Institut national des langues et civilisations orientales

Equipe de recherche Textes, Informatique, Multilinguisme

Discipline : Traitement automatique des langues, parcours Traductique et gestion de l'information

Pertinence de deux types d'analyse syntaxique dans le cadre d'un outil didactique d'aide à la lecture pour les apprenants serbes de FLE

Par : **Neda Lestarevic**

Sous la direction de : Mathieu Valette et François Stuck

Date de soutenance : 30 novembre 2018

Remerciements

Je tiens à remercier sincèrement toute l'équipe pédaogique du PluriTAL. Elle nous a offert une formation d'une grande qualité et nous a appris tellement de choses précieuses pendant ces deux années.

Je remercie mes deux directeurs de stage et de mémoire, François Stuck et Mathieu Valette, pour leur soutien infaillible tout au long de mon stage et de l'écriture de ce mémoire. Je les remercie également pour m'avoir donné l'occasion de travailler sur le dispositif Déjà Lu, qui était pour moi un projet très intéressant et très inspirant.

Finalement, je remercie les apprenants qui ont participé dans cette expérimentation pour le temps qu'ils ont accordé à ce sujet, ainsi que pour leurs idées et leurs conseils.

Table des matières

Introduction

Les systèmes d'information deviennent de plus en plus une partie intégrante de notre quotidien. Ils s'intègrent dans tous les domaines et dans toutes les professions.

Il y a quelques dizaines d'années, la profession de l'enseignant, autrefois une profession classique, se contentait d'un professeur au tableau en train de déclamer ses leçons aux élèves. Aujourd'hui elle est devenue un domaine que les enseignants partagent avec les ordinateurs.

La situation est la même dans le domaine de l'enseignement des langues. Nombreuses sont les plateformes disponibles sur l'internet, payantes ou libres d'accès, qui permettent aux élèves d'apprendre les bases d'une langue étrangère, puis d'améliorer leur niveau (DuoLingo, Babbel, LingQ...). Il s'agit de petites leçons suivies d'exercices et des tests qui permettent d'évaluer ce qui a été appris.

A part l'enseignement de la grammaire pure, il existe d'autres manières, plus pratiques, pour apprendre une langue – en écoutant, en parlant, en lisant. Aujourd'hui, nous avons également l'accès aux nombreux outils informatiques, les applications pour le smartphone, qui permettent le développement de ces compétences pragmatiques.

Ce mémoire se consacre à l'adaptation d'un tel outil, appelé «Déjà Lu », à la langue française.

Déjà Lu est un outil d'aide à la lecture. Il permet à l'utilisateur d'explorer un texte enrichi grâce à diverses informations linguistiques. Il intègre un certain nombre d'outils de Traitement Automatique des Langues (TAL), notamment un segmenteur de phrases en mots (tokenizer), un étiqueteur morphologique et un analyseur syntaxique.

Pour faire cette étude, nous nous sommes focalisés sur la fonctionnalité syntaxique de Déjà Lu. Le but de cette fonctionnalité est de faire réfléchir l'apprenant – si cet apprenant connaissait les fonctions syntaxiques des mots dans le texte, il pourrait deviner plus facilement le sens d'une phrase globalement, sans avoir forcément besoin de consulter un dictionnaire.

Dans l'objectif d'un apprentissage du Français Langue Etrangère (FLE), nous avons

cherché à comparer l'apport didactique de deux analyseurs syntaxiques du français en les intégrant dans l'outil "Déjà Lu" :

1) TALISMANE, qui suit la norme d'annotation du French Treebank (Abeillé et al., 2003) adaptée à la syntaxe et à la tradition du français ;
2) Stanford Parser intégrant les dépendances universelles (McDonald et al., 2013) – une forme d'annotation universelle applicable à des langues éventuellement distantes syntaxiquement.

L'enjeu consiste à découvrir quel type d'annotation est meilleur du point de vue de l'apprenant de FLE : le type d'annotation proche de la langue cible (TALISMANE) ou celui, plus universel, se rapprochant potentiellement de sa langue maternelle (Stanford Parser). Nous avons décidé d'évaluer cet apport didactique avec des apprenants locuteurs d'une langue syntaxiquement éloignée du français : une langue slave, le serbe, dont l'ordre des mots est libre et les cas grammaticaux expriment les fonctions syntaxiques.

Notre hypothèse est que l'influence de la syntaxe de la langue maternelle est importante pendant l'apprentissage des langues étrangères, et que les apprenants, surtout au début de l'apprentissage, se pencheront sur la syntaxe de leur propre langue pour comprendre celle d'une langue étrangère.

Étant donné que le niveau des connaissances en langue étrangère de l'apprenant peut être très bas, il risque de se tromper si l'analyse est incorrecte, de la même manière qu'une analyse correcte peut l'aider à comprendre le texte.

Nous tenons à préciser que cette étude représente une pré-étude, une réflexion sur le sujet proposé, et que les résultats obtenus sont une estimation. Une étude plus précise, selon un protocole d'évaluation plus rigoureux, sur un échantillon d'élèves plus important, et un corpus plus vaste, dépassait nos ressources actuels, et exigeait beaucoup plus de temps. Cependant, elle est envisageable pour l'avenir.

Dans la **première partie** de ce mémoire, nous allons présenter les acquis actuels dans les domaines de l'analyse syntaxique automatique pour le français, de l'apprentissage des langues assisté par ordinateur, et d'aide à la lecture. Cette partie pose les pré-requis à notre étude.

Dans le **deuxième chapitre** nous allons d'abord parler de la problématique de notre étude :

1. De la syntaxe de la langue serbe et ses différences avec la syntaxe de la langue française.

2. De la différence d'analyses entre deux analyseurs syntaxiques à comparer - Stanford Parser et TALISMANE.

La suite de ce chapitre sera dédiée à la présentation du corpus en français utilisé pour l'analyse syntaxique, et à la méthode que nous avons suivie pour faire cette étude.

Dans le **troisième chapitre** nous allons présenter le type d'évaluation utilisé, et les résultats obtenus sur un échantillon d'apprenants serbes.

Le **quatrième chapitre** aborde une discussion sur les résultats - à savoir la pertinence du choix de l'analyseur syntaxique pour l'apprenant.

Une **conclusion** clôturera ce mémoire.

Première partie

Etat de l'art

Chapitre 1

Apprentissage des langues assisté par ordinateur (ALAO)

L'ALAO est un domaine de recherche et développement qui regroupe plusieurs disciplines :
- la linguistique (la linguistique théorique, appliquée, la didactique des langues),
- la linguistique-informatique,
- l'informatique (en particulier l'intelligence artificielle (IA)),
- la psycholinguistique (Chanier, 1995).

L'intégration du numérique dans l'apprentissage en général et dans l'apprentissage des langues a commencé dans les années 80 et continue jusqu'à aujourd'hui. L'éventail des ressources est très large : supports matériels construits ou bruts, documents sonores ou vidéos disponibles sur Internet, ressources créées spécifiquement, applications diverses d'apprentissage. Ces ressources offrent à l'utilisateur la possibilité de travailler à son rythme.

Au début, la plupart du contenu, les applications ou les ressources étaient utilisées sur un support autonome, sur un CD-ROM ou une disquette. Aujourd'hui cela a changé - la plupart des ressources se trouvent en ligne, ce qui présente de nombreux avantages - la facilité de l'utilisation et de l'installation, la compatibilité et la pérennité, l'accès souvent gratuit, la sauvegarde des données personnelles sur le serveur, etc... (Desmet,2006)

La communauté ALAO compte de nombreuses associations qui traitent ce sujet : CALICO (Computer Assisted Language Learning and Instruction Consortium), EUROCALL (European Association for CALL), IAALT (International Association for Language Learning Technology), des manifestations scientifiques majeures (WorldCall, Eurocall, Jalt Call), des revues (Alsic, LLT, ReCall, Call, Calico, e-FLT, etc...)(Mkhitaryan,2014).

Parmi les divers types d'outils pour l'apprentissage des langues, par exemple les cours interactifs (Tell Me More, ALPCU - inalco), les exerciseurs, etc., nous nous s'intéresseront ici à un type d'outil particulier : l'aide à la lecture en L2, visant à favoriser la pratique de la lecture par les apprenants.

1.1 Outils d'aide à la lecture

Les outils d'aide à la lecture se focalisent sur les différents aspects du processus de la compréhension écrite. Il peuvent servir à l'explicitation linguistique, au choix thématique, à l'aide lexicale... Nous présentons une liste non-exhaustive de certains outils d'aide à la lecture.

1.1.1 NaviLire (Lundquist et al., 2006) et TextRay (Lundquist, 2007)

NaviLire est un logiciel créé par Lita Lundquist, Jean-Luc Minel et Javier Couto, qui permet l'affichage de textes et la navigation dans le texte. Il se base sur NaviTexte, qui est une plateforme informatique pour la navigation textuelle. La navigation dans ce cadre-ci représente la possibilité pour l'utilisateur d'afficher les informations supplémentaires sur le texte selon ses propres connaissances.

NaviLire est une adaptation de NaviTexte pour l'apprentissage des langues étrangères. Les textes sont ainsi annotés ou par l'enseignant, de façon semi-manuelle, c'est-à-dire, en s'appuyant sur certains outils TAL.

La perception de la structure et de la cohérence du texte est importante pour la compréhension du texte en général. En outre, le lecteur est censé cerner les référents et les relations entre eux, l'orientation du sujet qui est créée par la prédication. NaviLire se focalise sur la forme générale d'un texte - il permet au lecteur de repérer les pistes de lecture caractéristiques pour certains types de textes (Lundquist, 2013).

Le logiciel TextRay a été crée en 2005 en collaboration entre le Département pour la linguistique-informatique et FIRST, l'école de business de Copenhague, basé sur l'idée de Lita Lundquist. Ce logiciel vise 3 objectifs dans la lecture d'un texte :

1. Diviser le texte en unités de lecture plus petites
2. Aider le lecteur à identifier les constituants principaux d'une phrase
3. Aider l'utilisateur à reproduire ce qu'il entend sous forme écrite - avec la dictée.

NaviLire et TextRay prennent en compte les stratégies fondamentales qu'un lecteur emploie en lisant un texte écrit dans une langue étrangère, et de cette manière ils l'aident pendant le processus de la lecture et de la compréhension.

1.1.2 Contexto (Crispino et al., 1999)

En 1998, la plateforme Contexto a été créée par le laboratoire LaLICC de l'Université Paris–Sorbonne et le groupe de T.A.L. de l'Université de la République d'Uruguay. Cette

plateforme filtrait les textes sémantiquement (selon la thématique).

1.1.3 GLOSSER (Nerbonne et al., 1998)

GLOSSER est un logiciel qui permettait aux lecteurs de trouver les informations sur les mots dans le texte lu. Il fournit trois types d'information :

1. la forme choisie lemmatisée avec les étiquettes Part Of Speech - Partie du Discours (POS)
2. l'entrée de dictionnaire pour ce lemme
3. Les exemples de ce mot dans des corpus différents

Ce qui était un grand avantage de GLOSSER était le fait qu'il accède très vite au dictionnaire, le lecteur n'ayant alors plus besoin de trop interrompre sa lecture pour aller chercher par lui même un mot dans le dictionnaire (Mitkov, 2004).

1.1.4 COMPASS (Breidt & Feldweg, 1997)

COMPASS est un projet collaboratif entre plusieurs institutions : française, allemande et anglaise. COMPASS a des fonctionnalités proches de GLOSSER mais il donne aussi la traduction de l'extrait auquel appartient le mot sur lequel l'utilisateur a cliqué en tenant compte du contexte. Les couples de langue disponibles sont anglais - français et allemand - anglais (Mkhitaryan,2014).

1.2 Déjà Lu (ERTIM, 2014)

Déjà Lu est un projet en cours de développement au sein de l'Equipe de Recherche Textes, Informatique, Multilinguisme. Il vise à élaborer une application en ligne favorisant l'apprentissage des langues, et la pratique précoce et intensive de la lecture. La lecture est enrichie automatiquement d'informations linguistiques permettant au lecteur de pratiquer une lecture autonome et de développer ses compétences de la compréhension écrite, ses stratégies de lecture et de compréhension, sa réflexion métalinguistique.

Il s'agit d'une plateforme modulaire multilingue intégrant divers outils de TAL (segmenteur en phrases, analyseur en parties de discours, analyseur syntaxique), et proposant, selon les versions, diverses fonctionnalités d'exploration du texte, tant au niveau lexical, que morphologique ou syntaxique.

Sa première version, datant de 2014, proposait 4 langues : français, hongrois, hindi et thaï. Elle intégrait les segmenteurs de phrases et de mots (thaï) et les analyseurs morpho-

syntaxiques.

