoncé valide (dans un tour de parole lors d'un dialogue par exemple), ce qui n'est pas le cas de « **le Texas et* ». En conséquence, dans cet exemple précis, on

peut distinguer les unités syntaxiques suivantes : le segment « *et le Mexique* », les conjoints seuls « *le Texas* » et « *le Mexique* », mais aussi l'unité complète « *le Texas et le Mexique* ».

Le critère ii) concerne *l'identification de la tête syntaxique*[11]. Au sein des unités syntaxiques précédemment identifiées, il s'agit de déterminer quel mot en est la tête syntaxique. Cette étape s'appuie sur le *critère distributionnel avec effacement positif*, postulant que, dans une unité syntaxique U :

> Si U = AB, que A peut apparaître seul (c'est-à-dire que B peut être effacé), et que U et A ont la même distribution, alors A est la tête de U (Gerdes *et al.* 2024).

En comparant les unités « *le Texas »* et « *et le Mexique* », le critère distributionnel avec effacement positif est clairement vérifié (« *le Texas* » et « *le Texas et le Mexique* » ont la même distribution), d'où il ressort que « *le Texas* » est la tête.

L'identification de la tête dans l'unité « *et le Mexique* » est cependant problématique. En effet, la conjonction « *et* » ne peut fonctionner de manière autonome et la distribution de « *le Mexique* » diffère de celle de l'ensemble « *et le Mexique* », démontrant que les deux éléments en contrôlent la distribution. Par conséquent, il est nécessaire d'appliquer un critère alternatif, à savoir, *le critère distributionnel avec effacement négatif* qui stipule que :

> Si U = AB, que B peut apparaître seul, et que U et B n'ont pas la même distribution, alors A est la tête de U (Gerdes *et al.* 2024).

Selon ce critère, « *et* » est la tête syntaxique. Mais en même temps, la distribution du segment dépend également du conjoint « *le Mexique* ». Par exemple, « *et le Mexique* » peut être combiné avec un nom propre dans « *le Texas et le Mexique* », mais pas avec d'autres catégories comme un adjectif « **ancien et le Mexique* ». Cela suggère que les deux éléments pourraient être considérés comme têtes. Le choix final a été de privilégier la relation entre les conjoints « *le Texas* » et « *le*

Mexique ». Cette décision se justifie par le fait que cette relation existe même lorsque la conjonction de coordination est absente : « *le Texas, la Louisiane et le Mexique* ».

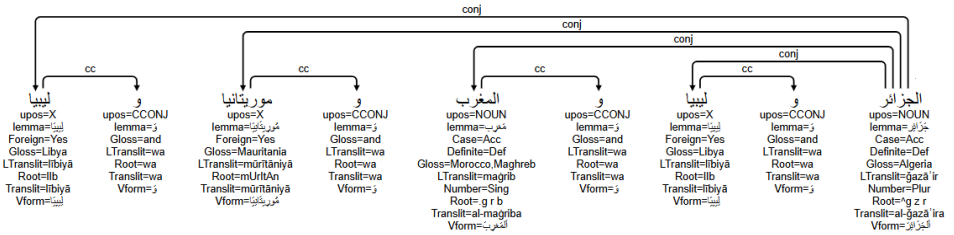
Notons qu'en arabe, la conjonction de coordination est systématiquement placée devant le deuxième conjoint et chacun des conjoints subséquents[12][1]. La construction arabe peut également appuyer cette analyse de la coordination, puisque seule l'analyse avec la conjonction de coordination comme dépendant peut faire de *ʾaḥmad* la tête du syntagme coordonné et fournir une analyse symétrique de *wa=ḥasan* et *wa=saʿd*.

[1] قابلتُ أحمد وحسن وسعد

qābaltu	*ʾaḥmad*	*wa=ḥasan*	*wa=saʿd*
rencontrer.PST.1SG	Ahmad	COORD=Hassan	COORD=Saad

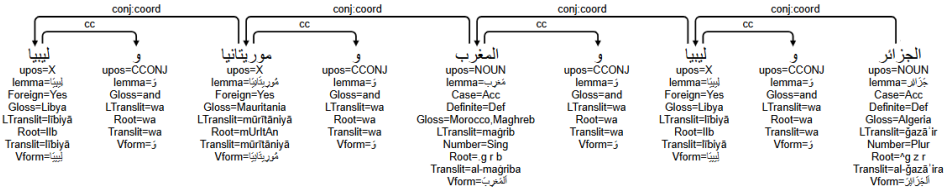
'J'ai rencontré Ahmad, Hassan et Saad'[13]

Dans le cas de la coordination itérée, SUD et UD présentent des analyses différentes. Alors que UD opte pour une analyse en structure de bouquet (figure 4), où tous les autres conjoints sont rattachés à la tête du premier conjoint, SUD adopte une analyse en chaîne (figure 5). Dans ce dernier cas, la tête de chaque conjoint est liée à celle du précédent (le deuxième au premier, le troisième au deuxième, etc.). Pour les concepteurs de SUD, l'argument principal en faveur de cette analyse est de réduire la longueur des dépendances, ce qui permet une meilleure compréhension du processus mental de la coordination (Gerdes *et al.* 2024).



'L'Algérie, la Libye, le Maroc, la Mauritanie et le Liban,

Figure 4 : UD_Arabic-PADT@2.15



‘L’Algérie, la Libye, le Maroc, la Mauritanie et le Liban’
Figure 5 : SUD_Arabic-PADT@2.15

En résumé, les données présentées, dans cette section, montrent que le choix d’annotation de la coordination standard est cohérent entre UD et SUD pour les deux langues, l’arabe et le français. La divergence entre les deux schémas apparaît uniquement lorsqu’il s’agit d’analyser la coordination standard itérée.

