Typologie automatique des langues à partir de treebanks

Philippe Blache[1], Grégoire de Montcheuil[1, 2], Stéphane Rauzy[1]
(1) Aix-Marseille Université & CNRS, LPL, Aix-en-Provence
(2) Equipex ORTOLANG

blache@lpl-aix.fr, stephane.rauzy@lpl-aix.fr,gregoire.montcheuil@lpl-aix.fr

Résumé. La typologie des langues repose sur l'étude de la réalisation de propriétés ou phénomènes linguistiques dans plusieurs langues ou familles de langues. Nous abordons dans cet article la question de la typologie syntaxique et proposons une méthode permettant d'extraire automatiquement ces propriétés à partir de treebanks, puis de les analyser en vue de dresser une telle typologie. Nous décrivons cette méthode ainsi que les outils développés pour la mettre en œuvre. Celle-ci a été appliquée à l'analyse de 10 langues décrites dans le *Universal Dependencies Treebank*. Nous validons ces résultats en montrant comment une technique de classification permet, sur la base des informations extraites, de reconstituer des familles de langues.

Abstract.

Automatic Linguistic Typology from Treebanks.
Linguistic typology studies different linguistic properties or phenomena in order to compare several languages or language families. We address in this paper the question of syntactic typology and propose a method for extracting automatically from treebanks syntactic properties, and bring them into a typology perspective. We present here the method and the different tools for inferring such information. The approach has been applied to 10 languages of the *Universal Dependencies Treebank*. We validate the results in showing how automatic classification corrélâtes with language families.

Mots-clés : Typologie, syntaxe, treebank, inférence de grammaire, Grammaire de Propriétés.
Keywords: Typology, syntax, grammar inference, Property Grammars.

Introduction

Les treebanks sont des ressources désormais indispensables pour l'analyse syntaxique automatique. Ils constituent de plus une source d'information précieuse pour la description, au sens linguistique du terme, des propriétés syntaxiques des langues. En associant des informations syntaxiques à des données naturelles sur une grande échelle, ils permettent en effet d'extraire des régularités générales, mais offrent en même temps la possibilité de décrire des réalisations spécifiques de certaines tournures syntaxiques. De plus, et dans la mesure où un treebank repose sur un guide d'annotation précis, il est possible d'extraire automatiquement un grand nombre d'informations, en vue par exemple d'appliquer des techniques d'apprentissage automatique ou encore d'étudier la distribution de certains phénomènes.

Cependant, les formats utilisés restent à un niveau de généralité élevé. Par exemple, les treebanks de constituants reposent sur une grammaire syntagmatique implicite, à laquelle s'ajoute éventuellement l'indication des principales fonctions syntaxiques. Il est ainsi possible d'extraire automatiquement une grammaire d'un treebank et de l'illustrer en fournissant l'ensemble des réalisations des règles syntagmatiques de cette grammaire. Mais dans une perspective linguistique, il est nécessaire d'identifier également des informations plus fines, du type de celles associées aux phénomènes de rection (ordre linéaire, cooccurrence etc.) : une langue est en effet caractérisée par ce type d'indices plus que par une grammaire à proprement parler. De plus, une grammaire de constituants (ou de dépendants) extraite de treebank ne permet pas de fournir des informations globales sur la langue, concernant par exemple le type d'ordre utilisé (libre, fixe), qui constituent cependant une information essentielle pour caractériser une langue.

Ces questions se posent de façon cruciale lorsque nous adoptons une perspective typologique : comparer plusieurs langues en comparant les grammaires extraites des treebanks, n'a pas grand sens. En revanche, la typologie s'attache à comparer les langues au travers des phénomènes plus spécifiques qui la caractérisent. Nous trouvons ainsi, pour ce qui concerne la syntaxe, des typologies s'appuyant sur les relations verbes/arguments, sur l'ordre tête/modifieurs, etc.

Nous proposons dans cet article une approche s'appuyant sur une représentation particulière de l'information syntaxique (les Grammaires de Propriétés) et visant à permettre la caractérisation de langues dans une perspective typologique. Cette approche repose sur un ensemble d'outils permettant l'inférence automatique de l'information grammaticale à partir de treebanks et leur utilisation dans une perspective typologique. Nous décrivons tout d'abord la méthode développée en illustrant son application au français et proposons dans un second temps une approche comparative entre une dizaine de langues : tchèque, allemand, anglais, suédois, espagnol, français, italien, finnois, hongrois et irlandais.

1 Les treebanks, sources de l'inférence grammaticale

Nous nous appuyons pour cette étude sur le *Universal Dependencies Treebank* (Nivre, 2015). Il s'agit d'un ensemble de treebanks, utilisant le formalisme des Grammaires de Dépendance, pour 10 langues différentes. La principale caractéristique de cette ressource unique est de s'appuyer sur un même jeu d'étiquettes pour les catégories grammaticales, le *Universal POS Tags*, jeu de 17 étiquettes[1] (Petrov et al., 2012). De même, les relations de dépendance ont été standardisées et sont regroupées en un ensemble commun, le *Universal Dependency Relations*, jeu de 40 étiquettes[2] (de Marneffe et al., 2014).

La table suivante détaille la taille des treebanks respectifs des différentes langues, en précisant les familles des langues ainsi que leurs principales caractéristiques typologiques. Dans la suite, nous indiquerons par UD_xx le treebank correspondant à la langue xx.

Code	Langue	Famille	Genre	#Arbres	#Tokens	Caractéristiques typologiques
cs	Tchèque	Indo-Européenne	Slave	87.913	1.482.147	SVO[3], accentuelle, ordre des mots libre
de	Allemand	Indo-Européenne	Germanique	15.918	297.985	V2 et SOV, flexionnelle, accusative, accentuelle, à accent d'intensité
en	Anglais	Indo-Européenne	Germanique	16.622	254.930	SVO, flexionnelle, accusative, accentuelle, à accent d'intensité
sv	Suédois	Indo-Européenne	Germanique	6.026	96.699	SVO, flexionnelle, accusative, accentuelle, à accent de hauteur
es	Espagnol	Indo-Européenne	Romane	16.006	430.764	SVO, syllabique
fr	Français	Indo-Européenne	Romane	16.418	398.964	SVO, flexionnelle, accusative, syllabique
it	Italien	Indo-Européenne	Romane	10.077	214.748	SVO, syllabique
fi	Finnois	Ouralienne	Fenique	13.581	181.022	SVO, ordre des mots libre
hu	Hongrois	Ouralienne	Ougrienne	1.299	25.064	SOV, ordre libre, agglutinante, accusative
ga	Irlandais	Indo-Européenne	Celte	1.020	23.686	VSO, flexionnelle, accusative, accentuelle, à accent d'intensité

Cette ressource, par l'effort de standardisation du jeu d'étiquettes et de relations ainsi que par la couverture multilingue, est unique en son genre. Elle s'avère parfaitement adaptée au projet de comparaison de langues sur la base de propriétés formelles acquises automatiquement.