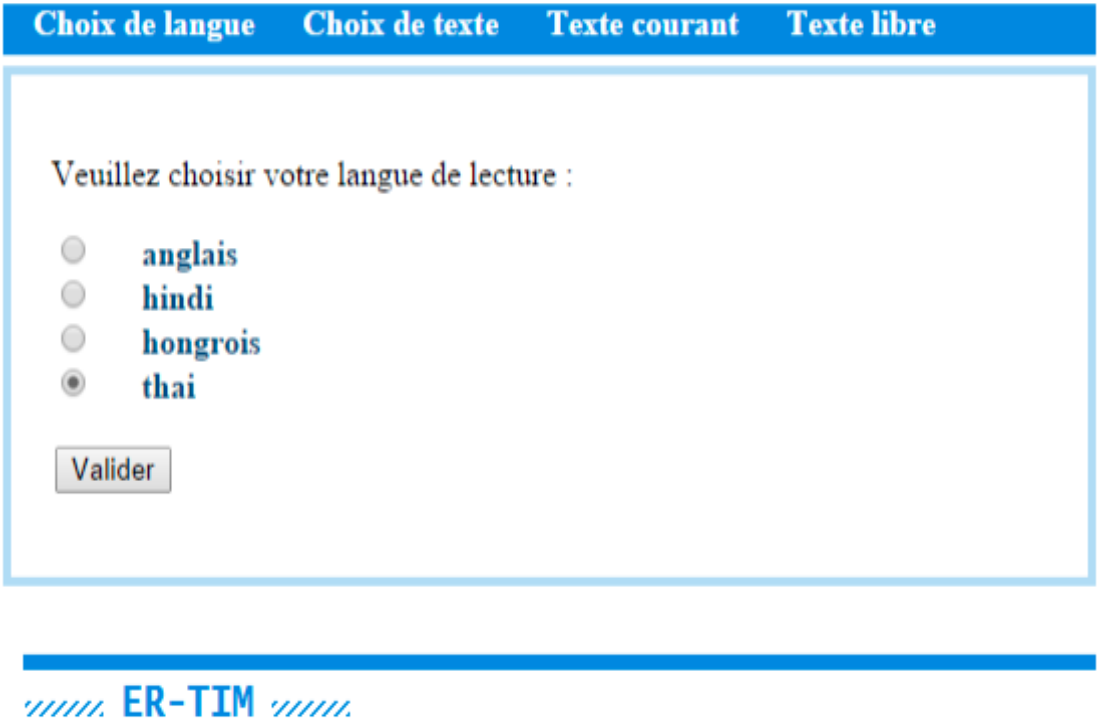

Figure 7. La première version de Déjà Lu, page d'accueil

La deuxième version de Déjà Lu propose quelques fonctionnalités supplémentaires. A l'analyse en parties du discours, elle ajoute l'analyse syntaxique pour 2 langues : français et hongrois, et certaines fonctionnalités lexicales qui sont en cours de développement (les emprunts, les mots composés, les toponymes, etc...).

Cette dernière version de l'application est personnalisée (l'utilisateur a son propre espace où il peut ajouter les textes de son choix, conformes à ses centres d'intérêt, à ses nécessités professionnelles et à son niveau) et autonome (la connexion internet n'est plus nécessaire une fois le texte choisi).

1.2.1 L'Interface Déjà Lu

Le dispositif Déjà Lu s'organise autour d'une architecture client-serveur. Les requêtes faites par l'utilisateur, via un module javaScript AJAX, sont envoyées au serveur sous forme de requêtes HTTP.

L'interface de l'outil fonctionne de la manière suivante :

1. Identification

L'utilisateur est invité à s'authentifier. Après l'identification il accède a son espace personnalisé où il peut choisir soit de lire les textes déjà ajoutés précédemment, soit d'y ajouter ses propres textes.

Figure 8. Déjà Lu Version 2, Identification

2. Ajout d'un texte

L'utilisateur peut ajouter son propre texte. Il peut le copier dans la partie de l'onglet prévue à cet effet, ou le charger depuis son disque dur ou d'un lien URL. Les types de textes qu'on peut ajouter sont plusieurs : pdf, doc, odt, txt, html... Le texte ajouté est ensuite pré-traité sur le serveur - on effectue le parsing, la normalisation, la pré-annotation...

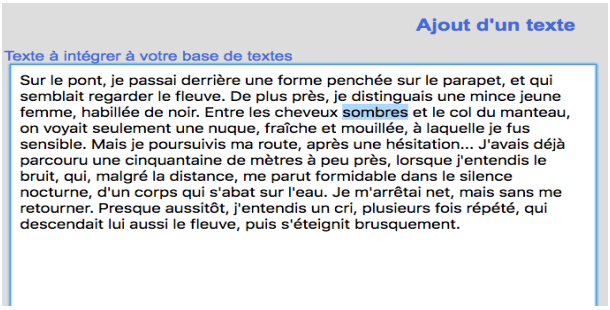

Figure 9. Déjà Lu Version 2, Ajout d'un texte

3. Choix du texte

L'utilisateur est invité à choisir un texte parmi les textes ajoutés. Dès que le texte est choisi, ce dernier est télécharge, ainsi que ses tables de pré-annotation, permettant l'accès au texte et toutes les fonctionnalités de Déjà Lu.

Figure 10. Déjà Lu Version 2, Choix d'un texte

4. Affichage du texte

L'utilisateur peut alors choisir les mots et les phrases qu'il veut interroger et mettre en valeur, ainsi que les fonctionnalités qu'il souhaite utiliser.

1.2.2 Les fonctionnalités

Les fonctionnalités disponibles pour la deuxième version de Déjà Lu se subdivisent en fonctionnalités lexicales, fonctionnalités syntaxiques et fonctionnalités morphologiques.

1. Fonctionnalités lexicales

- *Déjà lus* : Sélectionner toutes les occurrences du mot choisi qui sont apparus précédemment dans le texte.
- *Ailleurs* : Retrouver toutes les occurrences du mot choisi dans tous les textes ajoutés par l'utilisateur.
- *Mots empruntés* : Sélectionner tous les mots dans le texte qui sont empruntés aux autres langues.
- *Toponymes et autres noms* : Retrouver tous les noms des lieux et les noms propres dans le texte.
- *Mots composés* : Sélectionner tous les mots qui sont composés de deux ou plusieurs mots.

2. Fonctionnalités syntaxiques

- *Le verbe et ses actants* : Retrouver le verbe principal de la phrase choisie, ainsi que son sujet, objet et les compléments circonstanciels.
- *Les dépendants d'un mot* : Sélectionner tous les mots dans la phrase qui dépendent du mot sélectionné. Imaginons que nous avons sélectionné le mot "fille" dans la phrase "J'ai vu une belle fille." –> le déterminant "une" et l'adjectif "belle" dépendent du nom "fille", et ils seront affichés comme ses dépendants et colorés en bleu (*Figure 11. Un exemple de texte en hongrois*).
- *La tête d'un mot* : Sélectionner le mot dont dépend le mot choisi. Si nous prenons l'exemple précédent, en choisissant le mot "fille", sa tête dans cette phrase (le mot dont il dépend), qui est le verbe, va être affichée colorée en rouge.

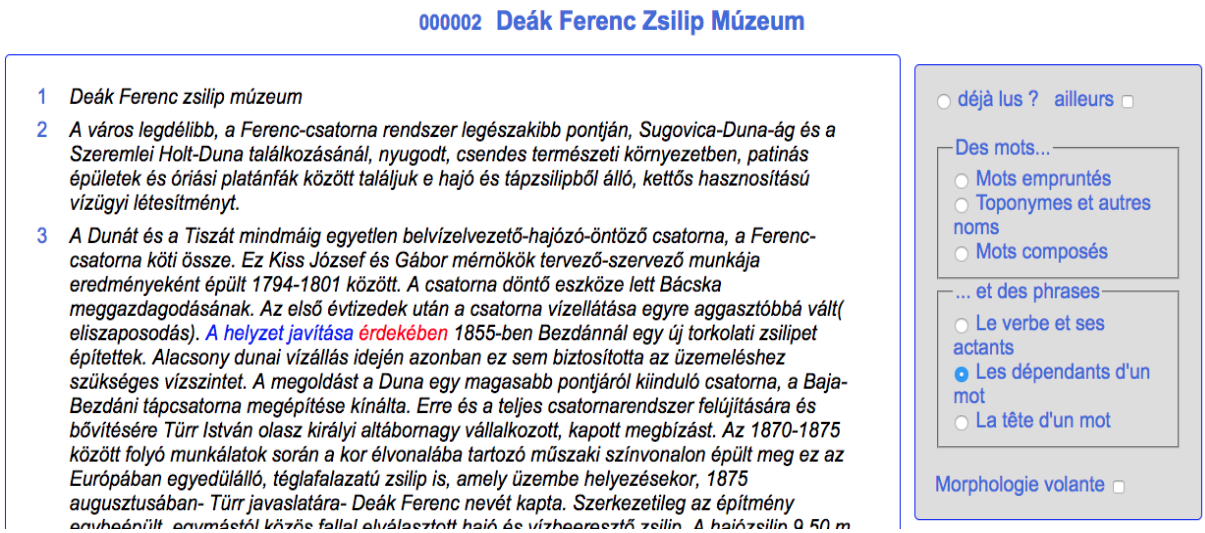

Figure 11. Déjà Lu Version 2, Affichage du texte, la fonctionnalité "Dépendants d'un mot"

3. Fonctionnalité morphologiques

- *Morphologie volante* : En survolant les mots avec la souris, nous pouvons visualiser leur partie du discours, leur forme canonique (pour les verbes - infinitif, pour les noms et adjectifs - masculin et singulier, etc...), et les détails morphologiques (le genre, le nombre, le mode, le temps) (*Figure 12.*).

Figure 12. Déjà Lu Version 2, Affichage du texte, la fonctionnalité "Morphologie volante"

Pour faire cette étude, nous nous sommes surtout focalisés sur les fonctionnalités syntaxiques de Déjà Lu. Nous les avons utilisées pour tester l'apport des analyseurs que nous avons comparés dans les chapitres qui suivent.

Chapitre 2

Analyse syntaxique automatique

L'analyse syntaxique est une tâche de la linguistique qui est assez précise lorsqu'elle est faite par l'humain qui a la connaissance de toutes les règles de la langue à appliquer et qui tient compte du sens de la phrase, du contexte, des possibles phrases complexes, etc. Pour une machine, cette tâche est plus complexe. Prenons l'exemple de cette phrase :

Exemple 1.
«L'artiste peint la nuit ».

Dans l'exemple 1, nous nous retrouvons devant une ambiguïté syntaxique - la nuit pourrait être un complément d'objet direct (c'est la nuit que l'artiste peint), ou un complément circonstanciel du temps (le moment où il peint). L'humain pourrait comprendre de quoi il s'agit seulement en voyant le contexte. Pour la machine, en effet, malgré les règles et les modèles que l'on puisse lui donner, si elle se retrouve devant une phrase ambiguë, il est possible qu'elle se trompe.

Pourtant, la machine reste essentielle pour le fonctionnement de nombreux systèmes (systèmes de traduction automatique, systèmes de questions-réponses, etc...)

2.1 Les débuts

Les premières analyses syntaxiques automatiques sont apparues lors des années 50 avec plusieurs algorithmes d'analyse développés par Victor Yngve, et par Kasami et Younger. Le plus connu était l'algorithme de Cocke-Younger-Kasami (CYK). Il permettait de déterminer si un mot faisait partie d'une grammaire. Si la réponse était oui, il générait l'arbre syntaxique pour ce mot. Ces algorithmes fonctionnaient à l'aide des grammaires formelles – hors contexte, décrites par Chomsky en 1956. (Harrison, 1978)

Dans les années 1960, suite aux travaux de Lucien Tesnière, un autre type de représentation de l'analyse syntaxique a été introduit sous la forme de grammaires de dépendances.

Dans les années 90, les recherches dans ce domaine ont été principalement focalisées sur les méthodes statistiques en utilisant des grammaires hors contexte probabilistes (Smith

& Johnson, 2007).

Aujourd'hui, le système statistique d'apprentissage supervisé est le plus utilisé, avec un corpus de données annotées qui sert de modèle pour l'analyse. Pourtant, les grammaires formelles sont utilisées encore dans les cas où il n'y a pas suffisamment de données annotées pour un apprentissage efficace.

2.2 Représentation syntaxique en constituants & représentation syntaxique en dépendances

L'analyse syntaxique d'un texte peut être représentée de deux manières - avec la représentation en constituants et la représentation en dépendances.

Pour les deux représentations, étant donné qu'il s'agit des structures hiérarchisées, la représentation graphique qui convient le mieux est celle d'un arbre (Gédéon, 2011).

2.2.1 Analyse syntaxique en constituants

La représentation en constituants comprend la division de la phrase en syntagmes. Les syntagmes sont des groupes de mots participant en même rôle syntaxique dans la phrase.

