CORELA: Cognition, représentation, langage, vol. HS-21, 2017, pp. 1-20.

Des corpus annotés au coeur d'une plate-forme, pour la formation linguistique des enseignants de FLE en Colombie.

Molina Mejia, Jorge Mauricio.

Corela
Cognition, représentation, langage

HS-21 | 2017
Linguistique de corpus : vues sur la constitution, l'analyse et l'outillage

Des corpus annotés au cœur d'une plate-forme, pour la formation linguistique des enseignants de FLE en Colombie

Jorge Mauricio MOLINA MEJIA

Édition électronique
URL : http://journals.openedition.org/corela/4857
DOI : 10.4000/corela.4857
ISSN : 1638-573X

Éditeur
Cercle linguistique du Centre et de l'Ouest - CerLICO

Référence électronique
Jorge Mauricio MOLINA MEJIA, « Des corpus annotés au cœur d'une plate-forme, pour la formation linguistique des enseignants de FLE en Colombie », *Corela* [En ligne], HS-21 | 2017, mis en ligne le 02 février 2017, consulté le 01 mai 2019. URL : http://journals.openedition.org/corela/4857 ; DOI : 10.4000/corela.4857

Des corpus annotés au cœur d'une plate-forme, pour la formation linguistique des enseignants de FLE en Colombie

Jorge Mauricio MOLINA MEJIA

Introduction

> « La linguistique de corpus a été développé pour extraire d'un corpus les connaissances linguistiques nécessaires à l'enseignement des langues ; un corpus étant un ensemble collecté et ordonné de données langagières réelles. » (Teubert, 2009, p. 1).

1 Selon les critères signalés par Teubert, et dans le cadre d'un projet de plate-forme pour la formation des futurs enseignants de FLE en Colombie (ELiTe-[FLE]²)[1] que nous avons élaboré (Molina Mejia, 2014 ; 2015 ; Molina Mejia & Antoniadis, 2014a ; 2014b ; 2014c ; 2015), nous avons jugé intéressant de commencer à partir de cette prémisse « d'enseignement des connaissances linguistiques à partir de données langagières réelles ». C'est pourquoi, nous avons décidé de constituer un petit corpus d'analyse linguistique fondé sur des données réelles ou authentiques. Or la plate-forme que nous proposons permet de créer des activités spécifiques pour la formation des formateurs, fondées sur l'approche théorique connue sous le nom de linguistique textuelle. Pour ceci nous avons constitué un échantillon d'analyse des phénomènes du langage issus de cette approche théorique en utilisant des textes authentiques tirés des anciens examens du DELF[2] et du DALF[3].

1. Linguistique de corpus et didactique du FLE

2 D'après certains auteurs, la notion de *corpus* ou de *linguistique de corpus* reste assez floue dans le domaine de la linguistique française (Souque, 2014 : 64). Pourtant, l'importance

acquise dans les dernières décennies par celle-ci (Williams, 2006), lui a permis de devenir un paradigme assez important dans les études du langage ; et le travail sur corpus a sû connaître un succès croissant dans le domaine de la linguistique et de la didactique des langues dans les dernières années, comme le montrent Guiliani et Hannachi (2010 : 145).

1.1. Linguistique de corpus et enseignement des langues

3 La notion de corpus électroniques n'est pas nouvelle en didactique des langues. En effet, les premiers travaux dans ce domaine datent des années 1990, et nous voyons des auteurs tels que G. Aston (1995) faire référence à l'utilité d'employer ce type de corpus dans les cours de langue. Aston considère que l'emploi de l'informatique fondée sur des corpus pour l'enseignement permet aux apprenants d'avoir un accès à des collections des textes libres ou à des matériaux très élaborés (Aston, 1995 : 257). Dans le domaine de la didactique des langues la linguistique de corpus peut être définie comme un champ permettant « d'enseigner tous types de langues, pour peu que l'on soit à même de constituer un corpus informatisé de la langue que l'on souhaite soumettre aux apprenants » (Guiliani & Hannachi, 2010 : 145).

4 Au début des années 1990 apparaît un courant que T. Johns a nommé DDL « *Data-Driven Learning* », ou en français ASC (Apprentissage sur corpus) (Boulton & Tyne, 2014 : 6). Le but de l'ASC serait, selon Johns, « d'éliminer l'intermédiaire [autrement dit l'enseignant] dans la mesure du possible, pour donner à l'apprenant un accès direct aux données » (Johns, 1991, cité et traduit par Boulton, 2007 : 38).

5 Un auteur tel que C. Landure (2011) propose la pratique de l'ASC comme une alternative dans l'enseignement et apprentissage des langues étrangères (Molina Mejia, 2015 : 146). Cet auteur signale que :

> « La démocratisation actuelle des nouvelles technologies offre d'autres perspectives pour l'enseignement et l'apprentissage d'une langue seconde et pourrait contribuer à changer en profondeur la façon de penser et de concevoir l'enseignement » (Landure, 2011 : 164).

6 En effet, d'après cet auteur l'idée est de passer d'un enseignement dit traditionnel vers un enseignement dont les enseignants auront un niveau de compromis beaucoup plus élevé avec leur métier, et pour ceci, ils pourraient acquérir un minimum des compétences dans le domaine de l'informatique (Landure, 2011 : 164). Nous considérons, également, qu'en plus de la formation dans le domaine de l'informatique, les enseignants des langues devraient se former à l'utilisation et à la constitution des corpus. Se former, en somme, à l'ASC en tant que nouvelle approche et paradigme de recherche, tel que le proposent Boulton et Tyne (2014 : 6).

1.2. Linguistique de corpus et formation des formateurs de FLE

7 Les corpus conçus spécifiquement pour la formation des formateurs en langues, et tout notamment dans le domaine du français langue étrangère, se font rares de nos jours (Molina Mejia, 2015). Il y a dans le domaine de la formation des apprenants de langue un nombre de plus en plus important de travaux dans le domaine de l'ASC, comme le signalent Boulton et Tyne (2014). Mais, pourquoi ce désintérêt pour le domaine spécifique de la formation des futurs enseignants de FLE en employant des corpus ? Nous considérons qu'il est plus « facile » de constituer des corpus d'apprenants pour un public

plus large, que pour un public beaucoup plus spécifique comme celui des futurs enseignants formés en milieu *exolingue*. En effet, les contraintes de la formation des futurs formateurs en milieu *exolingue* sont beaucoup plus complexes que lorsque des enseignants se forment dans un milieu *endolingue*[4]. Le premier type de formation requiert non seulement la formation dans le domaine de la didactique de la langue cible à enseigner, mais aussi la formation linguistique permettant la maîtrise de cette même langue (Arismendi & Colorado, 2015 ; Molina Mejia, 2015). Tout ceci demande un travail très axé sur la compétence linguistique, à un niveau élevé. C'est pourquoi, un corpus qui est destiné à la formation des futurs enseignants de FLE dont le français n'est pas la langue maternelle ni officielle du pays dans lequel ils suivent leurs études a des caractéristiques très spécifiques : le niveau de langue est très important (niveaux du CECRL[5], les futurs enseignants doivent atteindre le niveau B2 ou C1 du *Cadre*), les aspects grammaticaux doivent aussi êtres compris dans la formation (morphologie, syntaxe, orthographe, etc.).

