Un corpus arboré pour le français : le French Treebank

Anne Abeillé* — **Lionel Clément**** — **Loïc Liégeois*****

* *Laboratoire de Linguistique Formelle (LLF), Université Paris Diderot*
anne.abeille@univ-paris-diderot.fr
** *LaBRI , Université Bordeaux*
lionel.clement@u-bordeaux.fr
*** *CLILLAC-ARP et LLF, Université Paris Diderot*
loic.liegeois@univ-paris-diderot.fr

RÉSUMÉ. *Nous présentons un bilan du Corpus arboré du français, ou French Treebank (FTB) (1996-2016), qui est une ressource lexicale et syntaxique unique en son genre, richement annotée (et validée manuellement) pour les linguistes, et pour le TAL, avec environ 300 utilisateurs dans le monde. Après avoir exposé les principes de construction, et les principaux choix d'annotation, nous présentons l'état final du corpus, ses différents formats, et une première évaluation. Nous présentons aussi quelques ressources dérivées et des exemples d'interrogation.*

ABSTRACT. *We present a review of the French Treebank (FTB) (1996-2016), a lexical and syntactic resource with rich annotation and manual validation, which is usable by linguists and for NLP and has about 300 users in the world. We summarize the building principles and the main annotation choices, and describe the final version, the different formats and a first evaluation. We also present some derived resources and some query examples.*

MOTS-CLÉS : *corpus arboré, français, syntaxe.*

KEYWORDS: *treebank, French, syntax.*

1. Introduction

Nous présentons la version finale du French Treebank (FTB), projet initié à l'université Paris Diderot en 1996, avec le soutien de l'IUF, du CNRS, du LLF, de la DGLFLF et du CNRTL, et achevé en 2016.
Le projet a joué un rôle pionnier dans l'« outillage » de la langue française (Habert, 2004), car aucun corpus de référence n'existait à l'époque pour la syntaxe du français. Il consistait à annoter des textes écrits, sur le modèle de l'annotation du *Wall Street Journal* effectuée dans le cadre du Penn Treebank (Taylor *et al.*, 2003). Le journal *Le Monde* a été choisi pour sa disponibilité, son écriture soignée avec très peu de fautes typographiques, et la variété des sujets abordés.
Après avoir détaillé les différentes couches d'annotation (mots, syntagmes, fonctions), nous présentons la version finale du corpus avec ses différents formats, y compris la version en dépendances, et quelques chiffres concernant la distribution des catégories et des syntagmes. Dans un second temps, nous présentons une première évaluation, pour les différents niveaux d'annotation, ainsi que les principales utilisations du corpus, y compris les ressources dérivées, et enfin une comparaison avec d'autres corpus français plus récents annotés pour la syntaxe.

2. Les différentes couches d'annotation du FTB

Dans un premier temps ont été réalisées les annotations lexicales (catégories, sous-catégories, flexion, mots composés avec composants) (Abeillé et Clément, 1999b) ; puis les annotations syntaxiques (constituants majeurs, fonctions grammaticales) (Abeillé *et al.*, 2000 ; Abeillé *et al.*, 2003 ; Abeillé et Barrier, 2004). Elles ont été annotées par des outils dédiés (*taggeur*, *chunker*, étiqueteur fonctionnel) (Kinyon, 2001 ; Toussenel, 2001 ; Clément, 2001) et validées à la main, avec double correction par des annotateurs linguistes. Dans un dernier temps, ont été ajoutées des métadonnées pour chacun des articles : auteur, date, domaine.

2.1. *L'annotation lexicale*

L'annotation lexicale est particulièrement riche puisqu'elle comporte non seulement la catégorie (ou *part-of-speech*, POS) mais aussi la sous-catégorie, le lemme, la flexion et les catégories internes aux mots composés.
L'annotation automatique a été réalisée par des outils dédiés à cet effet, puis validée par des annotateurs linguistes. Une première étape a concerné la segmentation en phrases et en mots, avec validation manuelle pour tous les mots composés. Pour l'annotation des catégories, des sous-catégories et de la flexion, a été utilisé un *taggeur* dédié (Clément, 2001), fondé sur (Brill, 1993), avec plus 322 règles contextuelles écrites à la main, et un jeu réduit de 103 étiquettes, avec un taux d'erreur estimé à l'époque à 8 %. Les annotations ont été validées et enrichies par des annotateurs linguistes, puis complétées pour les lemmes avec des dictionnaires externes. Les composants de com-

posés (avec un jeu d'étiquettes réduit) ont fait l'objet d'une campagne d'annotation spécifique. Les outils informatiques et les procédures de validation sont décrits dans (Clément, 2001 ; Abeillé *et al.*, 2003) et les consignes d'annotation dans les guides associés au corpus (Abeillé et Clément, 1999a). Nous présentons ici les principaux choix d'annotation.

2.1.1. *L'annotation des catégories lexicales*

Le corpus compte 11 catégories lexicales, auxquelles s'ajoutent les étiquettes ET (mot étranger) et PONCT (ponctuation). La plupart des catégories, sauf Interjection, Verbe et Préposition, comportent des sous-catégories, par exemple subordination ou coordination pour Conjonction et propre ou commun pour Nom. L'ensemble des catégories avec leurs sous-catégories donne 41 étiquettes, et 218 une fois ajoutées les informations flexionnelles (voir tableau1). Par comparaison, un corpus comme le Penn Treebank comporte 36 étiquettes, en raison de la morphologie moins riche de l'anglais et d'un plus faible nombre de sous-catégories. À titre d'exemple pour le français, un corpus comme Frantext catégorisé comporte 22 étiquettes (voir section 6.1).

Ont été distinguées deux catégories : clitique (CL) pour les pronoms faibles, et PRO pour les pronoms forts. L'annotation morphologique prend en compte la flexion mais non la dérivation. L'étiquette PRE (pour préfixe) n'est utilisée que pour des préfixes détachables, écrits avec un trait d'union, parfois issus de mots comme *franco-*, pour *franco-allemand*.
Les mots étrangers intégrés à la syntaxe de la phrase sont étiquetés comme des mots français. Ceux qui ont l'étiquette ET sont ceux pour lesquels aucune catégorie ne peut être restituée, par exemple parce qu'ils sont en citation (*Errare humanum est* par exemple).

2.1.2. *L'annotation de la flexion*

Les informations de genre et de nombre sont ajoutées aux Adjectifs, Déterminants, Noms et Pronoms, plus la personne pour les possessifs et les Pronoms. Outre le nombre et la personne, les formes verbales sont aussi annotées pour le mode et le temps, ainsi que pour le genre pour les participes passés. Nous notons ces informations même si les formes ne sont pas distinctes. Ainsi, en contexte, la forme *rouge* sera notée comme masculin ou féminin, la forme *peux* comme 1re ou 2e personne du singulier, etc. Dans le cas des Pronoms, l'annotation se fait selon leur antécédent (*qui* relatif reçoit ainsi les mêmes genre, nombre et personne que son antécédent) ou leur référent (*je* reçoit l'information de genre correspondant à celui du locuteur). Dans le cas des noms propres, par exemple de marque ou de ville, l'information de genre n'est pas toujours disponible et reste alors non renseignée (Maurel et Belleil, 1996). Avec le recul, il aurait été plus facile de regrouper les numéraux sous une seule étiquette au lieu de quatre (Det, A, N, PRO). Il aurait aussi été intéressant de distinguer le *il* ou *ce* impersonnel ainsi que participe passé et participe passif, ce qui a été fait dans des versions ultérieures (Ribeyre *et al.*, 2014).

Catégorie	Étiquette	Sous-catégories	Flexion	Exemples
Adjectif	A	card, excl, indéf, inter, ord, poss, qual	genre, nomb, pers	*facile, mien, quelques, trois, troisième*
Adverbe	ADV	-, excl, inter, nég	-	*bien, heureusement, si*
Clitique	CL	objet, sujet, réfl	genre, nomb, pers	*je, toi, se*
Conjonction	C	coord, sub	-	*et, mais, que, si*
Déterminant	D	card, dém, déf, excl, indéf, inter, part, poss, nég	genre, nomb, pers	*ces, la, un, quel*
Mot étranger	ET	-	-	*and, uno*
Interjection	I	-	-	*hélas*
Nom	N	commun, propre	genre, nomb	*France, prix*
Pronom	PRO	card, dém, indéf, inter, nég, pers, poss, rel	genre, nomb, pers	*trois, tout, lequel, rien, eux*
Ponctuation	PONCT	fort, faible	-	. ! ? ,
Préfixe	PREF	-	-	*franco-, outre-*
Préposition	P	-	-	*à, de, sur*
Verbe	V	-	genre, nomb, pers, mode, temps	*avoir, montrait, pourra*

Tableau 1. *Les catégories morphosyntaxiques du FTB*

2.1.3. *L'annotation des lemmes*

Une fois les étiquettes morphosyntaxiques validées, les lemmes sont ajoutés automatiquement, avec un dictionnaire externe, et les rares cas restant ambigus (*suis* du verbe *suivre* ou du verbe *être*, *étaient* du verbe *étayer* ou du verbe *être*, par exemple) ont été résolus à la main.