Un syntagme est un intermédiaire entre l'ensemble global - la phrase et la division unitaire - le mot. Le principe de ce concept est de pouvoir subdiviser logiquement la phrase en groupes de plus en plus petits (Gédéon, 2011). Ce type d'analyse est adapté aux langues dont l'ordre de mots est plus ou moins fixe. Prenons l'exemple de la phrase suivante :

Exemple 2. Les petits ruisseaux font des grands rivières.

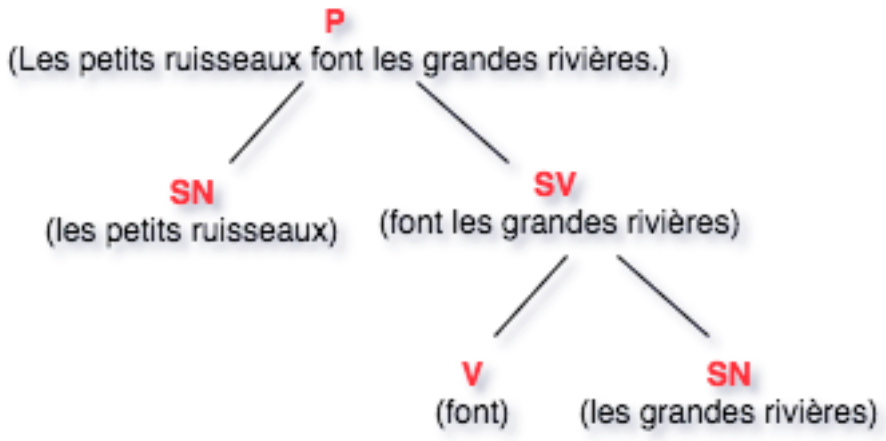

Figure 1. Représentation de l'analyse syntaxique en constituants sous forme d'un arbre

Comme nous pouvons le voir sur la *Figure 1.*, la plus grande unité, la phrase (P) est divisée en deux syntagmes :
1) Un syntagme nominal (SN), dont le noyau est le nom (ruisseaux), et
2) Un syntagme verbal (SV), dont le noyau est le verbe (font). Le syntagme verbal se subdivise en un verbe (V) et un syntagme nominal (les grandes rivières).

2.2.2 Analyse syntaxique en dépendances

La représentation en dépendances considère chaque mot de la phrase comme un élément indépendant - les mots ne sont pas regroupé en syntagmes. Chaque mot a un antécédent dont il dépend, qui est sa "tête". Le verbe principal est rattaché à une "racine" imaginaire ("root") de toute la phrase.

C'est Lucien Tesnière qui a introduit cette approche de syntaxe dans les années 60. Dans ses Eléments de syntaxe structurale (1959), il présente un modèle d'analyse syntaxique qui a pour but la description des connexions structurales entre mots en tant qu'éléments constitutifs de la phrase.

Une particularité de l'analyse en dépendances est la notion d'une racine « artificielle » de l'arbre, reliée au verbe principal de la phrase et à n'importe quel mot n'ayant pu être placé dans l'arbre de dépendance. Cette racine artificielle de l'arbre a été introduite pour des raisons techniques (Gédéon, 2011).

Rappelons-nous de la phrase dans l'exemple 2. Sa représentation en dépendances a la forme suivante :

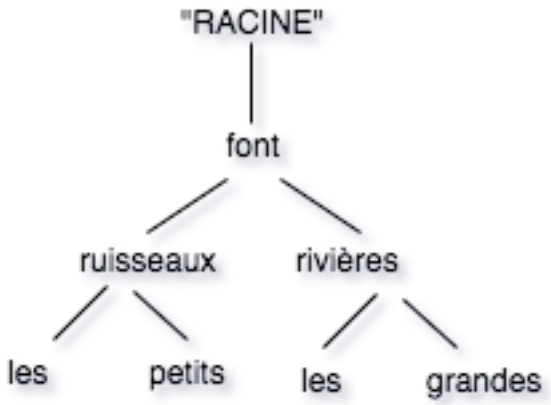

Figure 2. Représentation de l'analyse syntaxique en dépendances sous forme d'un arbre

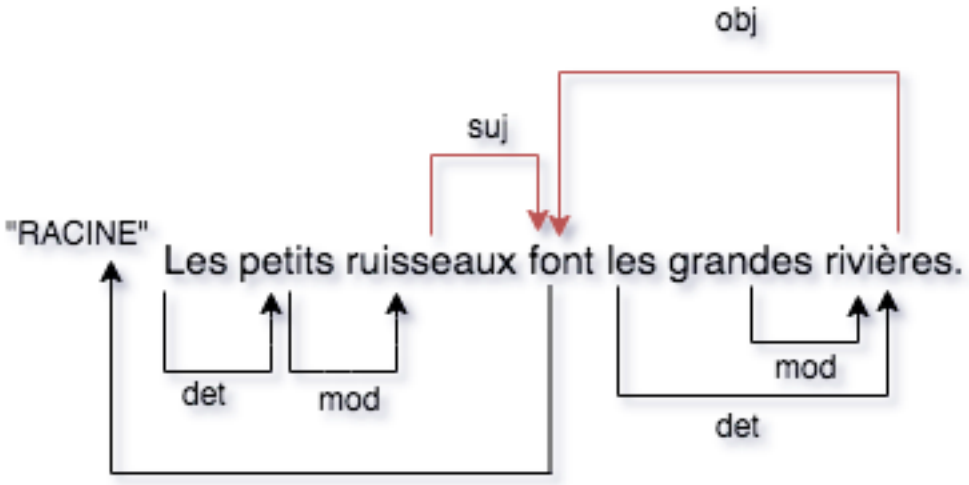

Figure 3. Représentation linéaire de l'analyse syntaxique en dépendances

La *Figure 2.* montre un arbre de dépendances de la phrase de notre exemple. Comme nous avons dit, le verbe est la "tête" de la phrase, les mots de la phrase qui dépendent directement du verbe se trouvent au premier niveau, ensuite, au niveau suivant sont représentés les mots qui dépendent des dépendants du verbe, etc.

La *Figure 3. (det = détermination ; mod = modifieur ; suj = sujet ; obj = objet)* présente une représentation linéaire des dépendances. Les flèches pointent sur la tête de chaque mot dans la phrase. Nous pouvons également voir les types des relations syntaxiques entre les éléments de la phrase.

2.3 Dépendances Universelles

Les Dépendances Universelles (UD) font partie d'un projet créé afin de fournir un cadre universel à l'analyse syntaxique de plusieurs langues (McDonald et al., 2013).

Elles représentent le résultat de différents initiatives pour une unification de traitements interlingues. Elles sont basées sur les dépendances universelles originales de Stanford Parser, un jeu de tags de Google et une la sortie au format CONLL-U.

La première version des guides d'annotation de dépendances universelles est sortie en Octobre 2014.

La première version des Universal Dependencies (UD) comptait des Treebanks pour 47 langues.

La dernière, la version 2.2. de juillet 2018 en compte desormais pour 71 langues.

Le but de ce système est d'annoter des phénomènes syntaxiques similaires de la même manière quelque soit la langue, mais tout en prenant soin de ne pas ajouter des fonctions qui n'existent pas dans une langue seulement parce qu'elles existent dans d'autres langues.

2.3.1 Quelques règles

Chaque mot dépend soit d'un autre mot de la phrase, soit de la racine artificielle - de "ROOT", comme décrit dans la section précédente.

L'analyse se base sur trois principes :

1. Les mots lexicaux (les mots porteurs du sens) sont reliés par des relations de dépendances ;

2. Les mots grammaticaux dépendent des mots lexicaux ;

3. La ponctuation dépend de la tête de la phrase ou de la subordonnée dans laquelle elle se trouve ;

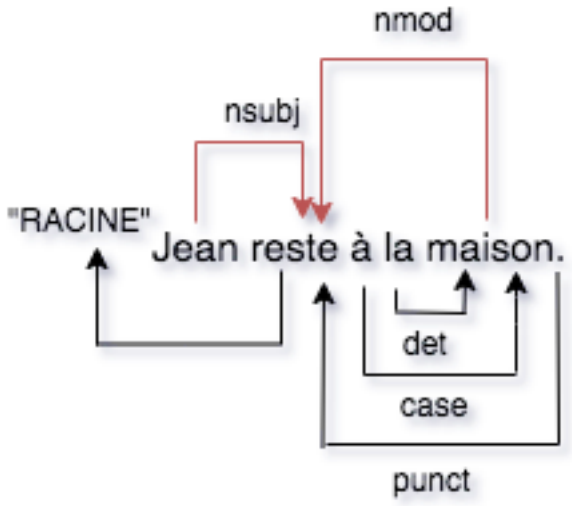

Figure 4. Analyse d'une phrase en dépendances universelles

La *Figure 4. (nsubj = sujet ; nmod = modifieur nominal ; det = détermination ; case = cas ; punct = ponctuation)* montre une analyse faite en dépendances universelles. Nous pouvons voir que c'est le nom "maison" (mot lexical, qui porte le sens) qui est rattaché directement au verbe, alors que la préposition "à" (mot grammatical, qui fait les relations entre les mots lexicaux) dépend du nom.

Les mots lexicaux en tant que gouverneurs de syntagmes permettent d'établir un parallélisme entre les langues, étant donné que les prépositions dans une langue peuvent être équivalentes aux terminaisons de flexion dans une autre langue. Prenons un exemple :

Exemple 3.
a) «Je vais à la maison». (en français)
b) «Idem kući». (en serbe)

Dans cet exemple, la préposition "à" dans la phrase a) est équivalente à la terminaison "i" (kuć - i) dans la phrase b) - avec la signification d'un but. En français, le but est exprimé avec la préposition "à", en serbe avec le cas datif. Étant donné que ce qui est commun entre les deux exemples du complément circonstanciel du but est le mot lexical "maison", il est choisi pour être le mot principal de ce syntagme, selon les dépendances universelles. Ainsi, les prépositions, les clitiques, les postpositions sont toujours dépendantes du mot auquel elles se réfèrent.

Le principe de voir les mots lexicaux rattachés directement au verbe principal de l'arbre syntaxique se différencie des tendances d'aujourd'hui où ce sont les mots grammaticaux qui représentent les noyaux de syntagmes. Cependant, cette tendance est en accord avec le travail de Tesnière (1959), qui pensait que les mots lexicaux sont les noyaux de la phrase

(McDonald et al., 2013).

Les dépendances universelles ont encore quelques spécificités d'analyse :

- Dans les relations de coordination le premier élément est celui dont dépendent d'autres éléments de cette coordination. Tous les autres éléments, ainsi que la conjonction de coordination (ou la ponctuation qui marque la coordination), dépendent de lui, ce qui montre la *Figure 5. (cc = conjonction de coordination ; conj = élément de coordination).*

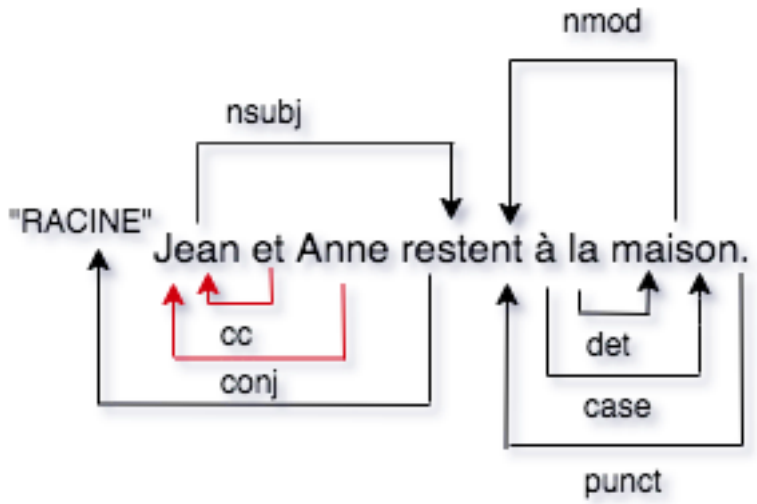

Figure 5. Gestion de la coordination avec les dépendances universelles

- Les verbes auxiliaires ou copules (dans les cas d'attribut) sont attachés au prédicat lexical.

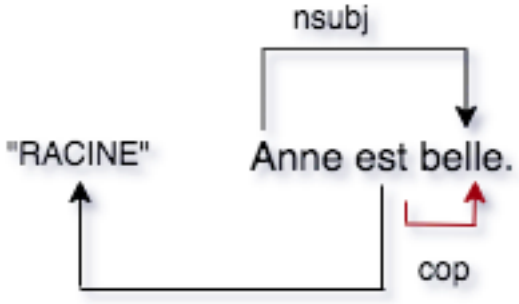

Figure 6. Gestion des verbes copulatifs par les dépendances universelles

Dans la *Figure 6. (cop = copule)*, c'est l'attribut qui est considéré comme la tête de la phrase, étant donné que c'est un mot lexical, avec du contenu. Le verbe copulatif est rattaché à l'attribut.

- Les clitiques et les mots contractés sont décomposés pendant l'analyse (du = de + le).

