Pourquoi se tourner vers le SUD
L'importance de choisir un schéma d'annotation en dépendance surface-syntaxique

Kim Gerdes[1] Bruno Guillaume[2]
Sylvain Kahane[3] Guy Perrier[2]
(1) Sorbonne Nouvelle, LPP (CNRS), Almanach (Inria)
(2) Université de Lorraine, CNRS, Inria, LORIA, Nancy
(3) Université Paris Nanterre, Modyco (CNRS)
kim@gerdes.fr, bruno.guillaume@inria.fr
sylvain@kahane.fr, guy.perrier@loria.fr

RÉSUMÉ

L'article défend le schéma d'annotation Surface-Syntactic Universal Dependencies (SUD) comme alternative au schéma standard des Universal Dependencies (UD) pour des projets d'annotation syntaxique, en particulier sur des textes oraux ou non-standard, menés dans un objectif comparatif et typologique.

ABSTRACT

Why you should turn SUD – The importance of choosing a Surface-Syntactic dependency annotation scheme. The article attempts to promote the Surface-Syntactic Universal Dependencies (SUD) annotation scheme to syntactic annotation projects, as an alternative to the standard Universal Dependencies (UD) scheme, particularly on oral or non-standard texts, conducted for comparative and typological studies.

MOTS-CLÉS : corpus arboré, treebank, syntaxe, dépendance, schéma d'annotation, critères distributionnels, tête fonctionnelle.

KEYWORDS: treebank, syntax, dependency, annotation scheme, distributional criteria, functional head.

1 Introduction

Reste-t-il toujours d'actualité d'affirmer que tous les schémas d'annotation syntaxique se valent, avec certains avantages et certains inconvénients ? Ces dernières années, avec l'expérience accumulée du développement de centaines de corpus arborés dans le monde, la question n'est toujours pas

close mais il y a du progrès. Par exemple l'annotation en constituants n'est plus à l'ordre du jour, en premier lieu pour des raisons d'efficacité du processus d'annotation ; les constituants, si jamais ils restaient nécessaires pour une analyse spécifique, peuvent être calculés automatiquement (et de manière plus cohérente) à partir d'un arbre de dépendance. Avec les dépendances, il convient de penser le choix de l'annotation, même si, avec des outils de transformation de graphes tel que Grew (Bonfante et al. 2018), il est possible de convertir un corpus annoté d'un schéma dans un autre. Nous présentons ici quelques usages de corpus arborés et discutons brièvement sur quelle base il est possible de comparer deux schémas d'annotation en dépendance. Nous présentons ensuite le schéma des Surface-syntactic Universal Dependencies (SUD, Gerdes et al. 2018, 2019, https://surfacesyntacticud.github.io/) et montrons quels sont ses atouts.

2 Critères de distinction entre schéma d'annotation

2.1 À quoi sert un corpus arboré ?

Jusqu'à aujourd'hui, une grande partie des corpus arborés sont développés dans un but applicatif : entraîner un analyseur qui, lui, sert indirectement dans un processus TAL de compréhension de textes. L'exploitation proprement linguistique d'un treebank en est encore à ses balbutiements. On peut lister comme buts du développement d'un treebank :

1. Exploiter avec un système TAL
2. Tester un schéma d'annotation théorique
3. Découvrir ou vérifier des tendances d'usage de constructions syntaxiques
4. Comparer l'usage de structures syntaxiques entre différentes langues

Au point 1, il convient d'ajouter que les systèmes statistiques et neuronaux actuels ont avant tout besoin d'une grande masse de données la plus cohérente possible. Ainsi, pour entraîner un système, il est souvent préférable d'encoder moins d'information de manière rapide et cohérente plutôt qu'une information plus riche et plus difficile à mettre sur les données linguistiques.

L'intégration d'une analyse théorique dans un guide d'annotation permet de tester à quel point l'analyse est réellement opérationnelle. Mais ceux qui s'avancent dans cette direction empirique se heurtent souvent à l'omniprésence de phénomènes syntaxiques qui d'une part n'ont aucun lien avec leur problème et d'autre part pour lesquels ils n'y a pas d'analyse facile et généralement acceptée (noms propres, entités nommées, expressions figées, verbes supports, dates, titres d'œuvres, mots étrangers, interjections, reformulations, phrases agrammaticales…). Pour de telles vérifications d'une analyse théorique sur corpus, il convient donc d'intégrer son analyse dans un schéma existant (éventuellement en le modifiant légèrement) plutôt que de commencer des analyses sur des phrases nues. Il est ainsi important que le schéma soit facile d'accès et proche des structures habituellement considérée en syntaxe.

