Les corpus arborés avant et après le numérique

Sylvain Kahane, Nicolas Mazziotta

Les corpus arborés avant et après le numérique

Sylvain Kahane* — **Nicolas Mazziotta****

* *Modyco, Université Paris Nanterre & CNRS*
** *U.R. Traverses, Université de Liège*

RÉSUMÉ. *Nous montrons comment, du* XVIII*e siècle à nos jours, l'annotation syntaxique de corpus a évolué de l'analyse exhaustive de phrases isolées à celle de listes d'exemples, puis à celle de textes entiers. Nous étudions l'évolution des visées de ces corpus arborés entre motivations pédagogique, théorique et ressources pour le TAL. Nous présentons quelques ouvrages clés, souvent peu connus de la communauté TAL comme de celle des linguistes : Buffier (1709), Beauzée (1765), Gaultier (1817), Clark (1847), Jespersen (1937) et Tesnière (1959). Nous concluons sur les liens actuels entre corpus arborés et TAL.*

MOTS-CLÉS : *corpus arboré, treebank, annotation syntaxique, analyse syntaxique, diagramme syntaxique.*

TITLE. *Treebanks before and after the digital technology*

ABSTRACT. *This paper explains how, from the 18th century to the present day, the syntactic annotation has evolved from the comprehensive analysis of isolated sentences to lists of examples, then to complete texts. We study the evolution of the aims of these treebanks between pedagogical and theoretical motivations and resources for NLP. We introduce some key works, often little known by the NLP community as well as by linguists: Buffier (1709), Beauzée (1765), Gaultier (1817), Clark (1847), Jespersen (1937), Tesnière (1959). We conclude on the current links between treebanks and NLP.*

KEYWORDS: *syntactic treebank, syntactic annotation, parsing, syntactic diagram.*

1. Introduction

Dans cet article[1], nous nous intéressons à l'histoire des corpus arborés en syntaxe. Par *corpus arborés*, nous entendons des corpus qui comprennent un certain nombre de phrases extraites de productions attestées auxquelles sont associées des analyses syntaxiques complètes. Ces analyses possèdent généralement une structure proche de celle d'un arbre de dépendance ou de constituants, d'où l'appellation commune de *corpus arboré* (angl. *treebank*). Si les corpus arborés se sont largement développés à l'âge numérique sous l'impulsion du traitement automatique des langues, nous allons montrer que ces ressources ont d'abord été développées à des fins pédagogiques, puis à des fins théoriques afin de valider les premiers modèles syntaxiques.

On situe généralement l'apparition des premiers corpus arborés dans les années 1970 avec le *Talbanken* du suédois (Einarsson, 1976), puis leur diffusion dans les années 1990 avec le *Penn Tree Bank* de l'anglais (Marcus *et al.*, 1993) et le *Prague Dependency Treebank* du tchèque (Hajič, 1998). Toutefois, il s'agit là des premières ressources au format numérique. Des ressources traditionnelles que nous considérons comme d'authentiques corpus arborés non numériques se sont en effet développées suite à l'apparition de démarches d'analyse syntaxique systématique d'exemples attestés au XVIII^e^ siècle avec Buffier (1709), puis de manière encore plus formalisée chez les encyclopédistes, Dumarsais (1754) et Beauzée (1765), cf. Kahane (2020). De nombreux ouvrages didactiques de grammaire du XIX^e^ siècle, à commencer par ceux du méconnu Louis Gaultier (1817), proposent de véritables collections d'exemples (et d'exercices corrigés) analysés systématiquement dans un même formalisme. Les analyses de Gaultier prennent la forme de diagrammes tabulaires qui représentent la structure de la phrase, dont l'un d'entre eux n'est pas sans rappeler le standard CoNLL (Buchholz et Marsi, 2006) (voir section 2). Au cours du XIX^e^ siècle se développent différentes conventions graphiques pour représenter la structure syntaxique et certains auteurs proposent des ouvrages entiers d'exemples analysés selon leurs conventions. Nous discuterons en particulier des ouvrages de Clark (1863) et de Reed et Kellogg (1889). C'est seulement au XX^e^ siècle que des linguistes davantage intéressés par les questions théoriques que didactiques s'emparent de la question et proposent des corpus d'exemples analysés dans le cadre théorique qu'ils défendent. C'est le cas en particulier de Jespersen (1937), de Tesnière (1959) et de Nida (1966).

Pour cadrer la discussion, nous commençons par un état de l'art succinct des corpus arborés à l'âge du numérique (section 2). Nous procédons ensuite chronologiquement. Nous commençons par l'étude des premières analyses syntaxiques systématiques proposées au XVIII^e^ siècle (section 3), car c'est cet intérêt pour l'exhaustivité qui a rendu possible la constitution de collections d'exemples analysés. Le cœur de l'article est consacré aux ouvrages pédagogiques du XIX^e^ siècle comportant d'importants corpus arborés dont les analyses sont présentées sous forme de diagrammes tabulaires ou hiérarchiques (section 4). Nous contrastons ces travaux avec ceux du XX^e^ siècle, qui en raison de leur visée théorique proposent généralement des corpus

1. Les deux auteurs ont contribué de manière équivalente à cette recherche.

multilingues (section 5). Notre conclusion se penche sur les pratiques actuelles et l'avenir des corpus arborés (section 6).

2. Les corpus arborés à l'âge du numérique

Une courte présentation des corpus arborés actuels permettra de mieux situer les travaux faits dans les siècles précédents. Si le premier corpus arboré électronique, le *Talbanken* du suédois (Einarsson, 1976), se développe dans les années 1970, c'est dans les années 1990 avec le *Penn Tree Bank* de l'anglais (Marcus *et al.*, 1993), que la communauté des linguistes et des talistes commence à s'intéresser vraiment à ce type de ressources. Le *Penn Tree Bank* a été développé sous l'impulsion de talistes avec l'utilisation d'outils de TAL pour la pré-annotation et dans l'objectif de développer des outils de TAL plus perfomants[2]. L'annotation syntaxique du corpus est une analyse en constituants encodée sous la forme d'un parenthésage du texte. Conformément au principe générativiste de mouvement, l'arbre syntaxique inclut des nœuds vides, comme le nœud *-NONE-*, dans la figure 1, qui indique la position du groupe syntaxique extrait *how many credit cards*, auquel il est lié par un index (cf. l'index *1* dans *WHNP-1* et **T*-1*)[3].

```
( (SBARQ
    (INTJ (UH So) )
    (WHNP-1
      (WHADJP (WRB how) (JJ many) )
      (, ,)
      (INTJ (UH um) )
      (, ,) (NN credit) (NNS cards) )
    (SQ (VBP do)
      (NP-SBJ (PRP you) )
      (VP (VB have)
        (NP (-NONE- *T*-1) )))
    (. ?) (-DFL- E_S) ))
```

Figure 1. *Exemple extrait du Penn Tree Bank :* So how many, um, credit cards do you have ?

2. On peut citer les premières phrases de l'introduction de Marcus *et al.* (1993) : « *There is a growing consensus that significant, rapid progress can be made in both text understanding and spoken language understanding by investigating those phenomena that occur most centrally in naturally occurring unconstrained materials and by attempting to automatically extract information about language from very large corpora. Such corpora are beginning to serve as an important research tool for investigators in natural language processing, speech recognition, and integrated spoken language systems, as well as in theoretical linguistics.* »

3. Exemple extrait de la page `catalog.ldc.upenn.edu/desc/addenda/LDC99T42.mrg.txt` du catalogue LDC qui distribue le Penn Tree Bank.

Le développement du *Prague Dependency Treebank* (Hajič, 1998) a débuté à la suite du *Penn Tree Bank*. Il s'agit d'un corpus de tchèque annoté en syntaxe de dépendance, avec une couche d'annotation en syntaxe de surface (appelée *analytical tree*) accompagnée d'une analyse en syntaxe profonde (appelée *tectogrammatical tree*), ainsi qu'un outil d'édition et de visualisation des arbres, comme le montrent les arbres de la figure 2 (Hajic *et al.*, 2001).

Plusieurs corpus arborés sont développés à la suite de ces premières expériences, avec une grande variété de schémas d'annotation et de formats d'encodage, jusqu'à la proposition du format tabulaire CoNLL[4]. Ce format particulièrement économique, inspiré du format proposé un an plus tôt par Hall et Nivre (2006), est aujourd'hui un standard pour l'encodage des analyses en dépendance sur des corpus de textes (Buchholz et Marsi, 2006). Il a contribué à populariser l'analyse en dépendance dans le domaine du TAL et tout particulièrement de l'analyse syntaxique automatique.

La figure 3 illustre l'encodage au format CoNLL d'un extrait du corpus arboré SUD_French-GSD[5]. Dans cet encodage, la structure est décrite dans un tableau généralement à 10 colonnes. Les mots de la phrase sont dans la colonne 2. La colonne 1 contient leur identifiant, la colonne 3 les lemmes, la colonne 4 les parties du discours, la colonne 6 les traits morphosyntaxiques standard et la colonne 10 des traits additionnels. L'arbre de dépendance est encodé dans les colonnes 7 et 8 : la colonne 7 contient l'identifiant du gouverneur de chaque mot (avec un 0 pour le mot 3 qui n'a pas de gouverneur) et la colonne 8 sa fonction syntaxique. Par exemple, le mot 1 est le *det* du mot 2. Les colonnes 5 et 9 restent vides (elles sont utilisées par les parsers) et la colonne 10 est un fourre-tout d'informations additionnelles.

De nombreux outils permettent le requêtage et l'affichage de fichiers au format CoNLL, comme Grew-match (Guillaume, 2021), sur lequel nous reviendrons dans la section 6, ou SETS (Luotolahti *et al.*, 2015). Certains outils d'annotation permettent de modifier dynamiquement un CoNLL à partir de sa forme graphique, comme ArboratorGrew (Gerdes, 2013 ; Guibon *et al.*, 2020), UD Annotatrix (Tyers *et al.*, 2017) ou ConnluEditor (Heinecke, 2019).

Maintenant que ce cadre est posé, nous pouvons entamer notre retour vers le futur en commençant par les analyses syntaxiques du début du XVIII[e] siècle.

4. D'après le colloque éponyme en apprentissage automatique, *Conference in Natural Language Learning*.

5. Le format SUD (Surface-Syntactic Universal Dependencies, `surfacesyntacticud.github.io`, (Gerdes *et al.*, 2018)) est une variante du format UD (Universal Dependencies, `universaldependencies.org`, (Nivre *et al.*, 2016 ; **?**)), où, contrairement à UD, les mots fonctionnels sont traités comme des têtes.