2.4 Quelques analyseurs syntaxiques pour le français

2.4.1 Stanford Parser (Green et al., 2011)

Stanford Parser (site officiel) est un logiciel d'analyse syntaxique automatique développé en 2003 par le groupe de recherche de l'Université de Stanford. Initialement il était censé annoter les textes en anglais, mais au fur et à mesure ses capacités ont été étendues vers d'autres langues (français, arabe, allemand, chinois. . .). L'outil est constamment amélioré. Cet analyseur analyse les textes en dépendances universelles. Cependant sa version française ne permet pas aujourd'hui la lemmatisation de tokens ce qui représente un défaut pour notre travail.

2.4.2 TALISMANE (Urieli & Tanguy, 2013)

TALISMANE (site web) est un analyseur syntaxique développé en 2013 dans le cadre de la thèse d'Assaf Urieli. Il a été entieèrement codé en Java. Il propose la tokenisation, la lemmatisation, l'analyse morphosyntaxique, la segmentation en phrases et l'analyse syntaxique en dépendances. Sa norme d'annotation est celle de French Treebank et elle est adapté à la syntaxe du français. TALISMANE combine l'approche statistique avec l'utilisation des grammaires dans le cas d'ambiguïtés syntaxiques.

2.4.3 FRMG (Thomasset & de La Clergerie, 2005)

Le FRench MetaGrammar (FRMG) est un analyseur syntaxique profond à large couverture pour le français (site web). Une description grammaticale de haut niveau sous forme de métagrammaire (MG) sert de point de départ pour la génération d'une grammaire d'arbres adjoints (TAG, Tree Adjoining Grammar) (Abeillé, 2002). Cette grammaire est transformée par le système DYALOG (de La Clergerie, 2005) en un analyseur syntaxique. Il fait la tokenisation, étiquetage morphologique, analyse syntaxique en dépendances et il a 4 modes d'utilisation : mode ligne de commande, mode shell, mode serveur, mode service web. Il utilise l'annotation au format EASY (Gendner et al., 2008). Le schéma d'annotation EASY n'est pas réellement un schéma d'annotation en dépendances syntaxiques, il s'agit plutôt d'une annotation hybride entre les dépendances et les constituants (Cerisara & Gardent, 2009).

2.4.4 MaltParser (Nivre et al., 2007)

MaltParser (site web) est un analyseur syntaxique statistique développé par Johan Hall, Jens Nilsson et Joakim Nivre à l'Université Växjö et à l'Université Uppsala en Suède. Il utilise un modèle, un corpus annoté - Treebank, pour analyser d'autres textes. C'est un analyseur en dépendances. Pour la version française, il utilise en tant que corpus de référence par défaut French Treebank.

2.4.5 Spacy (Honnibal & Johnson, 2015)

Spacy (site officiel) est un module Python qui représente une boîte à outils des différentes tâches de traitement automatique des langues. Sa première version est apparue en 2015. Il propose la tokenisation, lemmatisation, segmentation en phrases et l'analyse syntaxique en dépendances. Il permet le traitement de plus de 31 langues dans le monde. C'est un analyseur statistique. Pour la version française il utilise le corpus Sequoia (Candito & Seddah, 2012), annoté en dépendances universelles.

Deuxième partie

Problématique, Corpus & Méthode

Chapitre 3

Problématique

Les langues romaines et les langues slaves appartiennent à des familles de langues différentes. Ce qui veut dire qu'elles ne respectent parfois pas les mêmes règles syntaxiques, de grammaire, et de flexion. C'est pourquoi il peut paraître déstabilisant pour un apprenant serbe d'apprendre le français et inversement. Le processus d'apprentissage peut être facilité par un outil d'aide à la lecture informatisé intégrant les fonctionnalités grammaticales de la langue. Dans cette étude nous avons cherché à répondre à la problématique suivante : «l'utilisation d'un analyseur syntaxique est-elle pertinente pour des apprenants serbes de FLE ? Quel est l'analyseur le plus utile ? ».

Ce travail est dédié à la comparaison entre deux analyseurs syntaxiques en dépendances pour le français, TALISMANE et Stanford Parser. Cette comparaison s'effectue à l'aide de l'outil Déjà Lu, dans un cadre spécifique - l'apprentissage des langues assisté par ordinateur, avec un groupe d'apprenants du français de langue maternelle serbe.
Nous voulions découvrir :

- si l'analyse syntaxique intégrée dans l'outil d'aide à la lecture est bénéfique pour les apprenants ou, au contraire si elle peut les déconcerter étant donné qu'elle peut parfois contenir des erreurs ?

- quel est le type d'analyse préféré - celui plus universel ou celui plus centré vers le français ?

- quelles fonctionnalités de syntaxe sont les plus utiles pour les apprenants, celle de l'affichage des dépendances ou celles plus globales, par exemple l'affichage du verbe et de ses actants dans la phrase ?

3.1 Le choix de l'analyseur

Nous avons selectionné 2 analyseurs syntaxiques en dépendances pour faire la comparaison : TALISMANE et Stanford Parser. La raison pour laquelle nous avons choisi TALISMANE est le fait qu'il atteint une exactitude comparable aux autres parseurs pour le français (Urielli, 2013 ; Candito et al, 2010), mais il combine l'approche statistique et

l'utilisation des grammaires, ce qui pourrait apporter de meilleurs résultats dans les cas de phrases ambiguës.

De l'autre coté, Stanford Parser est un analyseur en dépendances universelles. Son analyse est plus universelle, alors que celle de TALISMANE correspond plus à la tradition du français.

Même s'il est préférable que l'analyse soit faite au mieux, l'exactitude de l'analyse n'était pas un critère que nous avons voulu interroger.

Nous avons trouvé intéressante cette opposition entre deux analyses, et nous avons cherché à l'étudier dans un cas pragmatique, celui de l'apprentissage de langues, où l'apprenant se trouve confronté entre sa langue maternelle et la langue qu'il veut apprendre. Dans certains cas, ces deux langues peuvent être syntaxiquement très éloignées, et dans ce cas, nous nous sommes demandés quelle type d'analyse pourrait être plus pertinente pour l'apprenant : l'analyse de la langue étrangère à la manière de sa propre langue maternelle, ou l'analyse de la langue étrangère selon la norme de cette langue.

3.1.1 Les différences entre Stanford Parser et TALISMANE

Les jeux d'étiquettes de Stanford Parser (dépendances universelles) et celui de TALISMANE Parser (FTBDep) adoptent deux approches différentes pour l'annotation de plusieurs phénomènes syntaxiques. Ils traitent différemment la coordination, les verbes copulatifs, les groupes prépositionnels, les clitiques.

La différence qui était intéressante pour notre étude était celle entre les analyses des groupes prépositionnels de ces deux parseurs.

Dans l'analyse faite par Stanford Parser toutes les prépositions sont liées à leur tête (le complément de la préposition) par la relation "case", alors que le complément de la préposition porte l'étiquette de la fonction exercée par le groupe prépositionnel dans la phrase. Comme expliqué dans la partie dédiée aux dépendances universelles, cette approche est universelle, puisqu'elle met en tant que gouverneurs des syntagmes les mots lexicaux (*Figure 4.*).

En revanche, avec TALISMANE, on annote la préposition avec la fonction du groupe prépositionnel entier, alors que le complément de la préposition est annoté en tant que "prep", et dépend de cette préposition. (*Figure 13.*)

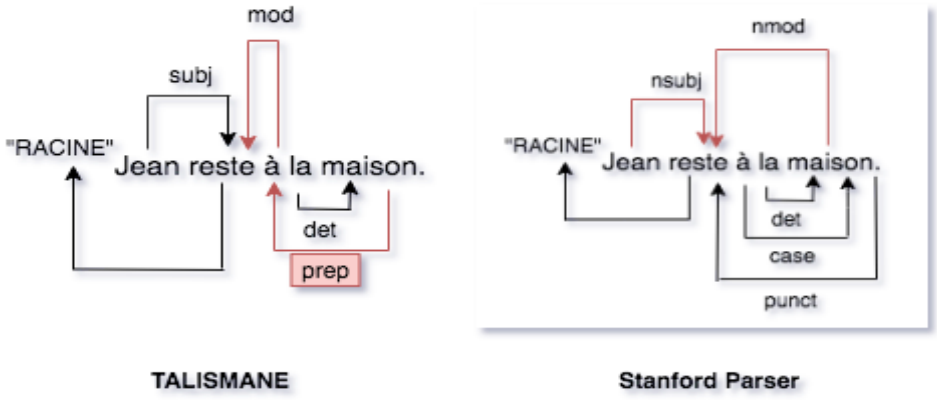

Figure 13. L'analyse des groupes prépositionnels avec TALISMANE et Stanford Parser

Nous avons estimé que cette différence entre les annotations pourrait être intéressante pour les apprenants de FLE locuteurs de langues slaves, étant donné que dans ces langues, les prépositions sont parfois remplacées par les flexions morphologiques - les cas. Pour ces apprenants, le fait d'attacher les prépositions (fait par TALISMANE) au verbe principal pourrait sembler étonnant.

3.2 Les particularités de la syntaxe de la langue serbe par rapport à la syntaxe du français

Nous avons choisi de comparer l'apport didactique des analyseurs syntaxiques avec un groupe d'apprenants de FLE de langue maternelle serbe.

Le serbe est une langue indo-européenne qui appartient à la famille des langues slaves méridionales et du sous-groupe occidental de ces langues.

Du point de vue sociolinguistique, c'est une langue commune aux Serbes, Croates, Bosniaques et Monténégrins, appelée « langue serbo-croate » à l'époque de l'ancienne Yougoslavie et aujourd'hui bosniaque-croate-monténégrin-serbe (BCMS). Cette langue commune compte environ 21 millions de locuteurs dans les Balkans et dans le monde (Klajn, 2015).

Le serbe, comme d'autres langues slaves, dispose d'une morphologie flexionnelle riche : les noms varient selon le genre (masculin, féminin ou neutre), le nombre (singulier, pluriel et paucal) et les cas (nominatif, génitif, datif, accusatif, vocatif, instrumental ou locatif). Les adjectifs ont aussi le degré de comparaison en plus des catégories mentionnées (positif,

comparatif ou superlatif) et de l'aspect (défini ou indéfini)(Miletic et al, 2015). Tous les pronoms se déclinent, et certains distinguent le genre, le nombre et la personne (première, deuxième ou troisième).

Certaines fonctions syntaxiques sont exprimées par les cas, ce qui rend la structure de la phrase très flexible.

L'ordre des constituants canonique est SVO (Sujet Verbe Objet), comme en français, mais les ordres SOV, VOS, VSO, OVS et OSV sont aussi corrects, et très fréquents.

Il est considéré que l'analyse syntaxique automatique des langues avec la structure de la phrase flexible doit être basée sur l'analyse en dépendances, et non pas sur l'analyse en constituants, étant donné que l'analyse en constituants ne dispose pas de mécanismes pour gérer la discontinuité des constituants (Buchholz, Marsi, 2006, Nivre et al., 2007). Cependant, la description syntaxique du serbe repose traditionnellement sur l'analyse en constituants (Stanojčić, Popović, 2011, Ivić, 2005).

Pour cette raison il n'existe pas encore de formalisme pour l'annotation du serbe en syntaxe de dépendances. Cela posait une difficulté pour notre travail, étant donné que les apprenants ne connaissaient pas forcément la notion des dépendances syntaxiques.

Chapitre 4

Corpus

Les textes utilisés pour l'expérimentation sont des textes proposés pour la compréhension écrite aux examens passés du Diplôme approfondi de langue française (DALF) et du Diplôme d'Etudes en Langue Française (DELF), tirés du site www.ciep.fr, pour les niveaux de A1 - C2. Ils ont été intégrés dans le dispositif Déjà Lu, adapté à cet effet. Le corpus compte 18 textes d'une taille moyenne de 300 mots. Les textes sont répartis en 6 groupes selon le niveau de l'apprenant : A1, A2, B1, B2, C1, C2. Chaque groupe contient 3 textes d'un même niveau.

4.1 Traitements

L'intégration des textes dans l'application a nécessite divers traitements. Tous les pré-traitements et post-traitements ont été faits automatiquement avec des programmes écrits en Python.