La caractérisation des propriétés d'une langue de même que la comparaison des caractéristiques syntaxiques de plusieurs langues ne peut se faire en effet directement sur la base de la comparaison des structures syntaxiques. Elle s'avère également complexe à partir de grammaires complètes (qu'il s'agisse de grammaires syntagmatiques ou de grammaires de dépendance). En revanche, il est possible de comparer des propriétés spécifiques, comme il est d'usage en typologie. Par exemple, une typologie classique consiste à étudier les relations verbes/arguments et leur linéarité.

[1] http://universaldependencies.github.io/docs/u/pos/all.html

[2] http://universaldependencies.github.io/docs/u/dep/all.html

[3] Les notations SVO, SOV ou VSO se réfèrent à l'ordre relatif entre sujet, verbe et objet ; et la notation V2 indique que le verbe est en seconde position.

Nous proposons d'extraire des treebanks ces propriétés élémentaires à partir desquelles il est possible d'établir des comparaisons entre les langues. Ces propriétés sont celles identifiées dans le cadre des *Grammaires de Propriétés*. Notre approche s'intéresse aux types de propriétés de base, telles que définies dans (Blache et al., 2012) :

– **linéarité** : deux composants (A,B) ont une relation de linéarité quand l'ordre d'apparition de ces composants est toujours le même. Nous noterons cette propriété *precede(A,B)* (i.e. A précède toujours B).
– **exigence** : deux composants (A,B) ont une relation d'exigence quand la présence de l'un exige la présence de l'autre. Nous noterons cette propriété *require(A,B)* (i.e. si A est présent, B est aussi présent, soit la formule logique $A => B$).
– **exclusion** : deux composants (A,B) ont une relation d'exclusion quand ils n'apparaissent jamais ensemble. Nous noterons cette propriété *exclude(A,B)* (i.e. si A est présent, B ne l'est pas). Contrairement aux 2 précédentes, cette dernière propriété est symétrique : *exclude(A,B)* ⇔ *exclude(B,A)*.
– **unicité** : un composant A répond à un propriété d'unicité s'il n'apparait jamais plusieurs fois dans la partie droite des règles. Nous noterons cette propriété *unicity(A)*.

L'inférence des propriétés à partir des treebanks, s'appuie sur un processus en deux étapes :

1. Extraction de la grammaire hors-contexte implicite dans les treebanks
2. Génération des propriétés à partir de ces grammaires

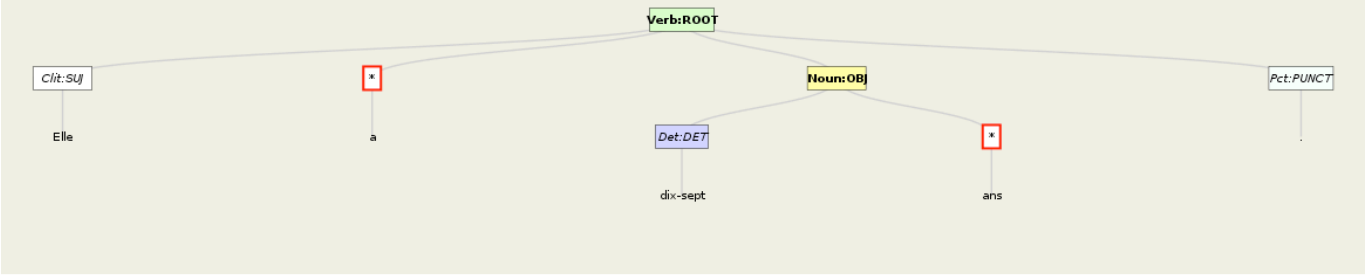
Le choix de s'appuyer sur un formalisme hors-contexte peut sembler paradoxal puisque nous utilisons en entrée des treebanks de dépendance. Il est cependant motivé par le fait qu'il est possible d'utiliser la même méthode pour tout type de treebank, quel que soit leur formalisme. Notre approche permet donc (même si cet aspect n'est pas décrit dans cet article) de traiter en entrée tout type de treebank.

1.1 Inférence de la grammaire CFG implicite

L'extraction d'une grammaire hors-contexte (CFG) à partir d'un treebank, dans le cas de formalismes syntagmatiques, repose sur méthode classique décrite dans (Charniak, 1996) : chaque nœud interne de l'arbre syntaxique correspond à une règle de réécriture dont la partie gauche est l'étiquette de ce nœud, et la partie droite la séquence d'étiquettes de ses descendants. La grammaire implicite est simplement formée de l'ensemble des différentes règles ainsi identifiées.

Dans le cas d'un treebank de dépendances, le principe est similaire : les arbres de dépendance ont une racine correspondant à la catégorie de la tête, celle-ci est indiquée de façon spécifique (par un étoile) parmi les descendants. L'application de la méthodologie décrite pour les constituants consiste à établir des règles dans lesquelles la partie gauche correspond à la tête de la relation de dépendance, la partie droite étant formée par les différents dépendants, dans leur ordre de projection, ainsi que de la projection de la tête indiquée par le symbole "*".

L'exemple suivant illustre un arbre de dépendance du corpus français ainsi que les deux règles de la CFG implicite qui en ont été extraites :



Elle a dix-sept ans .

```
Verb:ROOT -> Clit:SUJ * Noun:OBJ Pct:PUNCT
Noun:OBJ -> Det:DET *
```

1.2 Inférence des propriétés

À partir des CFG implicites, le calcul des propriétés est relativement aisé : pour chaque tête (partie gauche de règles), nous calculons la liste de ses différents dépendants (dans toutes les parties droites de règle correspondant à cette tête). Il est alors possible, à l'intérieur de chacun de ces sous-ensembles de catégories, d'identifier les propriétés en recherchant les patterns existants.