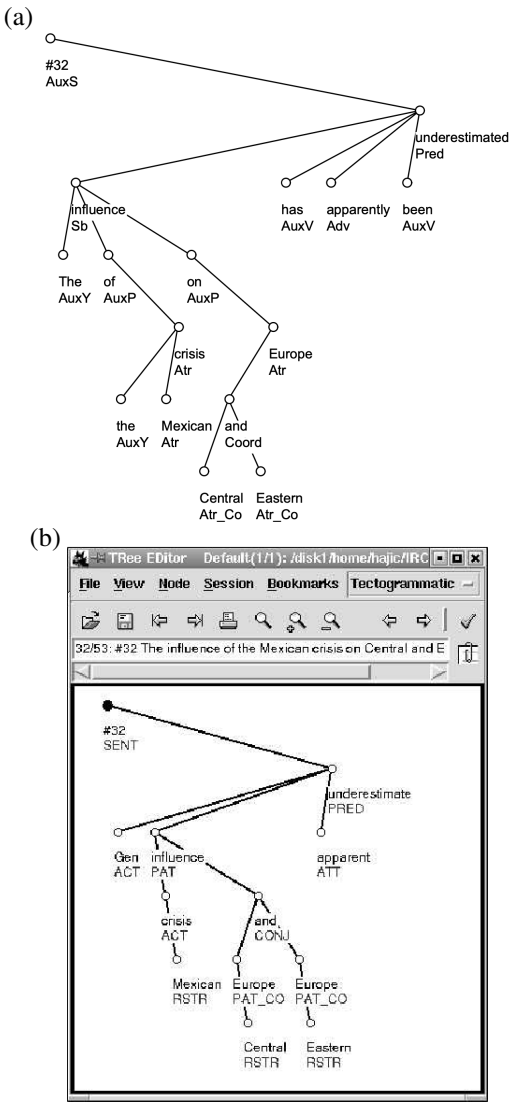


Figure 2. *Arbres analytique (a) et tectogrammatique (b) :* The influence of the Mexican crisis on Central and Eastern Europe has apparently been underestimated

3. Premières analyses syntaxiques exhaustives au XVIIIe siècle

Notre parcours historique sur les corpus arborés syntaxiques commence au XVIIIe siècle, où l'on trouve les premières analyses syntaxiques complètes de phrases

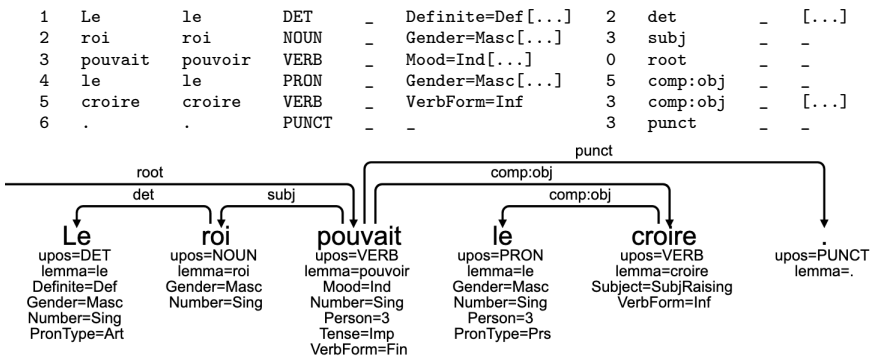


Figure 3. *Exemple extrait du SUD_French-GSD:* Le roi pouvait le croire *(fr-ud-train_01347; certaines informations omises sont indiquées par « [...] »)*

attestées[6]. Nous reproduisons ici l'intégralité d'une analyse du jésuite Claude Buffier (1661-1737), tirée de sa grammaire du français (1709, 84)[7] :

> « *Un homme qui étourdit les gens qu'il rencontre avec de frivoles discours a coutume de causer beaucoup d'ennui à tout le monde*. Je dis que dans ce discours, tous les mots sont pour modifier le nom *un homme*, et le verbe *a coutume*, et que c'est en cela que consiste tout le mystère et toute l'essence de la syntaxe des langues : 1° le nom *un homme*, est modifié d'abord par le *qui* déterminatif : car il ne s'agit pas ici d'un homme en général, mais d'un homme marqué et déterminé en particulier par l'action qu'il fait d'*étourdir* ; de même, il ne s'agit pas d'un homme *qui étourdit* en général, mais *qui étourdit* en particulier *les gens*, et non pas *les gens* en général, mais en particulier *les gens qu'il rencontre*. Or cet homme qui étourdit ceux qu'il rencontre, est encore particularisé par *avec des discours*, et *discours* est encore particularisé par *frivoles*. On peut voir le même dans la suite de la phrase : *a coutume* est particularisé par *de causer*, *de causer* est particularisé par ses deux régimes, par son régime absolu, savoir, *beaucoup d'ennui*, et par son régime respectif, *à tout le monde*. Voilà donc comment tous les mots d'une phrase quelque longue qu'elle soit, ne sont que pour modifier le nom et le verbe. »

Bien que la distinction entre la syntaxe et la sémantique ne soit pas encore aboutie, les termes *modifier*, *déterminer* et *particulariser* peuvent être compris comme « dé-

6. On trouve bien sûr des analyses syntaxiques avant cela, mais, à notre connaissance, jamais aussi complètes et systématiques. On pourra notamment consulter la remarquable grammaire de l'anglais (1653) que John Wallis (1616-1703) a rédigée en latin – traduction anglaise par Kemp (1972). Voir Imrényi et Mazziotta (2020) pour un historique des analyses en dépendance depuis Priscien.

7. L'orthographe et les mises en italiques sont modernisées.

pendre de » ou « être complément de ». Le terme *régime* correspond au terme *complément*, qui ne sera véritablement introduit que par Beauzée (voir plus loin).

Nous proposons de représenter notre interprétation de l'analyse de Buffier par le diagramme de la figure 4, où les flèches expriment les relations du type « est déterminé par », « est particularisé par » ou « est modifié par ». Notons que les deux termes de la relation sont à chaque fois assez clairement donnés par Buffier, chaque élément mentionné dans le texte particularisant le précédent. On peut voir que même si la terminologie comme l'argumentation ne distinguent pas clairement syntaxe et sémantique, la description est totalement compatible avec une analyse syntaxique actuelle.

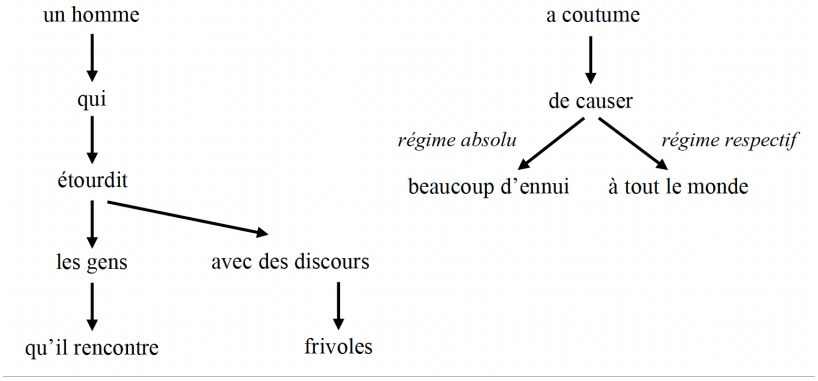


Figure 4. *Diagramme réalisé par nos soins à partir de l'analyse de Buffier (1709) :* Un homme qui étourdit les gens qu'il rencontre avec de frivoles discours a coutume de causer beaucoup d'ennui à tout le monde

On trouve d'autres analyses de phrases complexes chez d'autres auteurs français au XVIII[e] siècle, notamment chez Girard (1747) et dans l'*Encyclopédie* – dans les articles de Dumarsais (1754) et de Beauzée (1765). Voir Kahane (2020) pour une étude critique.

On doit à Beauzée la notion moderne de *complément* (Chevalier, 1968). Dans l'article de l'*Encyclopédie* qu'il consacre au terme *Régime*, Beauzée introduit un sous-article *Complément* (Beauzée n'a été en charge des articles de linguistique de l'*Encyclopédie* qu'à partir de la lettre *F*). Plus précisément, Beauzée distingue le *complément grammatical* ou *initial*, qui est un mot, du *complément logique* ou *total*, qui en est la projection, combinant ainsi les notions modernes de dépendance et de constituance[8] :

8. L'orthographe de la citation qui suit est modernisée.

> « Par exemple, dans cette phrase, *avec les soins requis dans les circonstances de cette nature* ; le mot *nature* est le complément grammatical de la préposition *de* : *cette nature* en est le complément logique : la préposition *de* est le complément initial du nom appellatif *les circonstances* ; et *de cette nature* en est le complément total : *les circonstances*, voilà le complément grammatical de la préposition *dans* ; et *les circonstances de cette nature* en est le complément logique. [...] » (Beauzée, 1765, 5)

Comme nous l'avons fait pour l'analyse de Buffier, nous pouvons proposer une diagrammatisation de l'analyse de Beauzée ou plus exactement des deux analyses superposées proposées par Beauzée. Dans la figure 5, nous représentons les relations « être le complément initial ou grammatical » par des bulles et « être le complément total ou logique » par des flèches.

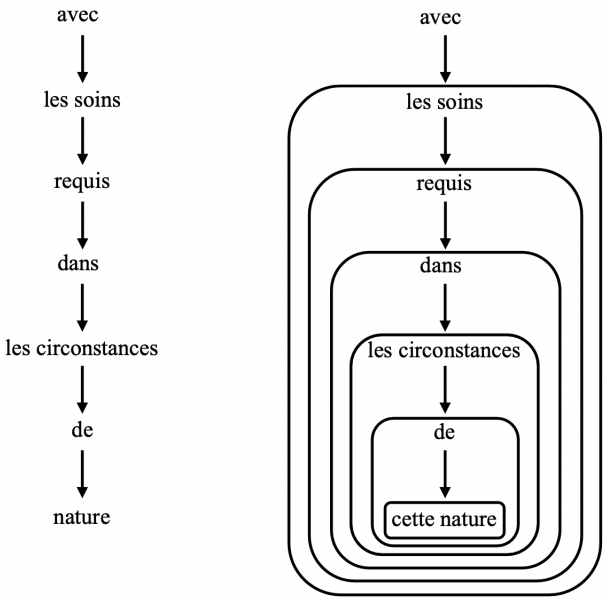


Figure 5. *Diagrammes réalisés par nos soins à partir de l'analyse double de Beauzée :* avec les soins requis dans les circonstances de cette nature

La démarche d'analyse syntaxique sous la forme d'un discours où tous les mots sont décrits successivement se développe particulièrement dans la grammaticographie anglo-saxonne, sous forme de parsing. À titre d'exemple, on peut citer cet extrait des exercices de Lindlay Murray (1745-1826), qui fait autorité dans le domaine de l'enseignement de la grammaire à la fin du XVIII^e^ siècle et au début du suivant :

> « "He who lives virtuously prepares for all events." *He* is a personal pronoun, of the third person singular number and masculine gender. *Who* is a

> relative pronoun, which has for its antecedent “he,” with which it agrees in gender and number [...]. *Lives* [is] a regular verb neuter, indicative mood, present tense, third person singular, agreeing with its nominative, “who” [...]. *Virtuously* is an adverb of quality. *Prepares* [is] a regular verb neuter, indicative mood, present tense, third person singular, agreeing with its nominative, “he.” *For* is a preposition. *All* is an adjective pronoun, of the indefinite kind, the plural number, and belongs to its substantive, “events,” with which it agrees [...]. *Events* is a common substantive of the neuter gender, the third person, in the plural number, and the objective case, governed by the preposition, “for” [...]. » (Murray, 1812, 18)

On voit que ces discours permettent de décrire une partie de la structuration syntaxique de phrases complètes, mais ils portent sur des phrases isolées et sont présentés d'une manière qui ne favorise pas les comparaisons. On peut encore ajouter, comme nous l'a fait remarquer un des relecteurs, que les exemples peuvent être éparpillés dans le texte et ne constituent pas, de ce fait, un corpus strict, c'est-à-dire un matériau textuel présenté indépendamment de son exégèse. Le recours ultérieur aux diagrammes va radicalement changer les choses.