8 Toutefois, il y a quelques travaux qui s'intéressent à la formation des futurs enseignants de langue, utilisant l'étude des corpus linguistiques et des corpus autres que linguistiques. Il s'agit dans le premier cas du travail sur corpus pour la formation des futurs enseignants de FLE de l'Université de Bretagne Sud, le projet *IntUne* (Giuliani & Hannachi, 2010), et pour les enseignants d'anglais au LANSAD[6] de l'Université Paris Diderot (Kübler, 2014). Dans ce premier type de corpus il y a un but de formation au travers des notions d'ordre linguistique (grammaire de la langue cible, par exemple). Pour le second cas il s'agit de l'analyse des corpus de journaux d'apprentissage (Cadet & Tellier, 2007), et de ceux qui analysent les interactions dans l'enseignement en ligne (Chanier & Ciekanski, 2010[7]). Le second type de corpus part de l'idée de la formation à partir des interactions entre les enseignants et leurs apprenants. Nous n'avons pas trouvé, pour l'instant, de corpus créé dans un but de formation linguistique des futurs enseignants de FLE en milieu *exolingue*.

9 C'est pour cela que notre idée est donc de travailler à l'annotation et à la mise en œuvre d'un tel corpus. Pour cela, nous nous fondons dans une approche d'apprentissage sur corpus, comme celle qui a été exposée dans la partie précédente. La section suivante présente, *grosso modo,* la constitution du corpus pour le système que nous avons conçu.

2. Constitution d'un corpus pour la formation des enseignants de FLE

10 Dès le début de la constitution de notre corpus, nous avons été confronté à la difficulté de trouver des documents authentiques classés par les niveaux du CECRL. La tâche s'est avérée plus compliquée lorsque nous cherchions un corpus non seulement classé par niveaux mais aussi déjà annoté de manière morphosyntaxique. Après une longue recherche nous avons trouvé sur le site du CIEP[8] un petit échantillon d'anciens examens classés par niveaux[9]. Le problème essentiel pour trouver des textes classés par niveaux est que ceux-ci soient libres de droit d'auteur. Il y a certains ouvrages qui classent les documents authentiques par niveaux, comme celui de Barthe et Chovelon (2009), mais ils sont soumis au droit d'auteur. Quant à des corpus incluant ces textes et déjà annotés suivant les caractéristiques signalées précédemment, nous n'avons rien trouvé. Il nous fallait donc trouver les textes et les annoter nous-même.

2.1. Quel corpus ?

11 Pour la création (automatique) des activités, le système informatique que nous avons conçu (*cf.* 3) s'appuie sur les annotations que seront portées par les textes. C'est ainsi que nous avons commencé par constituer un petit échantillon d'une quinzaine des documents, issus des journaux, des magazines, des œuvres littéraires et d'Internet. L'idée étant d'avoir des documents authentiques et variés, nous permettant d'offrir à notre public cible des textes représentatifs de la langue française. Nous avons 5 textes du niveau B1, 5 textes du niveau B2, et 5 textes du niveau C1[10]. Les textes comprennent : des extraits des romans, des articles journalistiques, des textes de vulgarisation scientifique, des extraits des essais, etc.

12 Notre corpus comporte autour de 5000 tokens pour l'ensemble des textes des trois niveaux du CECRL. Lors de l'annotation, les caractéristiques telles que le type de texte, la source (nom du magazine, journal, etc.), l'auteur, etc., ont été indiquées dans l'entête de chaque texte (*cf.* figure 3). En effet, ces métadonnées sont essentielles lors de la conception des séquences pédagogiques, afin de permettre aux enseignants d'avoir toutes les informations concernant les textes.

2.2. Quelles annotations ?

13 Suite au choix des documents, nous les avons annotés selon deux types d'approche : une première fondée sur le TAL (Traitement Automatique des Langues), avec Cordial Analyseur, qui permet ensuite d'en faire une deuxième manuelle au moyen d'un étiquetage en format XML[11] (figure 1, ci-dessous).

14 Dans la partie suivante nous expliquons brièvement les deux procédures d'annotation, nous nous appuyons pour ceci sur des captures d'écran qui montrent d'une part les étiquettes de l'analyseur morphosyntaxique Cordial, et d'autre part les annotations en XML. Nous finissons en montrant la DTD[12] qui possède la grammaire de base de notre corpus. Une fois le corpus annoté et vérifié grâce à la DTD, il sera stocké dans une BD[13] de corpus des textes.

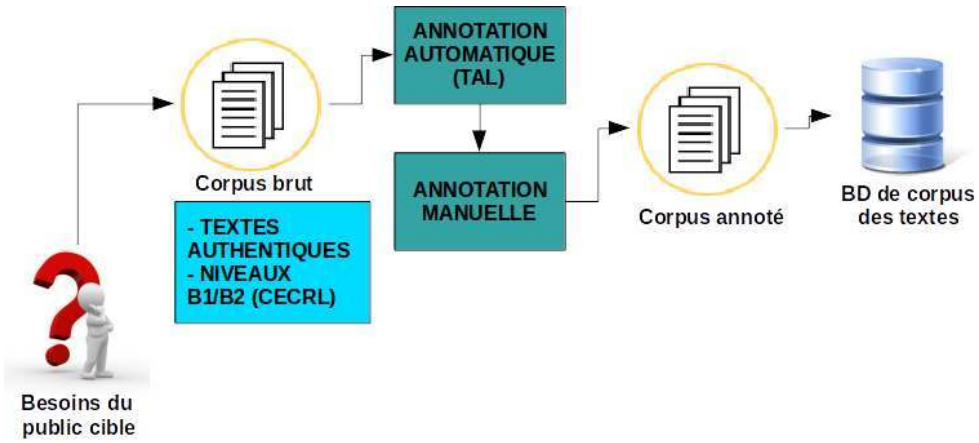

Figure 1: Annotations de corpus des textes (Molina Mejia, 2015 : 265).