Le projet Universal Dependencies (UD, De Marneffe et al. 2014, Nivre et al. 2019, universaldependencies.org) prévoit l'intégration d'analyses idiosyncratiques plus fines en proposant un jeu invariable de relations syntaxiques pour toutes les langues, mais ces relations principales peuvent être raffinées par des relations secondaires, séparées par deux points

(*relationPrincipale:relationSecondaire*). Similairement, le jeu des parties du discours est clos, mais des traits morpho-syntactiques peuvent être ajoutés à volonté à chaque token.
Par contre, la primauté des mots lexicaux dans l'arbre syntaxique UD amène à des structures très inhabituelles comme par exemple l'analyse des prépositions en tant que dépendant « casuel » du nom (appelée aussi l'analyse turque de l'anglais). En plus les analyse UD résultent dans des structures similaires entre langues même si, structurellement, les langues divergent dans la réalisation d'une construction. Ainsi, certaines mesures des différences typologiques ne sont pas possibles sur UD directement et nécessitent une transformation des treebanks (Gerdes et Kahane 2016, Osborne et Gerdes 2019).

2.2 Comment choisir un schéma d'annotation ?

On peut retenir les critères suivants. Le schéma doit

1. se baser sur des critères syntaxiques (et non sémantiques) si on veut :
   a. l'appliquer à des langues typologiquement différentes ;
   b. mesurer des différences syntaxiques entre langues ;
2. faciliter l'annotation par des critères distributionnels que l'annotateur peut appliquer de manière reproductible sans recourir à des lexiques extérieurs ;
3. distinguer d'une part une grille d'analyse obligatoire et universelle et d'autre part permettre des sous-spécifications et des raffinements idiosyncratiques des analyses (par langue ou par treebank) ;
4. s'intégrer dans les projets internationaux de développement de treebanks ;
5. se rapprocher des analyses classiques afin de faciliter des requêtes dans le treebank et des extensions du schéma ;
6. se limiter à un système de traits par tokens et de relations de dépendances hiérarchiques (c'est-à-dire un nœud domine l'autre) et binaires entre tokens, même si toutes les relations ne rentrent pas parfaitement dans ce schéma (e.g. les coordinations).

Le dernier point permet l'utilisation d'outils de visualisation, d'annotation et d'analyse automatique, – essentielle pour un processus d'annotation avec des boucles de bootstrapping.

3 Surface-syntactic Universal Dependencies

Le schéma Surface-syntactic Universal Dependencies (SUD, Gerdes et al. 2018, 2019, https://surfacesyntacticud.github.io/) est le résultat de l'expérience accumulée dans le développement de corpus arborés dans plusieurs projets (ANR Rhapsodie, ANR Orféo, ANR NaijaSynCor, ANR Profitérole, projet Procore « Semi-automatic Creation of a Parallel Treebank of Cantonese and Mandarin »). SUD se fonde sur des critères distributionnels et suit ainsi l'analyse dépendentielle classique (Hudson 1984, 1987, Mel'cuk 1988, Prague Dependency Treebank, Hajič & Hajičová 1997) avec des têtes fonctionnelles.
SUD est presque isomorphe à UD, dans le sens qu'une annotation SUD peut être transformée en annotation UD et vice versa avec peu de perte – les pertes peuvent principalement être attribuées à des analyses non-conformes avec les guides UD ou SUD (mais il y a aussi des pertes causées par la structure UD même, qui est plus plate que la structure SUD et ne contient pas tous les liens hiérarchiques entre dépendants) La transformation des treebanks UD en SUD permet de corriger

quelques liens problématiques pour les mesures comparatives et facilite ainsi des études typologiques sur les treebank UD.