Concrètement, pour chaque paire de composants distincts (A,B), nous classons les règles de la CFG où A et/ou B apparaissent (i.e. ensemble ou non). Les parties droites de ces règles constituent donc des suites de catégories à partir desquelles il est possible d'inférer une propriété. Par exemple, à partir des règles suivantes extraites de la CFG implicite du UD_fr concernant le sujet nominal, nous pouvons inférer la propriété de linéarité entre le déterminant et le nom :

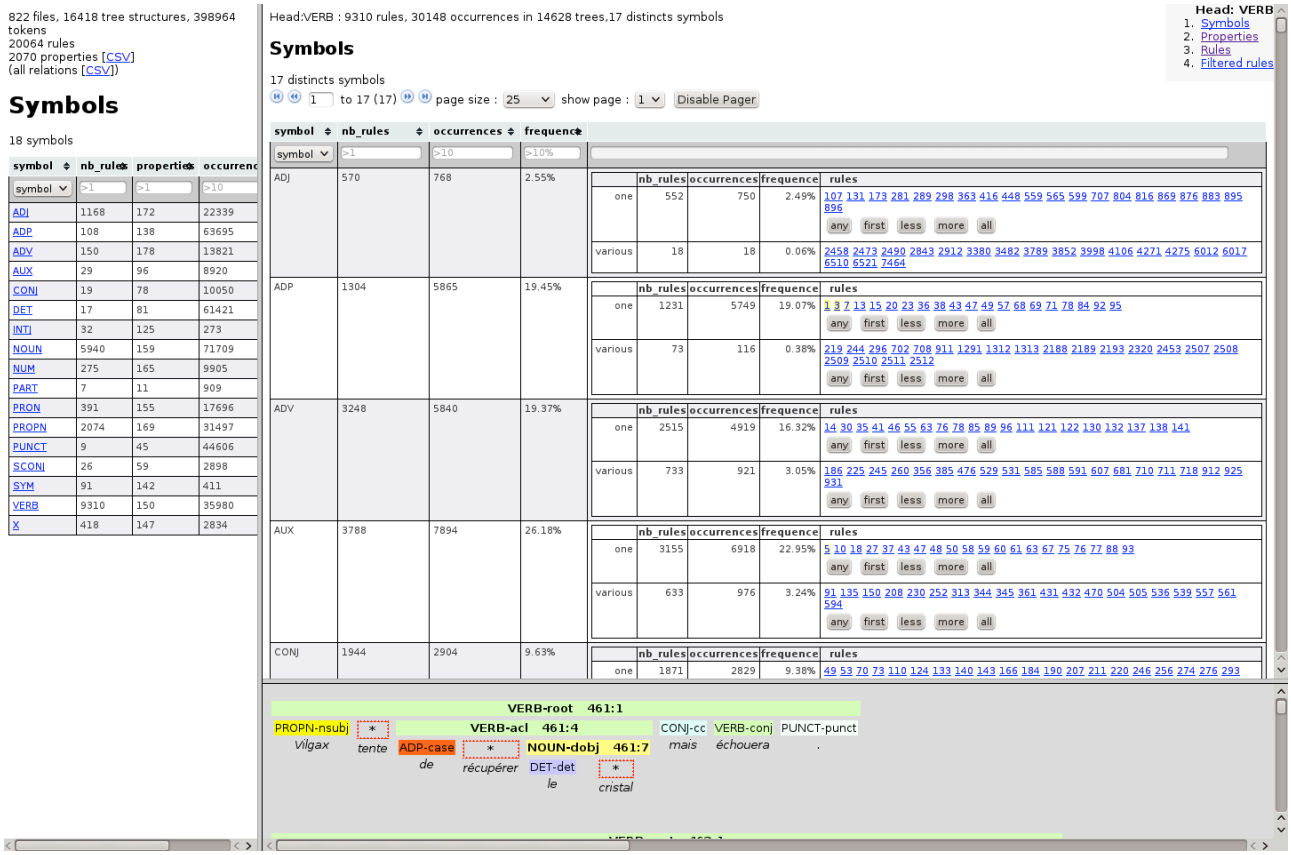
```
NOUN-nsubj -> DET-det *
NOUN-nsubj -> DET-det * NOUN-nmod      =>          Det < NOUN-nsubj
NOUN-nsubj -> DET-det * ADJ-amod
```

Ce même principe s'applique sur les autres types de propriétés comme l'exigence, l'exclusion ou l'unicité. Le principe consiste donc à prendre en considération, pour les propriétés binaires, tous les couples de catégories et vérifier dans quel ordre ils apparaissent, s'ils sont systématiquement en cooccurrence, ou encore systématiquement séparés. Dans l'exemple précédent, toutes les règles permettaient d'inférer la propriété considérée. Pour chacune des propriétés, les conditions de validation sont simplement exprimées, pour deux catégories A et B données :

- Linéarité : A et B sont réalisées et A apparaît avant B
- Exigence : A et B sont réalisées ensemble
- Exclusion A et B ne sont pas réalisées ensemble
- Unicité : A n'est réalisé qu'une fois

Nous avons développé un environnement permettant, à partir d'un treebank, d'inférer la grammaire CFG implicite et la grammaire de propriétés correspondante, et d'éditer le résultat sous la forme d'un navigateur HTML qui permet de visualiser les règles et les propriétés et d'explorer leurs occurrences dans le treebank. L'exemple suivant montre une page d'exploration dans laquelle la fenêtre se divise en 3 parties :

- Dans le cadre de gauche sont listés les différentes catégories du treebank avec les décomptes des occurrences, nombre de règles CFG associées et propriétés induites. Un lien permet de charger dans la partie principale la page décrivant le composant.
- La zone principale sert à présenter, pour chaque tête, les différents symboles qui le composent, les propriétés calculées et les règles CFG dont il est la partie gauche (dans la figure suivante, seule la description des catégories dépendantes est affichée)
- Le cadre du bas permet la visualisation des arbres de dépendance de la partie correspondante dans le treebank.