4. Des corpus arborés pour la didactique au XIX^e^ siècle

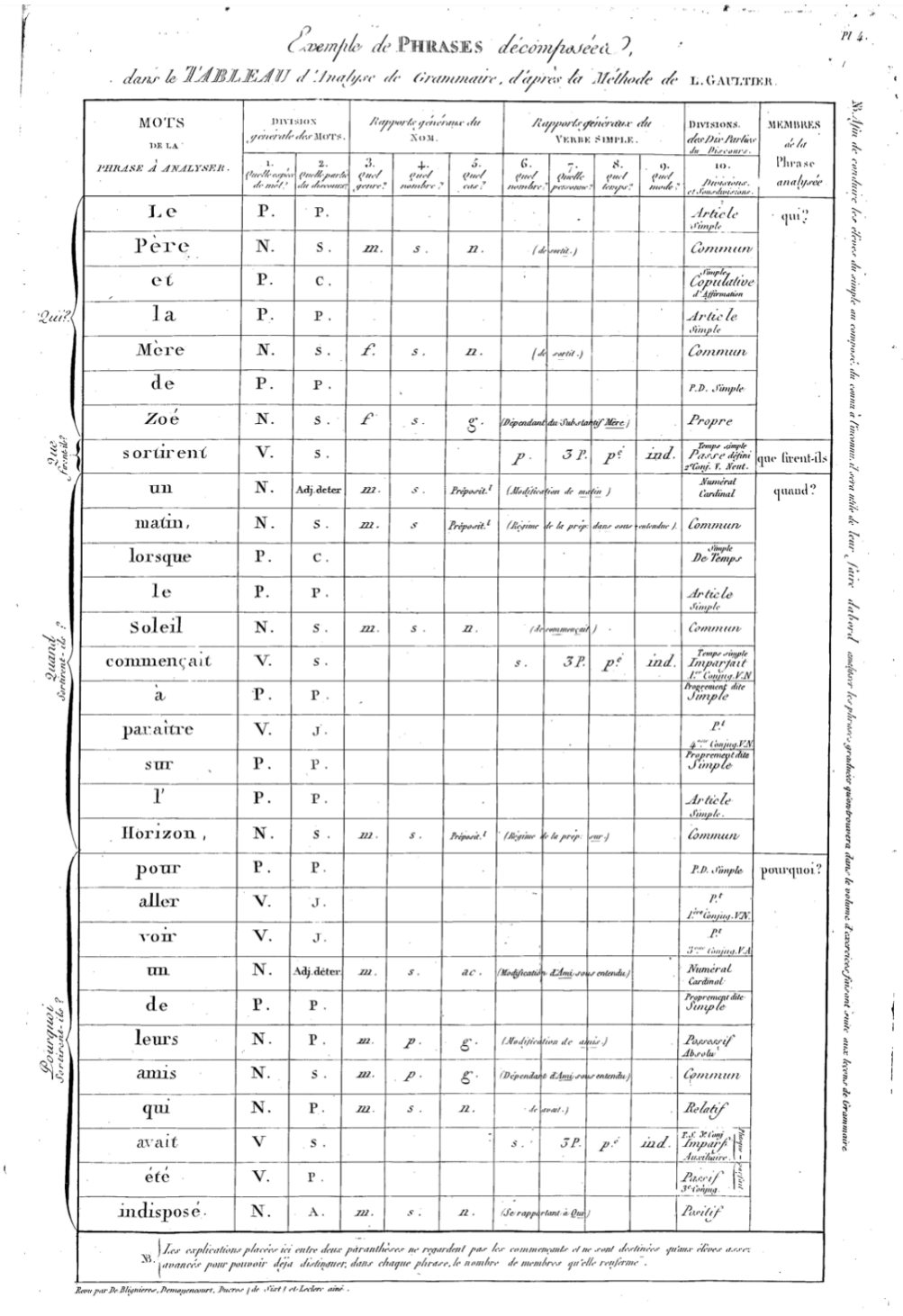
L'idée que l'on peut analyser des structures de manière exhaustive se développe au XIX^e^ siècle de deux façons : premièrement, on voit apparaître de grandes listes d'exemples analysés ; deuxièmement et probablement conséquemment, les analyses prennent la forme de diagrammes, les rendant à la fois plus concises et plus lisibles. Ces listes d'analyses, qui constituent indéniablement de véritables corpus arborés, sont développées au sein d'ouvrages didactiques monolingues visant à présenter les différentes constructions d'une langue donnée. Nous aborderons successivement le cas des diagrammes tabulaires, en particulier ceux de Louis Gaultier (sous-section 4.1) et des diagrammes hiérarchiques proposés par Clark et Reed et Kellogg (sous-section 4.2).

4.1. *Premiers diagrammes tabulaires*

À notre connaissance, le premier auteur à utiliser extensivement des diagrammes pour l'enseignement de la grammaire est l'abbé Louis Gaultier (1746-1818), dont l'*Atlas de grammaire* (1817) contient une grande variété de diagrammes à visée pédagogique[9]. Il comporte en particulier un diagramme tabulaire, qui encode un fragment d'analyse en dépendance (figure 6).

Dans l'introduction des *Élémens de grammaire* publié en 1829 par ses élèves, on trouve un diagramme similaire précédé de la description suivante :

9. Les numérotations de pages qui suivent renvoient à la copie pdf distribuée par `gallica.bnf.fr`, qui comprend différents feuillets numérotés séparément.



Exemple de PHRASES *décomposées ?,*

dans le TABLEAU *d'Analyse de Grammaire, d'après la Méthode de* L. GAULTIER.

Pl 4.

MOTS DE LA PHRASE À ANALYSER.	DIVISION générale des MOTS. 1. Quelle espèce de mot ?	2. Quelle partie du discours ?	Rapports généraux du NOM. 3. Quel genre ?	4. Quel nombre ?	5. Quel cas ?	Rapports généraux du VERBE SIMPLE. 6. Quel nombre ?	7. Quelle personne ?	8. Quel temps ?	9. Quel mode ?	DIVISIONS. des Div. Particul. du Discours. 10. Divisions et Subdivisions.	MEMBRES de la Phrase analysée
Le	P.	P.								Article Simple	qui ?
Père	N.	S.	m.	s.	n.	(de sorte.)				Commun	
et	P.	C.								Simple Copulative d'Affirmation	
la	P.	P.								Article Simple	
Mère	N.	S.	f.	s.	n.	(de sorte.)				Commun	
de	P.	P.								P.D. Simple	
Zoé	N.	S.	f	s.	g.	(Dépendant du Substantif Mère)				Propre	
sortirent	V.	S.				p.	3 P.	p.é	ind.	Temps simple Passé défini 2ème Conj. V. Neut.	que firent-ils
un	N.	Adj. déter.	m.	s.	Préposit.l	(Modification de matin)				Numéral Cardinal	quand ?
matin,	N.	S.	m.	s	Préposit.l	(Régime de la prép. dans sous entendue)				Commun	
lorsque	P.	C.								Simple De Temps	
le	P.	P.								Article Simple	
Soleil	N.	S.	m.	s.	n.	(de commençait)				Commun	
commençait	V.	S.				s.	3 P.	p.é	ind.	Temps simple Imparfait 1.er Conjug. V.N	
à	P.	P.								Proprement dite Simple	
paraître	V.	J.								P.t 4ème Conjug. V.N	
sur	P.	P.								Proprement dite Simple	
l'	P.	P.								Article Simple	
Horizon,	N.	S.	m.	s.	Préposit.l	(Régime de la prép. sur)				Commun	
pour	P.	P.								P.D. Simple	pourquoi ?
aller	V.	J.								P.t 1.er Conjug. V.N.	
voir	V.	J.								P.t 3ème Conjug. V.A.	
un	N.	Adj. déter.	m.	s.	ac.	(Modification d'Ami sous entendu)				Numéral Cardinal	
de	P.	P.								Proprement dite Simple	
leurs	N.	P.	m.	p.	g.	(Modification de amis)				Possessif Absolu	
amis	N.	S.	m.	p.	g.	(Dépendant d'Ami sous entendu)				Commun	
qui	N.	P.	m.	s.	n.	(de avait)				Relatif	
avait	V	S.				s.	3 P.	p.é	ind.	T.S. 2.e Conj. Imparfait Auxiliaire	Temps composés
été	V.	P.								Passif 3.e Conjug.	
indisposé.	N.	A.	m.	s.	n.	(Se rapportant à Qui)				Positif	

Qui ? — Que firent-ils ? — Quand ? Lorsque le Soleil commençait à paraître sur l'Horizon — Pourquoi ? Pour aller voir un de leurs amis qui avait été indisposé.

N.B. Les explications placées ici entre deux parenthèses ne regardent pas les commençants et ne sont destinées qu'aux élèves assez avancés pour pouvoir déjà distinguer, dans chaque phrase, le nombre de membres qu'elle renferme.

N.B. Afin de conduire les élèves du simple au composé, du connu à l'inconnu, il sera utile de leur faire d'abord analyser les phrases grammaticales que nous trouverons dans le volume d'exercices, placé ensuite aux leçons de grammaire.

Rom par De Bligniere, Demuyencourt, Pucros & de Siet et Leclerc aîné.

Figure 6. *Reproduction d'un tableau d'analyse grammaticale par Gaultier (1817, 11) :* Le Père et la Mère de Zoé sortirent un matin, lorsque le Soleil commençait à paraître sur l'Horizon, pour aller voir un de leur amis qui avait été indisposé.