2.1.4. *L'annotation des mots composés*

La segmentation en mots (tokens) considère tous les signes de ponctuation et les espaces comme des séparateurs, mais ne sépare pas *au* ou *du*. Ensuite, certains mots composés (*aujourd'hui, pomme de terre*) sont annotés comme composants et regroupés au sein d'un mot composé (« *compound* »). L'annotation des mots composés inclut les locutions et les « mots agglomérés » (Fradin, 2003), selon un ensemble de critères graphiques, morphologiques, syntaxiques et sémantiques définis dans le guide d'annotation (Abeillé et Clément, 1999a). La première phase d'annotation, automatique, a

été réalisée à l'aide des dictionnaires externes du LADL (Silberztein, 1993), puis enrichie et validée à la main. En effet, une séquence candidate n'est pas toujours un mot composé, comme pour *bien que* dans une phrase telle que *Juppé voudrait **bien que** quelqu'un l'aime.* Dans la version finale, les composants de composés ont tous une catégorie interne (catint) ; en revanche, ils n'ont ni sous-catégorie, ni lemme associé (voir figure 1). Ce niveau d'annotation permet d'étudier en tant que telle la formation des mots composés, mais aussi de dériver des versions du corpus en ignorant les mots composés (Schluter et van Genabith, 2007), ou en ne retenant que les mots composés grammaticaux et/ou irréguliers (Candito *et al.*, 2010).
Les mots composés reçoivent la même richesse d'annotation que les mots simples (POS, flexion, lemme) mais aussi une étiquette spécifique « *compound* » . En tant que mot composé, *bien que* reçoit l'étiquette CS et deux étiquettes internes (catint) : Adv pour *bien* et CS pour *que*.

Les discontinuités éventuelles sont notées avec les balises <next> et <prev>, comme pour *à cause, notamment, de*, où la préposition *à cause de* est coupée par un adverbe. Au total, le corpus compte 59 mots composés discontinus.

Le corpus comporte des mots composés grammaticaux (*peut-être, bien que*) mais aussi ce qu'on appelle aujourd'hui des entités nommées comme (*Parti socialiste*) (Sagot *et al.*, 2012). Certains sont très longs (par exemple *Direction de la consommation, de la concurrence et de la répression des fraudes* ou *Société des autoroutes du nord et de l'est de la France*).
Dans sa version actuelle, le corpus compte 32 546 mots composés, c'est-à-dire près de 1,5 mot composé par phrase. Avec le recul, certains noms composés auraient pu être considérés comme des combinaisons de mots simples, parce qu'ils respectent la syntaxe ordinaire, et certaines locutions verbales auraient pu être décomposées également. De plus, le critère de figement du nom pour les locutions verbales (*rendre compte* est figé mais pas *tirer (un bon) parti*) n'a pas toujours été bien suivi par les annotateurs, d'où certaines incohérences.

2.2. *L'annotation syntaxique*

L'annotation syntaxique comporte le découpage en constituants (syntagmes) et les principales fonctions grammaticales. L'annotation automatique a été réalisée par des outils dédiés à cet effet, puis validée par des annotateurs linguistes. Une première étape a concerné le découpage en syntagmes, avec validation manuelle pour les frontières de syntagmes et pour leur catégorie. L'étiqueteur syntaxique (*chunker*) fondé sur (Kinyon, 2001) identifiait les constituants majeurs (12 étiquettes), sans récursion. Évalué sur 500 phrases corrigées tirées au hasard, il avait 60 % de précision, 92,9 % de rappel et 94 % d'étiquettes correctes pour les bornes ouvrantes, et 60 % de précision, 58 % de rappel, 56.5 % d'étiquettes correctes pour les bornes fermantes (Toussenel, 2001 ; Clément, 2001 ; Abeillé *et al.*, 2003). L'étiqueteur fonctionnel (8 étiquettes) s'appuyait sur les syntagmes et 115 règles avec unification, écrites à la main. Il a été évalué sur un échantillon de 1 000 phrases corrigées, avec une préci-

sion moyenne de 89,69 % (max 99,47 % pour Sujet) et un rappel moyen de 89,27 % (max 95,48 % pour Modifieur) (Abeillé et Barrier, 2004).
Les consignes d'annotation sont présentées en détail dans les guides (Abeillé *et al.*, 1999 ; Abeillé, 2004). Nous présentons ici les principaux choix d'annotation, qui visaient à être compatibles avec plusieurs théories syntaxiques, de façon à ce que le corpus soit aisément convertible en différentes versions.

Catégorie	Étiquette	Fonctions	Exemples
Syntagme adjectival	AP	ATS, ATO, OBJ, MOD	*fier de lui, très grand*
Syntagme adverbial	AdP	OBJ, P-OBJ, MOD	*très bien, plus vite*
Syntagme coordonnant	COORD	SUJ, OBJ, A-OBJ, DE-OBJ, P-OBJ, MOD	*et la France, ni le Maroc ni la Tunisie*
Syntagme nominal	NP	ATS, ATO, SUJ, OBJ, A-OBJ, DE-OBJ, P-OBJ, MOD	*la France, plus de 30 %, tout cela*
Syntagme prépositionnel	PP	ATS, ATO, A-OBJ, DE-OBJ, P-OBJ, MOD	*à midi, en France*
Subordonnée	Sint, Srel, Ssub	ATS, ATO, SUJ, OBJ, A-OBJ, DE-OBJ, P-OBJ, MOD	*dont on parle, dit-on, quand il faudra*
Phrase racine	SENT	-	*Rien n'avance.*
Noyau verbal	VN	SUJ, OBJ, A-OBJ, DE-OBJ, P-OBJ, SUJ/OBJ, SUJ/A-OBJ, etc.	*j'ai vu, on parle, en avoir*
Syntagme verbal	VPinf, VPpart	ATS, ATO, SUJ, OBJ, A-OBJ, DE-OBJ, P-OBJ, MOD	*tout finir, de voir cela, en avançant*

Tableau 2. *L'annotation syntaxique du FTB*

2.2.1. *L'annotation des syntagmes*

Le choix a été fait de structures syntaxiques relativement plates. Au sein du syntagme nominal, le déterminant, le nom et ses dépendants sont au même niveau. Au sein de la phrase, sujet et compléments sont au même niveau également. La plupart des mots des catégories ouvertes (Adjectif, Nom, Verbe) projettent un syntagme même quand ils sont employés seuls : un nom propre ou un nom attribut correspond à un NP, un verbe intransitif à un VN (exemple 1a). Un adjectif attribut ou épithète postnominal correspond à un AP (exemple 1b), mais pas un épithète prénominal (exemple 1c) car il ne peut pas apparaître avec un dépendant avant le nom : *une facile *(à remporter) victoire* vs *une victoire facile (à remporter)* (Abeillé et Godard, 1999). Enfin, un adverbe ne correspond pas à un AdP quand il est seul, car les adverbes ont une distribution différente de celle des syntagmes adverbiaux (Abeillé et Godard, 2001). Un pronom faible (Clitique) appartient au VN, contrairement à un pronom fort (*lui*) qui projette un NP (exemple 1c) (Miller, 1992).

(1) a. [Paul $_{NP}$] [dort $_{VN}$] bien [dans [sa chambre $_{NP}$] $_{PP}$].
b. [La France $_{NP}$] [est $_{VN}$] [riche $_{AP}$].
c. [On a eu $_{VN}$] [un autre problème [important $_{AP}$] $_{NP}$] [avec [lui $_{NP}$] $_{PP}$].

La catégorie Syntagme verbal (VP) est réservée aux syntagmes subordonnés à l'infinitif (VPinf) ou au participe (VPpart). Les verbes conjugués projettent un noyau verbal (VN), qui comprend les auxiliaires, les participes et les clitiques, mais ne projettent pas de syntagme verbal (exemple 1c).

Le corpus ne comporte pas de catégories vides : les infinitifs n'ont pas de sujet implicite annoté, ni les impératifs. Nous avons une catégorie COORD pour les syntagmes coordonnés. Ainsi, ils peuvent être inclus dans un autre syntagme, ou être détachés (exemple 2c) (Abeillé, 2005). Il en résulte une structure symétrique pour les coordinations redoublées (exemple 2b), et asymétrique pour les autres (exemple 2a) (Mouret, 2007 ; Mouret, 2005).

(2) a. [le Maroc [et [la Tunisie $_{NP}$] $_{COORD}$] $_{NP}$]
b. [et [le Maroc $_{NP}$] $_{COORD}$] [et [la Tunisie $_{NP}$] $_{COORD}$]
c. [Il est parti $_{VN}$], [et vite $_{COORD}$].

Les frontières de syntagmes indiquent les enchâssements. Ainsi, un syntagme prépositionnel inclut généralement un syntagme nominal et un syntagme verbal un noyau verbal. Selon que le syntagme prépositionnel est inclus dans un syntagme nominal ou non, on distingue complément de verbe (exemple 3a) et complément de nom (exemple 3b). De même l'inclusion d'une subordonnée relative (Srel) dans un syntagme nominal indique son rattachement. Lorsque les deux analyses sont possibles, par exemple dans une construction à verbe support (Gross, 1976), l'annotation la plus plate a été retenue (exemple 3c).