3. La coordination enchâssée

La coordination enchâssée correspond à une construction dans laquelle un des éléments coordonnés est lui-même une coordination, du type [[A], [[B] ou [C]] et [D]]. Cette configuration crée une relation hiérarchique, introduisant ainsi un niveau d’organisation syntaxique supplémentaire (cf. Gerdes *et al*., 2024 ; Kahane, 2013). Dans l’exemple [2] en arabe, le second segment [*al-ʾinğilīziyyah ʾaw=al-faransiyyah*] (‘l’anglais ou le français’) est une coordination enchâssée à l’intérieur de la coordination principale [*al-ṣīniyyah* (‘le chinois) et […]].

[2] سوف يدرس أحمد [الصينية] و[[الإنجليزية] أو [الفرنسية]]

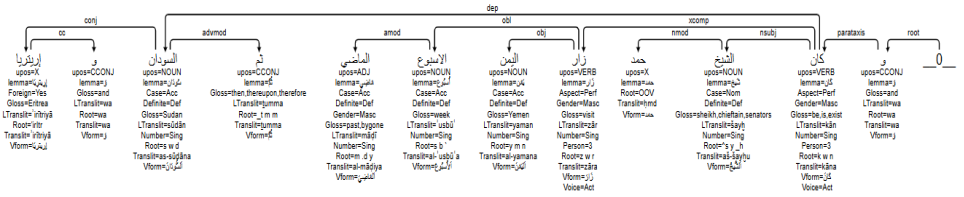
sawfa	*yadrisu*	*ʾaḥmad*	*[al-ṣīniyyah]*	*wa=[[al-ʾinğilīziyyah]*
FUT	étudier.PRS.3SG	Ahmad	DEF-chinois	COORD=DEF-anglais

ʾaw=[al-faransiyyah]]
COORD=DEF-français
‘Ahmad étudiera [le chinois] et [[l’anglais] ou [le français]]’

L’exemple attesté ci-dessous, tiré du *treebank* UD Arabic-PADT (figure 6) et sa représentation dans SUD (figure 7), illustre le cas de la coordination enchâssée. Nous pensons que le segment est mal analysé : la conjonction de coordination *ṯumma* (‘puis’, ‘ensuite’) est ici analysée comme modifieur adverbial `advmod`. De plus, la relation `dep` (unspecified dependency), qui est attribuée lorsqu’il est impossible de déterminer une relation plus précise, rattache le prédicat au gouverneur de la conjonction. L’analyse standard que UD propose consiste donc à relier la conjonction de coordination *ṯumma* par la relation `cc`. Le premier

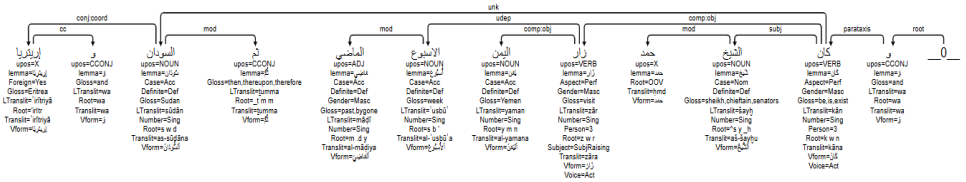
conjoint, *al-yaman* (‘le Yémen’), serait, lui, rattaché aux autres conjoints, *al-sūdān* (‘le Soudan’) et *ʾiritrīyā* (‘l’Érythrée’), par la relation `conj`.

Le fait que l’étiquette `conj` de SUD ne distingue pas les relations enchâssées des relations de surface, car elles forment une seule chaîne, le schéma SUD a introduit l’extension `@emb` (embedded) pour distinguer les coordinations enchâssées (Gerdes *et al*. 2024, p. 622), comme l’illustre l’exemple en anglais (figure 8).



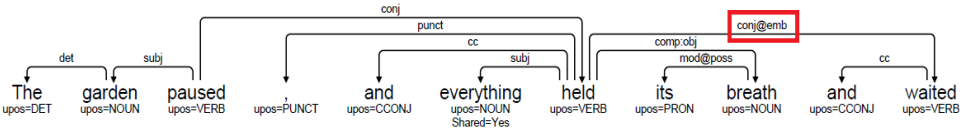
‘Le cheikh Hamad avait visité le Yémen la semaine dernière, puis le Soudan et l’Érythrée’.

Figure 6 : UD_Arabic-PADT@2.15



‘Le cheikh Hamad avait visité le Yémen la semaine dernière, puis le Soudan et l’Érythrée’.

Figure 7 : SUD_Arabic-PADT@2.15



‘Le jardin fit une pause, et toute chose retint son souffle et attendit’

Figure 8 : exemple en anglais dans la représentation SUD de la coordination enchâssée tiré de Gerdes *et al.* (2024)

La relation `conj@emb` sera donc utilisée pour relier les conjoints enchâssés *al-sūdān* (‘le Soudan’) et *ʾiritrīya* (‘l’Érythrée’) dans l’exemple en arabe.

4. La coordination de non-constituants

Une propriété fondamentale des structures coordonnées est leur capacité à relier deux ou plusieurs séquences de catégories non syntagmatiques. Ce phénomène est connu sous le nom de *coordination de*

non-constituants ou *de séquences*. La littérature existante identifie trois catégories de ce type de coordination (cf. Mouret, 2008 ; Gerdes et Kahane, 2015) : i) *la coordination de séquences de compléments* : il s'agit de la coordination de plusieurs compléments (qu'ils soient argumentaux ou modifieurs) qui suivent un même prédicat [3a] ; ii) *la coordination à montée du nœud droit* : où des séquences incomplètes sont coordonnées avant un constituant partagé à droite [3b]; iii) *la coordination elliptique* ou *à gapping* : où des phrases complètes sont coordonnées, avec une suppression d'un élément (généralement le verbe) dans la ou les propositions suivantes [3c].