Pour expliquer le pré-traitement nous prendrons le texte suivant en exemple :

```
Phrase 1. Salut Christophe,

Phrase 2. Nous organisons un barbecue dans 15 jours, le samedi 25.
Phrase 3. Nous vous attendons vers 18h, toi avec toute ta petite famille.
Phrase 4. Envoie-moi un mot pour confirmer que vous venez.

Phrase 5. A samedi, Claude
```

Figure 14. Texte avant le pré-traitement

1. Pré-traitement - normalisation des textes :
Suppression de lignes vides et marquage de paragraphes. L'analyseur syntaxique segmente les tokens en phrases, mais ne sait détecter les paragraphes. Pour pouvoir prendre en compte ces derniers, il importe préalablement d'insérer dans le texte des marques de paragraphes.

```
Phrase 1. Salut Christophe,
$Marquer$
Phrase 2. Nous organisons un barbecue dans 15 jours, le samedi 25.
Phrase 3. Nous vous attendons vers 18h, toi avec toute ta petite famille.
Phrase 4. Envoie-moi un mot pour confirmer que vous venez.
$Marqueur$
Phrase 5. A samedi, Claude
```

Figure 15. Texte après le pré-traitement

2. Parsing :

Chaque texte d'un même niveau a été étiqueté de trois manières différentes :

a) avec TALISMANE Parser

b) avec les seules parties du discours issues de TALISMANE étiqueteur - sans analyse syntaxique

c) avec Stanford Parser

Donc, certains textes proposent une analyse syntaxique, avec TALISMANE Parser ou avec Stanford Parser, et d'autres proposent seulement un étiquetage en parties du discours.

3. Post-traitement - la normalisation automatique de la sortie du parseur qui est au format tabulaire, avec un token par ligne :

- Pour TALISMANE : la rénumérotation absolue des tokens, la suppression de colonnes dupliquées

```
7   Il  il  CLS _   n=s|g=m|p=3 9   suj _   _
8   se  se  CLR _   n=p,s|p=3   9   aff _   _
9   produit produire    V   _   n=s|t=P|p=3 6   root    _   _
10  en  en  P   _   _   9   mod _   _
11  général général NC  _   n=s|g=m 10  prep    _   _
12  sur sur P   _   _   9   mod _   _
13  des des DET _   n=p 14  det _   _
14  informations    information NC  _   n=p|g=f 12  prep    _   _
15  mal mal ADV _   _   16  mod _   _
16  enregistrées    enregistrer VPP _   n=p|g=f|t=K 14  mod _   _
17  .   .   PONCT   _   _   16  ponct   _   _
```

Figure 16. La sortie normalisée de TALISMANE

- Pour Stanford Parser :

a) l'ajout de lemmatisation faite par TALISMANE (Stanford Parser ne propose pas la lemmatisation pour le français),

b) l'unification des étiquettes des parties du discours et des détails morphologiques avec celles TALISMANE,

c)la fusion des articles contractés qui ont été séparés pendant le parsing,

d) l'ajout de colonnes manquantes pour unifier les sorties de deux parseurs.

```
80      Le      le      DET     _       n=s|g=m 81      det     _       _
81      directeur       directeur       NC      _       n=s|g=m 85      nsubj   _       _
82      du      de      P+D     _       n=s|g=m 83      case    _       _
83      festival        festival        NC      _       n=s|g=m 81      nmod    _       _
84      a       avoir   V       _       n=s|t=P|p=3     85      aux     _       _
85      annoncé annoncer        VPP     _       n=s|g=m|t=K     79      ROOT    _       _
86      jeudi   jeudi   NC      _       n=s|g=m 85      nmod    _       _
87      dernier dernier ADJ     _       n=s|g=m 86      amod    _       _
88      la      la      DET     _       n=s|g=f 89      det     _       _
89      liste   liste   NC      _       n=s|g=f 85      dobj    _       _
90      officielle      officiel        ADJ     _       n=s|g=f 89      amod    _       _
91      des     de      P+D     _       n=p     92      case    _       _
92      films   film    NC      _       n=p|g=m 89      nmod    _       _
93      choisis choisir VPP     _       n=p|g=m|t=K     92      acl     _       _
94      pour    pour    P       _       _       96      case    _       _
95      le      le      DET     _       n=s|g=m 96      det     _       _
96      concours        concours        NC      _       g=m     93      nmod    _       _
97      :       :       PONCT   _       _       89      punct   _       _
```

Figure 17. La sortie normalisée de Stanford Parser

4. La création de la page au format html qui contient le texte reconstitué mais avec les identifiants uniques pour chaque mot. Ces identifiants permettent aisément l'annotation dynamique des mots du texte selon les diverses informations présentées dans la structure json décrite ci-après. La gestion de la ponctuation spécifique au français a également été effectuée.

```
<table>
<tr id='p1'>
<td class='noPar'>1</td>
<td class='paragraphe'><span class='phrase' id='s1'><span id='1'>Salut</span> <span id='2'>Christophe</span><span
id='3'>,</span> </span></td></tr>
<tr id='p2'>
<td class='noPar'>2</td>
<td class='paragraphe'><span class='phrase' id='s2'><span id='4'>L'</span><span id='5'>été</span> <span
id='6'>c'</span><span id='7'>est</span> <span id='8'>le</span> <span id='9'>21</span> <span id='10'>juin</span>
<span id='11'>et</span> <span id='12'>les</span> <span id='13'>beaux</span> <span id='14'>jours</span> <span
```

Figure 18. Une partie du code source de la page html du texte reconstitué, chaque mot a un identifiant unique

5. La création de la structure de dépendances

Les sorties normalisées après le parsing ont été transformés en une table de hashage. La clé de chaque token est son numéro unique et les valeurs sont toutes les informations pour ce token issues de l'analyse, comme les caractéristiques morpho-syntaxiques et syntaxiques, la définition des dépendances et le lemme.

```
{"71":["NPP","Claude","n=s","_","67"],"42":["P","avec","_","mod","39","46"],"58":
["PONCT",".","_","ponct","56","59"],"64":["VINF","confirmer","t=W","prep","63","66"],
```

Figure 19. Extrait de la table de hashage contenant les informations sur l'analyse de chaque mot

Dans le prochain chapitre nous allons présenter la méthode que nous avons suivie.

Chapitre 5

Méthode

Nous avons décidé d'évaluer l'apport didactique de deux types d'analyse syntaxique en dépendances avec des apprenants locuteurs d'une langue syntaxiquement éloignée du français : une langue slave, le serbe, dont l'ordre des mots est libre et les cas expriment les fonctions syntaxiques.

Nous avons choisi un groupe de 11 apprenants de FLE (niveaux A1-C2 du CERCL) de langue maternelle serbe.

Pour la première étape de cette étude, chaque apprenant avait pour tâche de lire, via le dispositif Déjà Lu, un jeu de textes adaptés à son niveau selon 3 modalités : textes sans aucune analyse syntaxique, juste annotés en parties du discours ; textes analysés par TALISMANE, et textes analysés par Stanford Parser.

Comme nous l'avons précisé précédemment, deux analyseurs que nous avons comparés ont deux approches différentes à l'analyse syntaxique. L'approche de Stanford est plus universelle, donc elle est comparable à la syntaxe du serbe, alors que l'approche de Talismane correspond à la tradition de l'analyse syntaxique du français.

Pour faciliter l'utilisation par des locuteurs serbes, l'interface a été entièrement localisée en serbe, ainsi que l'ensemble des étiquettes POS. Les noms de relations syntaxiques n'ont pas été présentées dans l'interface pour les raisons de simplicité de l'usage, seulement leur représentation graphique. Le jeu d'étiquettes et leurs traductions se trouvent dans l'annexe.

Nous avons vu les représentations graphiques des dépendances avec les flèches et sous la forme d'un arbre dans l'état de l'art. Déjà lu représente les dépendances d'une manière différente, en les colorant en bleu, avec les gouverneurs (têtes) en rouge.

Les fonctionnalités suivantes ont été proposées aux apprenants durant la lecture des textes :

a) Pour les textes qui ont été analysés syntaxiquement, soit avec Stanford Parser ou TALISMANE :

1. le verbe et ses actants (retrouver le verbe, son sujet, objet, et les compléments circonstanciels)

2. les dépendants d'un mot (retrouver les dépendants du mot choisi)

3. la tête d'un mot (retrouver le mot dont dépend le mot choisi)

4. La morphologie volante (afficher la partie du discours du mot survolé, son lemme, les détails morphologiques)

b) Pour les textes qui ont été analysés morphologiquement, donc sans analyse syntaxique effectuée :

1. La morphologie seule (afficher la partie de discours du mot survolé, son lemme, les détails morphologiques)

La deuxième étape comprend l'évaluation. Après avoir fini la lecture, chaque apprenant a répondu à un questionnaire portant sur sa compréhension des textes et ses préférences vis-à-vis des trois types d'annotation proposés.
Les résultats nous ont permis d'évaluer de façon empirique l'utilité d'un analyseur syntaxique intégré à notre dispositif d'aide à la lecture et le type d'analyse le mieux adapté pour des apprenants slaves de FLE.

Troisième partie

Évaluation & Résultats

Chapitre 6

Évaluation

Les questionnaires diffusés aux apprenants contenaient deux types de questions.

Le premier groupe de questions portait sur la compréhension des textes lus.

Ces questions étaient les questions des examens DALF et DELF dont les textes eux-mêmes ont été extraits.

Le deuxième groupe de questions portait sur l'utilisation des fonctionnalités d'aide pendant la lecture (quelle fonctionnalité était la plus utile, sur quel exemple, etc).

Nous avons conçu 6 questionnaires pour chaque niveau d'apprentissage du français.

Chaque questionnaire était divisé en 4 sous-parties :

- 3 parties étaient dédiées aux 3 textes respectivement ;

- une partie portait sur les impressions générales et sur l'évaluation générale.

Étant donné que nous avons voulu tester la pertinence d'un analyseur syntaxique intégré dans la plateforme Déjà Lu pour l des apprenants de langue maternelle serbe, nous nous sommes basés sur 4 critères lors de la conception du questionnaire.

6.1 Critères pour l'évaluation

1. L'avis de l'apprenant

Chaque apprenant a lu 3 textes, chacun étiqueté différemment (partie Problématique, Corpus & Méthode).

Ensuite, dans le questionnaire, il a répondu aux questions suivantes permettant d'évaluer ce critère (ce type de questions était formulé en serbe - nous allons citer leurs traductions) :

a) «Quel texte était le plus facile à lire à votre avis ? »

b) «Est-ce que vous l'avez trouvé le plus facile à lire grâce aux aides (fonctionnalités) disponibles ? »

c) «Si oui, quelles sont ces fonctionnalités et quels sont les exemples ou elles étaient utiles ? »

Ces questions nous ont permis d'évaluer le sentiment subjectif de l'apprenant vis-à-vis des types d'analyse et des aides disponibles. Il fallait également prendre en compte les réponses aux deux dernières questions - si ce sont les fonctionnalités qui ont facilité la lecture et ont influencé ce choix et quelles sont ces fonctionnalités - pour savoir s'il s'agit des fonctionnalités syntaxiques ou morphologiques. Si la réponse était oui et si c'étaient les fonctionnalités syntaxiques qui ont été les plus utiles - nous en avons déduit que c'est dû à l'analyse syntaxique. Nous avons ensuite noté l'analyseur qui se cache derrière le texte choisi.

2. La compréhension du texte

Le critère précédent était un critère subjectif. Il est possible que l'utilisateur pense qu'un texte était facile à lire, mais qu'il s'est trompé dans la réalité. C'est pourquoi il était nécessaire d'évaluer la compréhension des textes, ce qui était un critère objectif. Pour ce faire, nous avons posé à chaque apprenant quelques questions sur la compréhension de chaque de 3 textes que l'utilisateur avait lu précédemment (*Figure 19.*). Les questions étaient de types différents : ouvertes, fermées, à choix multiples... Leur nombre était de 5 environ, mais cela variait selon les niveaux. Ensuite nous avons choisi le texte pour lequel l'apprenant a eu le meilleur résultat sur le test de compréhension. Nous avons ensuite noté l'étiqueteur qui se cachait derrière ce texte.

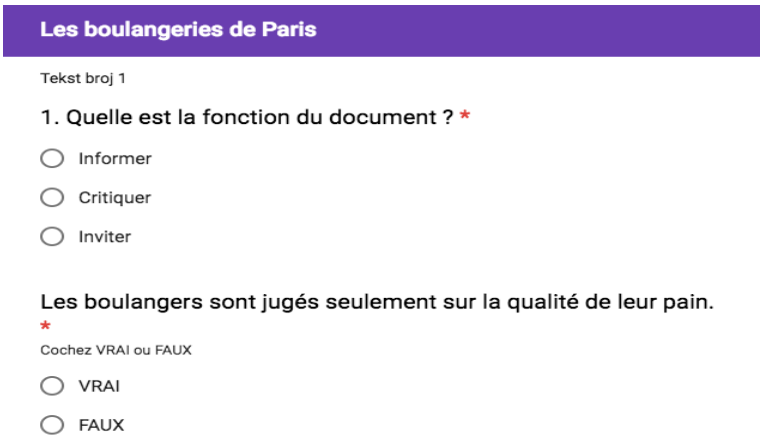

Figure 20. Un extrait du questionnaire portant sur la compréhension du texte pour le niveau A2

3. Fonctionnalités

Nous voulions également évaluer les fonctionnalités les plus utiles pour les apprenants. Nous avons conçu des questions portant sur les fonctionnalités qui les ont aidés pendant la lecture de chaque texte : ils pouvaient en choisir aucune, une ou plusieurs. Il ont ensuite décrit pourquoi elles étaient utiles ou pourquoi elles n'étaient pas utiles, avec des exemples.