(3) a. [Il y a $_{VN}$] [30 élèves $_{NP}$] [dans [cette classe $_{NP}$] $_{PP}$].
b. [On a arrêté $_{VN}$] [six [d'entre [eux $_{NP}$] $_{PP}$] $_{NP}$].
c. [Ce pays $_{NP}$] [a commis $_{VN}$] [des agressions $_{NP}$] [contre [ses voisins $_{NP}$] $_{PP}$].

2.2.2. *L'annotation des fonctions syntaxiques*

Les fonctions syntaxiques des syntagmes dépendant de verbes ont été annotées avec un outil dédié (Abeillé et Barrier, 2004) et corrigées à la main. Les syntagmes dépendant d'autres catégories ne portent pas de fonction. Celle-ci peut se déduire en partie du découpage en constituants : un syntagme nominal inclus dans un syntagme prépositionnel est complément de cette préposition ; une relative est modifieur du nom du syntagme nominal qui l'inclut, un adjectif aussi. Pour les syntagmes prépositionnels, la distinction entre modifieur et complément est toujours difficile ; elle a été faite

pour les dépendants de verbe (exemple 3a), pour lesquels existent des tests (obligatoire ou non, mobile ou non, remplaçable par un clitique ou non, etc.), mais non pour les dépendants de nom ou d'adjectif. Parmi les compléments prépositionnels de verbe, nous distinguons ceux en *à* (fonction A-OBJ), ceux en *de* (fonction DE-OBJ) et les autres (fonction P-OBJ). Un complément de lieu est noté P-OBJ, car la préposition pourrait être remplacée par une autre (exemple 2a). Il en résulte un jeu réduit de 8 étiquettes fonctionnelles, auxquelles s'ajoutent des combinaisons de fonctions dans le cas des séquences de clitiques (voir tableau 2), et 12 autres fonctions dans la version convertie en arbres de dépendance (voir section 3.1.4).
Quand un mot seul ne projette pas un syntagme, il ne reçoit pas de fonction. Avec le recul, le fait qu'un adverbe seul ne projette pas de syntagme empêche de distinguer adverbe complément (*aller bien*) et modifieur (*travailler bien*). Les clitiques sont inclus dans le noyau verbal, qui porte leur fonction. C'est le seul type de syntagme à pouvoir porter des fonctions combinées, par exemple Sujet et Objet (SUJ/OBJ), s'il contient plusieurs clitiques (exemple 4a). Les clitiques figés, qui ne correspondent pas à un complément, comme le *y* de *il y a*, ou les réfléchis intrinsèques ne portent pas de fonction. Le *il* impersonnel, en revanche, correspond à la fonction Sujet.

(4) a. [On leur a expliqué $_{\text{VN :SUJ/A\_OBJ}}$] [en détail $_{\text{PP :MOD}}$].
b. [Il y a $_{\text{VN :SUJ}}$] [des problèmes [de courant] $_{\text{NP :OBJ}}$] [à Moscou $_{\text{PP :P-OBJ}}$].
c. [Quelles solutions $_{\text{NP :OBJ}}$] [la France $_{\text{NP :SUJ}}$] [peut-elle $_{\text{VN :SUJ}}$] [proposer $_{\text{VP :OBJ}}$] ?
d. [La France $_{\text{NP :SUJ}}$] [en a perdu $_{\text{VN :OBJ}}$] [plusieurs $_{\text{NP :OBJ}}$].

Un verbe peut avoir deux sujets ou deux objets. Dans le cas de l'inversion complexe, le sujet nominal porte la fonction SUJ, tout comme le clitique inclus dans le noyau verbal (exemple 4c). Il n'y a pas de syntagmes discontinus et la même fonction Objet est annotée pour le clitique *en* et le pronom postverbal (exemple 4d). Les dépendances à distance ne sont pas notées non plus (Candito et Seddah, 2012a). Ainsi, dans l'exemple 4c, le syntagme initial est noté OBJ, mais sans préciser qu'il s'agit d'un complément de *proposer* et non de *peut-elle*.

2.3. *L'ajout des métadonnées*

En 2016, le FTB a été enrichi de métadonnées indiquant la date de parution, l'auteur et le domaine de chacun des articles, qui n'étaient pas disponibles à l'origine. Le corpus est constitué de 1 143 extraits d'articles du journal *Le Monde* pris aléatoirement dans les versions distribuées à l'époque, entre janvier 1990 et août 1993 : la moitié en 1992, un quart en 1990 et un quart en 1993. Cette période courte permet une bonne homogénéité et limite les variations diachroniques éventuelles. Chaque extrait contient en moyenne près de 19 phrases, 6 ne sont composés que d'une seule phrase tandis que l'extrait le plus long en comporte 28.

Un peu plus de la moitié des articles (579) ne sont pas signés et l'auteur est renseigné comme « LeMonde ». Les 564 autres ont 210 auteurs ou groupe d'auteurs différents. Si la majorité des auteurs (134) n'ont écrit qu'un seul article, certains sont surreprésentés, comme F. Renard (27 articles), A. Lebaude (23 articles) ou M. Colonna d'Istria (19 articles). L'annotation fondée sur le codage interne du *Monde* répartit les textes dans 14 domaines différents (Illouz *et al.*, 2000). Près de 80 % (912) des articles traitent de problématiques économiques : 737 viennent des pages « Économie » et 175 du supplément *Le Monde Économie*. Puis vient la politique étrangère (77 extraits) et le supplément *Le Monde Initiatives* (36 extraits).
Le corpus arboré French Treebank apparaît donc comme un corpus homogène : genre discursif, sources des articles de presse, années d'écriture des textes, etc. Avec le recul, le choix des articles aurait pu être plus équilibré.

3. La version 1.0 du FTB

Le Laboratoire de Linguistique Formelle (UMR 7110) a publié, en décembre 2016, la version finale du corpus « French TreeBank 1.0 », disponible en différents formats : XML d'origine, TIGER-XML, Penn TreeBank et CoNLL pour la version en dépendances (Candito *et al.*, 2009). Un site dédié permet de télécharger la ressource, recense la documentation disponible, et comporte une plateforme d'interrogation pour des requêtes lexicales et/ou syntaxiques (`ftb.linguist.univ-paris-diderot.fr`).
Le corpus est distribué gratuitement à des fins de recherche et avec une licence payante à des fins commerciales. Il compte plus de 300 utilisateurs dans le monde, dont la moitié sont inscrits depuis décembre 2016, avec en moyenne 5,6 nouveaux utilisateurs par mois.
Le corpus regroupe 21 550 phrases pour 644 595 tokens au total, signes de ponctuation compris, et 557 518 tokens, signes de ponctuation exclus. Ces chiffres sont calculés en comptant chaque composant de composé comme 1 token (par exemple, *la plupart* compte pour 2 tokens) et chaque amalgame (*au, du*) pour 2 également (*à* et *le* par exemple). Comme dans d'autres corpus journalistiques, la longueur moyenne des phrases est élevée, près de 26 mots par phrase.

3.1. *Les différents formats du FTB*

Afin de permettre son utilisation à la fois pour des études linguistiques et pour des tâches de TAL, le FTB a été structuré dans plusieurs formats standard. Le format XML d'origine a ainsi été progressivement converti aux formats TIGER-XML, Penn TreeBank et CoNLL.

3.1.1. *Le format XML d'origine*

Le corpus FTB a été au départ structuré au format XML (*Extensible Language Markup*), avec trois niveaux d'annotation : FTB-XML Texte Brut ; FTB-XML Constituants ; FTB-XML Fonctions. Le schéma XML est fondé sur la TEI (*Text Encoding*

Initiative), permettant ainsi de structurer les données par article, chaque balise <TEXT> correspondant à un extrait différent. Chaque extrait contient des métadonnées, structurées au moyen d'attributs de la TEI (comme pour la date ou le nom de l'auteur) et d'attributs spécifiques comme le domaine (« argument ») et l'identifiant unique de la phrase (« nb », voir figure 1).

Les déclinaisons « Constituants » et « Fonctions » de ce format d'origine sont également structurées en XML et sont les plus richement annotées. Pour les tokens, les balises <w> encadrent les annotations (catégorie, lemme, sous-catégorie, flexion), et sont enchâssées pour les composants de composés. Les constituants sont encadrés par une balise spécifiant leur nature (par exemple <PP> pour syntagme prépositionnel ou <NP> pour syntagme nominal). La fonction éventuelle du constituant est portée par un argument « fct » . Ainsi, dans la figure 1, le premier NP a la fonction Sujet (« SUJ »).

```
<text>
   <SENT argument="ETR" author="Minangoy Robert" date="1990-01-19"
      nb="1000" textID="456">
      <NP fct="SUJ">
         <w cat="PRO" ee="PRO-card-mp" ei="PROmp" lemma="six" mph="mp"
            subcat="card">Six</w>
         <PP>
            <w cat="P" compound="yes" ee="P" ei="P" lemma="d'entre">
               <w catint="P">d'</w>
               <w catint="P">entre</w>
            </w>
         <NP>
            <w cat="PRO" ee="PRO-3mp" ei="PRO3mp" lemma="eux" mph="3mp"
               subcat="pers">eux</w>
         </NP>
      </PP>
   </NP>
[...]
</text>
```