L'éditeur possède plusieurs options pour affiner le calcul des règles et des propriétés. Il est ainsi possible de choisir le niveau de finesse des symboles considérés, soit en ne conservant que la catégorie grammaticale, soit en utilisant la paire *<catégorie, fonction>*. Il est également possible de filtrer les règles (en fonction de leur nombre d'occurrences, ou de leur fréquence) pour considérer une propriété. L'outil est diffusé sur le site d'Ortolang/SLDR sous le nom de *MarsaGram*[4].

1.3 Identifier l'importance des propriétés par leur distribution

Notre hypothèse est qu'il est possible de caractériser les langues en fonction de la répartition des propriétés. Pour cela, nous proposons de prendre en compte les occurrences de chacune d'entre elles ainsi que leur distribution au sein des propriétés caractérisant la construction étudiée. Nous définissons ainsi un certain nombre de critères à partir desquels il sera possible d'analyser et comparer les caractéristiques de différentes langues. Ces critères s'appuient sur la fréquence des occurrences des règles CFG à partir desquelles les propriétés sont inférées. Par exemple, dans le treebank `UD_fr`, nous aurons les données suivantes concernant la propriété de précédence entre la tête verbale et le complément d'objet (`NOUN-dobj`) :

precede	*	NOUN-dobj		nb_rules	occurrences	frequence	rules
			precede	2	434	77.78%	0 1

Ces données indiquent que la propriété de linéarité est inférée à partir de 2 règles (dont les indices 0 et 1, qui sont des hyperliens, pointent vers les règles `VERB-root -> NOUN-nsubj * NOUN-dobj PUNCT-punct` et `VERB-root -> PRON-nsubj * NOUN-dobj PUNCT-punct`). Ces deux règles apparaissent dans le treebank 434 fois, ce qui représente 77,78% des occurrences des règles décrivant `VERB-root` (elles sont au nombre de 558).

[4] Accessible grâce à son identifiant pérenne : hdl:11041/ortolang-000917

Il est possible d'extraire plus d'informations de ces données. En particulier, une propriété sera d'autant plus importante qu'elle apparaît systématiquement. Nous avons un moyen direct d'évaluer cette importance : il suffit d'identifier les règles par lesquelles la propriété étudiée est activée, et de voir si cette propriété est vérifiée ou non. Ainsi, dans l'exemple précédent, toutes les règles dans lesquelles apparaissent les catégories `VERB` et `NOUN-dobj` comportent cet ordre linéaire. Nous considérerons donc que la propriété `VERB < NOUN-dobj` est importante et ne doit pas être transgressée. En revanche, l'exemple suivant illustre un cas pour lequel la propriété semble être moins stable. Il s'agit de la relation de précédence entre verbe et nom, tous deux dépendants d'un nom construit comme modifieur adverbial exprimé par la propriété : `VERB-cop < NOUN-nmod`. Cette propriété est validée dans la plupart des règles contenant ces deux catégories comme dans l'exemple suivant :

NOUN-advcl 7150:27					
					NOUN-nmod 7150:33
ADP-mark	PRON-nsubj	VERB-cop	DET-det	*	ADP-case / DET-det / *
comme	c'	était	le	cas	dans / le / livre

Elle est cependant non satisfaite dans certaines constructions inversant le complément nominal par rapport à la copule, ces exemples étant beaucoup plus rares :

NOUN-advcl 6147:4									
	NOUN-nmod 6147:6								
SCONJ-mark	ADP-case 6147:7		DET-det	*	ADJ-amod	PRON-nsubj	VERB-cop	ADV-neg	*	ADJ-amod
si	ADP-mwe	*	la	convention	fiscale	vous	êtes	non	résident	français
	de	par								

Il est possible de ne conserver que les propriétés vérifiées dans tous les cas (i.e. sans aucune règle contradictoire) ou de relâcher cette contrainte, comme dans l'exemple précédent, pour avoir des propriétés pondérées, obtenant ainsi des propriétés plus *fortes* (fréquentes et vérifiées dans tous les cas) et d'autres plus *faibles* (moins fréquentes ou moins souvent vérifiées). Nous proposons donc de fournir une première indication du poids associé à une propriété en s'appuyant sur cet indice, pouvant être pondéré par la fréquence de la propriété.

Nous notons, pour une propriété *p, Validating(p)* et *Violating(p)* l'ensemble des règles validant ou contredisant la propriété. Il est alors possible de calculer un premier poids noté w_0 correspondant au ratio entre le nombres d'occurrences des règles validant *p* par rapport à l'ensemble des règles liées à la propriété. On note par ailleurs *Occ(p)* la fonction retournant l'ensemble des occurrences de *p*, éventuellement contraintes par leurs satisfaction. Nous avons :

$$w0 = Occ(Validating(p)) / Occ(Validating(p)) + Occ(Violating(p))$$

On peut noter que si les occurrences des règles validant *p* sont plus nombreuses que celles qui la contredisent, alors w_0 est supérieur à 0,5. Dans le cas où la propriété est toujours vérifiée, alors $w_0=1$. Il est possible d'affiner ce poids en prenant en compte la fréquence des règles. Il s'agit plus précisément de pondérer w0 par le ratio entre le nombre d'occurrences des règles validant la propriété *p* et le nombre total d'occurrences des règles de la catégorie tête à laquelle *p* se réfère.

$$w1 = w0 * Occ(Validating(p)) / nb\ total\ d'occurrences$$

Le tableau suivant, extrait du navigateur, illustre cette répartition pour quelques propriétés de la catégorie `NOUN-nsubj` en français. Cette catégorie correspond dans le treebank à 5585 occurrences des 10 règles de la grammaire CFG qui la décrivent. Les deux premières propriétés se retrouvent en tant que pattern dans toutes les règles, avec donc une fréquence de 100%. Aucune règle ne correspondant à une violation, le poids `w0` est donc égal à 1, de même que le poids `w1`. En revanche, ce n'est pas le cas de la propriété d'exclusion entre le déterminant et le verbe modifieur du nom qui se trouve vérifiée dans la plupart des règles du corpus, à l'exception d'une, ayant une faible occurrence. L'indice `w0`, qui a dans ce cas la même valeur que la fréquence, conduit à un poids `w1` de plus faible valeur, indiquant la possibilité de relâchement de la propriété.