[3] a. يسافر أحمد إلي بيروت في الربيع وإلى باريس في الصيف

yusāfiru	*ʾaḥmad*	*ʾilā*	*bayrūt*	*fī*	*al-rabīʿ*
voyager.PRS.3SG	Ahmad	PREP	Beyrouth	PREP	printemps

wa=ʾilā	*bārīs*	*fī*	*as-ṣayf*
COORD=PREP	Paris	PREP	DEF-été

'Ahmad voyage à Beyrouth au printemps et à Paris en été'

b. هذه الفاتورة ضرورية لاسترداد أو لتبديل السلع

haḏihi	*al-fātūrah*	*ḍarūriyyah*	*li=istirdād*	*ʾaw*
DEM	DEF-facture	nécessaire	PREP=remboursement	COORD

li=tabdīl	*as-silaʿ*
PRP=échange	DEF-article.PL

'Cette facture est nécessaire pour un remboursement ou un échange des articles'

c. يدرس أحمد اللغة الفرنسية وحسن اللغة الألمانية

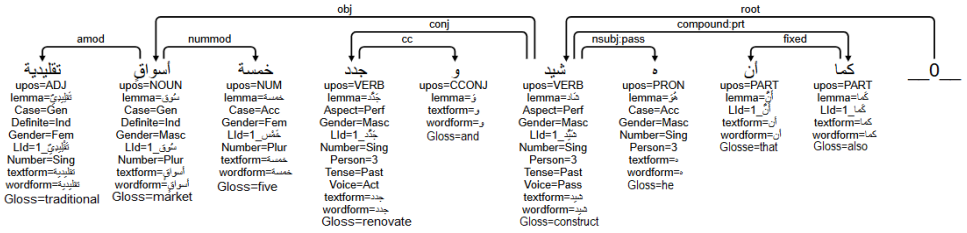
yadrisu	*ʾaḥmad*	*al-luġah*	*al-faransiyyah*	*wa=ḥasan*
étudier.PRS.3SG	Ahmad	DEF-langue	DEF-français	COORD-Hassan

al-luġah	*al-ʾalmāniyyah*
DEF-langue	DEF-allemand

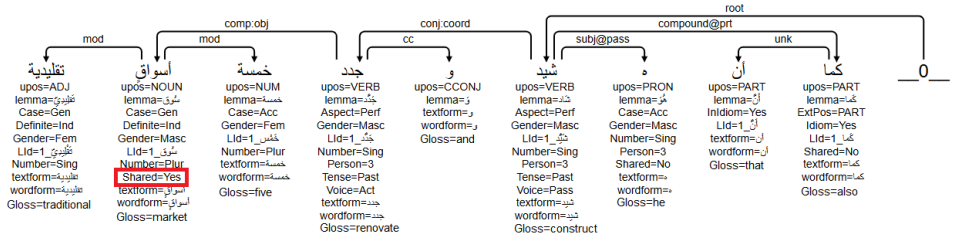
'Ahmad étudie la langue française et Hassan la langue allemande'

L'analyse de *la coordination à montée du nœud droit* dans UD Arabic-PUD (figure 9) choisit de rattacher le dépendant de la coordination, *ʾaswāq* ('matchés') dans l'exemple, à la tête de la coordination *šayyada* ('a construit'), qui est la tête du conjoint le plus à droite. Pour rester cohérent avec l'analyse de la coordination en tant que chaîne, SUD analyse le dépendant de la coordination en tant que dépendant de la tête du conjoint le plus à gauche *ǧaddada* ('a rénové'). Le problème avec cette analyse est qu'elle ne permet pas de distinguer entre un dépendant du premier conjoint et un dépendant de la structure entière. Pour différencier entre un dépendent uniquement au conjoint le plus

proche et un dépendant partagé par d'autres conjoints, un trait morphosyntaxique spécifique, *Shared=Yes*, a été introduit pour marquer ce dernier (Gerdes *et al.* 2024, pp. 622-623) (figure 10).



'Il a également construit et rénové cinq marchés traditionnels'
Figure 9 : UD_Arabic-PUD@2.15[14]

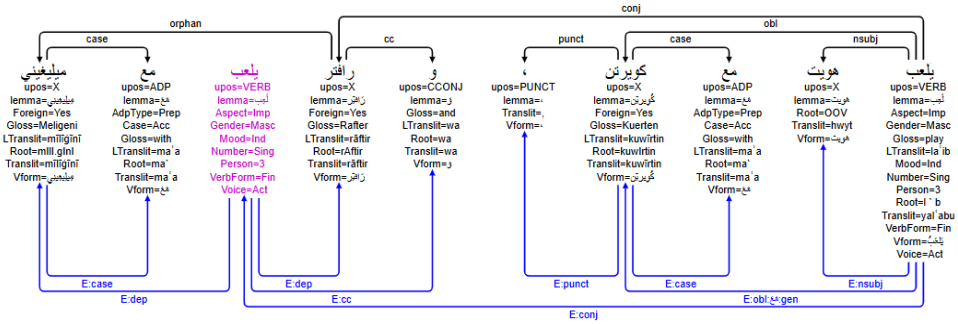


'Il a également construit et rénové cinq marchés traditionnels'
Figure 10 : SUD_Arabic-PUD@2.15

Dans le contexte de *la coordination elliptique*, l'ellipse du prédicat a été annotée en utilisant la relation `orphan`, dédiée à l'analyse des cas spécifiques d'ellipse du prédicat. Elle identifie donc un dépendant qui se retrouve sans lien de dépendance suite à la suppression du gouverneur dont il dépend initialement.

Dans la représentation UD (figure 11), l'argument principal du verbe elliptique *rāftar* ('Rafter') est promu pour devenir la tête de la seconde proposition coordonnée et se rattache donc au prédicat de la première proposition *yalʿab* ('joue') par la relation `conj`. Cet élément promu *rāftar* ('Rafter') gouverne lui-même la tête du complément verbal par la relation `orphan`. En plus de leur structure de dépendances de base, certains *treebanks* UD, dont Arabic-PADT, sont enrichis d'une strate d'annotation supplémentaire, appelée les dépendances enrichies (Enhanced Dependencies[15]). Celles-ci ont pour vocation de clarifier des liens syntaxiques et sémantiques qui sont sous-entendus ou non spécifiés dans l'annotation UD standard. Dans les cas d'ellipse, les dépendances enrichies peuvent ajouter des éléments vides « *nœuds nuls* » pour les

prédicats manquants. Ces nœuds, bien que ne correspondant pas à un mot de surface, peuvent se voir attribuer des valeurs (lemme, UPOS, trait, etc.) copiées de l'occurrence explicite du même prédicat. Les dépendances enrichies sont affichées en bleu sous la phrase. La même analyse UD avec ses principes de base est adoptée en français (figure 12).