Figure 1. *Le format d'origine du FTB, avec métadonnées et fonctions*

3.1.2. *Le format TIGER-XML*

Afin de pouvoir l'interroger à l'aide de TIGERSearch, le corpus FTB a été structuré au format TIGER-XML (König et Lezius, 2000). Sans entrer dans les détails du format (Liégeois et Abeillé, 2018), les annotations des nœuds terminaux (les tokens) sont distinguées de celles des nœuds non terminaux, qui incluent les mots composés, les constituants et leurs fonctions.
L'outil TIGERSearch permet d'interroger l'ensemble des annotations, en les combinant éventuellement : lemmes et catégories des tokens, catégories et fonctions des constituants... Les résultats obtenus peuvent être visualisés sous la forme d'arbres syntaxiques (figure 2). En outre, TIGERSearch étant intégré à la version portail

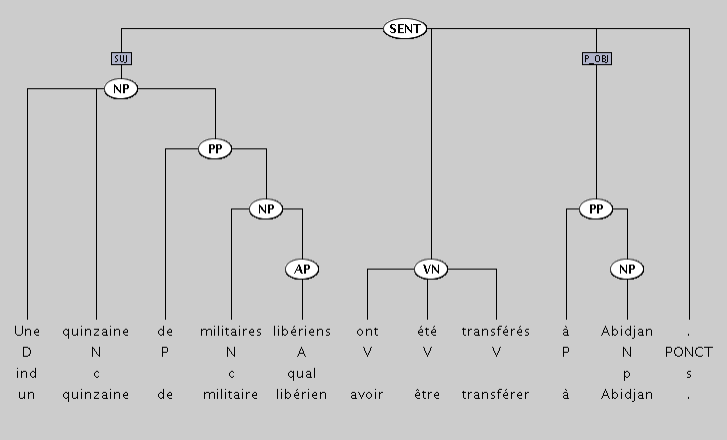

Figure 2. *Exemple d'affichage sous TIGERSearch*

de TXM (Heiden S., 2010), l'interrogation de l'ensemble des annotations du FTB est possible grâce à une plateforme en ligne sur le site du projet (`ftb.linguist.univ-paris-diderot.fr/`, section « Interroger »).

3.1.3. *Le format Penn TreeBank*

Le format Penn TreeBank (PTB) (Taylor *et al.*, 2003) caractérise la hiérarchie des constituants avec un système de parenthèses. Dans une phrase, une paire de parenthèses correspond à un niveau de l'arbre (exemple 5). Cette version du FTB utilise un jeu d'étiquettes simplifié pour les mots et des syntagmes, et certaines annotations sont perdues, comme le lemme ou la flexion, même si tous les composés sont gardés. Elle a toutefois l'avantage de permettre une visualisation graphique et l'interrogation du corpus à l'aide d'outils en ligne comme Tregex (Levy et Andrew, 2006) (figure 3).

(5) (SENT (NP-SUJ (D Une) (N quinzaine) (PP (P de) (NP (N militaires) (AP (A libériens))))) (VN (V ont) (V été) (V transférés)) (PP-P_OBJ (P à) (NP (N Abidjan))) (PONCT .))

3.1.4. *La version en dépendances (format CoNLL)*

Le quatrième format distribué est CoNLL, après conversion du corpus en arbres de dépendance (Candito *et al.*, 2010), selon le programme de conversion de Candito *et al.* (2010) légèrement modifié suite aux travaux de Seddah *et al.* (2013) pour la *SPMRL Shared Task*. La conversion est fondée sur le principe de définition d'une tête lexicale au sein de toute forme de syntagme. Les choix de conversion suivent les choix d'annotation syntaxique de la version en constituants (voir section 2.2), sauf que les

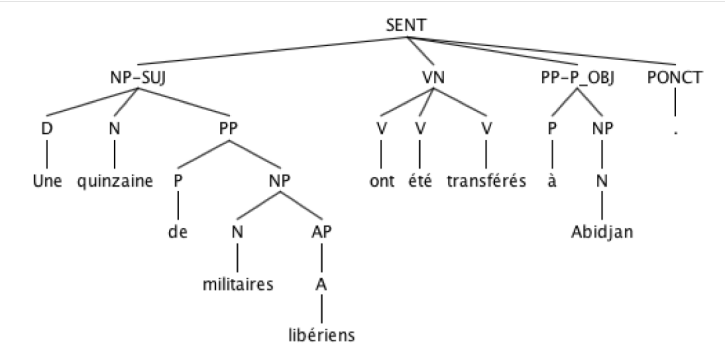

Figure 3. *Exemple d'affichage sous Tregex*

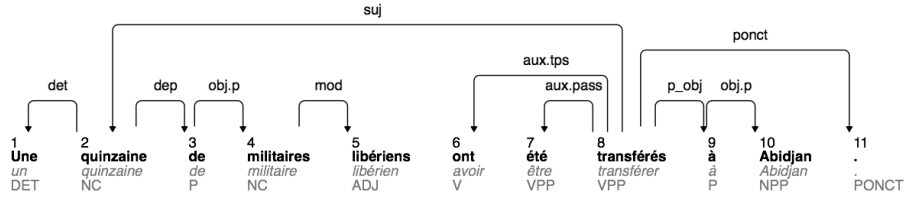

Figure 4. *Exemple d'affichage en dépendances*

prépositions sont toujours des têtes. Tous les mots composés ont été gardés, avec une étiquette de relation spéciale. Les fonctions étant associées aux mots et non aux syntagmes, seules ont été gardées les fonctions simples, et 12 fonctions ont été ajoutées automatiquement, en particulier pour les mots grammaticaux : *det* pour les déterminants, *obj-p* pour les compléments de préposition, *mod* pour les adjectifs épithètes et les relatives au sein du syntagme nominal, ou les adverbes au sein du syntagme adjectival, *dep* pour les autres dépendants des adjectifs et des noms, par exemple. Les auxiliaires de temps, passif et causatif reçoivent des étiquettes distinctes.
Cette version en dépendances (figure 4) est distribuée au format CoNLL, avec un token par ligne, et a donné lieu à un corpus dérivé au format "Universal dependencies" (Nivre *et al.*, 2016) (voir section 5.2).

3.2. *Le FTB (v.1.0) en quelques chiffres*

Nous présentons quelques chiffres concernant la distribution des mots, des catégories morphosyntaxiques, des syntagmes et des fonctions dans la version 1.0 du FTB.

3.2.1. *La répartition des catégories morphosyntaxiques*

Le corpus FTB comprend 644 595 tokens, signes de ponctuation inclus. Parmi ceux-ci, 553 024 sont des mots simples (avec trait « lemma » et/ou du trait « word »), pour un total de 30 688 types différents, et 78 634 des composants de composés. Le tableau 3 présente la distribution des catégories morphosyntaxiques pour les mots simples : certaines catégories sont surreprésentées (Nom, Préposition, Déterminant, Verbe) et d'autres, sous-représentées (Préfixe, Mot Étranger, Interjection). Si l'on compte en tokens, les mots les plus fréquents sont les Noms et les Prépositions, et les Clitiques sont d'un tiers plus nombreux que les autres Pronoms. Si l'on compte en types, les catégories les plus fréquentes sont les Noms et les Verbes, et les Pronoms sont 3,5 fois plus nombreux que les Clitiques.

Catégorie	Tokens	Types	Les 5 mots les plus fréquents
Nom	134 137	15 069	*pourcents, francs, M., milliards, millions*
Préposition	85 534	124	*de, à, des, d', du*
Déterminant	80 986	664	*la, l', les, le, un*
Ponctuation	79 862	18	, . ") (
Verbe	68 452	10 049	*est, a, ont, sont, été*
Adjectif	36 252	5 627	*français, deux, autres, dernier, premier*
Adverbe	22 510	731	*pas, plus, ne, n', aussi*
Conjonction	18 292	58	*et, que, ou, qu', mais*
Clitique	15 751	73	*s', se, il, on, en*
Pronom	10 461	234	*qui, dont, que, où, qu'*
Préfixe	423	58	*ex-, non-, vice-, quasi-, micro-*
Mot Étranger	299	208	*eiido, eiidos, bank, and, of*
Interjection	65	29	*hélas, bref, non, oui, attention*

Tableau 3. *Les catégories morphosyntaxiques dans le FTB (mots simples)*

Le tableau 4 présente la distribution des mots composés, selon leur catégorie morphosyntaxique : les plus fréquents sont les noms composés, et les adverbes composés, suivis de près par les prépositions complexes ou locutions prépositionnelles.

Pour savoir si ces chiffres sont représentatifs, nous manquons de données comparables pour d'autres corpus écrits, en particulier concernant les mots composés. L'on peut penser que la catégorie Clitique est plus représentée dans les corpus oraux, en raison de la fréquence des clitiques sujets (Blanche-Benveniste *et al.*, 1984), et la catégorie Interjection en registre informel.