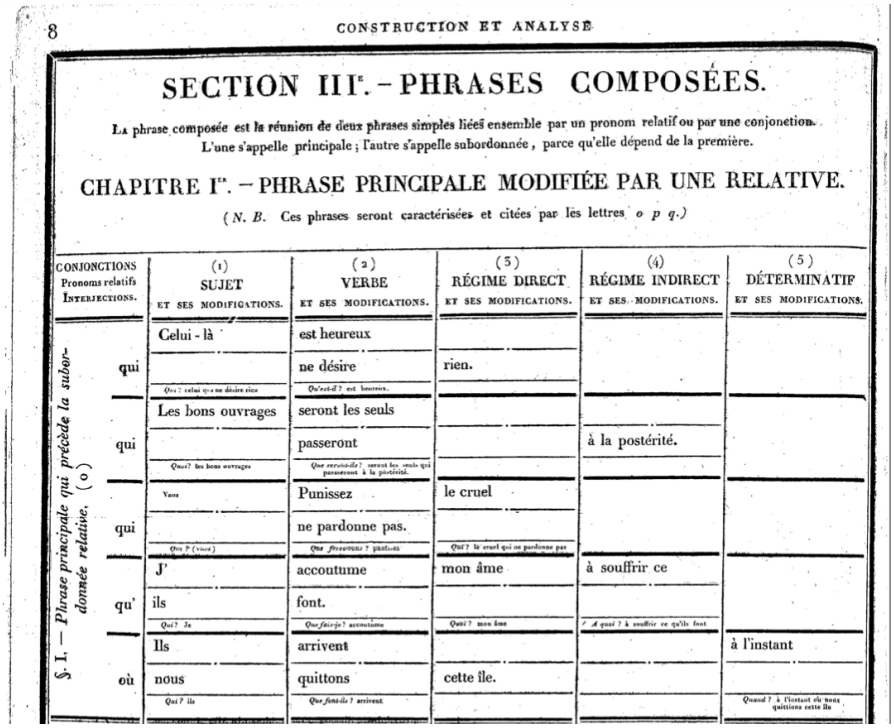
> « Pour faire l'analyse grammaticale, il faut avoir une feuille de papier, une ardoise ou un tableau noir partagé en dix colonnes. Dans une marge à gauche, on écrira les mots de la phrase à analyser les uns au-dessous des autres. Dans la première colonne, on indiquera à laquelle des trois parties primitives du discours, et dans la seconde à laquelle des dix parties secondaires du discours chaque mot appartient ; dans la troisième, la quatrième et la cinquième, on marquera le genre, le nombre et le cas des noms ; dans la sixième, la septième, la huitième et la neuvième, on indiquera le nombre, la personne, le temps en général et le mode du verbe personnel. Dans la dixième, on indiquera toutes les divisions et les subdivisions des dix parties du discours. »

Le diagramme de Gaultier de la figure 6 s'apparente fortement au format CoNLL utilisé aujourd'hui. On trouve en particulier, dans la colonne 5 intitulée *Quel cas ?*[10], un encodage des dépendances pour les noms de la phrase : ainsi *Père* et *Mère* sont analysés comme sujet (*n.* [pour nominatif] de *sortirent*), *Zoé* comme un dépendant génitival (*g.* de *Mère*), *matin* comme un complément (*régime de la préposition* dans *sous-entendue*), etc. La terminologie de Gaultier est plus traditionnelle que celles de ses prédécesseurs encyclopédistes (section 3), dont Beauzée, qui avait distingué précisément la notion morphosyntaxique de régime de la notion syntaxique de complément. Ainsi, les relations syntaxiques sont à nouveau encodées par des noms de cas (« nominatif », « génitif », etc.).

Même si on peut supposer qu'il a été utilisé à plusieurs reprises avec des élèves, ce premier type de diagramme reste sporadique dans les ouvrages de Gaultier et de ses étudiants. En revanche, un autre type de diagramme tabulaire totalise près de 200 exemples analysés dans Gaultier (1817, 17-36), et quelques-uns dans de Blignières *et al.* (1829, 228-244). Nous reproduisons dans la figure 7 quelques exemples de phrases comportant des propositions relatives. Dans cette analyse, six positions syntaxiques sont considérées : complémenteur, sujet, verbe, complément d'objet direct (régime direct), complément oblique (régime indirect) et complément circonstanciel (déterminatif)[11]. Chaque proposition est divisée en segments positionnés les uns à la suite des autres dans ces six positions. Par exemple, si l'on prend le dernier exemple de la figure 7, « Ils arrivent à l'instant où nous quittons cette île. » (Gaultier, 1817, 34),

10. La colonne 5 ne contient que le cas lui-même, mais celui-ci est complété par un texte entre parenthèses qui déborde sur les colonnes suivantes normalement consacrées aux catégories de la forme verbale. On notera, tout en bas du tableau, la mention suivante : *Les explications placées ici entre deux parenthèses ne regardent pas les commençants et ne sont destinées qu'aux élèves assez avancés pour pouvoir déjà distinguer dans chaque phrase le nombre de membres qu'elle renferme.* Les « membres » de la phrase en question sont indiqués dans la dernière colonne (non numérotée), où la phrase est découpée en quatre segments identifiés par autant de questions : *qui ? que firent-ils ? quand ? pourquoi ?* Cette segmentation et ces questions figurent aussi sur le côté gauche du tableau sous forme d'accolades.

11. On trouve déjà chez Girard (1747) des analyses syntaxiques de ce type – voir l'étude de Kahane (2020, 113-120, en particulier 118) –, mais elles ne sont pas utilisées de manière aussi systématique que chez Gaultier (1817).



8 CONSTRUCTION ET ANALYSE

SECTION III^e. – PHRASES COMPOSÉES.

La phrase composée est la réunion de deux phrases simples liées ensemble par un pronom relatif ou par une conjonction. L'une s'appelle principale ; l'autre s'appelle subordonnée, parce qu'elle dépend de la première.

CHAPITRE I^er. – PHRASE PRINCIPALE MODIFIÉE PAR UNE RELATIVE.

(*N. B.* Ces phrases seront caractérisées et citées par les lettres *o p q.*)

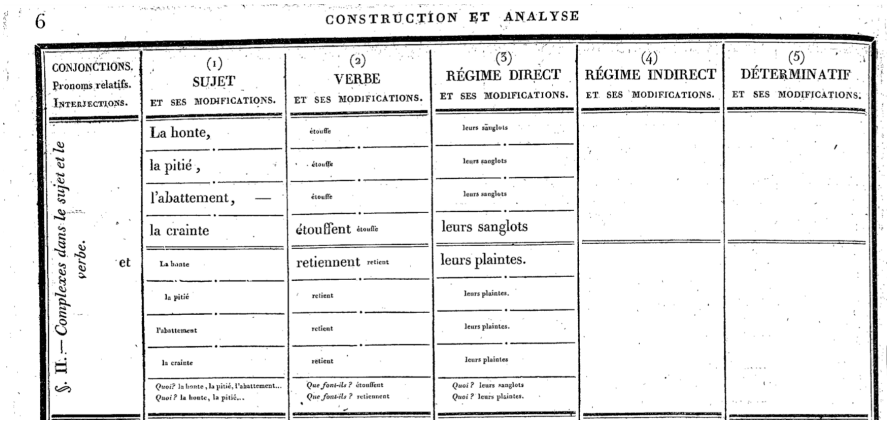
CONJONCTIONS Pronoms relatifs Interjections.	(1) SUJET ET SES MODIFICATIONS.	(2) VERBE ET SES MODIFICATIONS.	(3) RÉGIME DIRECT ET SES MODIFICATIONS.	(4) RÉGIME INDIRECT ET SES MODIFICATIONS.	(5) DÉTERMINATIF ET SES MODIFICATIONS.
§. I. — Phrase principale qui précède la subordonnée relative. (o)	Celui-là	est heureux			
qui		ne désire	rien.		
	Qui? celui qui ne désire rien	Qu'est-il? est heureux.			
	Les bons ouvrages	seront les seuls			
qui		passeront		à la postérité.	
	Quoi? les bons ouvrages	Que deviendra? seront les seuls qui passeront à la postérité.			
	Vous	Punissez	le cruel		
qui		ne pardonne pas.			
	Qui? (vous)	Que fera-t-on? punissez	Qui? le cruel qui ne pardonne pas		
	J'	accoutume	mon âme	à souffrir ce	
qu'	ils	font.			
	Qui? Je	Que fait-je? accoutume	Qui? mon âme	A quoi? à souffrir ce qu'ils font	
	Ils	arrivent			à l'instant
où	nous	quittons	cette île.		
	Qui? ils	Que fait-on? arrivent			Quand? à l'instant où nous quittons cette île

Figure 7. *Analyse de phrases complexes chez Gaultier (1817, 34)*

l'analyse indique que *ils* | *arrivent* | *à l'instant* se décompose en sujet-verbe-modifieur et la relative *où* | *nous* | *quittons* | *cette île* en complémenteur-sujet-verbe-objet. Le fait que la relative forme un constituant avec *à l'instant* est indiqué dans la troisième ligne de l'analyse (« *Quand ?* à l'instant où nous quittons cette île »).

Les analyses tabulaires que l'on rencontre pour la première fois chez Gaultier se développent dans différentes langues – en particulier en langue allemande par Becker (1829), puis sous une autre forme dans les éditions ultérieures selon Hudson, puis, sans doute sous l'influence de cette dernière, en langue anglaise, notamment dans les grammaires de Morell (1852) et de Meiklejohn (1886)[12]. La volonté des auteurs est toujours de « faire voir » la logique de l'analyse. Les colonnes formalisent le typage d'éléments récurrents. Elles permettent donc incidemment de retrouver des occurrences (tokens) de catégories grammaticales (types). L'analyse de la coordination qui en découle est particulièrement intéressante (voir figure 8) : elle rappelle les analyses en grille proposées par Blanche-Benveniste *et al.* (1979) avec la disposition

12. Nous remercions Richard Hudson pour les discussions au sujet de Gaultier, Becker, Morell et Meiklejohn. Des matériaux issus des grammaires de ces auteurs sont présentés sur son site (`dickhudson.com/uk/`).


6 CONSTRUCTION ET ANALYSE

CONJONCTIONS. Pronoms relatifs. Interjections.	(1) SUJET ET SES MODIFICATIONS.	(2) VERBE ET SES MODIFICATIONS.	(3) RÉGIME DIRECT ET SES MODIFICATIONS.	(4) RÉGIME INDIRECT ET SES MODIFICATIONS.	(5) DÉTERMINATIF ET SES MODIFICATIONS.
§. II. — Complexes dans le sujet et le verbe.	La honte,	étouffe	leurs sanglots		
	la pitié,	étouffe	leurs sanglots		
	l'abattement, —	étouffe	leurs sanglots		
	la crainte	étouffent étouffe	leurs sanglots		
et	La honte	retiennent retient	leurs plaintes.		
	la pitié	retient	leurs plaintes.		
	l'abattement	retient	leurs plaintes.		
	la crainte	retient	leurs plaintes		
	Quoi? la honte, la pitié, l'abattement... *Quoi ?* la honte, la pitié...	*Que font-ils ?* étouffent *Que font-ils ?* retiennent	*Quoi ?* leurs sanglots *Quoi ?* leurs plaintes.		