2.2.1. Annotations opérées automatiquement à l'aide de l'analyseur Cordial

15 Pour le premier type d'annotations, il s'agit d'associer à chaque unité textuelle des informations morphosyntaxiques (la catégorie grammaticale, le lemme, le genre, le nombre, le temps, etc.). Des vérifications semi-automatiques et manuelles ont permis de corriger les quelques erreurs de Cordial, et d'adapter les informations à nos objectifs. Cette première annotation nous a permis par la suite de mieux cibler les annotations manuelles. Dans l'image ci-dessous (figure 2, Molina Mejia, 2014 : 25), nous donnons un exemple des étiquettes que Cordial produit en sortie :

16 Si nous prenons par exemple la phrase qui est remarquée :

17 - « *C'est un phénomène unique dans l'histoire française, né d'une conjonction particulièrement favorable de la démographie et de l'économie plus rare que l'on sait déjà qu'il ne pourra pas durer* . »

18 Nous pouvons remarquer que chaque élément dispose d'une étiquette particulière selon le contexte dans lequel se trouve l'élément. Nous observons aussi que cette phrase composée (ou complexe), dispose dans la cinquième colonne d'un étiquetage selon les propositions qui se trouvent dans la phrase. Autre cette notion de proposition nous avons également des notions telles que les groupes nominaux, pronominaux et prépositionnels qui nous permettent de mieux cibler les coréférents textuels.

19 Comme nous pouvons apprécier dans ladite figure, à part les informations morphologiques et syntaxiques, Cordial propose aussi ces informations textuelles que nous venons d'analyser (division en phrases et en propositions), et des informations sémantiques pour chaque élément. À partir de ces informations les annotations manuelles se font plus aisément.

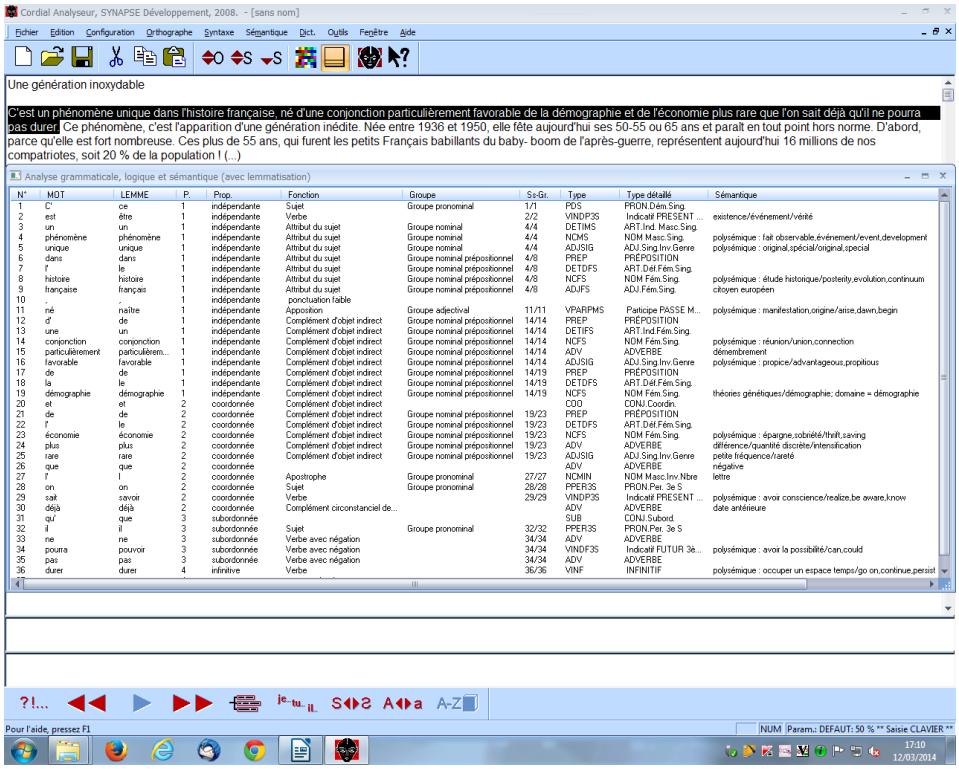

Figure 2: Composantes d'une séquence didactique (Molina Mejia, 2015).

2.2.2. Annotations rajoutées manuellement et balisant des caractéristiques issues de la linguistique textuelle

20 Aux annotations présentées dans la partie précédente nous avons ajouté d'autres annotations en utilisant le format XML. L'idée est de pouvoir annoter tous les phénomènes appartenant à la linguistique textuelle et qui ne peuvent pas être annotés de manière automatique : les coréférents textuels, les connecteurs et marqueurs logico-temporels et discursifs, etc.

21 La particularité du langage XML est justement de permettre de créer ses propres entités, en les annotant comme nous pouvons l'apprécier dans l'image ci-dessous (figure 3, Molina Mejia, 2014 : 25).

22 Pour les éléments nous avons gardé les étiquettes proposées par Cordial, par exemple « DETDMS » (Déterminants, masculins singuliers), et les « NCMS » (Noms communs, masculins, singuliers), etc. Cela nous a permis d'annoter par la suite les « GN » (Groupes Nominaux), lors de l'analyse des coréférents de type nominal. D'autre part, l'annotation des éléments de type « PPER » (Pronom personnel), nous a permis d'annoter plus aisément certains coréférents anaphoriques de type « pronominal ».

23 Aux annotations de Cordial nous ajoutons les annotations de la coréférence textuelle, telles que : « Référent », « Coréférent », « Anaphores », « Cataphores », les « id » (identifiants pour chaque référent ou coréférent, ou pour chaque chaîne de coréférence, etc.).