3.2.2. *La répartition des syntagmes et des fonctions*

Si l'on considère la répartition des syntagmes dans le corpus (tableau 5), le syntagme nominal (NP) est surreprésenté, et les relatives (Srel) sont plus nombreuses que les autres subordonnées. Les moins fréquents sont les syntagmes adverbiaux (AdP),

Catégorie	Tokens	Exemples
Nom	14 398	*Côte-d'Ivoire, Médecins Sans Frontière*
Adverbe	5 819	*un peu, à grands frais*
Préposition	5 215	*avant de, faute de*
Déterminant	3 633	*la plupart de, trois cents*
Conjonction	1 322	*alors que, depuis que*
Verbe	878	*avoir affaire, prendre la fuite*
Pronom	574	*celle-ci, lui-même*
Adjectif	564	*hors-cadre, est-allemand*
Clitique	52	*l'on*
Etranger	34	*New-Deal, open-market*
Interjection	6	*au secours, oh là là*

Tableau 4. *Les catégories morphosyntaxiques dans le FTB (mots composés)*

ce qui est en partie lié au choix de ne pas annoter de syntagme quand l'adverbe est seul.

Constituant	Nombre	Exemples
Syntagme nominal	151 840	*le rapport Delors, cette intention*
Syntagme prépositionnel	82 620	*du même genre, de Bruxelles*
Noyau verbal	50 780	*on devrait, il est*
Syntagme adjectival	24 040	*précieux, social-démocrate*
Syntagme coordonné	15 659	*et de la retenue, ou non*
Syntagme infinitif	12 568	*de prouver cette intention*
Syntagme participial	8 284	*approuvées par le conseil européen*
Subordonnée relative	6 636	*qui est envisagée dans le rapport Delors*
Autre subordonnée	6 100	*combien celle-ci restait incertaine*
Incise et parenthèse	3 527	*leur état d'esprit a changé*
Syntagme adverbial	1 349	*plus tôt, un peu vite*

Tableau 5. *Les catégories des constituants dans le corpus*

Si l'on considère la répartition des fonctions (tableau 6), la fonction Sujet est la plus représentée, et la fonction Modifieur (MOD) est quasiment au même niveau. Les attributs du sujet (ATS) sont beaucoup plus fréquents que les attributs de l'objet (ATO), qui est la fonction la moins représentée du corpus. Les 19 combinaisons de fonctions (SUJ/OBJ, etc.) sont associées aux VN avec clitiques, et notent leur ordre aussi bien que leur fonction.

Fonction	Nombre	Exemples
SUJ	34 878	*le deutschemark, plusieurs voies*
MOD	34 747	*ici, lui-même*
OBJ	27 590	*cette intention, la construction monétaire*
ATS	6 228	*une voie praticable, négatives*
A-OBJ	4 570	*à l'économie allemande, à reculer*
DE-OBJ	4 013	*du gouvernement, de toutes parts*
P-OBJ	3 173	*par une partie des militants*
ATO	486	*comme des privilèges, reprendre ses droits*
SUJ/OBJ	284	*il y en a, je l'ai appelé*
SUJ/A-OBJ	139	*il m'apparaît, il leur apprend*
SUJ/DE-OBJ	53	*on s'en doute, nous en séparer*
SUJ/MOD	28	*j'y vois, on en connaît*
OBJ/A-OBJ	15	*l'y autorise, ne l'y obligerait*
SUJ/P-OBJ	14	*on y reste, j'y suis*
SUJ/ATS	11	*il l'a été, on l'était*
OBJ/SUJ	11	*en faut-il, les soumettront-ils*
OBJ/DE-OBJ	7	*l'en empêcha, l'en pressent*
A-OBJ/SUJ	5	*nous semble-t-il, se demande-t-il*
DE-OBJ/SUJ	5	*en pâtira-t-elle, en sera-t-il*
A-OBJ/DE-OBJ	4	*lui en est donnée, nous en rebattent*
A-OBJ/OBJ	4	*se le procurer, vous le diront*
SUJ/OBJ/A-OBJ	3	*nous le leur apprendrons*
A-OBJ/MOD	2	*leur y ont enseigné*
OBJ/MOD	2	*nous en a notifié, s'en faire*
OBJ/P-OBJ	1	*s'y présente*
SUJ/A-OBJ/OBJ	1	*on nous l'a toujours dit*
SUJ/MOD/OBJ	1	*on nous l'a changé*

Tableau 6. *Les fonctions grammaticales dans le corpus*

4. Une évaluation du corpus FTB

Une évaluation quantitative n'a pas été effectuée lors des différentes campagnes d'annotation. À chaque fois, les annotateurs travaillaient sur des sorties annotées automatiquement (voir section 2.1) en consultant des guides d'annotation dédiés, avec des réunions régulières pour trancher les cas difficiles ou inattendus (qui venaient compléter les guides). Qu'il s'agisse des mots composés, des étiquettes morphosyntaxiques, des constituants ou des fonctions, chaque fichier (de 500 phrases) était corrigé par un premier annotateur, puis par un second (généralement plus expérimenté) qui vérifiait et améliorait le résultat. Les annotateurs étaient tous des étudiants avancés en linguistique (L3 ou master à Paris 7 ou Paris 10) et certains sont restés plusieurs années dans le projet. Il n'a pas été question à l'époque de faire annoter en parallèle le même fi-

chier et de mesurer l'accord interannotateur, comme on le ferait aujourd'hui (Artstein et Poesio, 2008). Pour combler cette lacune, nous avons réalisé une évaluation *a posteriori* en prenant la version finale distribuée comme référence.

4.1. *Une évaluation morphosyntaxique*

Une qualité du FTB est la richesse de ses annotations morphosyntaxiques, qui incluent de nombreuses sous-catégories ainsi que toutes les informations flexionnelles : ainsi *les* est annoté féminin ou masculin selon le nom qui suit, si c'est un Déterminant, et selon son antécédent si c'est un Clitique. Nous avons pris 100 phrases au hasard. En excluant les ponctuations (qui ne posent pas de problème d'annotation) et les composants de composés (qui ont des catégories plus sommaires), elles comprennent 2 168 tokens. Nous n'avons plus les fichiers de sortie de l'époque, donc nous avons pris une version sans annotation préalable, sauf les mots composés, ce qui est une tâche plus difficile qu'à l'époque. Un expert linguiste les a annotés (un mot par ligne), en utilisant les guides et le jeu des 122 étiquettes internes (attribut ei). En comparant avec le FTB, la valeur du kappa (Cohen, 1960) est 0,97 (pour un pourcentage d'accord de 97,3 %). Les désaccords concernent surtout le genre des Clitiques, des Pronoms et des Noms propres, difficile à renseigner hors contexte sans marque morphologique (*je*, *Air Inter*). Ils concernent aussi Adjectif ou Participe passé, Adverbe ou Préposition, Déterminant, Adjectif ou Pronom pour certains indéfinis (*l'un*, *l'autre*), *de* comme Préposition ou Déterminant. L'adjudication donne raison au corpus, selon les guides d'annotation. Si l'on se limite aux catégories lexicales (attribut pos), l'accord est quasi parfait (κ = 0,99, pour un pourcentage d'accord de 99 %), sur 9 catégories car Interjection, Mot étranger et Préfixe n'étaient pas représentés dans l'échantillon.

4.2. *Une évaluation en constituants*

Afin d'évaluer les constituants, nous avons pris les mêmes 100 phrases et demandé à l'expert de les annoter (format PTB) selon les guides du corpus. Il s'agit d'une tâche sans annotation automatique préalable, donc plus difficile que pour les annotateurs du projet. L'expert a annoté les principaux constituants, c'est-à-dire tous les noyaux verbaux et syntagmes verbaux, toutes les propositions (relatives, subordonnées, internes) et tous les syntagmes coordonnés, mais seulement les syntagmes nominaux, adjectivaux, adverbiaux et prépositionnels majeurs, c'est-à-dire recevant une fonction. Le fichier obtenu comporte 925 syntagmes, et nous l'avons comparé au FTB, avec le programme *evalb* (Sekine *et al.*, 2008), et obtenu une FMesure de 89,65. Les principaux cas de désaccord portaient sur le nombre de syntagmes (18 en plus ou en moins), soit un taux d'accord de 98,1%, sur les bornes ouvrantes (taux d'accord de 99,8 %), sur les bornes fermantes (taux d'accord de 97,5 %), et sur les étiquettes (taux d'accord de 99,1 %). Les désaccords sur les étiquettes portaient sur AP et VPpart ou NP et PP et étaient liés aux désaccords sur les catégories lexicales. Les désaccords sur les bornes

fermantes concernent l'annotation des appositions (dans le NP ou en dehors), et des coordinations.

4.3. *Une évaluation des fonctions*

Pour les fonctions grammaticales, nous avons reproduit la situation des annotateurs de l'époque, car nous avons gardé certaines sorties de l'annotateur fonctionnel automatique. Nous avons pris 100 phrases au hasard. Un expert linguiste a corrigé les fonctions associées aux 616 syntagmes concernés. Par rapport au FTB, κ est à 0,91 (pour un pourcentage d'accord de 93,2 %). Les principaux désaccords concernent les syntagmes prépositionnels, ajouts (MOD) ou compléments (A-OBJ, DE-OBJ, P-OBJ) et les fonctions associées aux Clitiques (pas de fonction pour les Clitiques figés). L'expert avait oublié la fonction MOD pour les constituants disloqués. Comme il ne pouvait pas corriger les syntagmes, contrairement aux annotateurs de l'époque, il y a quelques discordances (31 fonctions en plus ou en moins entre les deux versions), la version distribuée ayant fait l'objet de post-corrections en syntagmes supplémentaires (κ est à 0,86 si l'on en tient compte, pour un pourcentage d'accord de 88,9 %).