Toutes les parties du discours et une grande partie des relations (boîte orange de la Figure 1 ci-dessous) sont les mêmes en UD et en SUD. Les deux schémas se distinguent avant tout dans l'analyse des arguments verbaux et prépositionnels (boîte bleue ci-dessous, des liens très fréquents dans les corpus). Les relations spécifiquement SUD sont disposées dans une taxonomie, ce qui permet la sous-spécification d'une relation si une construction ne permet pas de choisir entre deux relations. SUD considère principalement 3 types de dépendants verbaux : sujet (*subj*), modifieurs (*mod*) et complément (*comp*). Les compléments peuvent être différenciés en cinq types : Les arguments obliques (*comp:obl*), les arguments directs (*comp:obj*) incluant l'argument d'une préposition, le lien entre auxiliaire et verbe lexical (*comp:aux*), le lien entre tête d'une clivée et le noyau de la phrase (*comp:cleft*) et finalement les attributs (*comp:pred*). Toute information référant à la relation sémantique entre deux unités est clairement séparée de la syntaxe mais peut optionnellement être ajoutée aux relations, séparées par le symbole arobase (@, boîtes bleues-claires ci-dessous).

Le schéma SUD a été développé et appliqué dans le contexte du développement de corpus non-standard, en particulier de l'oral. Des guides d'annotation SUD ont été développés pour le français, le naija et le chinois, démontrant ainsi la versatilité de l'approche SUD à l'annotation syntaxique.

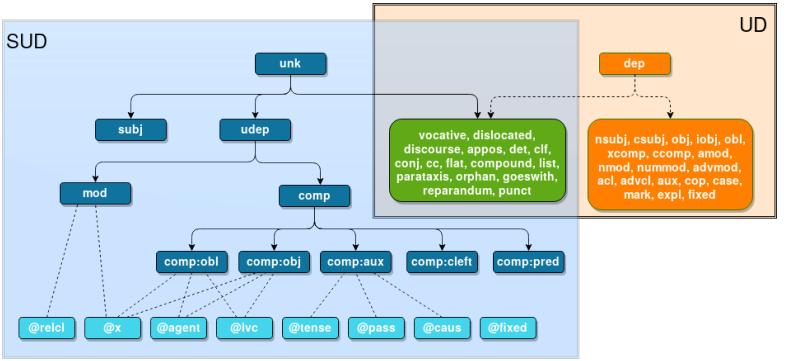

Figure 1 : Le schéma des relations SUD

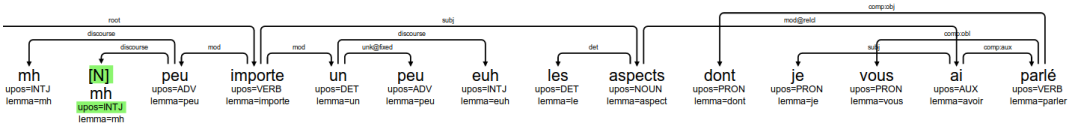

Figure 2 : exemple français d'une analyse en SUD

Références

Bonfante, G., Guillaume, B. and Perrier, G. (2018). *Application of Graph Rewriting to Natural Language Processing*. John Wiley & Sons, Incorporated.

De Marneffe, M. C., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J., & Manning, C. D. (2014). *Universal Stanford dependencies: A cross-linguistic typology.* In *LREC* (Vol. 14, pp. 4585-4592).

Gerdes, K. and Kahane, S. (2016). Dependency annotation choices: Assessing theoretical and practical issues of universal dependencies. In *Proceedings of the 10th Linguistic Annotation Workshop* held in conjunction with ACL 2016 (LAW-X 2016) (pp. 131-140).

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

Gerdes, K., Guillaume, B., Kahane, S., & Perrier, G. (2019). Improving Surface-syntactic Universal Dependencies (SUD): surface-syntactic relations and deep syntactic features. *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT 2019)* held at the Syntaxfest 2019.

Hajič, J., & Hajičová, E. (1997). Syntactic tagging in the prague tree bank. In Proceedings of the Second European Seminar "Language Applications for a Multilingual Europe (pp. 55-68).

Hudson, R.A. (1984). *Word grammar.* Oxford: Blackwell.

Hudson, R.A. (1987). Zwicky on heads. *Journal of linguistics,* 23(1), pp.109-132.

Mel'čuk, I. (1988). *Dependency Syntax: Theory and Practice*. The SUNY Press, Albany, N.Y.

Nivre J. et al. (2019). *Universal dependencies 2.4*. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Osborne, T., 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).