'Hewitt joue avec Kuerten et Rafter {joue} avec Meligeni'
Figure 11 : UD_Arabic-PADT@2.15

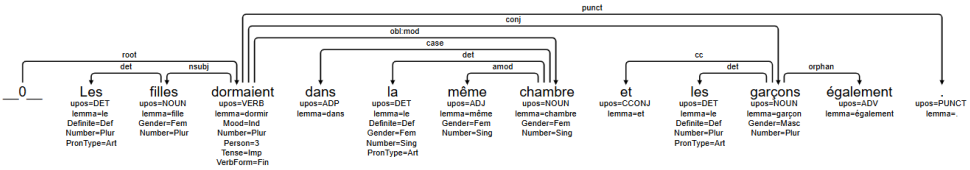
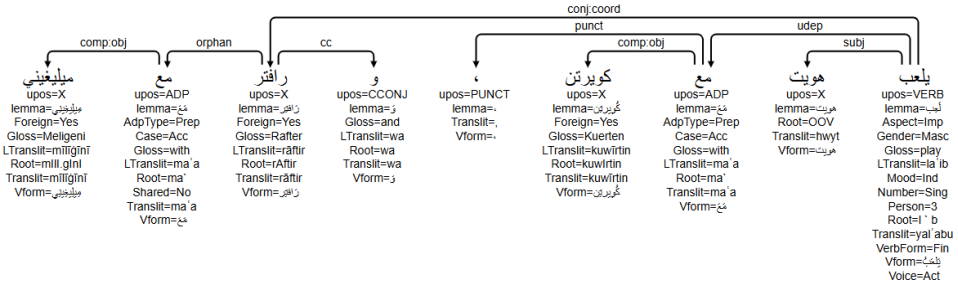


Figure 12 : UD_French-GSD@2.15

Par ailleurs, SUD adopte l'analyse standard de UD (figure 13). La seule différence est que la relation `orphan` rattache l'élément promu *rāftar* ('Rafter') avec la tête fonctionnelle du complément verbal dans la deuxième proposition (la préposition *maʿa* ('avec')), plutôt qu'au mot lexical *mīlīġīnī* (cf. les principes UD vs. SUD §1).



'Hewitt joue avec Kuerten et Rafter avec Meligeni'
Figure 13 : SUD_Arabic-PADT@2.15

Pour résumer, dans le cas de *la coordination à montée du nœud droit*, UD et SUD divergent : UD annote le dépendant de la coordination à la tête du conjoint le plus à droite, alors que SUD l'attribue à la tête du conjoint le plus à gauche. SUD introduit la trait *Shared=Yes* pour distinguer les dépendants partagés. En revanche, les deux schémas s'accordent sur l'utilisation de la relation `orphan` pour *la coordination elliptique*.

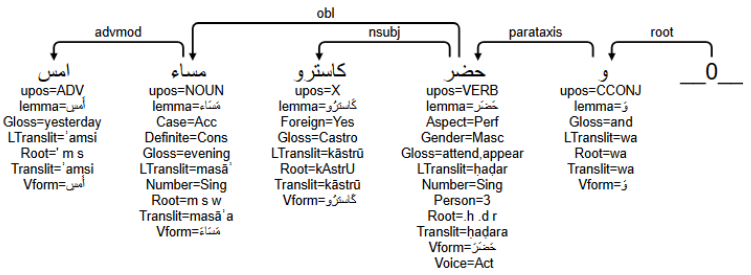
5. Les cas idiosyncrasiques de l'arabe

Nous allons nous pencher ici sur des emplois particuliers du *wa* en arabe. Dans ces cas, le *wa* ne sert plus de conjonction de coordination, mais joue un rôle syntaxique différent au sein de la phrase.

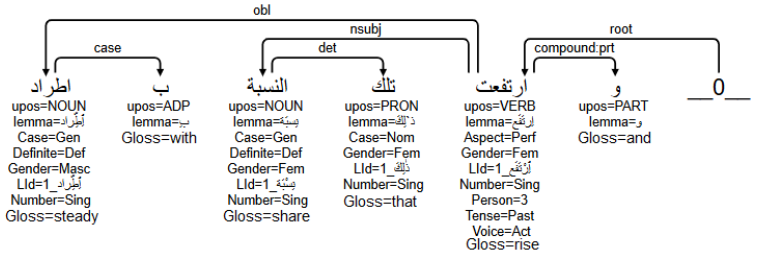
5.1 Le *wa* initial

En arabe, il est fréquent d'employer les conjonctions de coordination, notamment le *wa* ('et') et le *fa* ('et', 'puis', 'donc', etc..), comme connecteur discursif à la position initiale d'une phrase[16] pour établir une certaine cohésion discursive. Il s'agit dans ce cas de ce que la grammaire traditionnelle arabe appelle le *wa/fa ʾal-ʾistiʾnāfiyya ʾaw al-ʾibtidāʾiyya* ('le *wa*/*fa* de reprise ou initial) (Yaʿqūb, 2006, p. 404). Contrairement aux conjonctions de coordination canoniques qui relient deux éléments de même nature et fonction, ces unités n'établissent nécessairement pas une coordination syntaxique stricte avec la phrase ou la proposition précédente.

Dans le cas de *wa* initial, les *treebanks* arabes adoptent des analyses différentes. Dans UD Arabic-PADT, le *wa* initial prend la catégorie CCONJ (conjonction de coordination). Il est la racine de la phrase et gouverne le verbe par la relation `parataxis`, qui sert à analyser les constructions où des propositions sont combinées par des relations moins contraignantes que la coordination standard (figure 14). UD Arabic-PUD le catégorise comme PART (particule). Il est rattaché au verbe par la sous-relation `compound:prt`, destinée aux particules verbales affixées ou non à un verbe (figure 15).