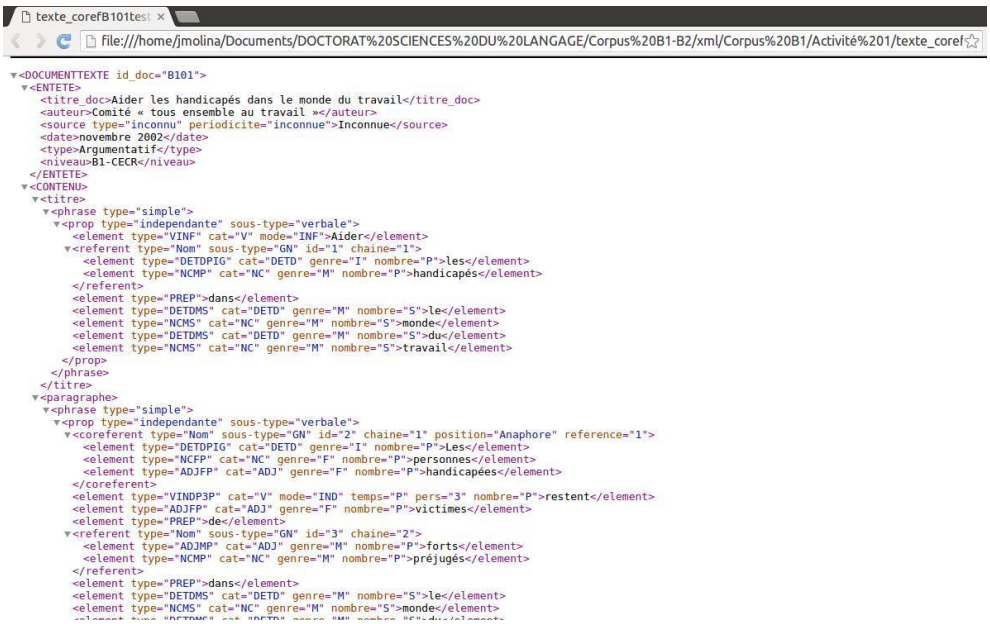

Figure 3: Capture d'écran du système ELiTe-[FLE]².

24 En effet, ces annotations nous permettent d'isoler, comme dans le cas ci-dessous, les référents principaux, les coréférents anaphoriques ou cataphoriques, les différents types d'anaphores (fidèles, infidèles, associatives, etc.). Il nous permet également d'identifier et de signaler les chaînes de coréférence (une ou plusieurs selon le texte annoté).

2.2.3. DTD du corpus

25 La DTD est le document qui normalise toutes les annotations d'un corpus. Elle « définit et contraint les balises et les attributs qu'ils [les documents XML] doivent utiliser pour être valides » (Souque, 2014 : 85). Pour Habert *et al.* (1998 : 64), la DTD précise le nombre de fois et l'ordre dans lequel les éléments balisés peuvent apparaître dans le document normalisé.

26 Dans notre cas, l'annotation globale des textes est décrite et normalisée par la DTD correspondante. En effet, la DTD spécifique que nous avons constituée, et qui nous a permis de vérifier la cohérence de nos annotations, grâce à « une grammaire de base qui contient tous les éléments étiquetés de manière automatique et ensuite balisés manuellement » (Molina Mejia, 2015 : 288). Ce document permet d'établir combien de fois un élément doit apparaître on non dans le texte, l'ordre des éléments et la manière comme ils ont été constitués.

27 La DTD a permis, également, la mise en forme informatique des activités pédagogiques. C'est-à-dire, la DTD a été utilisée afin de permettre l'élaboration du système informatique à partir de son contenu. C'est ainsi que les programmeurs ont pu retrouver les éléments à être employés lors des activités.

28 L'image ci-dessous (figure 4, Molina Mejia, 2015) nous permet d'apprécier les différents éléments qui ont été annotés et la manière globale comme fonctionne le système à partir du corpus.

29 La DTD contient trois types d'informations importantes permettant son exploitation informatique : les éléments (ELEMENT) qui ont été annotés et décrits dans le document XML (*cf.* figure 3), les attributs de chaque élément (ATTLIST), et les valeurs de ces attributs qui sont insérés à l'intérieur des « ATTLIST ».

30 Exemple : Pour l'élément « paragraphe » nous voyons qu'il est constitué des « phrases » ou « d'éléments ». Pour les « phrases » ils peuvent avoir l'attribut « type », et les valeurs de type de phrase sont « simple » ou « complexe ».

```
<!ELEMENT DOCUMENTTEXTE (ENTETE, CONTENU)>
        <!ATTLIST DOCUMENTTEXTE id_doc ID #REQUIRED>
        <!ELEMENT ENTETE (titre_doc, auteur, source, date, type, niveau)>
                <!ELEMENT titre_doc (#PCDATA)>
                <!ELEMENT auteur (#PCDATA)>
                <!ELEMENT source (#PCDATA)>
                        <!ATTLIST source type (magazine | journal | livre | internet | roman | inconnu) #REQUIRED>
                        <!ATTLIST source periodicite (hebdomadaire | journaliere | mensuelle | bimensuelle | annuelle | inconnue) #IMPLIED>
                <!ELEMENT date (#PCDATA)>
                <!ELEMENT type (#PCDATA)>
                <!ELEMENT niveau (#PCDATA)>
                <!ELEMENT activite (#PCDATA)>
        <!ELEMENT CONTENU (titre?, (sous-titre?, paragraphe)*)>
                <!ELEMENT titre (phrase+)>
                <!ELEMENT sous-titre (phrase+)>
                <!ELEMENT paragraphe (phrase+ | element)*>
                <!ELEMENT phrase (#PCDATA | prop | element)*>
                        <!ATTLIST phrase type (simple | complexe) #REQUIRED>
                <!ELEMENT ellipse EMPTY>
                <!ELEMENT prop (#PCDATA | referent | coreferent | element | prop)*>
                <!ELEMENT referent (#PCDATA | element)*>
                <!ELEMENT coreferent (#PCDATA | element)*>
                <!ELEMENT element (#PCDATA)*>
                        <!ATTLIST prop type (principale | independante | subordonnee | averbale | incise) #REQUIRED>
                        <!ATTLIST prop sous-type (principale | participiale | completive | coordonnee | relative | juxtaposee | verbale |
  infinitive | conjonctive | averbale) #IMPLIED>
                        <!ATTLIST element type (DETIFS | NCFS | ADJSIG | PDS | VINDP3S |DETIMS | NCMS | ADV |VINF | DETDPIG | NCMP | PREP
  | DETDMS | NCFP | ADJFP | VINDP3P | ADJMP | PCTFORTE | DETDFS | ADJORD | DETPOSS | PPER3P | PPER3S | ADJFS | NCSIG | PCTFAIB | ADJNUM |
  NCMIN | PRI | NPFS | COO | ADJMS | NPI | NCHMS | SUB | NCFIN | ADJIND | VCONP3S | VPARPMS | NCPIG | NPMS | DETDEM | PIPIG | VPARPMP |
  VPARPFS | VPARPRES | PIMS | VINDF3S | ADJPIG | NPFIN | NPSIG | NPHSIG | VINDI3P | NPMIN | VINDI3S |PIMP | VPARPFP | PPER2P | VINDP2P |
  PIFP | NCI | PDP | ADJINT | PPER1P | VINDF2P | ADJINV | PII | VIMPP2P | VSUBP3S | VINDF3P | VINDPS3S | INT | PISIG | VINDPS3P | PPER2S |
  PPER1S | VINDI1P | VINDP2S | VINDP1S | ADJMIN | VIMPP1P | PRFS | PRMS | VIMPP2S | VCONP3P | PIFS | VINDP1P | ADJHFS | ADJHMS | VINDI2S |
  ELLIP) #REQUIRED>
                        <!ATTLIST element cat (NC | V | DETD | ADJ | PPER | DETI | NP | PD | PI | PR) #IMPLIED>
                        <!ATTLIST element mode (INF | IND | CON | PAR | IMP | SUB) #IMPLIED>
                        <!ATTLIST element temps (P | F | PRES | I | PS) #IMPLIED>
                        <!ATTLIST element pers (1 | 2 | 3) #IMPLIED>
                        <!ATTLIST element genre (M | F | I) #IMPLIED>
                        <!ATTLIST element nombre (S | P | I) #IMPLIED>
                        <!ATTLIST referent type (Nom) #IMPLIED>
                        <!ATTLIST referent sous-type (GN | NP | NC) #IMPLIED>
                        <!ATTLIST referent idn ID #REQUIRED>
                        <!ATTLIST coreferent type (Nom | Pron) #IMPLIED>
                        <!ATTLIST coreferent sous-type (GN | COD | Pers | Rel | N | Compl | Pronom | COI | NP | NC) #IMPLIED>
                        <!ATTLIST coreferent classe (fidele | infidele | determinative | totale | partielle | integrante | associative)
  #IMPLIED>
                        <!ATTLIST coreferent idn ID #REQUIRED>
                        <!ATTLIST coreferent chaine IDREF #REQUIRED>
                        <!ATTLIST coreferent position (Anaphore | Cataphore) #IMPLIED>
                        <!ATTLIST coreferent reference IDREF #REQUIRED>

                        |
```