Property	1st cat	2nd cat	Freq	w0	w1				
Unicity	DET-det	-	100%	1.0000	1.0000	nb_rules	occ	freq	rules

						Validating	10	5585	100.00%	0 1 2 3 4 5 6 7 8 9
Precede	DET-det	*	100%	1.0000	1.0000		nb_rules	occ	freq	rules
						Validating	10	5585	100.00%	0 1 2 3 4 5 6 7 8 9
Exclude	DET-det	VERB-acl:relcl	98.57%	0.9857	0.9716		nb_rules	occ	freq	rules
						Validating	9	5505	98.57%	0 1 2 3 4 5 6 7 8
						Violating	1	80	1.43%	9

2 Caractériser les langues

Notre hypothèse est que la distribution des propriétés ainsi que leur importance relative permet d'établir une forme de caractérisation de la langue décrite dans le treebank. Cette caractérisation s'appuie sur un ensemble d'éléments qui, réunis, permettent de donner une image globale des caractéristiques syntaxiques.

Taille de la grammaire : une première indication porte sur le nombre de propriétés qu'il est possible d'identifier, en rapport avec la taille du *tagset*. Il s'agit d'un élément d'information régulièrement utilisé dans la description typologique des langues, notamment dans la perspective de l'étude de leur complexité (Dahl, 2004). Cependant, les propriétés fournissent d'autres types d'information. Par exemple, une langue comportant un très grand nombre de propriétés aura des formes de surface contraintes, avec une variabilité limitée. En effet, les propriétés réduisent par leur application l'espace de recherche définissant les formes possibles. Si les catégories utilisées sont soumises à un grand nombre de contraintes, leur combinatoire s'en trouvera donc limitée. Une conséquence directe sera alors la présence dans la langue de constructions nombreuses, mais peu variables. Le tableau suivant récapitule ces données pour les langues du corpus considéré :

CS	DE	EN	ES	FI	FR	GA	HU	IT	SV
598	683	755	708	523	716	547	448	750	547

Ces données confirment la prédiction émise : les langues considérées comme étant à ordre des mots libre (le tchèque, le finnois et le hongrois) sont décrites par des grammaires de propriétés de taille significativement plus réduite que les autres.

Caractérisation des têtes : toutes les catégories têtes sont caractérisées par l'ensemble des propriétés qui relient leurs dépendants. Il est donc possible pour chaque catégorie d'extraire plusieurs types d'information : la couverture des contraintes (le nombre de catégories qu'elles affectent) et leur importance. De la même façon que pour la langue, ces deux indications permettent de décrire ses réalisations possibles d'une construction. Nous proposons dans ce qui suit une analyse pour chaque type de propriété de l'influence de ces facteurs. Nous définissons pour cela un indice de cohésion qui est une fonction de ces facteurs :

On note C une catégorie, D_c l'ensemble des catégories dépendantes de C, et P_c l'ensemble des contraintes s'appliquant aux catégorie de D_c. Nous appelons dans ce qui suit *construction* tout ensemble formé d'une catégorie tête et de ses dépendants.

- Taille (t_c) : c'est la taille du graphe (le nombre de sommets) formé par les contraintes de P_c. Il s'agit en d'autres termes du nombre de catégories de D_c affectées par les contraintes de P_c.

Densité ($Dens_c$): c'est la densité du graphe formé par les contraintes de P_c. Un graphe dense contient un grand nombre de sommets (ici des catégories) connectés par des arrêtes (des contraintes). Un graphe dense peut être complet : tous les sommets sont connectés entre eux. Le nombre maximal de relation pour un type de propriété binaire donné étant de $t_c*(t_c-1)$, nous avons donc l'indicateur de densité suivant :

$$Dens_c = nb\_prop / t_c*(t_c-1)$$

L'intuition est qu'un ensemble de dépendants est plus ou moins fortement contraint et donc plus ou moins variable. Si la taille du graphe de contrainte P_c s'approche du nombre total de catégorie de D_c, si sa densité est élevée ainsi que la moyenne des poids des contraintes de P_c, alors la réalisation de la construction est très fortement limitée. L'interprétation de cette mesure est différente en fonction des types de contraintes. Par exemple, pour ce qui concerne la

linéarité, une densité indique le nombre de catégories affectées par un ordre. Dans ce cas, la densité est indicatrice de la liberté de l'ordre des mots dans le constituant. Une densité faible indique un ordre des mots libre. À l'opposé, un graphe de contraintes complet à l'ordre près (toutes les catégories sont contraintes 2 à 2), indique un ordre des mots fixe à l'intérieur de la construction.

Le tableau suivant illustre l'application de cette mesure au français et au finnois :

	ADJ	ADP	ADV	AUX	CONJ	DET	INTJ	NOUN	NUM	PART	PRON	PROPN
FR	0,40	0,33	0,33	0,32	0,53	0,24	0,30	0,86	0,43	0,27	0,24	0,67
FI	0,04	0,30	0,08	0,36	0,42	0,00	0,46	0,06	0,07	0,00	0,06	0,14

	PUNCT	SCONJ	SYM	VERB	X	Moy
FR	0,35	0,75	0,60	0,46	0,40	0,42
FI	0,29	1,00	0,06	0,05	0,15	0,22

Ce tableau indique pour chaque catégorie tête la densité de linéarité calculée selon la formule précédente. Il révèle des différences importantes à la fois entre les catégories elles-mêmes, mais également au niveau global. La densité de linéarité moyenne du français est très supérieure à celle du finnois, prédisant ainsi correctement l'observation typologique selon laquelle le finnois a un ordre des mots libre, ce qui n'est pas le cas du français.