‘[et] Castro a assisté hier soir à […]’
Figure 14 : UD_Arabic-PADT@2.15



‘[et] ce pourcentage a augmenté de manière constante’
Figure 15 : UD_Arabic-PUD@2.15

En français écrit, cet usage est assez rare, contrairement au français parlé où il est fréquent. Dans les *treebanks* du français parlé (UD French-Rhapsodie, UD French-ParisStories), l’analyse adoptée est celle de rattacher le *et* initial, catégorisé comme CCONJ, au verbe par la relation `cc`, sans établir de relation `conj` (figure 16).

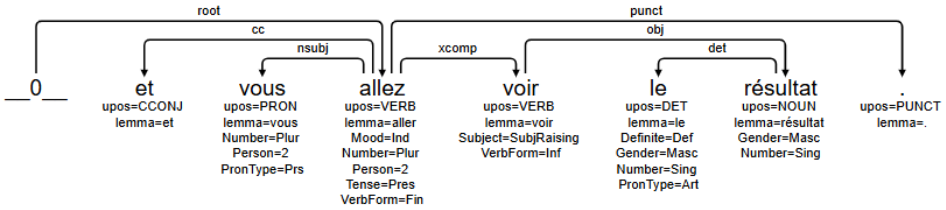


Figure 16 : UD_French-Rhapsodie@2.15

Nous suggérons d’appliquer au *wa* initial en arabe l’analyse déjà employée pour la conjonction initiale en français. En effet, cette approche présente plusieurs avantages : i) elle maintient la catégorie CCONJ pour le mot. Cela confirme son rôle en tant que connecteur grammatical, même s’il ne s’agit pas d’une coordination syntaxique stricte avec la phrase ou la proposition précédente. ii) En évitant d’établir la relation `conj`, on reconnaît que ces conjonctions en début de phrase fonctionnent comme marqueur discursif. Leurs fonctions sont de structurer le discours, de

marquer une transition ou une continuité, plutôt que de relier des éléments syntaxiquement équivalents. iii) Le choix systématique de la relation `cc` plutôt que `parataxis` dans ce contexte précis est un impératif pour une annotation optimale et conforme aux principes de (S)UD. La relation `cc` est spécifiquement destinée aux coordinations explicites avec une conjonction. En revanche, la relation `parataxis` est mieux réservée aux cas où il n'y a pas de conjonction explicite ou lorsque le lien est purement discursif et non marqué par un connecteur grammatical formel. Comme le souligne le guide UD, `parataxis` couvre les juxtapositions de phrases, les interjections, le discours direct, les reformulations ou les liens sémantiques lâches entre propositions indépendantes.

5.2 Le *wa* circonstanciel

Un autre emploi particulier de la conjonction de coordination *wa* en arabe est le *wa* dit *wāw al-ḥāl* (le '*waw* de l'état ou le '*waw* circonstanciel'). Cette unité introduit une proposition subordonnée exprimant les circonstances (état ou manière). Il permet donc de joindre une action à la circonstance ou à l'état dans lequel elle se déroule (cf. As-Sāmurrāʾiy, 2000, pp. 296-307). Typiquement, une proposition circonstancielle introduite par *wāw al-ḥāl* est soit une construction prédicative [4a], soit une construction verbale avec un verbe à l'inaccomplie, notamment avec un pronom se référant au sujet ou à l'objet de la proposition principale [4b], ou à l'accompli précédé par la particule *qadd* ('déjà') [4c].

[4] a. غادر أحمد والشوارع مزدحمة

ġādara *ʾaḥmad* *wa=ʾš-šawāriʿ* *muzdaḥimah*
partir.PST.3SG Ahmad COORD=DEF-rue.PL bondée.PL
Litt. Ahmad est parti et les rues {étaient} bondées
'Ahmad est parti alors que les rues étaient bondées'

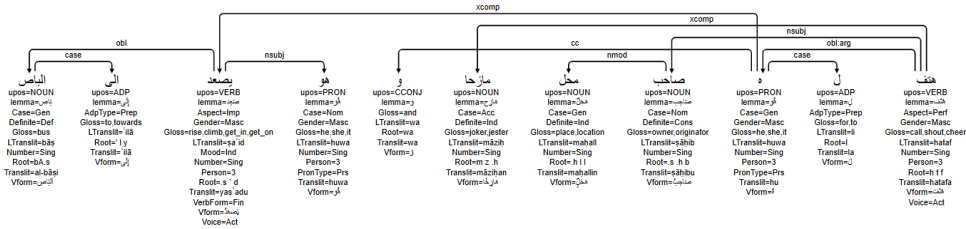
b. أقبل أحمد وهو يبتسم

ʾaqbala *ʾaḥmad* *wa=hūwa* *yabtasim*
s'approcher.PST.3SG Ahmad COORD=PRO sourir.PRS.3SG
Litt. Ahmad s'est approché et il souriait.
'Ahmad s'est approché alors qu'il souriait/Khaled s'est approché en souriant'

c. رأيت أحمد وقد فاز

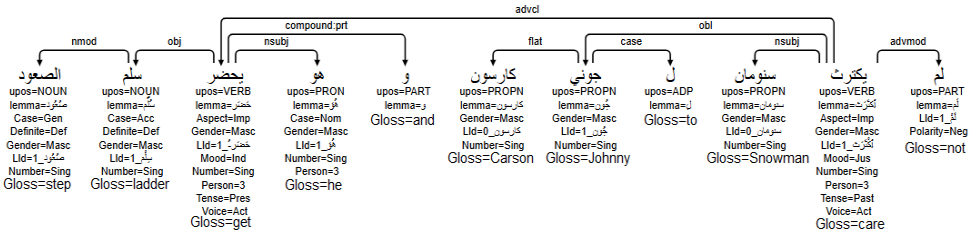
raʾaytu *ʾaḥmad* *wa=qadd* *fāz*
voir.PST.1SG Ahmad COORD=PART gagner.PST.3SG
Litt. J'ai vu Ahmad et il a déjà gagné.
'J'ai vu Ahmad alors qu'il avait gagné'