Figure 4: Module de gestion des séquences didactiques (Molina Mejia, 2015).

2.3. Pour quelles activités ?

31 Une fois les textes annotés et la DTD constituée, nous avons pu mettre en œuvre des séquences pédagogiques de formation, divisées en séries d'activités ou d'exercices. Les activités élaborées mettent en œuvre des phénomènes liés à la linguistique textuelle. Elles concernent :

32 – **La structure logique du texte :** l'analyse des phrases principales dans des textes ; l'étude de la structuration des textes ; l'analyse des différents types de texte et/ou des séquences textuelles.

33 – **La cohérence et cohésion textuelles :** l'étude de la coréférence textuelle (anaphores et cataphores, référents et coréférents, chaînes de coréférence, etc.) ; l'analyse des connecteurs et des marqueurs logico-temporels et discursifs dans les textes.

34 – **La progression thématique :** travail autour des notions de *thème* et de *rhème*, et des différents types de progression thématique (linéaire, à thème constant ou dérivée).

35 Pour un niveau donné, les textes associés à des activités sont interchangeables ; ils font partie du paramétrage des activités. Les séquences des activités (figure 5) sont constituées des 4 phases (exercices de repérage, explications au travers des supports théoriques, exercices de systématisation et exercices d'application de connaissances) que l'enseignant peut également paramétrer (*cf.* figure 6 § figure 7).

36 Un aspect intéressant du système est le fait de proposer le stockage du support théorique (d'un corpus théorique, phase 2). Pour l'instant les enseignants peuvent stocker sur la « BD de corpus des textes » des documents en PDF ou en traitement du texte (Word,

LibreOffice Writer, etc.), mais l'idée est de proposer une interface en HTML ou Javascript permettant l'explication des phénomènes issus de la linguistique textuelle.

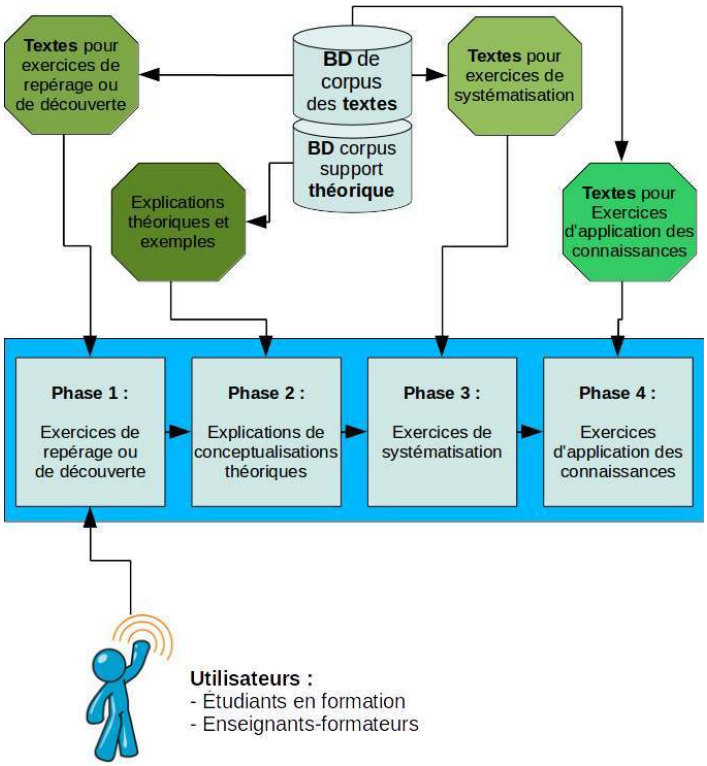

Figure 5: Exercice de repérage 1.

3. Un système informatique fondé sur des corpus

37 Le système d'ALAO (Apprentissage des Langues Assisté par Ordinateur), ELiTe-[FLE]², que nous proposons est fondé sur un travail fait sur des corpus. Il fonctionne à partir des corpus étiquetés et annotés grâce à des techniques issues du TAL (Traitement Automatique des Langues) et des traitements effectués de manière manuelle, comme nous l'avons présenté dans la section précédente.

3.1. ELiTe-[FLE]² et la formation des formateurs

38 L'idée de posséder un système d'ALAO comme ELiTe-[FLE]² est d'offrir, aux enseignants-formateurs et à leurs étudiants en formation, un environnement informatique pouvant être utilisé dans la formation en présentiel ainsi qu'en autonomie guidée. Ce système devrait servir dans le processus de formation linguistique des futurs formateurs de langue française (Molina Mejia, 2015).

39 ELiTe-[FLE]² fonctionne à partir d'activités fondées sur le corpus linguistique que nous avons présenté dans la partie précédente. Certaines activités permettent d'appréhender des notions telles que la cohérence et la cohésion textuelles et discursives, les notions d'anaphores, de cataphores, la progression thématique, etc. Nous considérons que la maîtrise de ces notions passe par l'apprentissage d'autres notions telles que : GN, COD,

COI, etc. En somme, des notions que ces futurs formateurs devront enseigner par la suite à des apprenants de langue.