Caractérisation des langues : Nous proposons de synthétiser pour chaque langue les informations concernant toutes les catégories à l'intérieur d'un *ensemble de dépendances* (i.e. l'ensemble des dépendants d'une même tête). Nous regroupons ainsi, pour chaque propriété p affectant une catégorie c d'un tel ensemble (notée p_c) les informations suivantes :

- `props` : nombre de propriétés p_c
- `%props` : ratio entre le nombre de propriétés p_c et le nombre total de propriétés de type p pour la tête
- `occ` : nombre d'occurrences des règles (de la grammaire CFG) validant la propriété p_c
- `%occ` : ratio entre le nombre d'occurrences des règles validant p_c et le nombre total des règles validant une propriété de type p
- `mean(w0)` : moyenne des w0 pour p_c
- `mean(w1)` : moyenne des w1 pour p_c

L'exemple suivant propose une synthèse des informations concernant les propriétés de précédence affectant le déterminant pour le tchèque et l'allemand :

lang	head	prop_type	cat	props	%props	occ	%(occ)	mean(w0)	mean(w1)
cs	NOUN	precede	DET	6	15,38%	41151	7,34%	0,96	0,02
de	NOUN	precede	DET	9	23,08%	73508	53,99%	0,96	0,16

Ce tableau indique que pour le tchèque, parmi les dépendants du nom, le déterminant intervient dans 6 propriétés de linéarité, soit 15,38% des propriétés de linéarité des dépendants du nom. Ces propriétés correspondent à des règles dont la somme des occurrences est 41.151 (7,3% des occurrences des règles liées à une propriété de linéarité). La moyenne des w0 de ces 6 propriétés de linéarité est de 0,96, celle de w1 de 0,02. Une comparaison avec les valeurs comparables pour l'allemand montre des différences significatives : le déterminant a un ordre linéaire beaucoup plus contraint en allemand qu'en tchèque, ce qui est révélé d'une part par le ratio du nombre de propriétés de linéarité concernant le déterminant par rapport à toutes les propriétés de linéarité du nom, mais également (et surtout) par la distribution des occurrences des règles correspondantes (représentant 53,93% des règles validant une linéarité pour le nom).

Le même type de comparaison peut être opéré à un niveau un peu plus général, en prenant en compte simultanément tous les types de propriétés.

lang	head	prop_type	cat	props	%props	occ	%(occ)	mean(w0)	mean(w1)
cs	NOUN	_ALL_	DET	22	13,41%	893059	12,26%	0,98	0,13
de	NOUN	_ALL_	DET	19	13,97%	393535	31,83%	0,97	0,42

3 Classer les langues

Le projet d'approcher la typologie des langues par classification automatique sur la base de traits morpho-syntaxiques a été explorée dans différentes études, avec des perspectives différentes : traits lexicaux (Enright et al., 2011 ; Barbancon et al., 2007 ; Ellison et al. 2006), interférence langue maternelle/seconde (Nagata et al. 2013), mesures de distance (Batagelj et al., 1992; Kita, 1999). Certaines approches, plus rares, s'appuient sur des informations syntaxiques (Sidorov et al., 2013 ; Abramov et al., 2011). Nous proposons dans cet article d'appliquer des méthodes de classification à partir de paramètres syntaxiques précis, obtenus à grande échelle à partir du UDT.

Il est possible, grâce aux informations produites par les propriétés, de comparer les langues en vue de leur classification. Il s'agit ici de vérifier la pertinence de ces informations, mais également la possibilité d'établir un modèle prédictif de regroupement, voire de typologie. Cette classification repose sur l'identification des propriétés communes à plusieurs langues. Cette opération est rendue possible par le fait que le même jeu d'étiquettes est utilisé pour les différentes langues du *Universal Treebank*, rendant ainsi les propriétés directement comparables. L'hypothèse est que, à la différence des règles syntagmatiques ou de dépendance, les propriétés peuvent à la fois représenter des types d'information très spécifiques, entre deux sous-catégories particulières, mais également être regroupées par type.

Chaque propriété dans notre approche est représentée par un quadruplet *<C, tp, A, B>* où *C* est le contexte de la propriété (la partie gauche des règles sur laquelle celle-ci est calculée), *tp* est le type de propriété (*unicity*, *precede*, *require* ou *exclude*) et *A* et *B* sont les composants de la propriété. Étant donné que le jeu d'étiquettes auquel appartiennent *A, B* et *C* est le même pour toutes les langues, nous pouvons calculer si une propriété *p = <C, tp, A, B>* présente pour une langue l'est également dans une autre.

Le tableau suivant présente les occurrences des propriétés communes entre l'italien et d'autres langues :

properties	lang	cs	de	en	es	fi	fr	ga	hu	it	sv
ALL	it	706	740	1022	972	755	1023	566	505	-	630
precede	it	114	142	192	225	111	186	97	86	-	142

L'italien partage le plus de propriétés avec le français (1.023, dont 186 de linéarité), puis l'anglais (1.022, dont 192 de linéarité) et c'est avec le hongrois qu'il partage le moins de propriétés. Du point de vue de la linéarité, c'est avec l'espagnol qu'il y a le plus d'occurrences de propriétés communes (225).

Nous proposons d'affiner cette estimation par la définition d'une fonction de similarité. Soit *P(lg)* un ensemble de propriétés calculées pour la langue *lg*, une mesure de similarité entre deux langues $simil(lg_1, lg_2)$ peut être obtenue par le rapport des propriétés communes aux deux langues relativement à l'ensemble des propriétés de l'une et l'autre de ces langues :

$$simil(lg1,lg2) = card(P(lg1) \cap P(lg2)) / card(P(lg1) \cup P(lg2))$$

	#properties	cs	de	en	es	fi	fr	ga	hu	it
cs	1665									
de	1695	0,645								
en	2267	0,656	0,634							
es	2001	0,675	0,562	0,588						
fi	1299	0,724	0,756	0,655	0,701					
fr	2070	0,652	0,621	0,517	0,528	0,662				
ga	1228	0,716	0,649	0,723	0,671	0,780	0,673			
hu	969	0,702	0,707	0,759	0,738	0,732	0,741	0,726		
it	1679	0,687	0,669	0,595	0,581	0,596	0,575	0,696	0,717	
sv	1250	0,626	0,637	0,690	0,663	0,747	0,654	0,652	0,655	0,681

La distance associée compte les propriétés présentes dans seulement une des deux langues (différence symétrique) :

$$dist(lg1,lg2) = card(P(lg1) \oplus P(lg2) / card(P(lg1) \cup P(lg2)$$
$$= 1-simil(lg1,lg2).$$