Les *treebanks* UD Arabic-PADT et UD Arabic-PUD présentent des analyses divergentes pour le *wāw al-ḥāl*, même si les deux le catégorisent toujours comme une conjonction de coordination (CCONJ). Dans UD Arabic-PADT (figure 17), le *wāw al-ḥāl* est systématiquement classé comme CCONJ. L'objet référent du sujet de la proposition subordonnée gouverne le *wa* via la relation `cc`, et le verbe de cette même subordonnée est rattaché par une relation `xcomp` (open clausal complement). En revanche, dans le *treebank* UD Arabic-PUD (figure 18), le *wa* est catégorisé comme une particule (PART). Il dépend du verbe de la proposition subordonnée, tandis que le verbe de la proposition principale gouverne le verbe de la subordonnée par la relation `advcl` (adverbial clause modifier). Il est à noter que cette relation `advcl` est simplement désignée par `mod` en SUD.



'[…] le propriétaire d'un magasin lui a crié en plaisantant alors qu'il montait dans le bus'

Figure 17 : UD_Arabic-PADT@2.15



'Snowman ne prêta aucune attention à Johnny Carson alors qu'il préparait l'échelle d'accès'

Figure 18 : UD_Arabic-PUD@2.15

Nous proposons d'adopter cette deuxième analyse du *treebank* UD Arabic-PUD pour le *wāw al-ḥāl*, avec une modification de rattacher toujours le *wa* par la relation `cc` et non `compound:part`. Cette approche permet de reconnaître la nature conjonctive du *wa*, tout en capturant la relation adverbiale `advcl` entre la proposition principale et la proposition de l'état. Cette combinaison garantit une meilleure

représentation de la structure sémantique et syntaxique de ces constructions complexes.

5.3 Le *wa* d'accompagnement

En se basant sur l'analyse de 260 langues, Stassen (2000) a établi une distinction typologique entre deux catégories de langues : les *AND-languages* et les *WITH-languages*. Dans les langues de la première catégorie, la conjonction de coordination *and* qui relie deux SN et la préposition comitative *with* sont lexicalement distincts. Alors que les langues de la deuxième catégorie fusionnent ces fonctions en un seul et même terme. L'arabe appartient à cette dernière catégorie, où la conjonction *wa* sert à indiquer l'accompagnement [5].

[5] سارَ أحمدُ والنهرَ

sāra	*ʾaḥmad-u*	*wa=l-nahr-a*
marcher.PST.3SG	Ahmad-NOM	et=DEF-rivière-ACC

'Ahmad a marché avec/le long de la rivière'

Le *wa* dans ce cas, appelé *wāw al-maʿiyyah* ('le *wa* d'accompagnement') dans la grammaire traditionnelle arabe (Al-Hamd et Al-Zaaby, 1993, p. 352), n'est pas une conjonction de coordination, mais plutôt un connecteur comitatif. En effet, dans l'exemple en [5], les deux éléments, que le *wa* relie, ne partagent pas les mêmes propriétés sémantiques ni le même statut syntaxique (El Kassas, 2005, p. 164). Le nom qui dépend du *wa* d'accompagnement (appelé *mafʿūl maʿahu* 'complément comitatif' dans la grammaire traditionnelle) porte systématique le cas accusatif.

Le *treebank* UD Arabic-PADT parvient à rendre compte avec précision de la distinction entre le *wa* coordonnant et le *wa* d'accompagnement. Dans le segment (figure 19), tiré de l'exemple en [6], l'élément qui suit le *wa* d'accompagnement *raʾīsa* ('le président') se rattache au verbe par la sous-relation `obl:arg` (oblique argument) qui rattache les arguments sélectionnés par le verbe, mais qu'il ne s'agit ni d'un sujet `nsubj`, `csubj` ni d'un objet direct `obj`. Dans les dépendances enrichies, l'extension `:acc` a été ajoutée pour marquer le cas accusatif. Le *wa* d'accompagnement lui-même est rattaché à ce complément comitatif par la relation `case`, consacré aux adpositions (prépositions ou postpositions), tout en gardant la catégorie grammaticale

CCONJ. Une étude approfondie pourrait se pencher sur le statut catégoriel de *wa* d'accompagnement en arabe : s'agit-il d'une préposition ou d'une conjonction de subordination ?

[6] وكان الأمين العام للحزب الوطني وزير الإعلام صفوت الشريف التقى عصر يوم 10/27 الحالي ورئيس حزب الوفد الدكتور نعمان جمعة

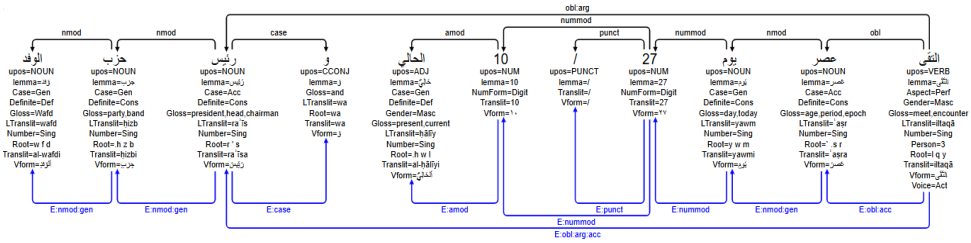
wa=kāna	*al-ʾamīn*	*al-ʿām*	*li=l-ḥzb*	*al-waṭanī*
COORD=AUX	DEF-Secrétaire	DEF-général	PREP=DEF-Parti	DEF-National

wazīr	*al-ʾiʿlām*	*ṣafwat al-šarīf*	*iltqā*	*ʿaṣra*
Ministre	DEF-Information	Safwat Al-Sherif	rencontrer.PST.3SG	après-midi