Figure 8. *Analyse de coordinations par Gaultier (1817, 20)*

verticale des paradigmes entre conjoints, à la différence que Gaultier complète l'analyse pour mettre en évidence que les paradigmes de 4 et 2 éléments se combinent pour donner $4 \times 2 = 8$ propositions élémentaires[13].

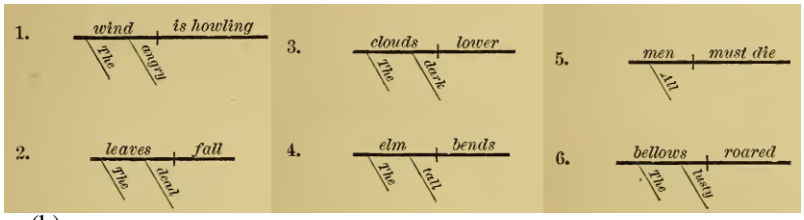
4.2. *Les premiers diagrammes hiérarchiques et les* Keys

Les premiers diagrammes hiérarchiques apparaissent dans les années 1830 dans une grammaire du latin (1832) par le grammairien allemand Johann Gustav Freidrich Billroth (1808-1836) (unique diagramme connu d'un auteur mort prématurément) et dans une grammaire de l'anglais à destination des sourds (1836) par le savant américain Frederick A. P. Barnard (1809-1889)[14]. Les données n'y sont pas représentées sous la forme de tableaux, mais sous celle d'un réseau d'éléments hiérarchisés. À partir de 1847, Stephen W. Clark propose une série de grammaires de l'anglais comportant un grand nombre d'exemples analysés par des diagrammes arborescents originaux. La naissance de ces diagrammes syntaxiques hiérarchiques ouvre en effet la possibilité de collectionner des listes d'analyses exhaustives. Les grammaires de Stephen W. Clark (1810-1901) (Clark, 1847 ; Clark, 1855) et d'Alonzo Reed (?-1899) et Brainerd Kellogg (1834-1920) (Reed et Kellogg, 1876 ; Reed et Kellogg, 1877) sont ainsi accompagnées d'ouvrages qualifiés de « *Keys* », c'est-à-dire de solutions aux exercices (Clark, 1863 ; Reed et Kellogg, 1889). Dans ces ouvrages, les phrases

13. Comme remarqué par Kahane (2012), Tesnière (1959) propose, comme dans l'analyse en grille, de traiter les relations entre conjoints orthogonalement aux relations de subordination et il met en évidence la combinaison des paradigmes dans ses stemmas 265 et 266, p. 345.

14. Voir la thèse peu diffusée de Brittain (1973), ainsi que l'étude de Mazziotta et Kahane (2017), qui décrit les propriétés des premiers diagrammes d'analyse en constituants.

(a)



(b)

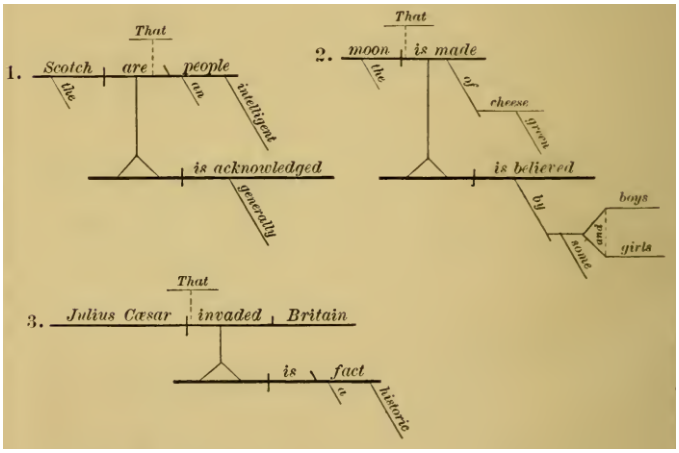


Figure 9. *Liste d'analyses diagrammatiques : (a) Reed et Kellogg, 1889 : 1 ; (b) Reed et Kellogg, 1889 : 30*

d'exercices de la grammaire correspondante (qui peuvent être inventées ou tirées de la littérature) sont accompagnées d'une analyse exhaustive sous la forme d'un diagramme[15,16]. La figure 9 donne deux exemples tirés de la *Key* de Reed et Kellogg (1889), qui s'appuie sur les grammaires des auteurs (Reed et Kellogg, 1876 ; Reed et Kellogg, 1877).

15. Nous renvoyons à Mazziotta (2016) concernant les systèmes de Clark et à Gleason (1965, 142-161) pour une présentation succincte du système de Reed et Kellogg, encore en usage de nos jours – voir, par exemple, le manuel de Otto et Bauer (2019). Chez Reed et Kellogg, chaque mot correspond à un trait, horizontal ou oblique selon qu'il s'agit d'un nom ou verbe ou d'un modifieur. Clark associe, quant à lui, chaque mot à une bulle. La structure de la phrase est la combinaison directe de ces traits ou de ces bulles, sans que les relations entre les mots ne soient représentées par un signe discret.

16. La démarche est ici radicalement opposée aux habitudes traditionnelles comme celle de Murray (1799), qui est plutôt une correction d'exercices d'identification d'erreurs orthographiques ou grammaticales. La *Key* de Murray correspondant à la 16e édition (Murray, 1812) ne comporte pas de correction des exercices d'analyse morphosyntaxique que l'édition précédente du livre d'exercices comporte (section 3).

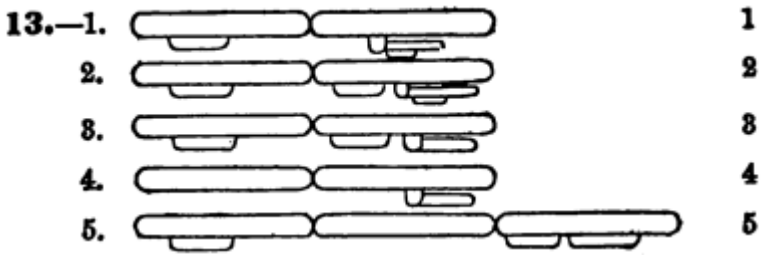


Figure 10. *Liste d'analyses diagrammatiques sans étiquetage des entités réifiant les mots (Clark 1863 : 10) : 1.* The sun rose on the sea*, 2.* A mist rose slowly from the lake*, 3.* The night passed away in song*, 4.* Morning retrurned in joy *et 5.* The mountains showed their gray heads *(Clark 1863 : 11, cf. Clark 1855 : 13)*

Les diagrammes sont regroupés en fonction du point de grammaire que chaque exemple illustre. La mise en série a comme effet de faire ressortir graphiquement la partie commune partagée par tous les diagrammes. Dans la figure 9a, on observe un trait horizontal et un petit trait vertical qui l'interrompt, ainsi que des traits obliques sous la première partie du trait horizontal ; dans la figure 9b, on repère le dessin en forme de « Y » inversé. La mise en série de diagrammes met en évidence un invariant graphique qui correspond à l'illustration d'un point de grammaire identifié dans les ouvrages auxquels se réfère la *Key*. Dans la figure 9a, il s'agit de la construction des « modified subjects » (Reed et Kellogg, 1876, (1880) 27-30). Dans la figure 9b, il s'agit des « noun clauses » (Reed et Kellogg, 1876, (1880) 81-82). L'exemple de Clark (1863), qui s'appuie sur Clark (1855), est plus précoce, mais plus complexe. Il n'étiquette pas les mots dans ses listes de diagrammes. Ce sont des phrases réelles qui figurent en vis-à-vis de la figure 10, qui représente l'analyse.

De la même manière qu'expliqué pour la *Key* de Reed et Kellogg, les diagrammes ont tous comme point commun de comporter un invariant (ici l'agencement horizontal d'au moins deux bulles). Tant Clark que Reed et Kellogg ont conscience que l'analyse syntaxique complète force à prendre position (Clark, 1863, 3) et pousse à trancher dans certains cas naturellement ambigus (Reed et Kellogg, 1889, s.n.). Il s'agit là d'une conséquence directe de la recherche d'exhaustivité, dont on sait les implications pour la constitution des corpus à l'heure actuelle. La perspective des auteurs est en quelque sorte renversée par rapport à la nôtre : eux voient effectivement leurs *Keys* comme des solutions aux exercices proposés dans d'autres ouvrages (dont ils sont la « clé »), alors que nous considérons ces *Keys* comme des corpus annotés dont les grammaires constituent le point d'entrée. Il est évident que le contraste principal avec les corpus modernes est que ces derniers permettent aux utilisateurs de définir eux-mêmes des requêtes, alors que les anciens ouvrages (format papier oblige) constituent une sorte d'index d'un nombre limité de requêtes. Il en résulte que la sélection des requêtes représentées correspond à ce que les pédagogues ont jugé pertinent de mettre à disposition de leurs lecteurs, selon une progression pédagogique qui correspond à

celle de leur grammaire et non dans la visée théorique des auteurs qui suivront (section 5) ou dans la visée exploratoire des outils d'exploitation des corpus actuels (6). Une autre différence se situe au niveau de la saillance de ce qui est représenté. Les *Keys* ne mettent en évidence les structures que de manière indirecte (il faut chercher l'invariant graphique entre les exemples). De leur côté, les outils modernes permettent la mise en évidence d'un pivot focal dans les résultats de la requête (par exemple en surlignant les mots).

La démarche d'accumulation d'exemples à des fins pédagogiques ouvre la voie à l'exploitation théorique de listes similaires.

5. Valider une théorie par des corpus arborés au XX^e^ siècle

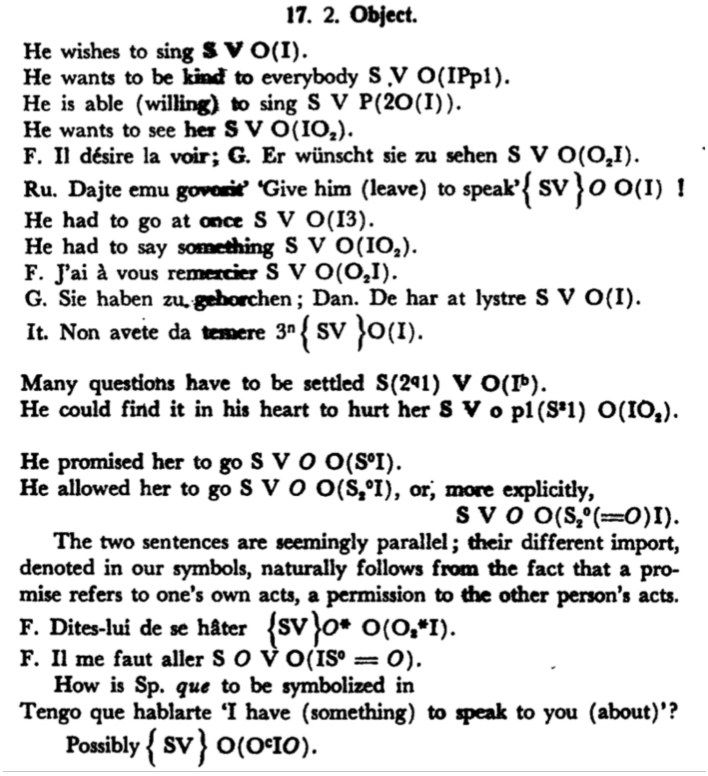
Les ouvrages du XIX^e^ siècle comportant des corpus d'exemples arborés sont des grammaires à visée pédagogique. Elles portent sur une langue unique, le français pour Gaultier, l'anglais pour Clark ou Reed et Kellogg. Les structures sont utilisées pour présenter les différentes constructions de la langue qu'il s'agit d'apprendre, sans qu'un cadre théorique général, applicable à différentes langues, ne soit dégagé. C'est au XX^e^ siècle que se développent les premiers ouvrages de syntaxe générale, dont une particularité importante est qu'ils comportent des exemples de plusieurs langues[17]. Nous observons les démarches d'Otto Jespersen (sous-section 5.1) et de Lucien Tesnière (5.2) pour illustrer notre propos.