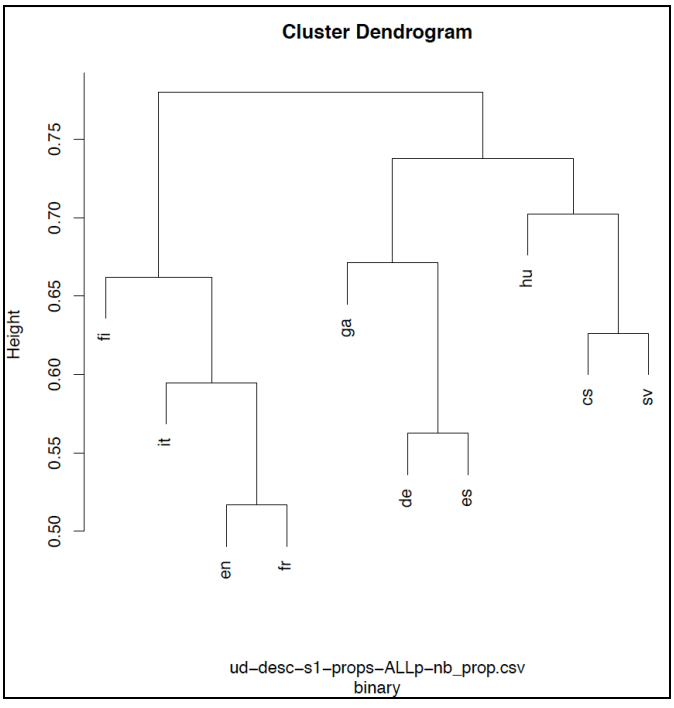
Nous obtenons grâce à ces mesures un calcul de distance entre les langues. La première table indique les distances entre langues sur la base de la prise en compte de toutes le propriétés. On remarque ainsi que l'anglais, l'espagnol et l'italien sont les langues les plus proches du français, ce qui correspond aux attentes, y compris pour l'anglais compte tenu de la situation très particulière de cette langue germanique, mais proche syntaxiquement des langues romanes. Un rapprochement peu attendu mais semblant robuste apparaît également entre l'allemand et l'espagnol.

Le même calcul peut être fait en prenant en compte un seul type de propriété. Nous présentons dans le tableau suivant ces informations calculées pour les propriétés de linéarité. On remarque que dans ce cas, l'espagnol et l'italien sont plus nettement proches du français. De même, un rapprochement entre l'allemand et le suédois peut-être observé.

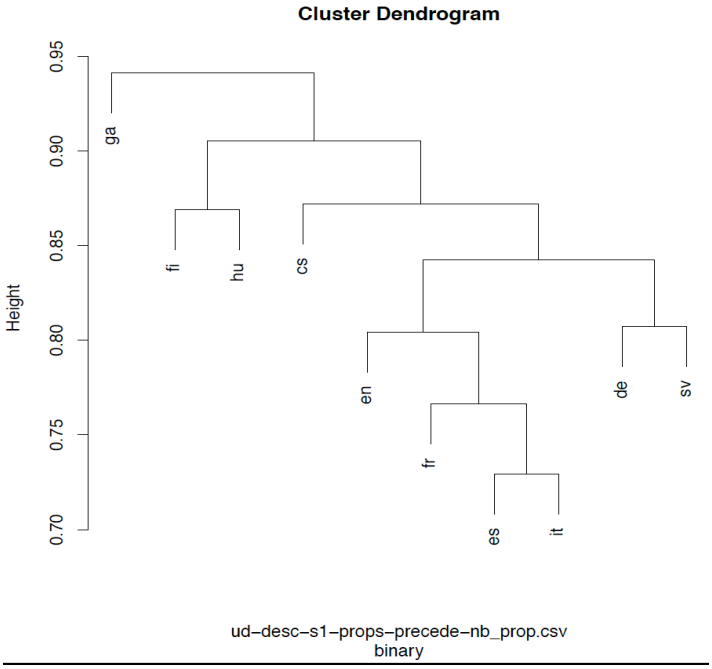
	#properties	cs	de	en	es	fi	fr	ga	hu	it
cs	375									
de	445	0,867								
en	562	0,850	0,811							
es	563	0,856	0,797	0,782						
fi	334	0,885	0,886	0,835	0,862					
fr	489	0,872	0,843	0,804	0,762	0,888				
ga	318	0,907	0,873	0,892	0,878	0,942	0,893			
hu	271	0,882	0,867	0,905	0,882	0,869	0,903	0,929		
it	493	0,849	0,822	0,778	0,729	0,845	0,766	0,864	0,873	
sv	341	0,830	0,807	0,833	0,795	0,895	0,841	0,862	0,860	0,795

À l'aide de ces mesures de distance entre toutes les langues, nous proposons de réaliser un clustering hiérarchique des différentes langues du *Universal Treebank*. Ce traitement a été effectué sous R à l'aide de la fonction *hclust*. Celle-ci réalise (par défaut) un groupement agglomératif à liens complets (*bottom-up, complete-linkage*).

Les figures suivantes présentent les dendrogrammes obtenus. La première figure prend en considération tous les types de propriétés (*unicity*, *precede, require* et *exclude*), les données étant celles de la table 9.



La classification dans ce cas ne permet pas de dégager des regroupements très nets, quelle que soit la typologie : ordre des mots fixe ou libre, position des arguments par rapport au verbe, etc. Ainsi que nous l'avions remarqué sur la base des données, des rapprochements ponctuels sont cependant significatifs entre le français, l'anglais et l'italien. À noter ici que l'anglais se retrouve très éloigné des autres langues germaniques. L'application du même type de clustering en se limitant aux propriétés de linéarité est en revanche très concluant, comme indiqué dans la figure suivante :



Les classes dégagées correspondent en effet à des attentes typologiques. Nous retrouvons en effet au sein d'une même classe les langues romanes (`es, it, fr`), desquelles se rapproche l'anglais (comme décrit précédemment). De même, nous retrouvons les langues germaniques au sein d'une même classe (`de, sv`). Le finnois et le hongrois se retrouvent également dans une classe, ce qui est également une attente typologique pour les langues finno-ougriennes. L'irlandais enfin, se retrouve dans une position éloignée étant seule représentante de sa classe dans ce corpus.