4. Critère de pertinence

Finalement, en combinant les avis des apprenants (critère 1), l'évaluation de leur compréhension (critère 2), et leurs réponses quant aux fonctionnalités (critère 3) nous avons estimé la pertinence, du moins pour les apprenants serbes, d'un analyseur syntaxique intégré dans le dispostif d'aide à la lecture.

Chapitre 7

Résultats

Les résultats obtenus nous ont permis d'estimer l'utilité de l'intégration d'un analyseur syntaxique dans la plateforme, et le type d'analyse qui convient le mieux aux apprenants slaves de FLE, ainsi que les fonctionnalités syntaxiques qui aident le plus dans la compréhension des textes. Nous avons établi des critères pour juger l'apport de chaque analyseur.

1. L'avis des apprenants

Les apprenants ont indiqué le texte le plus facile à lire parmi ceux disponibles pour leur niveau. Ils ont ensuite précisé si leur choix a été influencé par l'utilisation des fonctionnalités d'aide à la lecture, et si oui, par quelles fonctionnalités concrètement. Parmi les textes choisis comme les plus faciles à lire avec ou sans analyseur par les apprenants, figurent 4 textes étiquetés avec Stanford Parser, 2 textes sans analyse syntaxique et 1 texte étiqueté avec TALISMANE, ce que nous pouvons voir dans la *Table 1 (S.A.S. = Sans Analyse Syntaxique)*.

Textes choisis						
Niveau	A1	A2	B1	B2	C1	C2
Parseur	Stanford	Stanford & S.A.S.	S.A.S.	TALISMANE	Stanford	Stanford
% Apprenants	100 %	50 % & 50 %	100 %	100 %	100 %	100 %

Table 1. Textes choisis par les utilisateurs comme étant les plus faciles à lire groupés par niveau

Au niveau global, 54,5 % des apprenants ont choisi les textes étiquetés avec Stanford Parser comme les plus faciles, 18,2 % ceux étiquetés par TALISMANE, et 27,3 % les textes sans analyse syntaxique.

	Stanford	TALISMANE	Sans analyse
Apprenants	54,5%	18,2%	27,3%

Table 2. Pourcentage total d'apprenants qui ont choisi les textes étiquetés avec Stanford, Talismane et sans analyse

Il faut également prendre en compte que 72,7 % des apprenants ont déclaré que c'est grâce à l'utilisation des fonctionnalités d'aides syntaxiques que le texte choisi était le plus facile à lire. 27,3% des apprenants ont déclaré que l'utilisation d'aides n'a pas influencé leur choix, et que c'étaient d'autres facteurs (ce sont en même temps les utilisateurs qui ont choisi le texte sans analyse).

2. La compréhension du texte

Chaque apprenant a lu trois textes. Nous avons sélectionné le texte avec les meilleurs résultats sur le test de compréhension pour chaque apprenant (le plus grand nombre de réponses correctes de l'apprenant par rapport au nombre de questions). Dans le cas d'égalité, nous avons noté tous les textes avec les mêmes résultats. Parmi les textes que les apprenants ont le mieux compris, d'après leur score sur les tests de compréhension, figurent :

5 textes étiquetés avec Stanford Parser
3 textes étiquetés avec TALISMANE
4 textes sans analyse syntaxique

Les résultats ressemblent à ceux dans la Table 1., mais il diffèrent un peu.

Nous pouvons voir que les apprenants des niveaux A2,B1,C1 et C2 ont bien compris les textes étiquetés soit avec Stanford Parser, soit sans analyse syntaxique - ce qui coïncide partiellement avec leur choix du critère 1. Ceux du niveau A1 ont bien fait les tests portants sur les textes étiquetés avec Stanford Parser et Talismane, ce qui correspond partiellement à leurs choix. Finalement, ceux du niveau B2 ont le mieux fait le texte étiqueté avec TALISMANE, ce qui coïncide complètement avec leur choix du critère 1(*Table 3.*).

La troisième ligne de la table représente le pourcentage d'apprenants dont le meilleur score aux tests de compréhension était pour le texte étiqueté avec le parseur donné. Il est également possible qu'un apprenant ait les mêmes résultats maximales pour plusieurs textes.

	Stanford	TALISMANE	Sans analyse
Nb. de textes avec le meilleur score	5	3	4
Niveaux	A1,A2,B1,C1,C2	A1,A2,B2	A2, B1,C1,C2
% Apprenants	63,6 %	54,5%	36,3 %

Table 3. Le nombre de textes avec les meilleurs résultats de compréhension

3. Critère de fonctionnalités :

Nous avons voulu découvrir quelles fonctionnalités syntaxiques, parmi celles qui ont été

proposées, étaient les plus utiles pour les apprenants, et pour quel groupe d'apprenants, présenté dans la *Table 4. (1 = utile ; 0 = pas utile ; PDD = partie du discours).*

Apprenant	Verbe et ses actants	Dép. d'un mot	Tête d'un mot	PDD	Niveau
J.S.	1	1	0	1	A1
A.A.	1	1	0	1	A1
J.A.	1	1	1	1	A1
D.A.	1	1	1	0	A2
A.F.	1	1	1	1	A2
L.S.	0	0	0	1	B1
M.L.	0	1	0	1	B1
A.S.	1	0	0	1	B2
P.J.	0	1	0	1	B2
R.S.	0	1	0	1	C1
U.M.	1	0	0	1	C2
Moyenne	63,6 %	72,7%	27,3 %	90,9 %	

Table 4. Utilité des fonctionnalités syntaxiques par apprenant

Parmi les fonctionnalités syntaxiques les plus utiles selon notre échantillon d'apprenants, étaient les "dépendants d'un mot", puis la fonctionnalité "verbe et ses actants" (*Table 4.*).

En tenant compte de la colonne "Niveau", nous pouvons remarquer que, plus le niveau est élevé, l'utilisation des fonctionnalités syntaxiques diminue, et voir clairement la différence entre deux moitiés de la table. Le pourcentage d'apprenants trouvant les fonctionnalités utiles dans les niveaux de B1 - C2 est 33,3 % pour "Verbe et ses actants", 50 % pour les "Dépendants d'un mot", et 0% pour la "Tête d'un mot".

La fonctionnalité "Morphologie volante", ou PDD, était la plus utile parmi les élèves, mais il s'agit d'une fonction morphologique, donc nous ne la prenons pas en compte pour nos résultats.

4. Critère de pertinence :
Pour chaque niveau d'apprentissage, au moins 1 texte avec le score maximal a été étiqueté avec un analyseur syntaxique, ce qui pourrait être une piste significative, dans le sens où nous pouvons voir que l'analyse syntaxique automatique est utile pour la compréhension

des textes (*Table 3.*). Le nombre significatif d'apprenants avait des bons résultats de compréhension de textes étiquetés avec les deux analyseurs syntaxiques. (*Table 3.*). Un grand nombre d'élèves a également trouvé les fonctions syntaxiques utiles, ce qui appuie cette théorie (*Table 4.*). Cependant, ce sont surtout les apprenants avec des niveaux bas des connaissances qui ont utilisé les aides syntaxiques. Les apprenants des niveaux de B1-C2 ont utilisé beaucoup moins d'aides. Cela peut être dû, soit à leurs compétences développés, soit au fait qu'un analyseur syntaxique est plus susceptible de se tromper en analysant les textes plus compliqués.

Chapitre 8

Discussion

Comme nous l'avons déjà précisé, cette étude représente une version préliminaire de l'expérimentation conçue à partir de la réflexion théorique sur la méthodologie de l'évaluation des analyseurs syntaxiques automatiques, et des fonctionnalités de la plateforme Déjà Lu. Il est par conséquent probable que la démarche scientifique a besoin d'être perfectionnée. Pour la perfectionner, il faudrait reprendre la méthodologie en questionnant un plus grand nombre d'apprenants, selon un protocole de test plus rigoureux (groupe test/groupe de contrôle).

Cependant, les résultats obtenus après l'expérimentation effectuée sur notre échantillon réduit d'apprenants du FLE nous ont donné quelques informations importantes.

Premièrement, le choix des apprenants et les meilleurs scores sur les tests coïncident - les deux montrent que l'analyseur avec les meilleurs résultats est Stanford Parser. Cela montre que les apprenants serbes ont désigné comme les plus faciles, et qu'ils ont le mieux compris, les textes étiquetés par l'analyseur dont l'analyse correspond plus à la tradition de leur langue maternelle. Il s'agit d'une piste qui pourrait être développée ultérieurement, avec des tests plus précis qui permettraient d'évaluer la compréhension de syntagmes spécifiques, annotées de deux manières, et non pas du texte en son intégralité.

Deuxièmement, parmi les textes que les apprenants ont le mieux compris, selon le score sur les tests de compréhension, figure, pour chaque niveau d'apprentissage, au moins 1 texte proposant une analyse syntaxique, ce qui montre que l'analyse syntaxique pourrait être pertinente pour les apprenants de FLE dans le cadre d'un outil d'aide à la lecture.

Finalement, parmi les fonctionnalités syntaxiques désignées comme les plus utiles par les apprenants se trouvent "Les dépendants d'un mot", et "Verbe et ses actants". Pourtant, les fonctionnalités syntaxiques étaient surtout pertinentes pour les apprenants des niveaux plus bas de FLE. Ce pourrait être une information utile pour mettre en place les améliorations des fonctionnalités syntaxiques des outils d'aide à la lecture, et aussi une piste de réflexion sur l'adaptation des fonctionnalités au niveau et à la langue maternelle de l'apprenant.

Les pistes fournies par les résultats de cette étude peuvent servir de base pour une étude plus grande, avec un groupe d'élèves plus important et les questions plus précisément conçues.

Conclusion & Perspectives

Ce travail représente une première version d'une étude comarative entre deux types d'analyseurs syntaxiques en dépendances du français, et sur la pertinence didactique d'une telle analyse, dans le cadre d'un outil d'aide à la lecture, pour les apprenants de FLE de la langue maternelle slave, et casuelle - le serbe.

Nous avons rassemblé un corpus de 18 textes repartis en 6 niveaux d'apprentissage. Pour chaque niveau, nous avons étiqueté 3 textes, chacun selon les modalités différents : 1 texte avec Stanford Parser, 1 texte avec TALISMANE, et 1 texte en parties de discours, sans analyse syntaxique effectuée. Stanford Parser et TALISMANE traitent les dépendances de deux manières différentes : Stanford Parser d'une manière universelle adaptée notamment à la syntaxe des langues morphologiquement riches, et TALISMANE d'une manière propre à la syntaxe et à la tradition du français. Donc Stanford Parser analyse les textes d'une manière plus proche à la langue maternelle de l'apprenant, et TALISMANE les analyse d'une manière plus proche à la langue cible, la langue qu'il apprend.

Un groupe de 11 apprenants de la langue maternelle serbe a lu ces textes intégrés dans le dispositif Déjà Lu et adaptés au niveau de chacun, en utilisant les fonctionnalités d'aide syntaxiques proposées. Ensuite il a répondu aux questions de la compréhension et des impressions sur les options proposées.

Les résultats nous ont permis d'obtenir les pistes préalables qui correspondent en partie avec nos hypothèses et attentes (l'influence de la langue source (celle de l'apprenant) lors du décodage syntaxique d'une autre langue). Les textes étiquetés avec Stanford Parser était désignés comme étant les plus faciles, et les apprenants avaient les meilleurs résultats sur les tests de compréhension de ces textes.

Un analyseur syntaxique intégré dans la plateforme a été pertinent pour notre groupe d'apprenants.

Certaines options syntaxiques étaient plus utiles que les autres selon les apprenants - dépendants d'un mot et verbe et ses actants.

Enfin, ce sont surtout les apprenants avec des niveaux plus bas qui ont le plus utilisé les fonctionnalités syntaxiques.

Ces résultats nous ont également permis de voir les limites de notre étude, comme le nombre restreint d'apprenants, ou la généralité des questions posées dans les questionnaires.

Bibliographie

Sources primaires

Abeillé A., L. Clément et F. Toussenel, « Building a treebank for French », *in* : *Treebanks, Kluwer, Dordrecht* (2003), p. 165–187.

Candito, Marie et al., « Benchmarking of Statistical Dependency Parsers for French », *in* : *23rd International Conference on Computational Linguistics - COLING* (2010), p. 108–116.

Chanier, Thierry, *Acquisition des Langues Assistée par Ordinateur (ALAO)*, Clermont-Ferrand : Université Blaise Pascal - Clermont-Ferrand II, 1995.

Christophe Cerisara, Claire Gardent, « Analyse syntaxique du français parlé », *in* : *Journée ATALA* (2009).