40 Le système proposé permet de préparer des séquences didactiques de formation et pour ceci il se sert des textes stockés dans la base de données du corpus textuel. L'image ci-dessous (figure 6, Molina Mejia, 2015) nous donne un aperçu général du système permettant la création des séquences didactiques de la part des enseignants-formateurs.

41 Dans cette figure nous pouvons voir les possibilités qui offre la plate-forme :

42 Soit le menu en haut, soit le menu en face de chaque séquence créée permettent aux enseignants de faire plusieurs actions : Un enseignant peut créer une séquence didactique en utilisant les options offertes par le système (*cf.* figure 7). Il peut tester la séquence qu'il a créée, c'est-à dire la développer afin d'observer comme la séquence fonctionne. L'enseignant peut aussi modifier ou supprimer une séquence déjà existante, seulement si celle-ci lui appartient. Finalement, l'enseignant peut envoyer la séquence de son choix aux adresses e-mail des étudiants en formation.

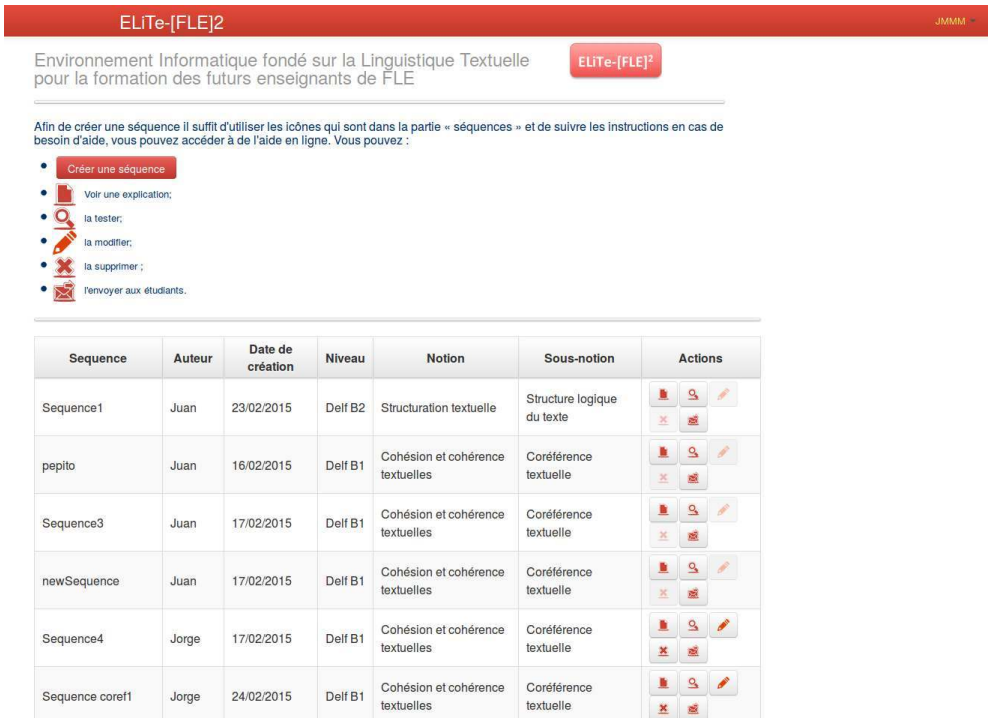

Figure 6: Feed-back correspondant à l'exercice de repérage 1 (extrait du texte).

43 Les enseignants peuvent ainsi créer leurs propres séquences, utiliser les séquences créées par leurs collègues, ou créer des nouvelles séquences en modifiant les séquences existantes. Pour tout ceci, ils utilisent un module de gestion des séquences (figure 7, Molina Mejia, 2015).

44 Ce module permet aux enseignants de choisir une des trois notions à étudier et une sous-notion (cf. 2.3). Ils peuvent choisir le niveau par rapport à l'activité (B1, B2 ou C1), selon ce niveau il y a certains textes proposés par la plate-forme. Les textes changent selon le type d'activité (exercice de repérage et exercice de systématisation). L'enseignant peut aussi décider d'employer ou non du support pédagogique. Il peut donner les consignes pour l'activité d'application de connaissances (qui apparaîtra une fois toutes les autres activités seront développées). Finalement, l'enseignant peut aussi décrire la séquence afin

de partager cette information avec ses collègues ou pour se rappeler lui-même des activités qu'il a déjà créées.

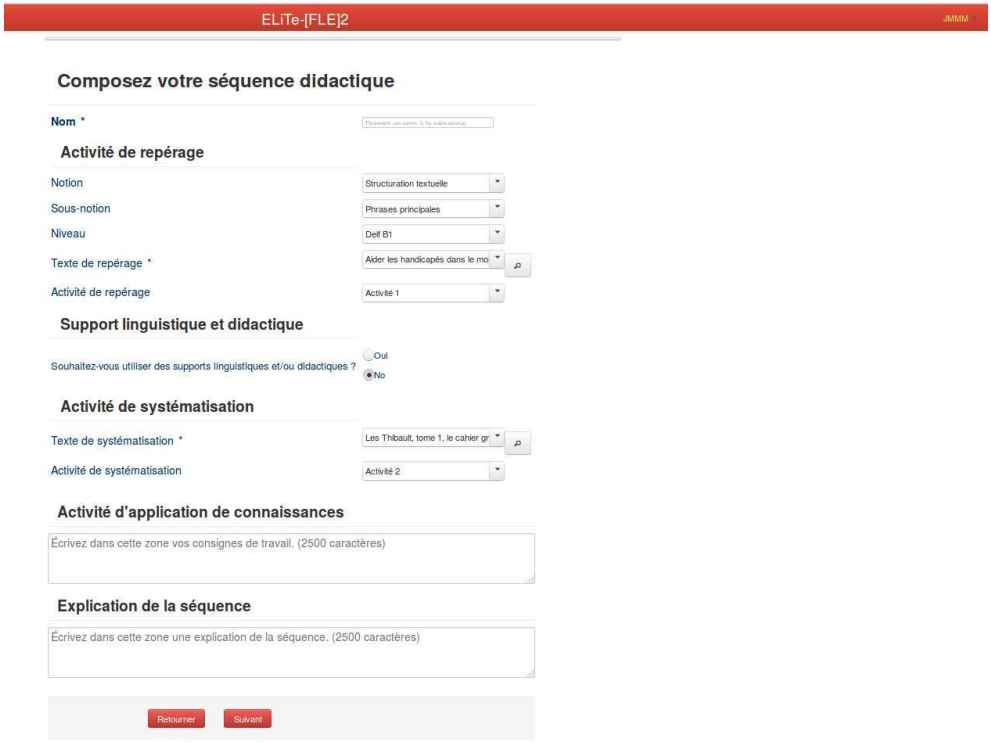

Figure 7: Activité de repérage 2.