Annotation	Tokens	Nombre d'étiquettes	Taux d'accord (kappa)	Pourcentage d'accord
cat. lexicales	2 168	11	0,99	99 %
souscat + morpho	2 168	122	0,97	97,3 %
syntagmes	925	10	0,99	98,1 %
fonctions	616	11	0,91	93,2 %

Tableau 7. *Les taux d'accord sur les étiquettes du FTB (sur 100 phrases)*

Nous concluons que les annotations du FTB sont de très bonne qualité, même s'il reste toujours des erreurs qui sont corrigées régulièrement, et que les choix d'annotation sont reproductibles.

5. Exemples d'utilisation du FTB

Depuis sa création, le corpus FTB a été utilisé pour des applications variées, dans le cadre d'études linguistiques ou psycholinguistiques, ou en traitement automatique des langues.

5.1. *Utilisation du FTB pour des études linguistiques*

Le corpus FTB a été exploité pour réaliser diverses études linguistiques. Sans être exhaustif, nous pouvons citer les études sur les phrases sans verbe de Laurens (2008) et les relatives sans verbe de Bîlbîie et Laurens (2009), les coordinations itératives

(Mouret, 2005), (Mouret, 2007), les coordinations de phrases avec ellipse (Abeillé et Mouret, 2010), les relatives en *dont* (Abeillé *et al.*, 2016) et l'inversion du sujet dans les relatives en *que* (Pozniak *et al.*, 2019). Ces études s'appuient sur les annotations réalisées au niveau des catégories morphosyntaxiques et des types de constituants, mais aussi sur l'annotation fonctionnelle. Abeillé *et al.* (2016) ont trouvé que dans la majorité des cas, *dont* est utilisé pour le complément du sujet (dont la majorité, dont le directeur). En comparant avec un grand corpus littéraire (Frantext aux XIXe et XXe siècles), Abeillé et Winckel (2018) ont trouvé une proportion du même ordre, ce qui est un indice de la représentativité du FTB pour les questions de syntaxe, du moins à l'écrit en registre formel.

À l'aide de TIGERSearch, nous avons cherché les catégories associées à la fonction Sujet, la plus fréquente dans le corpus. Sans surprise, les syntagmes nominaux sont les plus nombreux avec 26 789 occurrences au total, soit près de 77 % (voir tableau 8), suivis par les Clitiques (dans ce cas, la fonction Sujet est portée par le noyau verbal (VN)). Quand un syntagme coordonné (COORD) porte la fonction, il s'agit d'une coordination itérative (*ni la France ni l'Angleterre*). Au total, 5 types de constituants sont employés comme sujet. Des requêtes plus fines permettent par exemple d'étudier l'inversion du sujet ou l'accord sujet verbe. Ainsi, Pozniak *et al.* (2019) ont trouvé que 50 % des sujets nominaux sont inversés dans les relatives en *que*. Nous avons trouvé 6 cas d'infinitifs sujets inversés (sur 99), dont 5 avec *vaut / vaudrait mieux* :

(6) a. `[...]` *mieux vaut* [*s'entraîner* $_{VPinf\text{-}SUJ}$].
b. *Reste* [*à savoir quelle en sera l'ampleur* $_{VPinf\text{-}SUJ}$].

Mouret (2007) a trouvé que les sujets infinitifs coordonnés permettent aussi bien l'accord singulier que l'accord pluriel, et ce dans deux phrases d'un même article :

(7) a. [*Ne pas savoir se servir d'un ordinateur (66,5 %), travailler à temps réduit (64,9 %) et être âgé de plus de quarante-cinq ans (52,3 %)* $_{VPinf\text{-}SUJ}$] *constitue un handicap, lors d'une promotion.*
b. *À l'inverse,* [*être un homme (48,6 %) et avoir des diplômes (85,6 %)* $_{VPinf\text{-}SUJ}$] *sont manifestement des atouts* `[...]`.

Constituant	Sujet	Exemples
NP	26 789	*lequel, Mr Hans Modrow*
VN	7 946	*on devrait, il y avait*
VPinf	99	*assister à un tel bouleversement*
Ssub	24	*qu'il soit employé comme nom ou comme adjectif*
COORD	20	*ni système de prix ni marché*

Tableau 8. *La fonction Sujet dans le corpus*

5.2. *L'utilisation du FTB pour la constitution d'autres ressources*

Le corpus FTB a permis la création de ressources dérivées.

5.2.1. *Les lexiques dérivés*

Un certain nombre de lexiques ont été dérivés comme TreeLex (Kupść, 2009 ; Kupść et Abeillé, 2008). Ce dictionnaire de valence, dont les entrées on été extraites automatiquement du corpus, comporte les adjectifs (2 200 entrées) et les verbes (2 000 entrées) du corpus. La valence, automatiquement extraite du corpus, convertie pour le passif, le réfléchi, etc. et corrigée manuellement, est indiquée pour chacun des lemmes. Un dictionnaire de valence d'adjectifs a été étendu par Fabre et Kupść (2009). Le lexique Nomage (Balvet *et al.*, 2011) contient quant à lui un ensemble de noms déverbaux (morphologiquement dérivés d'un verbe) pour lesquels, à partir des exemples présents dans le corpus FTB, les auteurs ont ajouté une couche d'annotation sémantique (arguments et classe aspectuelle).

5.2.2. *Les corpus dérivés*

Plusieurs corpus ont été dérivés à partir des versions antérieures du FTB, par exemple le corpus de Dublin (2007) qui comporte 4 741 phrases du FTB (*Modified FTB*) (Schluter et van Genabith, 2007), le corpus d'Aix-en-Provence qui comporte 1 471 phrases du FTB, en ajoutant des annotations des grammaires de propriétés (FTB-LPL) (Blache et Rauzy, 2012). Le Dependency Corpus d'Alpage (FTB-DEP) comporte 12 500 phrases du FTB converties au format CoNLL (Candito *et al.*, 2009). Un corpus de référence (*gold*) a été produit par Seddah *et al.* (2013) pour la *Shared Task* de SPRML 2013 : 38 fichiers du FTB ont été convertis au format CoNLL. Enfin, le corpus U-FTB a été obtenu après après conversion automatique (Seddah *et al.*, 2018) en suivant les étiquettes "Universal Dependencies" (http ://universaldependencies.org/) (McDonald *et al.*, 2013).

D'autres projets ont ajouté des annotations supplémentaires : ainsi Sagot *et al.* (2012) y ajoutent les entités nommées, Candito et Seddah (2012a), Ribeyre *et al.* (2014) ajoutent des relations de syntaxe profonde, pour les dépendances à distance, le passif, les constructions impersonnelles, etc., Djemaa *et al.* (2016) ajoutent des informations sémantiques (de type framenet) sur un sous-ensemble de prédicats, et Danlos *et al.* (2015) ajoutent un étiquetage des connecteurs de discours et des relations de discours (French Discourse Treebank) sur l'ensemble du corpus.
Du côté de la psycholinguistique, Pynte *et al.* (2009) ont ajouté les temps de lecture et les mouvements oculaires sur 52 173 tokens du FTB (qui constituent la partie française du Dundee Corpus), et (Rauzy et Blache, 2012) ont annoté avec temps de lecture et mouvements oculaires 198 phrases (6 572 tokens) du FTB (corpus physiologique du LPL). Hale (2014) a, quant à lui, calculé un modèle de surprise fondé sur le FTB, très utilisé en psycholinguistique computationnelle.

Les guides d'annotation ont par ailleurs été réutilisés dans le projet d'évaluation et d'annotation Easy (Paroubek *et al.*, 2007), pour les corpus Passage (Villemonte de La Clergerie *et al.*, 2008) et Sequoia (Candito *et al.*, 2014), ainsi que plus récemment pour des corpus de questions (Seddah et Candito, 2016), de médias sociaux (Seddah *et al.*, 2012), ou un corpus oral de radio (Abeillé et Crabbé, 2013), avec des adaptations nécessaires.

6. Comparaison avec d'autres corpus français annotés

6.1. *Comparaison avec des corpus taggés*

Depuis la création du FTB, de nombreux corpus taggés ont vu le jour. Les plus gros, pour le français écrit, sont le frWaC (2,6 milliards de mots) et le frTenTen (10 milliards de mots) issus de pages Web (Baroni *et al.*, 2009). Ils utilisent un jeu d'étiquettes réduit (33 étiquettes). Ils sont annotés automatiquement avec Treetagger, sans correction, ce qui engendre de nombreuses erreurs. Il en va de même du corpus littéraire Frantext (253 millions de mots), entièrement catégorisé depuis 2018, qui utilise un jeu de 22 étiquettes, récemment mis à jour pour correspondre au jeu de la version en dépendances du FTB (Candito *et al.*, 2009).