4 Conclusion

Nous avons présenté dans cet article une méthode ainsi qu'un ensemble d'outils pour acquérir automatiquement des informations en vue d'une description typologique des langues. La méthode présentée consiste à inférer automatiquement à partir de treebanks la grammaire hors-contexte implicite puis d'en extraire les propriétés telles que définies dans les Grammaires de Propriétés. Les outils développés pour cela constituent une plateforme de navigation dans le treebank en même temps qu'un outil de visualisation de données. Cet environnement, de même que les données produites sont disponibles via l'entrepôt de données « *anonyme* ».

Cette première opération permet donc d'inférer automatiquement deux types de grammaires, dans deux formalismes différents. La méthode proposée est générique et peut s'appliquer à des treebanks de dépendance (comme c'est le cas dans cet article), mais également à des treebanks en constituants. En termes de perspectives, nous appliquons actuellement cette méthode à trois grands treebanks en constituants : le Penn, l'Arabic et le Chinese treebank). À terme, nous serons ainsi en mesure d'inclure dans notre étude un très grand nombre de langues, proposant ainsi une technique automatique pour la typologie à grande échelle.

Nous avons montré dans cet article comment, à partir d'une représentation de l'information syntaxique sous la forme de propriétés, il était possible de dégager ou vérifier les grandes propriétés typologiques à partir desquelles nous avons proposé une technique de classification permettant de retrouver automatiquement, sur la base des propriétés de linéarité, les familles de langues sur la base de leur distance. Ce résultat permet de valider la pertinence de la représentation des informations syntaxiques sous la forme de propriétés en vue d'une description typologique. En particulier, plusieurs études portant sur la complexité linguistique ont montré l'intérêt d'une représentation de ce type. Il devient ainsi possible d'envisager le développement d'un outil de comparaison de la complexité des langues, s'appuyant sur des bases formelles.

Par ailleurs, cette méthode présente l'avantage de produire automatiquement des ressources de haut niveau (treebanks hybrides, ajoutant les propriétés syntaxiques explicites aux informations de dépendance et/ou de constituance ainsi que les informations dérivées comme la densité) à partir desquelles de nombreuses applications reposant sur des techniques

d'apprentissage automatique peuvent être appliquées. Ces ressources permettront le développement d'une plateforme d'analyse syntaxique automatique multilingue, reposant sur les Grammaires de Propriétés.

Remerciements

Ce travail réalisé dans le cadre du Labex BLRI (ANR-11-LABX-0036) et de l'Equipex ORTOLANG (ANR–11–EQPX–0032)), ayant ainsi bénéficié d'une aide de l'État au titre du projet A*MIDEX (ANR-11-IDEX-0001-02).

Références

ABRAMOV O., MEHLER A. (2011) "Automatic Language Classification by means of Syntactic Dependency Networks", in *Journal of Quantitative Linguistics*, 4:291-336

BATAGELJ V., PISANSKI T., AND KERZIC D. (1992), "Automatic clustering of languages", in *Computational Linguistics*, 18(3):339–352.

BARBANCON F. WARNOW T., EVANS S., RINGE D., NAKHLEH L. (2007), "An experimental study comparing linguistic phylogenetic reconstruction methods", ms#732, *Department of Statistics, University of California, Berkeley*.

BLACHE P., RAUZY S. (2012). « Enrichissement du FTB : un treebank hybride constituants/propriétés », in Actes de la conférence *JEP-TALN-RECITAL 2012*, volume 2, 307-320.

BLACHE P. (2005). Property grammars: A fully constraint-based theory. In H. Christiansen et al., editor, *Constraint Solving and Language Processing*, volume LNAI 3438. Springer.

CHARNIAK E. (1996). « Tree-bank Grammars ». In proceedings of 13th *National Conference on Artificial Intelligence,* 10311036.

DAHL O. (2004) *The Growth and Maintenance of Linguistic Complexity*, John Benjamins

DE MARNEFFE M.-C., DOZAT T., SILVEIRA N., HAVERINEN K., GINTER F., NIVRE J., MANNING C. (2014). Universal Stanford Dependencies: A cross-linguistic typology. In proceedings of 9th *International Conference on Language Resources and Evaluation (LREC'14)*.

ELLISON M., KIRBY S. (2006), "Measuring Language Divergence by Intra-Lexical Comparison", in *Proceedings of COLING-ACL-2006*

ENRIGHT J., KONDRAK G. (2011), "The application of chordal graphs to inferring phylogenetic trees of languages", in Proceedings of *5th IJCNLP*

FERRER I CANCHO R., SOLE R., KÖHLER R. (2004), "Patterns in syntactic dependency networks" in *Phys. Rev. E*, 69:5

KITA K. *(1999),* "Automatic Clustering of Languages Based on Probabilistic Models", in *Journal of Quantitative Linguistics*, 6(2):167-171.

NIVRE J., BOSCO C., CHOI J., DE MARNEFFE M.-C., DOZAT T., FARKAS R., FOSTER J., GINTER F., GOLDBERG Y., HAJIC J., KANERVA J., LAIPPALA V., LENCI A., LYNN T., MANNING C., MCDONALD R., MISSILÄ A., MONTEMAGNI S., PETROV S., PYYSALO S., SILVEIRA N., SIMI M., SMITH A., TSARFATY R., VINCZE V., ZEMAN D. (2015). Universal Dependencies 1.0. http://hdl.handle.net/11234/1-1464.

PETROV S., DAS D., MCDONALD R. (2012). A Universal Part-of-Speech Tagset. In proceedings of the 8th *International Conference on Language Resources and Evaluation (LREC'12).Nagata R. and Whittaker E. (2013), "Reconstructing an Indo-European Family Tree from Non-native English Texts", in Proceedings of the 51st Annual Meeting of the ACL*

RAMA, T., SINGH K. (2009), "From Bag of Languages to Family Trees From Noisy Corpus", in *Proceedings of RANLP-2009*

SIDOROV G., VELASQUEZ F., STAMATATOS E., GELBUKH A., CHANONA-HERNANDEZ L. (2013) "Syntactic Dependency-Based N-grams as Classification Features", in *Advances in Computational Intelligence*, LNCS-7630, Springer

SINGH K. AND SURANA H. (2007), "Can Corpus Based Measures be Used for Comparative Study of Languages?", in *Proceedings of 9th Meeting of the ACL SIG in Computational Morphology and Phonology*