3.2. ELiTe-[FLE]² et des activités fondés sur un corpus de textes

45 Dans un premier temps nous avons testé le travail développé au niveau du corpus pour des activités sur la coréférence textuelle. Nous montrons une activité développée à partir du système ELiTe-[FLE]², telle que nous l'avons décrite dans notre travail (*cf.* Molina Mejia, 2015 § chapitre 11) et dans un article publié récemment (Molina Mejia & Antoniadis, 2015).

46 Dans la partie qui suit nous montrons seulement les activités qui se servent du corpus, dont les activités de repérage ou de découverte et les activités de systématisation.

3.2.1. Exercices de repérage ou de découverte

47 Le corpus annoté permet le développement d'activités dans lesquelles les étudiants en formation peuvent repérer dans un texte des éléments généraux à partir d'une notion à travailler. L'exemple ci-dessous (figure 8, Molina Mejia & Antoniadis, 2015) nous montre la manière dont les étudiants, dans une démarche inductive vont découvrir la notion de coréférence textuelle avec le guidage de l'enseignant-formateur. Dans cette première activité, il s'agit de glisser depuis le texte qui se trouve dans la partie supérieure vers la seconde case les unités textuelles (dans ce cas spécifique des anaphores), et vers la troisième case le référent principal qui complète la série. L'idée est de découvrir les référents et coréférents textuels faisant partie de la notion de coréférence textuelle.

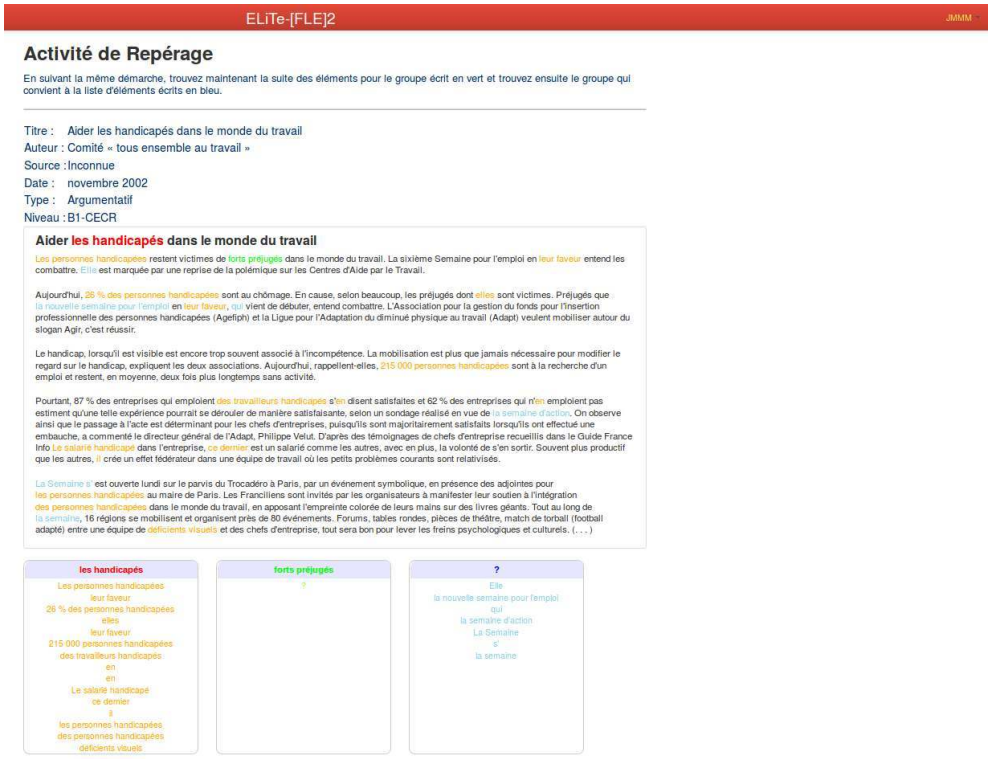

Figure 8: Feed-back pour l'activité de repérage 2.

48 Grâce aux annotations du corpus, un feed-back apparaît dans chaque activité. Dans ce cas, nous avons un feed-back qui nous permet de donner aux étudiants en formation certains éléments d'ordre morphosyntaxique, permettant d'acquérir les notions liées à la formation linguistique (figure 9, Molina Mejia & Antoniadis, 2015). À gauche nous voyons un feed-back dont les notions qui apparaissent sont GN et qui demande une autre unité textuelle (Pronom, par exemple). L'idée est de faire réfléchir les futurs enseignants, toujours dans une approche inductive, à l'emploi des certains notions liées à la linguistique textuelle, telles que : « coréférents de type nominal » ou « coréférents de type pronominal », et ainsi de suite.

49 Sur la droite nous apercevons le cas d'un feed-back de type « positif », pour l'instant il ne donne pas d'informations de type métalinguistique, mais l'idée est d'améliorer cet aspect pour l'avenir en donnant aux apprenants le type de coréférent trouvé (GN ou Pron).

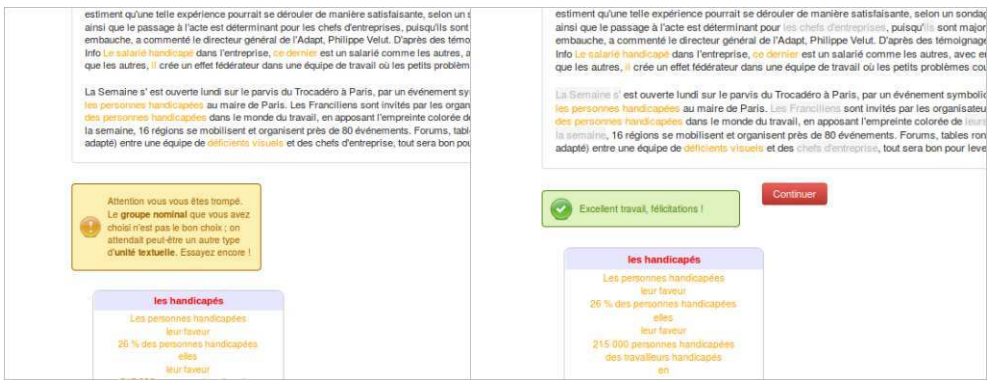

Figure 9: Exercice de systématisation 1.