6.2. *Comparaison avec d'autres corpus arborés*

La plupart des corpus arborés qui ont vu le jour depuis, se sont inspirés du FTB, et ont eu pour but d'être plus équilibrés et d'ajouter des relations pour prendre en compte l'oral ou pour être plus proches de la sémantique. Les corpus écrits Passage et Séquoia sont plus équilibrés que le FTB puisque, outre des textes journalistiques (*Est Républicain*), ils incluent aussi des textes de Wikipédia, du Parlement européen et de l'Agence européenne des médicament. Le corpus Passage (Villemonte de La Clergerie *et al.*, 2008) comprend 2 millions de mots, dont 4 000 phrases corrigées à la main. Le corpus Séquoia (Candito et Seddah, 2012b) comporte 3 204 phrases et 69 246 tokens, annotés automatiquement et corrigés, disponibles aux formats PTB ou CoNLL, avec un jeu de 28 étiquettes lexicales. Pour les syntagmes, et pour les dépendances, il utilise le même jeu d'étiquettes que le FTB (format PTB et format CoNLL). Dans sa dernière version (Candito *et al.*, 2014), ont été ajoutées des annotations « profondes » plus proches de la sémantique, pour le passif ou l'impersonnel, le sujet implicite des infinitifs ou les cas d'ellipse. Le nombre de relations est donc plus important avec 28 étiquettes différentes.

Le corpus French GSD (Google Stanford Dependencies) a été créé en 2015 au sein du projet multilingue de Universal Dependency Treebanks (McDonald *et al.*, 2013). Il a été ensuite modifié pour se conformer au schéma d'annotation Universal Dependencies (Nivre *et al.*, 2016), et est intégré aux différentes versions du projet Universal Dependencies, depuis la version 2.0. Il comporte 16 342 phrases (389 363 tokens), à partir de blogs et de pages Wikipédia, sans métadonnées. Il s'appuie sur 17 étiquettes

lexicales universelles, enrichies par 37 traits morphosyntaxiques. Les clitiques ne sont pas distingués des pronoms forts. Les phrases sont segmentées de telle sorte que les seuls tokens contenant des espaces sont des nombres. Des expressions polylexicales ont été annotées (essentiellement des mots composés grammaticaux) en utilisant une représentation plate, et une étiquette de relation spécifique. Pour la syntaxe, il s'appuie sur un jeu de 34 relations universelles, avec 16 sous-types. Le choix est de toujours avoir pour tête une catégorie majeure (ainsi les prépositions ne sont jamais têtes, ni le verbe *être*). Pour les groupes prépositionnels dépendant de verbes, il ne distingue pas systématiquement complément et modifieur, ni objet et attribut pour les infinitifs ou les complétives. Il a fait l'objet de corrections manuelles et automatiques, mais n'a pas été évalué pour les catégories lexicales ni les traits morphosyntaxiques. Pour une évaluation des relations sur 100 phrases, voir (Guillaume *et al.*, 2019).
Le Corpus d'Étude pour le Français Contemporain (CEFC) (Debaisieux *et al.*, 2016), qui compte 10 millions de mots dont 6 millions de textes écrits, se veut représentatif des variétés du français oral et écrit. Il mêle des transcriptions et des textes variés (interviews, oral en interaction, oral spontané, parole avec variation régionale, parole d'enfant, documents administratifs, journaux, romans). Seule une sous-partie (172 000 mots) est annotée pour la syntaxe, selon un schéma en dépendances, et corrigée manuellement. Elle utilise 21 étiquettes lexicales, et 12 relations fonctionnelles, 9 pour la « microsyntaxe » (par exemple sujet, spécifieur et dépendant, qui ne distinguent pas entre complément et modifieur) et 3 pour la « macrosyntaxe » (par exemple périphérique et parenthétique).

Corpus	Tokens	Tokens corrigés	Étiquettes mots	Étiquettes syntagmes	Relations
FTB	644k	644k	218	12	8/20
Sequoia	69k	69k	28	12	8/28
CEFC	10M	172k	21	-	12
U-GSD	389k	389k	17/74	-	50

Tableau 9. *Les principaux corpus arborés pour le français*

Le FTB reste une ressource unique par la finesse de ses annotations lexicales, comme par le volume de ses données corrigées.

7. Conclusion

Le Corpus arboré du français (French Treebank) de l'université Paris Diderot a joué un rôle majeur dans l'outillage du français, et a inspiré de nombreux projets dérivés. C'est un corpus homogène, par le choix des textes (journal *Le Monde*) et par leurs dates (1990-1993). Il est de taille moyenne comparé à certains corpus annotés plus récents, et non équilibré, mais avec une richesse d'annotation inégalée, tant par le jeu d'étiquettes (218 étiquettes morphosyntaxiques, 20 étiquettes syntaxiques), que par le nombre de mots composés. Les nombreuses phases de validation manuelle en

font une ressource de qualité, même si des erreurs résiduelles peuvent toujours subsister, et que certains choix d'annotation sont aujourd'hui datés. Sa version finale, qui est disponible en plusieurs formats, permet à la fois des utilisations variées en traitement automatique des langues (format en dépendances CoNLL par exemple) et des requêtes à l'aide d'outils génériques (TXM, TIGERSearch, Tregex) pour des utilisateurs linguistes. Disponible sur un site dédié, il est toujours d'actualité comme en témoigne le nombre croissant de nouveaux utilisateurs. Il n'a pas la prétention d'être représentatif des variétés du français, mais présente l'avantage de permettre des études syntaxiques et lexicales fondées sur « le bon usage » du français écrit contemporain.

Remerciements

Le projet a été soutenu par l'IUF, chaire junior 1996-2001 et chaire senior 2007-2012, d'Anne Abeillé, par l'université Paris Diderot, le LLF, le CNRTL et la DGLFLF. Nous tenons à remercier les relecteurs de TAL ainsi que, pour leur travail sur les premières versions du corpus, Nicolas Barrier (annotation en fonctions), Martine Cheradame (coordination des stagiaires et guides d'annotation), Alexandra Kinyon, Jacques Steinlin et François Toussenel (annotation en constituants), Rodrigo Reyes (annotation morphosyntaxique) et pour leur contribution à la version finale : Marie-Hélène Candito et Benoit Crabbé (détection d'erreurs, conversion au format CoNLL, annotation en fonctions), Vanessa Combet (correction), Achille Falaise, Clément Plancq, Alexandre Roulois et Johan Ferguth (site du projet, plateforme d'interrogation et maintenance des différents formats).

8. Bibliographie

Abeillé A., *Corpus arboré pour le français : guide d'annotation en fonctions*, Université Paris Diderot, Paris, 2004.

Abeillé A., « Les syntagmes conjoints et leurs fonctions syntaxiques », *Langages*, vol. 160, p. 42-66, 2005.

Abeillé A., Barrier N., « Enriching a French treebank », *4th LREC*, Lisbonne, p. 2233-2236, 2004.

Abeillé A., Clément L., *Corpus arboré pour le français : guide d'annotation morphosyntaxique*, Université Paris Diderot, Paris, 1999a.

Abeillé A., Clément L., « A reference tagged corpus for French », *in* T. Brants H. U. (dir.), *LINC, EACL*, Bergen, p. 17-24, 1999b.

Abeillé A., Clément L., Toussenel F., « Building a Treebank for French », *in* Abeillé A. (dir.), *Treebanks : Building and Using Parsed Corpora*, Kluwer, Dordrecht, p. 165-188, 2003.

Abeillé A., Crabbé B., « Vers un treebank du français parlé », *20ème Conférence TALN*, 2013.

Abeillé A., Godard D., « La position de l'adjectif épithète en français : le poids des mots », *Recherches linguistiques de Vincennes*, vol. 28, p. 9-32, 1999.

Abeillé A., Godard D., « A Class of 'lite' Adverbs in French », *in* Camps J., Wiltshire C. (dir.), *Romance Syntax, Semantics and their L2 Acquisition*, John Benjamins, p. 9-25, 2001.

Abeillé A., Hemforth B., Winckel E., « Les relatives en dont : études empiriques », *5ème CMLF*, ILF, Tours, p. 26-43, 2016.

Abeillé A., Kinyon A., Clément L., « Building a treebank for French », *2d LREC*, Athènes, 2000.

Abeillé A., Mouret F., « Quelques contraintes sémantiques et discursives sur les coordinations elliptiques », *Revue de sémantique et de pragmatique*, vol. 24, n° 3, p. 177-206, 2010.

Abeillé A., Toussenel F., Chéradame M., *Corpus arboré pour le français : guide d'annotation en constituants*, Université Paris Diderot, Paris, 1999.

Abeillé A., Winckel E., « Dont and de qui relatives in written French », *Grammar and Corpora*, 2018.

Artstein R., Poesio M., « Inter-Coder Agreement for Computational Linguistics », *Computational Linguistics*, vol. 34, n° 4, p. 555-596, 2008.

Balvet A., Barque L., Condette M.-H., Haas P., Huyghe R., Marin R., Merlo A., « La ressource Nomage. Confronter les attentes théoriques aux observations du comportement linguistique des nominalisations en corpus », *TAL*, vol. 52, n° 3, p. 129-152, 2011.

Baroni M., Bernardini S., Ferraresi A., Zanchetta E., « The WaCky wide web : a collection of very large linguistically processed web-crawled corpora », *Language ressources and evaluation*, vol. 43, n° 3, p. 209-226, 2009.

Blache P., Rauzy S., « Enrichissement du FTB : un treebank hybride constituants/propriétés », *19ème Conférence TALN*, 2012.

Blanche-Benveniste C., Deulofeu J., Stéfanini J., van den Eynde K., *Pronom et syntaxe : l'approche pronominale et son application au français*, SELAF, 1984.