5.1. *Otto Jespersen*

Après avoir publié son grand ouvrage de linguistique générale, *Philosophy of grammar* (Jespersen, 1924), Otto Jespersen (1860-1943) propose l'ouvrage intitulé *Analytic syntax* (Jespersen, 1937), qui est une collection organisée d'exemples dans plusieurs langues européennes (anglais, latin, danois, français, espagnol, portugais, italien, finnois, allemand, russe et grec). Il y développe un système de notation original permettant de diagrammatiser par une formule la structure syntaxique de chaque exemple étudié (sur ce système, voir Cigana, 2020, 232-234).

Il s'agit essentiellement d'une analyse en constituants, comme le souligne James D. McCawley dans sa préface de l'édition de 1984 publiée par l'University of Chicago Press. Par exemple, dans la figure 11, la phrase *He wants to see her* est analysée S V O(IO_2), indiquant que dans cette structure de type SVO, O se décompose lui-même en un infinitif I et un objet O_2. Dans d'autres analyses de ce même extrait,

17. Nous devons mentionner ici la thèse de Weil (1844), exceptionnelle à de nombreux égards. Dans ce travail consacré à l'ordre des mots, Henri Weil (1818-1909) s'intéresse à plusieurs langues (latin, grec ancien, français, anglais, allemand, turc et chinois) et montre que la linéarisation est guidée par trois types de facteurs : la structure syntaxique (il considère une structure de dépendance à la suite de Beauzée et distingue les langues et constructions à têtes finales *vs* initiales), la prosodie et la structure thème-rhème.



17. 2. Object.

He wishes to sing S V O(I).
He wants to be kind to everybody S V O(IPp1).
He is able (willing) to sing S V P(2O(I)).
He wants to see her S V O(IO_2).
F. Il désire la voir; G. Er wünscht sie zu sehen S V O(O_2I).
Ru. Dajte emu govorit' 'Give him (leave) to speak' { SV } O O(I) !
He had to go at once S V O(I3).
He had to say something S V O(IO_2).
F. J'ai à vous remercier S V O(O_2I).
G. Sie haben zu gehorchen; Dan. De har at lystre S V O(I).
It. Non avete da temere 3^n { SV } O(I).

Many questions have to be settled S(2^q1) V O(I^b).
He could find it in his heart to hurt her S V o p1(S^s1) O(IO_2).

He promised her to go S V O O(S^0I).
He allowed her to go S V O O($S_2{}^0$I), or, more explicitly,
S V O O($S_2{}^0$(=O)I).
The two sentences are seemingly parallel; their different import, denoted in our symbols, naturally follows from the fact that a promise refers to one's own acts, a permission to the other person's acts.
F. Dites-lui de se hâter {SV} O* O(O_2*I).
F. Il me faut aller S O V O(IS^0 = O).
How is Sp. *que* to be symbolized in
Tengo que hablarte 'I have (something) to speak to you (about)'?
Possibly { SV } O(O^cIO).

Figure 11. *Section sur les infinitifs en position objet de Jespersen (1937, 48-49)*

Jespersen introduit des éléments vides, comme S^0 (pour *He promised her to go*), ou la coindexation d'un trace, notée *, avec une autre position (pour *Dites-lui de se hâter*). On pourra consulter, en contraste, l'interprétation dépendancielle de Cigana (2020) des annotations de Jespersen.

Dans son introduction, Jespersen souligne l'importance de l'exemple par la maxime suivante (non attribuée) : *Nur das Beispiel führt zum Licht; Vieles Reden tut es nicht* (Seul l'exemple permet de faire la lumière ; Les longs discours ne suffisent pas). Il poursuit en précisant la nature des exemples étudiés :

> « Certains sont empruntés à mes lectures, d'autres aux grammaires que l'on peut trouver couramment ; mais un grand nombre sont en fait des phrases préfabriquées. […] je n'ai peut-être pas réussi à éviter une certaine monotonie […] Ma seule excuse est que cet ouvrage n'a aucune prétention pédagogique […] il importait surtout de trouver des phrases courtes et non ambiguës qui permettent d'illustrer l'emploi et l'utilité des formules dans un grand nombre de constructions grammaticales. » (Jespersen, 1937, (1971) 16-17)

Le travail à visée non pédagogique de Jespersen reste isolé dans cette première moitié du XX^e siècle. Il faut attendre Nida (1966) ou Ross (1967) pour voir à nouveau des ouvrages comprenant de longues listes d'exemples syntaxiquement analysés dans une perspective théorisante.

5.2. *Lucien Tesnière*

Les *Éléments de syntaxe structurale*, ouvrage posthume de Lucien Tesnière (1893-1954) commencé en 1932 et publié en 1959, sont connus pour introduire un modèle théorique complet d'analyse en dépendance. Le livre contient également des matériaux procédant de la même démarche. Plusieurs analyses exhaustives de textes sous forme de diagrammes (dits « stemmas ») figurent à la fin de l'ouvrage (1959, 638-653)[18] : deux poèmes – *La cigale et la fourmi* de La Fontaine (voir la figure 12)[19] et *Le vase brisé* de Sully Prudhomme –, une longue phrase en grec de Platon et une autre de Tacite en latin, des extraits du *Polyeucte* et du *Cid* de Corneille, d'*Athalie* de Racine, de *Booz endormi* de Victor Hugo et du *Crime de Sylvestre Bonnard* d'Anatole France. Si nous citons la liste de ces textes, c'est qu'il s'agit, à notre connaissance, du premier exemple de corpus arborés basé sur des textes suivis attestés. Tous les travaux mentionnés jusque-là contenaient uniquement des analyses de phrases isolées, souvent construites ou simplifiées. Ces exemples sont précédés d'un chapitre intitulé « Le stemma intégral » (Tesnière, 1959, 629-32), dont voici quelques extraits :

> « 1. – Si nous faisons usage de toutes les possibilités que la stemmatisation d'une phrase peut nous offrir pour en représenter graphiquement l'infinie complication structurale, nous aboutissons à un stemma d'une complexité telle que nous n'y avons pratiquement à peu près jamais recouru au cours de cet ouvrage.
>
> 2. – Mais à côté des stemmas partiels et fragmentaires que nous avons utilisés pour faire comprendre telle ou telle partie de la syntaxe structurale, il est possible, au moins théoriquement, de concevoir un stemma intégral faisant état de tous les éléments structuraux rencontrés dans une phrase, ou tout au moins de se rapprocher de cet idéal. [...]
>
> 5 – Pratiquement nous n'avons guère eu l'occasion de présenter de stemmas de cette nature, le souci de la clarté de notre exposé nous ayant au

18. La tradition philologique de ces diagrammes n'est pas claire : dans l'introduction des *Éléments*, Fourquet indique qu'ils « ont été redessinés par M. Georges Bichet » (Tesnière, 1959, iv).

19. Indiquons quelques conventions utilisées par Tesnière dans ses stemmas. Les dépendances sont indiquées par des traits pleins, obliques pour les relations tête-dépendant et horizontaux pour la coordination. Les traits hachurés indiquent des relations de coréférence. Les symboles en forme de « T » indiquent un cas particulier de combinaison que Tesnière nomme la translation : ainsi dans le deuxième stemma, *de* translate *mouche* pour lui permettre d'occuper une position normalement dévolue à un adjectif. Les ronds pointés indiquent un translatif zéro. On pourra consulter Kahane et Osborne (2015) pour une analyse critique de l'ouvrage de Tesnière.