Couto, Javier, *Une plate-forme informatique de Navigation Textuelle : modélisation, architecture*, Paris : Université Paris-Sorbonne - Paris IV, 2006.

Crispino, Gustavo, *Une plate-forme informatique de l'Exploration Contextuelle : modélisation, architecture et réalisation (ContextO)*, Paris : UNIVERSITE DE PARIS IV – SORBONNE, 2003.

Elisabeth, Breidt et Feldweg Helmut, « Accessing Foreign Languages with COMPASS », *in* : *Machine Translation* 12 (1997), p. 153–174.

Gédéon, Paul, *Fusion d'analyseurs syntaxiques pour la production d'une analyse syntaxique robuste*, Montréal : Université de Montréal - Ecole Polytechnique de Montréal, 2011.

Gendner, Véronique et al., « Les annotations syntaxique de référence peas », *in* : *Technical report, Projet ANR Passage,* 1.11. (2008).

Green, Spence et al., « Multiword Expression Identification with Tree Substitution Grammars : A Parsing tour de force with French », *in* : *EMNLP* (2011).

Harrison, Michael, *Introduction to formal language theory*, Reading : Addison-Wesley, 1978.

Honnibal, Matthew et Johnson Mark, « An Improved Non-monotonic Transition System for Dependency Parsing », *in* : *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (2015), p. 1373–1378.

Lundquist Lita Minel Jean-Luc, Couto Javier, « NaviLire, Teaching French by Navigating in Text », *in* : *Conference : The 11th International Conference. IMPU, Information Processing and Management of Uncertainty in Knowledge-based Systems* 11 (2006).

M Candito, D Seddah, « Le corpus Sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical », *in* : *19e conférence sur le Traitement Automatique des Langues Naturelles* (2012).

McDonald, Ryan et al., « Universal dependency annotation for multilingual parsing », *in* : *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* 2 (2013), p. 92–97.

Miletic Aleksandra Fabre Cécile, Stosic Dejan, « Construction du jeu d'étiquettes pour le parsing du serbe », *in* : *22e journées du Traitement Automatique des Langues Naturelles, Caen* (2015), p. 2–4.

Mkhitaryan, Satenik, *La détection des prédicats complexes hindi dans le cadre d'un outil d'aide à la lecture*, Paris : Institut des langues et civilisations orientales, 2014.

Nerbonne, John et al., « Reading more into foreign languages », *in* : *Proceedings of the Fifth Conference on Applied Natural Language Processing. Association for Computational Linguistics* (1997), p. 135–138.

Nivre, Joakim et Johan Hall, « Maltparser : A language-independent system for data-driven dependency parsing », *in* : *Proc. of the Fourth Workshop on Treebanks and Linguistic Theories* (2005), p. 13–95.

Pierre, Desmet, « L'enseignement/apprentissage des langues à l'ère du numérique : tendances récentes et défis », *in* : *Revue française de linguistique appliquée* 11(1) (2015), p. 119–138.

Piper Predrag, Ivić Milka, *Sintaksa savremenoga srpskog jezika : prosta rečenica*, Belgrade : Beogradska knjiga, 2005.

Smith Noah, Johnson Mark, « Weighted and probabilistic context-free grammars are equally expressive », *in* : *Computational Linguistics* 33(4) (2007).

Stanojčić Živojin, Popović Ljubomir, *Gramatika srpskog jezika za gimnazije i srednje škole*, Belgrade : Zavod za udžbenike, 2016.

Tesnière, Lucien, *Eléments de syntaxe structurale*, Paris : Klincksieck, 1959.

Thomasset, F. et E. de La Clergerie, « Comment obtenir plus des métagrammaires », *in* : *Actes de la Conf´erence sur le Traitement Automatique des Langues Naturelles* (2005).

Urieli, Assaf, *Robust French syntax analysis : reconciling statistical methods and linguistic knowledge in the Talismane toolkit*, Toulouse : Université Toulouse le Mirail - Toulouse II, 2013.

Glossaire

BCMS bosniaque-croate-monténégrin-serbe. 29

clitique En linguistique, une copule est un mot dont la fonction est de lier l'attribut au sujet d'une proposition - Wikipedia plural. 20, 22

copule En linguistique, une copule est un mot dont la fonction est de lier l'attribut au sujet d'une proposition - Wikipedia plural. 21

DALF Diplôme approfondi de langue française. 31

datif Cas, utilisé en latin, en grec, etc., exprimant généralement la fonction de complément d'objet secondaire (ou complément d'attribution). - Larousse. 20, 29

DELF Diplôme d'Etudes en Langue Française. 31

FLE Français Langue Etrangère. 1, 2, 29

FRMG FRench MetaGrammar. 22

génitif Cas des langues à déclinaison exprimant le plus souvent un rapport de subordination entre deux noms (possessif, partitif, etc.). - Larousse. 29

lemmatisation La lemmatisation (en anglais : lemmatization) est l'analyse lexicale d'un texte dans le but de regrouper les mots d'une même famille. Les mots d'une même famille détectés dans un texte sont donc réduits en une unique entité que l'on appelle un « lemme » ou la « forme canonique d'un mot ». Par exemple, tous ces mots ont le même lemme "Définir" : Définir, définition, définitions, définissons - Yakaferci. 22, 23, 32

paucal représente un petit nombre d'unités en breton pour quelques noms ou encore en russe et d'autres langues slaves pour les nombres dont l'unité est inférieure à 5 dans les rares noms où il n'a pas fusionné avec le génitif singulier - Wikipedia. 29

POS Part Of Speech - Partie du Discours. 9

Stanford Parser analyseur de dépendances récent basé sur des transitions et intégrant un réseau de neurones - Wikipedia. i, 2, 3, 19, 22, 28, 32, 35

syntagme En linguistique structurale, groupe d'éléments formant une unité dans une organisation hiérarchisée. (On distingue le syntagme nominal, le syntagme verbal, le syntagme prépositionnel, le syntagme adjectival.) - Larousse. 16, 17, 20, 28, 53

TAL Traitement Automatique des Langues. 1, 9

TALISMANE L'outil Talismane est un analyseur syntaxique développé par Assaf Urieli dans le cadre de sa thèse au sein du laboratoire CLLE-ERSS, sous la direction de Ludovic Tanguy - http://redac.univ-tlse2.fr/applications/talismane.html. i, ii, 2, 3, 22, 27–29, 32, 35, 43, 44, 49

tokenizer l'outil informatique qui transforme un texte en plusieurs mots séparés par des espaces - Wiktionary. 1

Treebank Corpus de référence annoté syntaxiquement et sémantiquement. 2, 19, 22, 23

UD Universal Dependencies. 19

Annexes

8.1 Annexe 1 - Questionnaire A1

Upitnik - Nivo A1

olimo odgovorite na pitanja vezana za razumevanje pročitanih teksova i za dostupne funkcije.
Hvala!

*Обавезно

Lettre à Cristophe

Tekst broj 1.

1. **A qui écrit Claude ?** *

2. **Il lui écrit pour :** *
   Означите само један овал.
   - ◯ son anniversaire
   - ◯ lui féliciter de son travail
   - ◯ l'inviter à une soirée

3. **Claude propose de retrouver ses amis le :** *
   Означите само један овал.
   - ◯ 15 juin
   - ◯ 21 juin
   - ◯ 25 juin

4. **Où a lieu la fête ?**
   Означите само један овал.
   - ◯ À la montagne
   - ◯ À la campagne
   - ◯ À la mer

5. **Que doivent apporter ses amis?** *
   Означите само један овал.
   - ◯ un dessert
   - ◯ les boissons
   - ◯ les jeux

6. **Da li su Vam neki delovi teksta bili teški za razumevanje? Koji?** *

7. **Da li ste naišli na nepoznate reči u tekstu? Koje su to reči?** *

8. **Da li su Vam opcije pomogle da razumete nepoznate delove teksta i reči? Ako jesu, koje su to opcije ?** *
Изаберите све што важи.

- [ ] Analiza rečenice
- [ ] Zavisne reči
- [ ] Glavna reč u sintagmi
- [ ] Nijedna od ponuđenih
- [ ] Vrste reči

9. **Ako Vam nijedna opcija nije pomogla, molimo objasnite zašto:** *

10. **Ako su Vam neke opcije pomogle, na kojim konkretnim primerima su Vam bile korisne ?** *

Que faire à Paris ?

Tekst broj 2

11. **L'après-midi, s'il y a du soleil, dans l'article on vous conseille...** *
Означите само један овал.

- ◯ de visiter un musée.
- ◯ d'aller dans un parc.
- ◯ de faire un tour de bateau sur la Seine.

12. **Quel musée on vous conseille de visiter s'il pleut ?** *

13. **Le soir , s'il ne fait pas beau , on vous propose... ***

Означите само један овал.

- ◯ d'aller au théâtre
- ◯ de faire un tour de bateau sur la Seine
- ◯ d'aller au restaurant

14. **Le restaurant proposé dans l'article est à côté ... ***

Означите само један овал.

- ◯ de la Tour Eiffel.
- ◯ des Champs-Élysées.
- ◯ du musée Carnavalet.

15. **Où pouvez-vous réserver ? ***

_____

16. **Da li su Vam neki delovi teksta bili teški za razumevanje? Koji? ***

_____

_____

_____

_____

_____

17. **Da li ste naišli na nepoznate reči u tekstu? Koje su to reči? ***

_____

_____

_____

_____

_____

18. **Da li su Vam opcije pomogle da razumete nepoznate delove teksta i reči? Ako jesu, koje su to opcije ? ***

Изаберите све што важи.

- ☐ Vrste reči
- ☐ Nijedna opcija

19. **Ako Vam nijedna opcija nije pomogla, molimo objasnite zašto: ***

_____

_____

_____

_____

_____

20. **Ako su Vam neke opcije pomogle, na kojim konkretnim primerima su Vam bile korisne ?** *

Annonce

Tekst broj 3

21. **Cette annonce concerne quel travail ?** *

22. **Quel jour devez-vous travailler ?** *

23. **Il faut être disponible...** *

Означите само један овал.

- le matin
- le midi
- le soir

24. **Combien est-on payé pour une soirée ?** *

25. **Si vous êtes intéressé par l'annonce, que devez-vous faire ?** *

26. **Da li su Vam neki delovi teksta bili teški za razumevanje? Koji?** *

27. **Da li ste naišli na nepoznate reči u tekstu? Koje su to reči?** *

28. **Da li su Vam opcije pomogle da razumete nepoznate delove teksta i reči? Ako jesu, koje su to opcije ?** *

Изаберите све што важи.

- [ ] Analiza rečenice
- [ ] Zavisne reči
- [ ] Glavna reč u sintagmi
- [ ] Vrste reči

29. **Ako Vam nijedna opcija nije pomogla, molimo objasnite zašto:** *

30. **Ako su Vam neke opcije pomogle, na kojim konkretnim primerima su Vam bile korisne ?** *

Opšti utisci

31. **Koji od ova 3 teksta Vam je bio najlakši za čitanje ?** *

Означите само један овал.

- ◯ Lettre à Cristophe
- ◯ Que faire à Paris ?
- ◯ Annonce

32. **Da li su na to uticale dostupne opcije ? Ako da, koje su to opcije ?** *

Hvala!