yawm	*27/10*	*al-ḥālī*	*wa=rʾīsa*	*ḥizb*	*al-wafd*
jour	27/10	courant	avec=président	Parti	DEF-Wafd

ad-duktūr	*nuʿmān*	*ǧumuʿah*
DEF-docteur	Numan	Gomaa

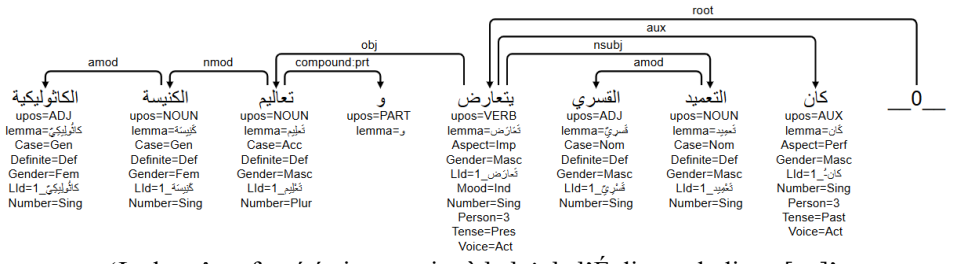
'Le Secrétaire général du Parti National et Ministre de l'Information, Safwat Al-Sherif, a rencontré dans l'après-midi du 27/10 courant le président du Parti Al-Wafd, Dr. Numan Gomaa'



'[…] a rencontré dans l'après-midi du 27/10 courant le président du Parti Al-Wafd […]'
Figure 19 : UD_Arabic-PADT@2.15

Parmi les 710 occurrences du morphème *wa* dans le corpus UD Arabic-PUD, un seul cas de *wa* d'accompagnement a été identifiée (voir figure 20). On observe de nouveau une divergence de traitement entre les différents *treebanks* arabes. Dans ce cas précis, le *wa* d'accompagnement est classé comme une PART (particule). Il dépend du complément comitatif via la relation `compound:part`, lequel est lui-même régi par le verbe via la relation d'objet direct `obj`. Nous nous opposons à cette analyse considérant le nom figurant après le *wa* d'accompagnement comme `obj`. Bien que ce nom porte systématiquement le cas accusatif, un trait partagé avec les objets directs, il ne remplit pas pour autant les critères d'un argument central du verbe. L'élément introduit par le *wa* d'accompagnement ne participe pas activement à l'action verbale de la même manière que le sujet ou l'objet direct. Si l'objet direct est le patient

de l'action, l'élément après le *wa* d'accompagnement ne subit pas directement l'action du verbe. Son rôle est davantage celui d'une circonstance qui accompagne l'action principale.



'Le baptême forcé était contraire à la loi de l'Église catholique […]'

Figure 20 : UD_Arabic-PUD@2.15

L'analyse que nous suggérons d'adopter pour la construction introduite par le *wa* d'accompagnement en arabe est donc celle proposée par le *treebank* UD Arabic-PADT, telle qu'elle a été décrite précédemment.

Conclusion

Notre contribution dans cette étude a consisté à explorer l'annotation syntaxique du phénomène de la coordination en arabe et en français, en exploitant les *treebanks* des projets *Universal Dependencies* (UD) et *Surface Syntactic Universal Dependencies* (SUD). La recherche a mis l'accent sur certains cas non canoniques et idiosyncrasiques des structures coordonnées dans les deux langues étudiées. L'objectif était d'évaluer dans quelle mesure les deux schémas d'annotation proposent des solutions pertinentes pour ces structures complexes.

Il ressort de l'étude que les deux schémas d'annotation présentent des divergences quant à la manière de rendre compte de ces phénomènes (cf. la coordination standard itérée, la coordination enchâssée, la coordination de non-constituants). Ces écarts résultent des divergences fondamentales dans les principes sur lesquels reposent ces deux schémas. SUD s'avère, d'ailleurs, capable de proposer des solutions plus nettes et solides pour les structures où les règles de UD deviennent complexes, voire ambiguës.

En outre, la présente étude a révélé une double réalité pour l'annotation syntaxique de la coordination, qui vient entièrement à l'appui

des résultats d'une étude antérieure (Galal, à paraître) : une cohérence pour les constructions standards, mais des divergences notables pour les cas atypiques et idiosyncrasiques. Cette incohérence ne se limite pas aux comparaisons entre les *treebanks* de l'arabe et du français, mais elle est également présente au sein des *treebanks* d'une même langue (cf. le *wa* initial, le *wa* circonstanciel, le *wa* d'accompagnement). Ces incohérences représentent un défi majeur, tant pour la linguistique que pour le domaine du traitement automatique des langues (TAL). En effet, elles entravent la comparabilité des *treebanks* à des fins des études typologiques et limitent l'exploitation de ces données pour une analyse automatique standardisée.

L'étude a proposé des corrections d'annotation pour certains cas idiosyncrasiques de l'arabe, visant à accroître la cohérence des *treebanks* arabes entre eux et avec d'autres langues. Cependant, un effort de correction et de standardisation plus vaste des *treebanks* arabes reste essentiel pour optimiser les futures versions du projet.

Notes

[1] Le terme *treebank* est un anglicisme courant en linguistique informatique et en traitement automatique des langues (TAL). Il désigne un corpus de textes authentiques, enrichis d'une analyse syntaxique approfondie. Cette analyse est généralement formalisée sous la forme d'un arbre de dépendance ou de constituant (Kahane et Mazziotta, 2022, p. 64).

[2] https://universaldependencies.org/.

[3] https://surfacesyntacticud.org/.

[4] UD Arabic-PADT, UD Arabic-NYUAD et UD Arabic-PUD.

[5] UD French-GSD, UD French-Sequoia, UD French-ParTUT, UD French-PUD, UD French-Rhapsodie, UD French-FQB, UD French-ParisStories et UD French-ALTS.