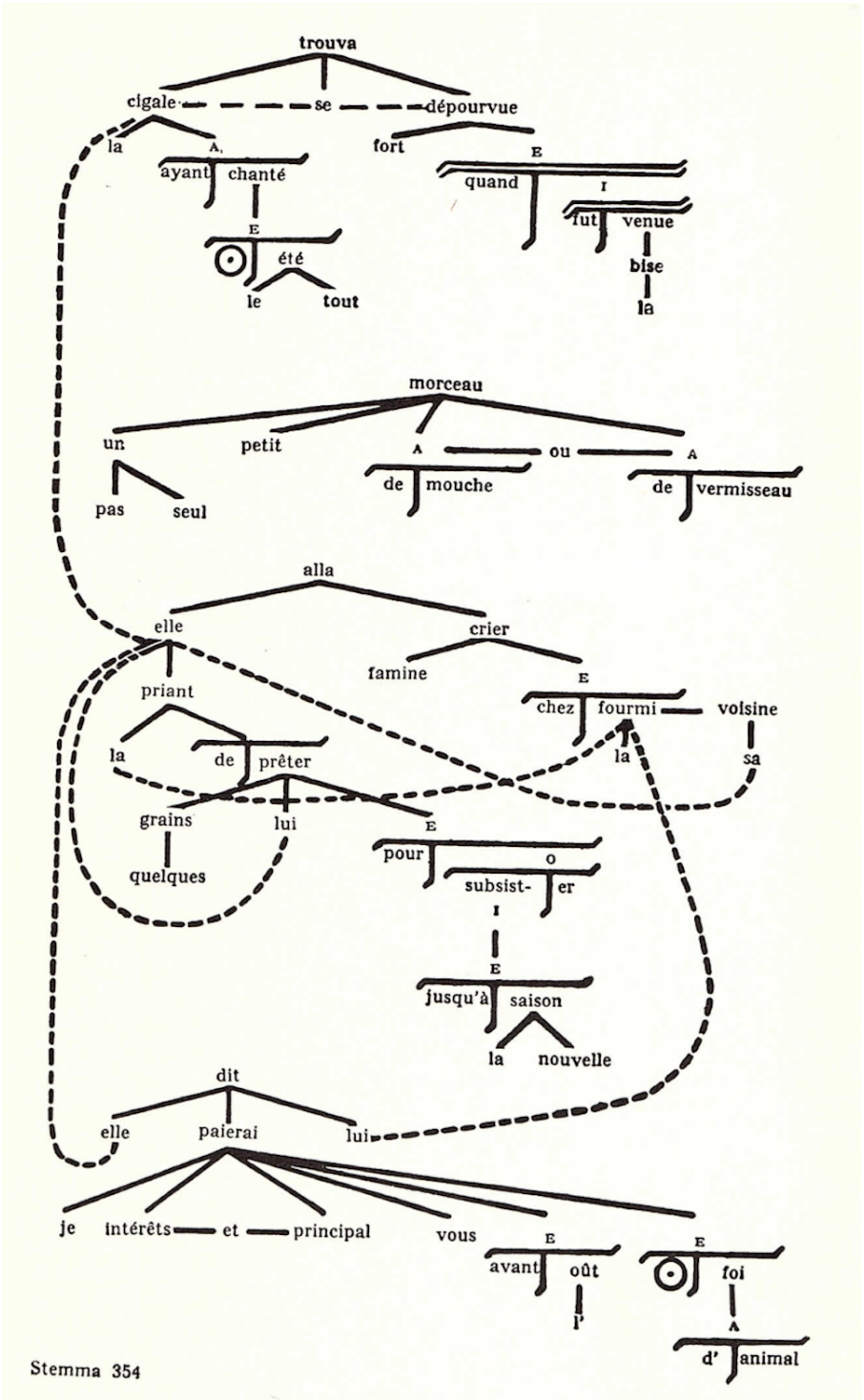


Figure 12. *Première moitié de l'analyse de* La cigale et la fourmi *par Tesnière (1959 : 638)*

> contraire toujours fait une loi de ne pas compliquer le fait à faire comprendre de faits superfétatoires. [...]
>
> 10. – Mais en réalité le langage ne se présente pas à nous comme une succession de phrases isolées. La normale est au contraire le plus souvent une suite de phrases qui expriment en principe des idées agencées entre elles de façon à former un tout organisé en vue d'exprimer, soit oralement, soit par écrit, une pensée plus ou moins complexe.
>
> 11. – [...] Le stemma intégral d'une interminable conversation ou d'un long discours comporte celui de toutes les phrases qui les composent. »

Tesnière souligne à la fois l'utilité d'adopter des exemples simples pour présenter le cadre théorique et la nécessité, au moment voulu, de confronter la théorie avec de vrais textes dans toute leur complexité.

Nous allons maintenant revenir sur la période actuelle à la lumière de notre parcours des deux siècles précédents.

6. Conclusion : des corpus arborés au XXIe siècle pour quel objectif

Les analyses de textes telles que celles proposées par Tesnière à la fin de son ouvrage sont difficiles à exploiter pour extraire de l'information. Elles servent à illustrer la théorie et, bien que l'on ne doutera pas de la visée pédagogique de l'auteur, elles nous paraissent surtout utiles à l'auteur, qui vérifie que son système lui permet d'appréhender n'importe quel énoncé. Des décennies après cet auteur, les systèmes de requêtes et de concordanciers, que le passage au numérique a permis, rendent possible de sélectionner toutes les constructions d'un type donné. Il devient alors également possible de combiner les avantages d'une analyse suivie d'un texte complet avec un classement pédagogique des phénomènes, comme dans les grammaires du XIXe siècle (section 4). Le corpus arboré peut ainsi être appréhendé selon différentes facettes, chaque requête offrant une vue particulière sur les données.

On a vu que les corpus analysés en syntaxe ont une longue histoire avant le numérique. Ils apparaissent sous une double pression : une pression pédagogique, pour donner aux apprenants d'une langue des exemples des différentes constructions de la langue (sections 3 et 4) ; une pression théorique (section 5), pour vérifier que les modèles de langue proposés ont une couverture exhaustive des phénomènes rencontrés. Avec l'apparition du numérique, les corpus sont devenus électroniques et l'annotation syntaxique a subi de nouvelles contraintes : paradoxalement, la dématérialisation de l'encodage a entraîné une simplification des annotations. Il est plus simple sur un texte informatique d'ajouter des parenthèses que de tracer des traits et des flèches (voir l'exemple du *Penn Tree Bank* de la figure 1). Il faudra attendre les années 2000 pour que se développe un moyen simple d'encoder des arbres de dépendance, le format tabulaire CoNLL (voir l'exemple de la figure 4). Mais un tel format est peu iconique, peu ergonomique pour des humains qui doivent parcourir la table en passant d'un identifiant à l'autre et n'est véritablement lisible qu'avec une interface proposant une repré-

sentation graphique sous forme d'un diagramme arborescent. Chacun de ces formats d'encodage, parenthésage ou format tabulaire, contraint fortement les utilisateurs : même si toutes sortes d'informations peuvent être encodées quelque part, certaines le sont plus facilement que d'autres, ce qui a conduit à un certain appauvrissement des représentations utilisées[20]. D'un autre côté, la numérisation a permis à l'annotation de connaître un nouvel essor : les données servent à présent à la fois d'input et d'output aux outils de TAL.

Aujourd'hui, le lien entre traitement automatique des langues et corpus arborés est complexe. Si au moment de l'avènement des corpus arborés numériques, l'apprentissage d'analyseurs syntaxiques automatiques sur des corpus arborés[21] apparaissait comme l'outil le plus puissant pour obtenir des modèles de langue, le développement actuel des méthodes d'analyse distributionnelle sur de très grands corpus, comme les *transformers* BERT (Devlin *et al.*, 2019), rend l'analyse syntaxique souvent dispensable. Les corpus arborés se trouvent alors renvoyés à leur usage initial, qui est celui d'une source d'exemples pour la pédagogie et pour l'étude théorique et la recherche de constructions nouvelles. Notons tout de même que, même s'ils ne jouent plus nécessairement un rôle central dans le développement à proprement parler des outils de traitement automatique des langues, les corpus arborés restent utiles pour l'évaluation des outils et pour vérifier que ceux-ci ont bien saisi la structure des énoncés.

À l'heure actuelle, ce sont donc davantage les développeurs et utilisateurs de corpus arborés qui ont besoin du traitement automatique des langues que l'inverse. Les méthodes d'apprentissage permettent un développement accéléré des corpus arborés : après avoir annoté manuellement quelques phrases, il est déjà possible d'apprendre un analyseur donnant des résultats suffisamment bons pour pré-annoter automatiquement le reste du corpus. En répétant régulièrement cette opération (procédure dite de *bootstrapping* ; Breiman, 1996 ; Seraji *et al.*, 2012), le processus d'annotation devient à chaque itération plus performant. On peut donner pour exemple les résultats d'une expérience faite par Guiller (2020) sur le corpus arboré SUD-Naija_NSC avec un parser bi-affine utilisant un BERT mutilingue[22]. Comme le montre la figure 13, avec 10 phrases, on dépasse 80 % de précision pour la reconnaissance des parties du discours (POS) et avec 100 phrases on a plus de 85 % d'étiquettes fonctionnelles correctes (LUS) et 75 % de gouverneurs reconnus (UAS)[23]. Ces résultats sont rendus possibles par l'utilisation d'un transformer BERT et donc d'un apprentissage préalable sur un

20. Parallèlement, les théoriciens du langage, dans un souci de formalisation, se sont eux-mêmes astreints à utiliser des structures mathématiques bien définies, ce qui les a amenés aussi à privilégier des structures d'arbres au détriment de structures plus ambitieuses. Voir par exemple, les arbres de constituants à la base de tous les modèles génératifs depuis Chomsky (1957).

21. La littérature sur l'apprentissage automatique d'analyseurs à partir de corpus arborés est immense. On pourra consulter Kübler *et al.* (2009) pour les principes de base.

22. Le naija est un pidgin-créole de l'anglais et le corpus Naija_NSC utilise l'orthographe standard de l'anglais pour les mots lexicaux. Les résultats seraient certainement moins bons avec une langue sans lien avec les langues ayant servi à l'entraînement du BERT multilingue.

23. Le LAS (*Labelled Attachment Score*) calcule le nombre de mots qui ont à la fois le bon gouverneur (UAS) et la bonne étiquette fonctionnelle (LUS).

modèle	POS	LUS	UAS	LAS
1 phrases	51.32	41.18	16.14	10.07
10 phrases	82.13	72.95	39.53	33.72
100 phrases	93.38	86.17	75.28	68.15
1000 phrases	97.29	93.44	90.92	86.46
5000 phrases	97.89	94.73	93.39	89.48

Figure 13. *Performances atteintes par un parser bi-affine basé sur un BERT multilingue et entraîné sur des tailles différentes du SUD-Naija_NSC (Guiller 2020, 51)*

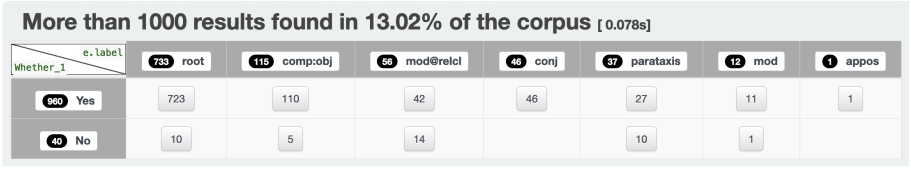


Figure 14. *Requête Grew-match avec double clustering*

grand corpus brut – voir également les expériences de de Lhoneux *et al.* (2022). Il est aussi possible d'utiliser des méthodes d'apprentissage par transfert si l'on possède des corpus arborés d'autres langues ayant des constructions similaires (Aufrant, 2018).

Concernant l'utilisation des corpus arborés, les systèmes de requêtes permettent d'accéder aux données et de faire différents tris. La plateforme Grew-match (Guillaume, 2021) permet notamment de regrouper les résultats d'une requête selon différentes clés. La figure 14 donne le résultat d'une requête[24] qui nous dit si oui ou non le sujet `S` d'un verbe `V` se trouve avant `V` (`whether_1`) en fonction de la position syntaxique de `V` (`e.label`). On voit par exemple que quand le verbe `V` est la tête d'une relative (`mod@relcl`), il y a 14 sujets inversés contre 42 dans l'ordre standard. En cliquant sur le 14, on obtient les exemples correspondants.