8.2 Annexe 2 - Quelques textes utilisés dans l'expérience

Texte A1

Bonjour,
Je recherche une personne pour garder mes enfants de 1 et 7 ans. Il faut être disponible pour travailler les jeudis, vendredis et samedis soirs après 17 heures.
Vous devez habiter dans le centre de Limoges ou avoir une voiture.
Tarifs : 45 € pour une soirée.
Expérience avec les enfants souhaitée.
Si vous êtes intéressé, appelez-moi au 06 38 46 27 11.
Anna Lemaître

Texte A2

Madame, Monsieur,
C'est avec grand plaisir que je vous envoie ce message pour vous décrire votre voyage à Paris en notre compagnie.
À votre arrivée, le 7 juin prochain, nous allons vous accueillir à l'aéroport pour vous conduire à votre hôtel.
Puis, du 8 au 17 juin, vous allez visiter la capitale de notre pays accompagnés de notre guide, Hervé, qui va vous faire découvrir les musées, les rues, les boutiques mais aussi les restaurants de Paris.
N'oubliez pas votre appareil photo !
Les températures en ce moment à Paris sont comprises entre 20 et 24 degrés. Un climat délicieux.
À très bientôt.
L'équipe de voyagemalin.com

Texte B1

« Il est difficile d'imaginer un accent trop fort pour présenter un journal national. »
M. Aphatie a été le premier à animer une émission dite « sérieuse » à la radio puis à la télévision. L'entrée du journaliste basque dans le monde audiovisuel n'avait donc rien d'évident. « J'ai longtemps travaillé en presse écrite », indique-t-il. « C'est en représentant mon journal, Le Parisien, à l'émission Res Publica, à la radio, que j'ai rencontré Jean-Luc Hess, qui dirigeait alors la station. En 1999, il m'a proposé de devenir chef du service politique. »
Curieusement, l'actuel président du groupe Radio-France ne se souvient pas du débat après l'arrivée de M. Aphatie :
« Son arrivée n'a pas été critiquée, car il était évident que Jean-Michel Aphatie avait beaucoup de présence à la radio. Son accent est si naturel que cela n'a posé aucun problème. »

Texte B2

(...) Et, en France, la classe politique pourrait bien ressembler, tout du moins en nombre d'années, à un « pouvoir gris».
Plus de la moitié des sénateurs ont plus de 60 ans, l'âge moyen des maires de communes de plus de 3 500 habitants est de 54 ans. Idem pour ceux qui les élisent. Alors qu'ils représentent 20 % de la population, les retraités forment déjà 30 % des électeurs. Bref, la démocratie a les tempes blanchies. Or, quand les papy-boomers seront à la retraite, ils auront encore plus de temps. Ne vont-ils pas accaparer les affaires publiques à leur avantage ? Faut-il redouter une gérontocratie à la française ? « Fantasme ! » Jean-Philippe Viriot-Durandal, maître de conférences en sociologie à l'université de Franche-Comté, en est convaincu. Vieux ne veut pas forcément dire réac. Il en veut pour preuve deux exemples, décortiqués dans son livre. D'abord, la dernière élection présidentielle, où le vote pour le Front national fut autant le fait des jeunes électeurs que de leurs aînés. La preuve aussi avec l'adoption de l'euro. « Six mois après l'introduction définitive de la monnaie unique européenne, les nostalgiques du franc étaient à peine plus nombreux chez les plus de 65 ans que dans l'ensemble de la population - 53 %, contre 48 % », écrit le sociologue. Qui ajoute que « l'opposition entre les jeunes générations, ouvertes sur le monde et sur l'avenir, et les anciennes, rétives au changement et réfractaires à la nouveauté », ne serait que caricature.

Texte C1

Il est évident que des écoles si différentes ont dû employer des procédés de composition absolument opposés. Le romancier qui transforme la vérité constante, brutale et déplaisante, pour en tirer une aventure exceptionnelle et séduisante, doit, sans souci exagéré de la vraisemblance, manipuler les événements à son gré; les préparer et les arranger pour plaire au lecteur, l'émouvoir ou l'attendrir. Le plan de son roman n'est qu'une série de combinaisons ingénieuses conduisant avec adresse au dénouement. Les incidents sont disposés et gradués vers le point culminant et l'effet de la fin, qui est un événement capital et décisif, satisfaisant toutes les curiosités éveillées au début, mettant une barrière à l'intérêt, et terminant si complètement l'histoire racontée qu'on ne désire plus savoir ce que deviendront, le lendemain, les personnages les plus attachants.
Le romancier, au contraire, qui prétend nous donner une image exacte de la vie, doit éviter avec soin tout enchaînement d'événements qui paraîtrait exceptionnel. Son but n'est point de nous raconter une histoire, de nous amuser ou de nous attendrir, mais de nous forcer à penser, à comprendre le sens profond et caché des événements. A force d'avoir vu et médité il regarde l'univers, les choses, les faits et les hommes d'une certaine façon qui lui est propre et qui résulte de l'ensemble de ses observations réfléchies. C'est cette vision personnelle du monde qu'il cherche à nous communiquer en la reproduisant dans un livre. Pour nous émouvoir, comme il l'a été lui-même par le spectacle de la vie, il doit la reproduire devant nos yeux avec une scrupuleuse ressemblance.
Il devra donc composer son oeuvre d'une manière si adroite, si dissimulée, et d'apparence si simple, qu'il soit impossible d'en apercevoir et d'en indiquer le plan, de découvrir ses intentions.

Texte C2

Que restera-t-il à la rentrée des acquis de l'année scolaire ? Cerveau et seau de plage jouent-ils les vases communicants ? Concentrés toute l'année sur les résultats scolaires de leurs enfants, les parents ne sont pas prêts à laisser les bains de mer ou de soleil délaver les tables de multiplication, les plus-que-parfait de l'indicatif et autres théorèmes de géométrie si chèrement acquis. Pour conjurer les deux mois de vacances, ils achètent donc leur potion en librairie pour un investissement moyen de 7 euros. Cela s'appelle cahier de vacances, et 4,5 millions d'écoliers et d'élèves n'ont d'autre solution que de lui trouver une place dans leur valise. Après, c'est une autre affaire. A tel point que, parmi les parents qui ont investi dans ces produits, 4,4 % déclarent que leur enfant ne l'a jamais ouvert... et 72,2 % qu'il ne l'a utilisé qu'en partie.
Des chiffres tirés d'une des rares enquêtes sur le sujet, réalisée en 2000 (publiée en 2001) par l'Institut de recherche sur l'économie de l'éducation (Irédu) auprès des parents de 2 500 enfants de l'académie de Dijon (Côte-d'Or). Qu'est-ce qui peut bien poser problème dans ces petits livrets pourtant plutôt attractifs pour que seuls 23,4 % de leurs détenteurs arrivent au bout ? Jeune retraité de l'éducation nationale, Roger Rougier a sa réponse. « Je les ai subis, et je les ai fait subir à mes enfants, jusqu'à ce que j'ose m'en affranchir, résume cet inventeur de produits plus ludiques. Je les ai abandonnés le jour où j'ai commencé à faire tenir des cahiers d'été à mes enfants et à créer des jeux avec eux. Je me souviens d'un été durant lequel nous nous étions beaucoup déplacés en caravane. En quittant chaque étape, mes enfants dessinaient le lieu ou une personne qu'ils y avaient rencontrée. A la fin des vacances, nous avons assemblé ces vignettes et créé ensemble un jeu de l'oie. Comme support d'échange, ça a été fantastique. Nous ne pouvions pas faire une partie sans que l'un d'eux ne raconte un épisode ou un personnage », rappelle celui qui, ni vu ni connu, a fait, par ce biais, travailler la narration à ses enfants.

8.3 Annexe 3 - Les jeux d'étiquettes de Stanford Parser et de Talismane

Nous présentons les jeux de tags utilisés dans ce mémoire. En ce qui concerne les dépendances universelles, seules les étiquettes qui avaient apparues dans nos textes sont ici, nous ne les avons pas citées toutes étant donné que les dépendances universelles parfois

couvrent les phénomènes syntaxiques qui ne concernent pas le français. Le jeu d'étiquettes de parties du discours ont été traduites en serbe pour les besoins de ce mémoire, et leurs traductions se trouvent dans la table 7.

Relations pour gouverneurs verbaux (cf. annotation du French Treebank + ajouts)	
suj	Sujet
obj	objet
de_obj	SP argumental en de, non locatif
a_obj	SP argumental en à, non locatif
p_obj	autre SP argumental
ats	Attribut du sujet
ato	Attribut de l'objet
mod	Modifieur
auux_tps	auxiliaires de temps
aux_pass	auxiliaires du passif
aux_caus	verbe causatif (en cas de complexe causatif + inf)
aff	clitiques figés
Relations pour gouverneurs non verbaux	
mod	Modifieurs repérés structuralement (par exemple adjectifs épithètes), autres que les relatives
mod_rel	Relatives adnominales
coord	Relation portée par un coordonnant, avec comme gouverneur le coordonné immédiatement précédent
arg	Utilisé dans le cas de prépositions « liées » : dans *de Charybde en Scylla*, parallèlement au traitement de la coordination, *en Scylla* est dépendant de type *arg* de la première préposition *(de)*
dep_coord	Relation portée par un coordonné (sauf le premier), avec comme gouverneur le coordonnant immédiatement précédent
det	Relation portée par les déterminants
ponct	Relation portée par tout dépendant typographique, sauf pour les virgules jouant le rôle de coordonnant (qui porte la relation ***coord***)
dep	Relation sous-spécifiée, pour les dépendants prépositionnels (pas de gestion de la distinction argument / ajout pour les gouverneurs non verbaux)

Table 5. TALISMANE - Relations pour les gouverneurs verbaux et non-verbaux (Candito et al., 2011b)

Etiquette	Description
acl	proposition relative adjectivale
advcl	proposition relative modifiant l'adverbe
advmod	modifieur adverbial
amod	modifieur adjectival
appos	modifieur appositionnel
aux	verbe auxiliaire
case	cas
cc	conjonction de coordination
ccomp	proposition relative
compound	relation entre les éléments des expressions complexes
conj	relation entre deux éléments coordonnés
cop	copule
det	determination
discourse	élément du discours
expl	mot expletive
mark	marqueur
nmod	modifieur nominal
nsubj	sujet nominal
nummod	modifieur numérique
obj	objet
obl	oblique nominal
parataxis	éléments coordonnés sans une marque de coordination explicite
punct	punctuation
root	racine
vocative	vocative
xcomp	proposition relative

Table 6. Jeu de tags de dépendances universelles

Etiquette	Description	Traduction en serbe
ADJ	adjectif	pridev
ADV	adverbe	prilog
ADVWH	adverbe interrogatif	upitni prilog
CC	conjonction de coordination	sastavni veznik
CLO	clitique objet	lična zamenica u funkciji objekta
CLR	clitique réflexive	povratna zamenica
CLS	clitique sujet	lična zamenica
CS	conjonction de subordination	zavisni veznik
DET	determinant	determinant
DETWH	determinant interrogatif	upitni determinant
ET	mot étranger	strana reč
I	interjection	uzvik
NC	nom commun	zajednička imenica
NPP	nom propre	vlastita imenica
P	préposition	predlog
P+D	préposition + déterminant	sažeti član
P + PRO	préposition + pronom	predlog + zamenica
PONCT	ponctuation	interpunkcija
PRO	pronom	zamenica
PROREL	pronom relatif	relativna zamenica
PROWH	pronom interrogatif	upitna zamenica
V	verbe (indicatif)	glagol (indikativ)
VIMP	impératif	imperativ
VINF	infinitif	infinitiv
VPP	participe passé	particip perfekta
VPR	participe présent	particip prezenta
VS	subjonctif	subžonktiv

Table 7. Jeu de tags des parties du discours de French Treebank, utilisé dans ce mémoire (Crabbé and Candito, 2008)

8.4 Annexe 4 - Extraits des programmes utilisées pour cette étude

```
###Fonctions###

#Creation d'une structure JSON sous forme d'un dictionnaire en Python
def creationStructureJson(fiParse):
        lecture = ouverture(fiParse)
        lecture = lecture.split("\n")
        #initialisation de la structure JSON
        structure = {}

        """Pour chaque ligne dans le fichier parsé et normalisé on crée un dictionnaire contenant le numéro du
token et
        toutes les informations issues du parsing"""
        for token in lecture:
                #S'il ne s'agit pas d'une fin de phrase ou de paragraphe, c'est la phrase courante
                if re.match("^(§?)$", token) is None:
                        listeTok = token.split("\t")
                        #numéro unique du token
                        numToken = listeTok[0]
                        #liste des résultats du parsing
                        listeParse = []
                        listeParse.append(listeTok[3]) #POS
                        listeParse.append(listeTok[2]) #LEMME
                        listeParse.append(listeTok[5]) #Détails du POS
                        listeParse.append(listeTok[7]) #Role syntaxique
                        #verifier avec le prof
                        #Si le role syntaxique est égale à 'root', changement du numéro de tete à 0
                        # if listeTok[6] == "0":
                        #       listeTok[7] = "root"
                        if listeTok[7] == "root":
                                listeTok[6] = "0"

                        listeParse.append(listeTok[6]) #N° de tete
                        structure[numToken] = listeParse

        #Ajout des dépendances
        for token in lecture:
```

Figure 21. Extrait du programme qui crée la structure des dépendances

```
#La rénumérotation absolue
def numerotationAbsolue(fichIn, fichOut):
        #renumerotation absolue des tokens
        lecture=ouverture(fichIn)
        lecture=lecture.split("\n")
        cmpt=0
        offsetPhrase=0
        f = open(fichOut, "w", encoding="utf-8")

        for line in lecture:
                line = line.strip()
                #S'il ne s'agit pas de fin de paragraphe ou de phrase, il s'agit de la phrase courante
                if re.match("^(§?)$", line) is None:     # ^(§?)     $1 == "
                        cmpt+=1
                        listeTok = line.split("\t")
                        listeTok[0] = str(offsetPhrase + int(listeTok[0]))
                        listeTok[6] = str(offsetPhrase + int(listeTok[6]))
                        line = "\t".join(listeTok)
                else:
                        offsetPhrase=cmpt

                f.write(line +"\n")
        f.close()


#Normalisation du fichier parsé
def normalisationParsage(fichierParse):
        #1. Deparsage de signes de paragraphes "§""
        deparse = ouverture(fichierParse)
        deparse = re.sub("\n1\t§.*?\n\n", "§\n", deparse);
        ecriture(deparse, fichierParse)
```

Figure 22. Extrait du programme qui normalise les fichiers issus du parsing