[6] Dans cette étude, une attention particulière sera accordée à la construction coordonnée introduite par la conjonction de coordination *wa* ('et') en arabe, étant donné que le *wa* est considéré comme le marqueur coordonnant par excellence, et que tous les autres marqueurs ont le même comportement syntaxique que le *wa* dans leur emploi en tant que conjonction de coordination.

[7] 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.

[8] Nous utiliserons l'outil *Grew-match* (Guillaume, 2021) pour interroger les *treebanks*. Cette plateforme est accessible via ce lien : https://match.grew.fr/.

[9] Afin d'affiner l'analyse du phénomène étudié, nous avons délibérément omis certains segments des phrases, ce qui a entraîné la suppression de relations syntaxiques qui figuraient dans la structure d'origine.

[10] Une unité syntaxique est « *une portion de l'énoncé qui forme un signe linguistique et*

qui commute librement dans l'énoncé ou qui est analogue à une telle unité » (Kahane et Gerdes, 2022, p. 262).

[11] 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).

[12] Le français possède, d'ailleurs, une construction particulière, la coordination à redoublement (cf. Mouret, 2007), où la conjonction de coordination précède le premier terme conjoint et obligatoirement chacun des termes conjoints qui le suivent :

i) Paul ira [et à Londres, et à Venise, et à Rome] en janvier (Mouret, 2007, p.2).

[13] Les exemples arabes sont translittérés selon la norme DIN-31635.Les gloses respectent les *Règles de Glosage de Leipzig* :
https://www.eva.mpg.de/lingua/resources/glossing-rules.php.

[14] Comme la plupart des phrases du *treebank* Arabic-PUD manquaient de gloses, nous avons choisi de les ajouter manuellement. Pour assurer une cohérence avec les autres *treebanks* UD, ces gloses ont été insérées en anglais.

[15] https://universaldependencies.org/u/overview/enhanced-syntax.html#ellipsis.

[16] Il faut bien noter que le morphème *wa*, bien que qualifié d'initial, ne se place pas obligatoirement en tête de phrase. Il peut relier deux propositions qui n'ont ni lien de sens ni lien grammatical :

ii) ثم قضى أجلا وأجل مسمى (le Coran, Sourate Al-ʾanʿām, verset 2)

ṯumma	*qaḍā*	*ʾağalan*	*wa=ʾağalun*	*musamā*
puis	décrérer.PST.3SG	terme	coord=terme	fixé

'puis Il vous a décrété un terme, et il y a un terme fixé auprès de Lui' (Traduction de Hamidullah, 1977)

Je remercie le relecteur de la revue pour avoir attiré mon attention sur cette remarque et cet exemple.

Remerciements

Je remercie Sylvain Kahane pour ses nombreuses corrections et suggestions, que je n'ai pu intégrer qu'imparfaitement. Mes remerciements vont également aux deux relecteurs de la revue pour leurs remarques pertinentes.

Références bibliographiques

Abeillé, A. (2005). Les syntagmes conjoints et leurs fonctions syntaxiques. *Langages*, *160*(4), 42-66.

Abeillé, A., et Mouret, F. (2010). Quelques contraintes sur les coordinations elliptiques en français. *Revue de sémantique et de pragmatique*, 24(3), 177-206.

Al Khalaf, E., Mashaqba, B., et Aldiqs, R. (2024). The syntax of non-canonical coordination in Jordanian Arabic: An experimental investigation. *Open Linguistics*, *10*(1), 1–14. https://doi.org/10.1515/opli-2024-0001.

Al-Hamd, A. T. et Al-Zaaby, Y. G. (1993). *al-muʿğamu al-wāfī fī adawāti an-naḥw al-ʿarbī* (العربي المعجم الوافي في أدوات النحو) (2e éd.). Irbid : Dār Al-ʾamal.

As-Sāmurrāʾiy, F. S. (2000). *maʿānī an-naḥw* (معاني النحو). Amman : Dār Al-Fikr.

Biskri, I., et Bensaber, B. A. (2008). The Categorial Annotation of Coordination in Arabic. In *Proceedings of the Twenty-First International Florida Artificial Intelligence Research Society Conference*. 462-467.

Boukedi, S., et Haddar, K. (2014). HPSG Grammar Treating of Different Forms of Arabic Coordination. *Research in Computing Science*, *86*, 25-41.

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

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

Galal (à paraître), Évaluation des schémas Universal Dependencies et Surface Syntactic UD pour l'annotation syntaxique de constructions complexes en arabe et en français. *TJHSS*, *6*(4).

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.

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.

Gerdes, K., et Kahane, S. (2015). Non-constituent coordination and other coordinative constructions as dependency graphs. In *Proceedings of the 3rd International Conference on Dependency Linguistics (Depling)*.

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. (2019). 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. (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.

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.

Hamidullah, M. (Tard.). (1977). *Le Saint Coran et la traduction en langue française du sens de ses versets*. Revue et corrigée par le complexe du roi Fahd. Version électronique : 1.2 (04/13). Tiré de [www.lenoblecoran.fr]

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. (2013). Protocole de codage microsyntaxique, Guide d'annotation du treebank Rhapsodie de français parlé (Avec la participation de K. Gerdes, P. Pietrandrea, C. Benzitoun, et R. Bawden). http://www.projet-rhapsodie.fr/

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.

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

Mouret, F. (2007). *Grammaire des constructions coordonnées. Coordinations simples et coordinations à redoublement en français contemporain*. (Doctoral dissertation, Université Paris-Diderot-Paris VII).

Mouret, F. (2008). Les coordinations de termes dissemblables sont-elles elliptiques ? In *Actes du 2e Congrès Mondial de Linguistique Française (CMLF)*. 2563-2575.

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.

Stassen, Leon. 2000. "AND-languages and WITH-languages". *Linguistic Typology* 4, 1-54. https://doi.org/10.1515/lity.2000.4.1.1

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

Yaʿqūb, E. B. (2006). *mawsūʿatu ʿulūmi al-luġati al-ʿarabiyyah* (موسوعة علوم اللغة العربية). Beyrouth: Dār Al-kutub Al-ʿilmiyyah.