Associés à de tels outils, les corpus arborés deviennent des instruments puissants pour l'analyse et la description d'une langue. Le développement linéaire de la base Universal Dependencies depuis 2014 (environ 25 treebanks et 15 langues supplémen-

24. La requête est `pattern { e: G−>V; V −[subj]−> S; S[upos=NOUN] }`. Autrement dit, on recherche un motif (`pattern`) du graphe avec deux nœuds `V` et `S`, où `S` est un sujet nominal de `V`. Le lien `e` est introduit pour pouvoir interroger la fonction de `V`, c'est-à-dire la relation qui le lie à son gouverneur `G`. Les résultats sont triés d'abord en fonction de la position `S` par rapport à `V` (`S << V`; réponse « Yes » ou « No ») et de l'étiquette de la relation entre `V` et son gouverneur `G` (`e.label`). (`universal.grew.fr/?custom=62cdcb8a55a6b`).

taires chaque année), ainsi que la diversité toujours plus grande des langues qui bénéficient d'un corpus arboré, montre un intérêt soutenu de l'ensemble de la communauté des linguistes, des linguistes de terrain aux talistes, en passant par les théoriciens. Le développement de corpus arborés, initié par des linguistes pour la compréhension de la grammaire, l'enseignement et la validation des théories linguistiques, puis boosté par la communauté TAL de l'avènement du numérique à aujourd'hui, se trouve ainsi à nouveau investi par les linguistes avec des possibilités de développement et d'exploitation démultipliées par les travaux en TAL.

Remerciements

Nous souhaitons remercier Loïc Grobol pour ses commentaires sur la première version de ce travail, Richard Hudson pour les échanges à propos des grammaires anciennes, ainsi que les trois relecteurs de la revue pour leurs très nombreuses remarques.

7. Bibliographie

Aufrant L., Training parsers for low-resourced languages : improving cross-lingual transfer with monolingual knowledge, PhD thesis, Université Paris Saclay, 2018.

Barnard F. A. P., *Analytic grammar, with symbolic illustration*, French, New York, 1836.

Beauzée N., « Régime », *in* D. Diderot, J. L. R. D'Alembert (eds), *Encyclopédie*, vol. 14, p. 5-11, 1765.

Becker K. F., *Deutsche Grammatik*, J. C. Hermann'sche, Frankfurt, 1829.

Billroth J. G. F., *Lateinische Syntax für die obern Klassen gelehrter Schulen*, Weidmann, 1832.

Blanche-Benveniste C., Borel B., Deulofeu J., Durand J., Giacomi A., Loufrani C., « Des grilles pour le français parlé », *Recherches sur le Français Parlé*, n° 2, p. 163-206, 1979.

Breiman L., « Bagging predictors », *Machine learning*, vol. 24, n° 2, p. 123-140, 1996.

Brittain R. C., A critical history of systems of sentence diagramming in English, PhD thesis, University of Texas, Austin, 1973.

Buchholz S., Marsi E., « CoNLL-X shared task on multilingual dependency parsing », *Proceedings of the tenth Conference on Computational Natural Language Learning (CoNLL)*, p. 149-164, 2006.

Buffier C., *Grammaire françoise sur un plan nouveau*, Le Clerc-Brunet-Leconte & Montalant, Paris, 1709.

Chevalier J.-C., *Histoire de la syntaxe : Naissance de la notion de complément dans la grammaire française (1530–1750)*, Droz, Paris, 1968.

Chomsky N., *Syntactic structures*, Mouton, 1957.

Cigana L., « Some aspects of dependency in Otto Jespersen's structural syntax », *in* A. Imrényi, N. Mazziotta (eds), *Chapters of Dependency Grammar : A historical survey from antiquity to Tesnière*, John Benjamins, Amsterdam/Philadelphia, p. 215-251, 2020.

Clark S. W., *The science of the English grammar : A practical grammar in which words, phrases, and sentences are classified to their offices, and their relation to each other, illustrated by a complete system of diagrams*, H. W. Barnes & Company, Cincinnati, 1847.

Clark S. W., *The science of English language. A practical grammar [...]. Revised edition*, A.S. Barnes & Co., Derby, Bradley & Co., New York, 1855.

Clark S. W., *Key to Clark's grammar : in which the analyses of the sentences of the grammar are indicated by diagrams*, A.S. Barnes & Burr, New York, 1863.

de Blignières, Demoyencourt, Ducrot (de Sixt), Le Clerc ainé, *Élemens de grammaire française, extraits de la grammaire de l'abbé Gaultier*, Jules Renouard, Paris, 1829.

de Lhoneux M., Zhang S., Søgaard A., « Zero-Shot Dependency Parsing with Worst-Case Aware Automated Curriculum Learning », *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, Association for Computational Linguistics, Dublin, Ireland, p. 578-587, 2022.

Devlin J., Chang M.-W., Lee K., Toutanova K., « Bert : Pre-training of deep bidirectional transformers for language understanding », 2019.

Dumarsais C. C., « Construction », *in* D. Diderot, J. L. R. D'Alembert (eds), *Encyclopédie*, vol. 4, p. 73-92, 1754.

Einarsson J., « Talbankens skriftspråkskonkordans », 1976.

Gaultier L., *Atlas de grammaire, ou tables propres à exciter et à soutenir l'attention des enfans dans l'étude de cette science*, Jules Renouard, Paris, 1817.

Gerdes K., « Collaborative dependency annotation », *Proceedings of the second international conference on dependency linguistics (DepLing)*, p. 88-97, 2013.

Gerdes K., Guillaume B., Kahane S., Perrier G., « SUD or surface-syntactic universal dependencies : An annotation scheme near-isomorphic to UD », *Proceedings of the second Universal Dependencies Workshop (UDW)*, Association for Computational Linguistics (ACL), 2018.

Girard G., *Les vrais principes de la langue françoise ou la parole réduite en méthode*, Le Breton, Paris, 1747.

Gleason H. A. J., *Linguistics and English grammar*, Holt, Rinehart and Winston, New York, Chicago, San Francisco, Toronto and London, 1965.

Guibon G., Courtin M., Gerdes K., Guillaume B., « When collaborative treebank curation meets graph grammars », *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*, 2020.

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

Guiller K., *Analyse syntaxique du pidgin-créole du Nigéria à l'aide d'un transformer (BERT) : Méthodes et résultats*, Université Sorbonne Nouvelle, 2020. Mémoire de master.

Hajič J., « Building a syntactically annotated corpus : The prague dependency treebank », *in* E. Hajičová (ed.), *Issues of valency and meaning : Studies in honour of Jarmila Panevová*, Karolinum, p. 106-132, 1998.

Hajic J., Vidová-Hladká B., Pajas P., « The prague dependency treebank : Annotation structure and support », *Proceedings of the IRCS workshop on linguistic databases*, p. 105-114, 2001.

Hall J., Nivre J., « A generic architecture for data-driven dependency parsing », *Proceedings of the 15th Nordic Conference of Computational Linguistics (NODALIDA 2005)*, p. 47-56, 2006.

Heinecke J., « ConlluEditor : A fully graphical editor for Universal Dependencies treebank files », *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest)*, p. 87-93, 2019.

Imrényi A., Mazziotta N., *Chapters of Dependency Grammar : A historical survey from antiquity to Tesnière*, nº 212 in *Studies in Language Companion Series*, John Benjamins, Amsterdam/Philadelphia, 2020.

Jespersen O., *The philosophy of language*, Allen & Unwin, Londres, 1924.

Jespersen O., *Analytic syntax*, Allen & Unwin, Londres, 1937.

Kahane S., « De l'analyse en grille à la modélisation des entassements », *in* S. Caddeo, M.-N. Roubaud, M. Rouquier, F. Sabio (eds), *Penser les langues avec Claire Blanche-Benveniste*, Presses Universitaires de Provence, p. 101-116, 2012.

Kahane S., « How dependency syntax found its modern form in the French Encyclopedia : From Buffier (1709) to Beauzée (1765) », *in* A. Imrényi, N. Mazziotta (eds), *Chapters of Dependency Grammar : A historical survey from antiquity to Tesnière*, John Benjamins, Amsterdam/Philadelphia, p. 85-131, 2020.

Kahane S., Osborne T., « Translators' introduction », *Elements of structural syntax*, John Benjamins, Amsterdam/Philadelphia, p. xxix-lxxiv, 2015.

Kemp J. A., *John Wallis's grammar of the English language*, Longman, London, 1972.

Kübler S., McDonald R., Nivre J., « Dependency parsing », *Synthesis Lectures on Human Language Technologies*, vol. 1, nº 1, p. 1-127, 2009.

Luotolahti J., Kanerva J., Pyysalo S., Ginter F., « SETS : Scalable and efficient tree search in dependency graphs », *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL) : Demonstrations*, p. 51-55, 2015.

Marcus M., Santorini B., Marcinkiewicz M. A., « Building a large annotated corpus of English : The Penn Treebank. 1993 », *Computational linguistics*, 1993.

Mazziotta N., « Drawing sentences before syntactic trees : Stephen Watkins Clark's sentence diagrams (1847) », *Historiographia linguistica*, vol. 43, nº 3, p. 301-342, 2016.

Mazziotta N., Kahane S., « To what extent is immediate constituency analysis dependency-based ? A survey of foundational texts », *Proceedings of the fourth international conference on Dependency Linguistics (Depling)*, ACL, p. 116-126, 2017.

Murray L., *A key to the exercises : adapted to L. Murray's English grammar*, Longman & Rees, Darton & Harvey et Wilson, Spence & Mawman, London, 1799.

Murray L., *English exercises, adapted to Murray's English grammar : [...]*, 16 edn, Collins & Co., New York, 1812.

Nida E., *A synopsys of English syntax*, Mouton and Co, London/The Hague, 1966.

Nivre J., De Marneffe M.-C., Ginter F., Goldberg Y., Hajic J., Manning C. D., McDonald R., Petrov S., Pyysalo S., Silveira N. *et al.*, « Universal dependencies v1 : A multilingual treebank collection », *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*, p. 1659-1666, 2016.

Otto J., Bauer S. W., *The diagramming dictionary : A complete reference tool for young writers, aspiring rhetoricians, and anyone else who needs to understand how English works*, Well-Trained Mind Press, Charles City, Virginia, 2019. OCLC : 1104510563.

Reed A., Kellogg B., *Graded lessons in English. An elementary English grammar [...]*, Clark and Maynard, New York, 1876.

Reed A., Kellogg B., *Higher lessons in English. A work on grammar and composition [...]*, Clark and Maynard, New York, 1877.

Reed A., Kellogg B., *A key containing diagrams of the sentences given for analysis in Reed and Kellogg's Graded lessons in English and Higher lessons in English*, Effingham Maynard & Co., New York, 1889.

Seraji M., Megyesi B., Nivre J., « Bootstrapping a Persian dependency treebank », *Linguistic Issues in Language Technology*, 2012.

Tesnière L., *Éléments de syntaxe structurale*, Klincksieck, Paris, 1959.

Tyers F., Sheyanova M., Washington J., « UD Annotatrix : An annotation tool for Universal Dependencies », *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories (TLT)*, p. 10-17, 2017.

Wallis J., *Grammatica Linguae Anglicanae*, Leon Lichfield, Oxford, 1653. [Grammar of the English Language].

Weil H., De l'ordre des mots dans les langues anciennes comparées aux langues modernes, PhD thesis, Sorbonne, Paris, 1844.