Brill E., A corpus-based approach to language learning, Thèse de Doctorat, University of Pennsylvania, 1993.

Bîlbîie G., Laurens F., « A Construction-based Analysis of Verbless Relative Adjuncts in French and Romanian », *in* Müller S. (dir.), *Proceedings 16th HPSG Conference*, CSLI Publications, Stanford, p. 5-25, 2009.

Candito M.-H., Crabbé B., Denis P., « Statistical French dependency parsing : treebank conversion and first results », *7th LREC*, La Valletta, 2010.

Candito M.-H., Crabbé B., Denis P., Guérin F., « Analyse syntaxique du français : des constituants aux dépendances », *16ème Conférence TALN*, Senlis, 2009.

Candito M.-H., Perrier G., Guillaume B., Ribeyre C., Fort K., Seddah D., de la Clergerie E., « Deep Syntax Annotation of the Sequoia French Treebank », *9th LREC*, Reykjavik, 2014.

Candito M.-H., Seddah D., « Effectively long-distance dependencies in French : annotation and parsing evaluation », *TLT11*, 2012a.

Candito M.-H., Seddah D., « Le corpus Sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical », *19e Conférence TALN*, Grenoble, France, 2012b.

Clément L., Construction et exploitation d'un corpus syntaxiquement annoté pour le français, Thèse de Doctorat, Université Paris 7, 2001.

Cohen J., « A Coefficient of Agreement for Nominal Scales », *Educational and Psychological Measurement*, vol. 20, n° 1, p. 37-46, 1960.

Danlos L., Colinet M., Steinlin J., « FDTB1 : Repérage des connecteurs de discours dans un corpus français », *Discours*, 2015.

Debaisieux J.-M., Benzitoun C., Deulofeu H.-J., « Le projet ORFEO : Un corpus d'études pour le français contemporain », *Revue Corpus*, vol. 15, p. 91-114, 2016.

Djemaa M., Candito M.-H., Muller P., Vieu L., « Corpus annotation within the French FrameNet : a domain-by-domain methodology », *LREC*, 2016.

Fabre C., Kupść A., « Large and noisy vs small and reliable : combining 2 types of corpora for adjective valence extraction », *5th Corpus Linguistics conference*, Liverpool, 2009.

Fradin B., *Nouvelles approches en morphologie*, PUF, 2003.

Gross M., « Sur quelques groupes nominaux complexes », *in* Chevalier J.-C., Gross M. (dir.), *Méthodes en grammaire française*, Klincksieck, Paris, p. 97-119, 1976.

Guillaume B., de Marneffe M.-C., Perrier G., « Conversion et amelioration de corpus du francais annotes en Universal Dependencies », *TAL*, vol. 60, nº 2, p. 42-66, 2019.

Habert B., « Outiller la linguistique : de l'emprunt de techniques aux rencontres de savoirs », *Revue française de linguistique appliquée*, vol. IX, nº 1, p. 5-24, 2004.

Hale J. T., *Automaton Theories of Human Sentence Comprehension*, University of Chicago Press, 2014.

Heiden S. Magué J.-P. P. B., « TXM : Une plateforme logicielle open-source pour la textométrie conception et développement », *in* Bolasco S. (dir.), *Proc. 10th JADT*, p. 1021-1032, 2010.

Illouz G., Habert B., Folch H., Prévost S., « TyPex : Generic Feature for Text Profiler », *Computer-Assisted Information Retrieval*, RIAO, 2000.

Kinyon A., « A Language-Independent Shallow-Parser Compiler », *39th ACL and 10th EACL,Toulouse*, p. 322-329, 2001.

Kupść A., « TreeLex Meets Adjectival Tables », *International Conference RANLP*, Borovets (Bulgarie), 2009.

Kupść A., Abeillé A., « Treelex : a subcategorization lexicon automatically extracted from a French Treebank », *ICGL*, Workshop on syntactic annotations, Hong-Kong, 2008.

König E., Lezius W., The TIGER language A Description Language for Syntax Graphs, Formal Definition, Technical report, IMS, University of Stuttgart, 2000.

Laurens F., « French predicative verbless utterances », *in* Müller S. (dir.), *Proceedings 15th HPSG Conference, Keihanna*, CSLI Publications, Stanford, p. 152-172, 2008.

Levy R., Andrew G., « Tregex and Tsurgeon : tools for querying and manipulating tree data structures », *5th LREC*, 2006.

Liégeois L., Abeillé A., *Corpus arboré pour le français : guide d'interrogation*, Université Paris Diderot, Paris, 2018.

Maurel D., Belleil C., « Un dictionnaire électronique relationnel des noms propres liés à la géographie », *LINX*, vol. 34-35, p. 77-88, 1996.

McDonald R., Nivre J., Quirmbach-Brundage Y., Goldberg Y., Das D., Ganchev K., Hall K., Petrov S., Zhang H., Täckström O., Bedini C., Castell N. B., Lee J., « Universal dependency annotation for multilingual parsing », *51st ACL Meeting*, Sofia, p. 9297, 2013.

Miller P. H., *Clitics and Constituents in Phrase Structure Grammar*, Garland, 1992.

Mouret F., « La syntaxe des coordinations corrélatives du français », *Langages*, vol. 39, nº 160, p. 67-92, 2005.

Mouret F., Grammaire des constructions coordonnées. Coordinations simples et coordinations à redoublement en français contemporain, Thèse de Doctorat, Université Paris Diderot, 2007.

Nazarenko A., Habert B., Salem A., *Les linguistiques de corpus*, Armand Colin, 1997.

Nivre J., de Marneffe M.-C., Ginter F., Goldberg Y., Hajic J., Manning C. D., McDonald R. T., Petrov S., Pyysalo S., Silveira N., Tsarfaty R., Zeman D., « Universal Dependencies v1 : A Multilingual Treebank Collection », *10th LREC*, Portoroz, 2016.

Paroubek P., Vilnat A., Robba I., Ayache C., « Les résultats de la campagne EASY d'évaluation des analyseurs syntaxiques du français », *14ème Conférence TALN*, 2007.

Pozniak C., Abeillé A., Hemforth B., « French relatives and subject inversion : what's your preference ? », *in* Crysmann B., Sailer M. (dir.), *One-to-Many Relations in Morphology, Syntax and Semantics*, Language Science Press, Berlin, 2019.

Pynte J., New B., Kennedy A., « On-line syntactic and semantic influences in reading revisited », *Journal of Eye Movement Research*, vol. 3, n° 1, p. 1-12, 2009.

Rauzy S., Blache P., « Robustness and processing difficulty models. A pilot study for eye-tracking data on the French Treebank », *COLING*, 2012.

Ribeyre C., Candito M.-H., Seddah D., « Semi-Automatic Deep Syntactic Annotations of the French Treebank », *TLT13*, Tubingen, 2014.

Sagot B., Richard M., Stern R., « Annotation référentielle du Corpus Arboré de Paris 7 en entités nommées », *19ème Conférence TALN*, vol. 2, Grenoble, p. 535-542, 2012.

Schluter N., van Genabith J., « Preparing, Restructuring and Augmenting a French Treebank : Lexicalised Parsing or Coherent Treebanks », *10th PACLING*, Melbourne (Australie), 2007.

Seddah D., Candito M.-H., « Hard Time Parsing Questions : Building a QuestionBank for French », *10th LREC*, 2016.

Seddah D., Clergerie E. D. L., Sagot B., Alonso H. M., Candito M., « Cheating a Parser to Death : Data-driven Cross-Treebank Annotation Transfer », *11th LREC*, ELRA, Miyazaki, Japan, 2018.

Seddah D., Sagot B., Candito M.-H., Mouilleron V., Combet V., « The French social media bank : a treebank of noisy user generated content », *COLING*, Mumbai, 2012.

Seddah D., Tsarfaty R., Kübler S., Candito M.-H., Choi J. D., Farkas R., Foster J., Goenaga I., Gojenola K., Goldberg Y., Green S., Habash N., Kuhlmann M., Maier W., Nivre J., Przepiórkowski A., Roth R., Seeker W., Versley Y., Vincze V., Woli M., Wróblewska A., de la Clergerie E. V., « Overview of the SPMRL 2013 Shared Task : Cross-Framework Evaluation of Parsing Morphologically Rich Languages », *4th Workshop on Statistical Parsing of Morphologically Rich Languages*, ACL, Seattle, p. 146-182, 2013.

Sekine S., Collins M., Brooks D., Ellis D., *Evalb software*, 2008. `nlp.cs.nyu.edu/evalb`.

Silberztein M., *Dictionnaires électroniques et analyse automatique de textes : le système Intex*, Masson, Paris, 1993.

Taylor A., Marcus M., Santorini B., « Treebanks : Building and using parsed corpora », *in* Abeillé A. (dir.), *Treebanks*, Kluwer, Dordrecht, p. 5-22, 2003.

Toussenel F.*, « Marquage de constituants sur un corpus français, résultats et exploitation linguistiques »*, Mémoire de Master, Université Paris Diderot, 2001.

Villemonte de La Clergerie É., Hamon O., Mostefa D., Ayache C., Paroubek P., Vilnat A., « PASSAGE : from French Parser Evaluation to Large Sized Treebank », *6th LREC*, Marrakech, 2008.