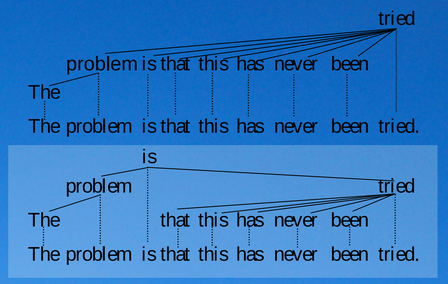
Pourquoi se tourner vers le SUD

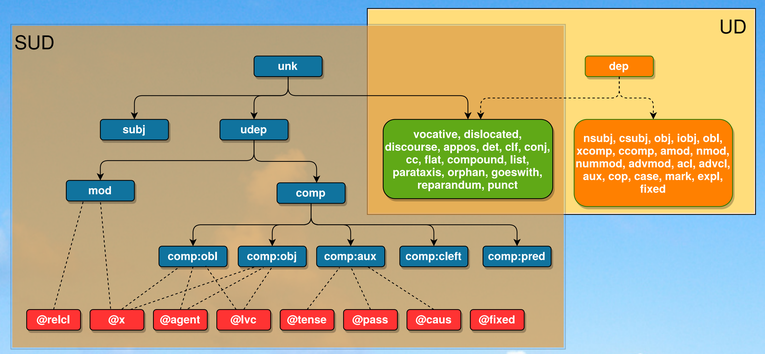
L'importance de choisir un schéma d'annotationen en dépendance surface-syntaxique

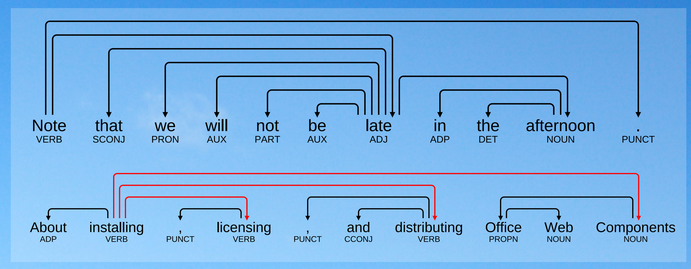
Kim Gerdes[1] Bruno Guillaume[2] Sylvain Kahane[3] Guy Perrier[2]
kim@gerdes.fr bruno.guillaume@inria.fr sylvain@kahane.fr guy.perrier@loria.fr
(1) Sorbonne Nouvelle, LPP (CNRS), Almanach (Inria) (2) Université de Lorraine, CNRS, Inria, LORIA, Nancy
(3) Université Paris Nanterre, Modyco (CNRS)

Le projet UD est génial (pour la syntaxe, la typologie, le TAL).
Mais la primauté des mots lexicaux résulte dans
- des structures très inhabituelles (par ex. prépositions en tant que dépendant « casuel » du nom appelée aussi l'analyse turque de l'anglais).
- des difficultés d'encoder des expressions figées
- des structures similaires entre langues même si, structurellement, les langues divergent dans la réalisation d'une construction
- des incohérences :

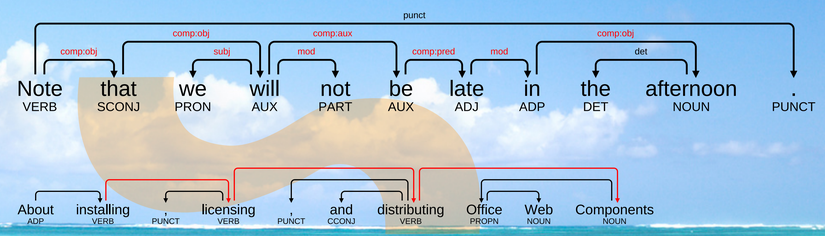


Surface-Syntactic Universal Dependencies **- même tokenisation, mêmes parties du discours, mêmes traits morphosyntaxiques**
- nouvelle définition de la tête, nouveau jeu de relations



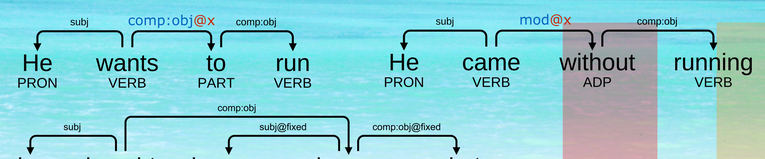


Critères distributionnels (Bloomfield, Hudson, Mel'čuk)
→ têtes fonctionnelles. ADP, AUX, SCONJ sont des têtes :

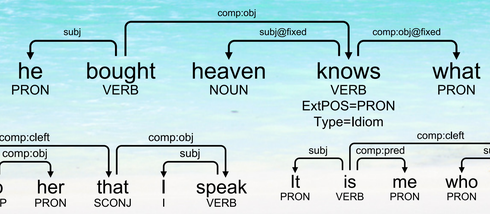


→ coordinations en chaîne (pas en bouquet) :

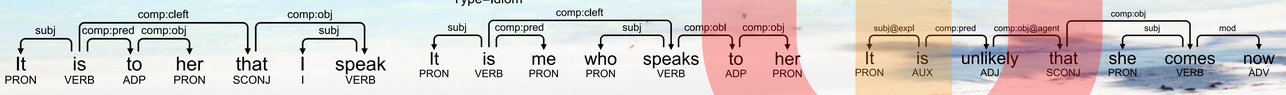
→ traits de syntaxe profonde :



→ encodage de figement :



→ clivée et explétif :



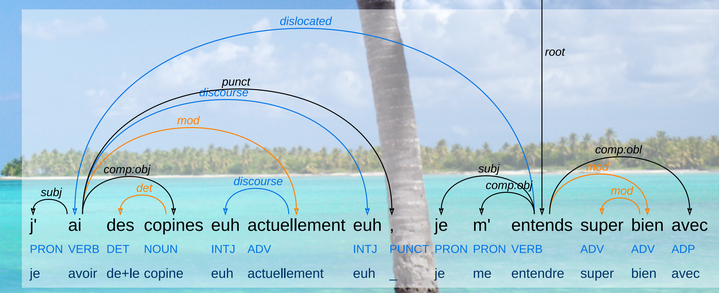
À quoi sert un corpus arboré ?

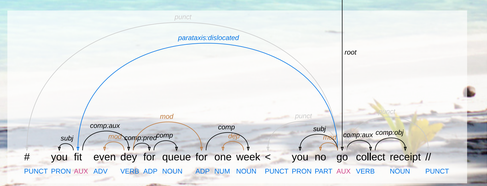
1. Exploiter avec un système TAL
2. Tester un schéma d'annotation théorique
3. Découvrir ou vérifier des tendances d'usage de constructions syntaxiques
4. Comparer l'usage de structures syntaxiques entre différentes langues

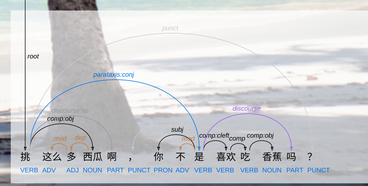
Conseils pour choisir un schéma d'annotation

Le schéma doit
1. se baser sur des critères syntaxiques pour
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) par example : *relationPrincipale:relationSecondaire* ;
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) ;
7. Prévoir différents niveaux de granularité

Annotations natives pour le français parlé, le naija, le chinois parlé :







Tous les treebanks UD ont été convertis en SUD et sont disponsibles sur https://surfacesyntacticud.github.io/