50 Dans une autre activité les étudiants en formation sont conduits à découvrir les notions liées à la langue cible. Dans le cas que nous voyons dans l'image ci-dessous (figure 10, Molina Mejia & Antoniadis, 2015), les étudiants travaillent sur la notion d'anaphores nominales (GN, N), et pronominales (Pron). L'activité consiste à déplacer les éléments qui se trouvent dans la partie supérieure vers la partie inférieure, il faut les séparer en deux parties dans la partie supérieure il faut laisser seulement les GN et N et dans la partie inférieure tous les Pron.

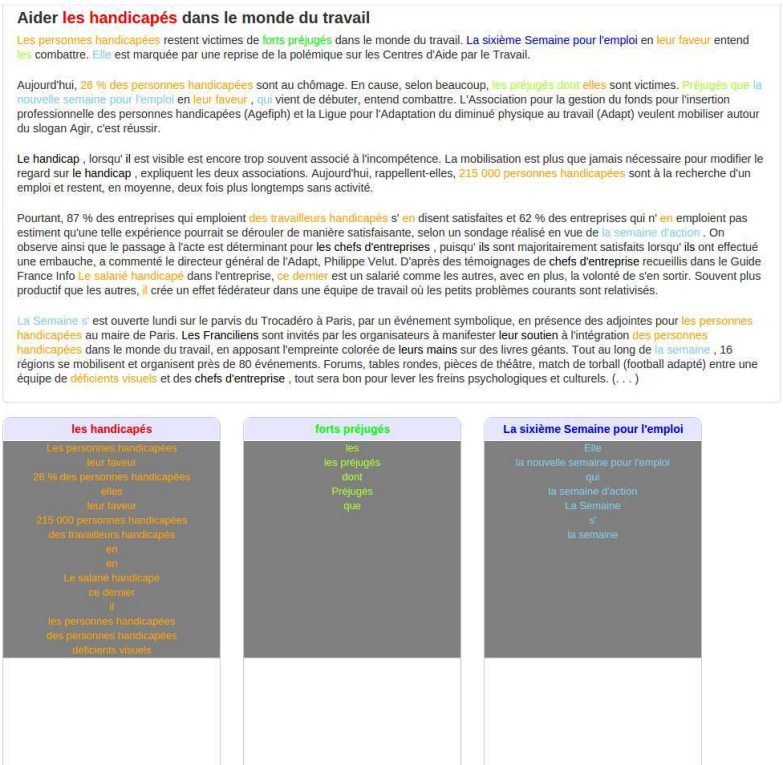

Figure 10: Exercice de systématisation 2.

51 Une fois les éléments déplacés, nous avons un feed-back dans lequel une zone est grisé jusqu'à ce que l'étudiant trouvera les éléments semblables GN et N dans la partie supérieure et Pron dans la partie inférieure (figure 11, Molina Mejia & Antoniadis, 2015). L'idée est de lui faire découvrir que certains coréférents sont d'ordre nominal et d'autres de type pronominal. Dans d'autres activités il est possible de faire repérer des anaphores de type fidèle, infidèle, etc. Toutes les zones doivent rester en blanc afin que l'activité soit validée.

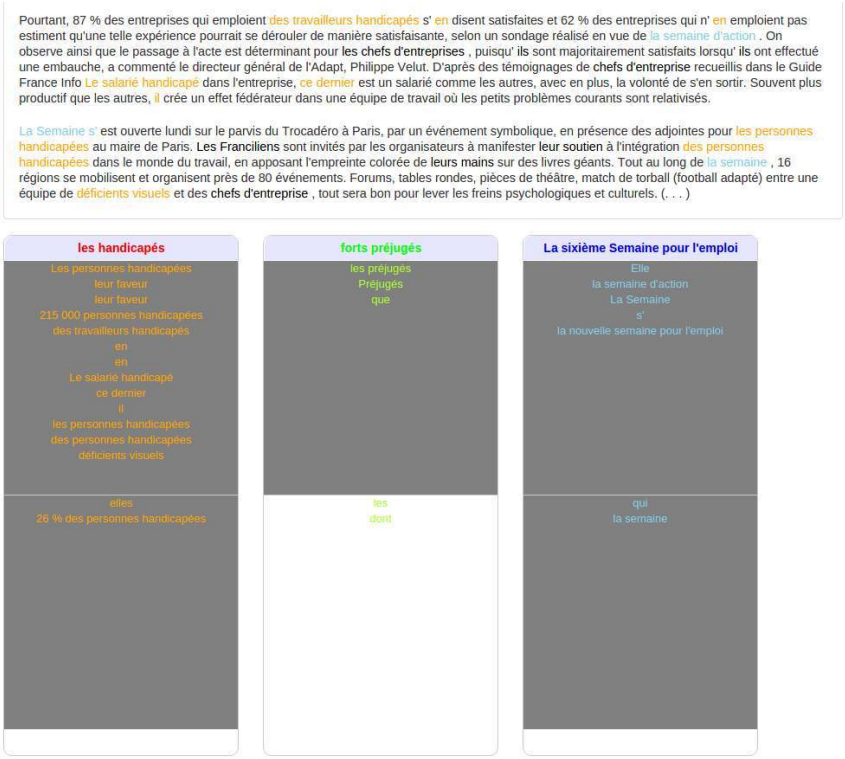

Figure 11: Exercice de systématisation 3.

3.2.2. Exercices de systématisation

52 Ce type d'exercices doit permettre aux étudiants de mieux maîtriser les notions apprises grâce à la réalisation des activités leur permettant de mettre en œuvre les connaissances acquises. Autrement dit, une fois que les étudiants auront fait les activités de repérage et eu des explications de la part des enseignants-formateurs, ils pourront s'exercer à leur maîtrise.

53 Dans la figure ci-dessous (figure 12, Molina Mejia & Antoniadis, 2015), nous avons un exemple de ce type d'exercice. Dans cette activité il s'agit de conduire les étudiants à faire des inférences à partir de la notion de chaîne de coréférence. Ils devront, tout d'abord, choisir les couleurs en fonction du contexte dans lequel se trouve l'élément à trouver. Pour ceci, ils disposent d'un menu déroulant pour chaque élément. Les couleurs des coréférents sont choisies en fonction des référents principaux.

54 Il y a trois couleurs qui correspondent aux trois chaînes de coréférence (orange pour les coréférents de la chaîne 1, bleu pour les coréférents de la chaîne 2 et vert pour les coréférents de la chaîne 3). L'idée est que pour chaque élément il y a toujours ces trois choix.

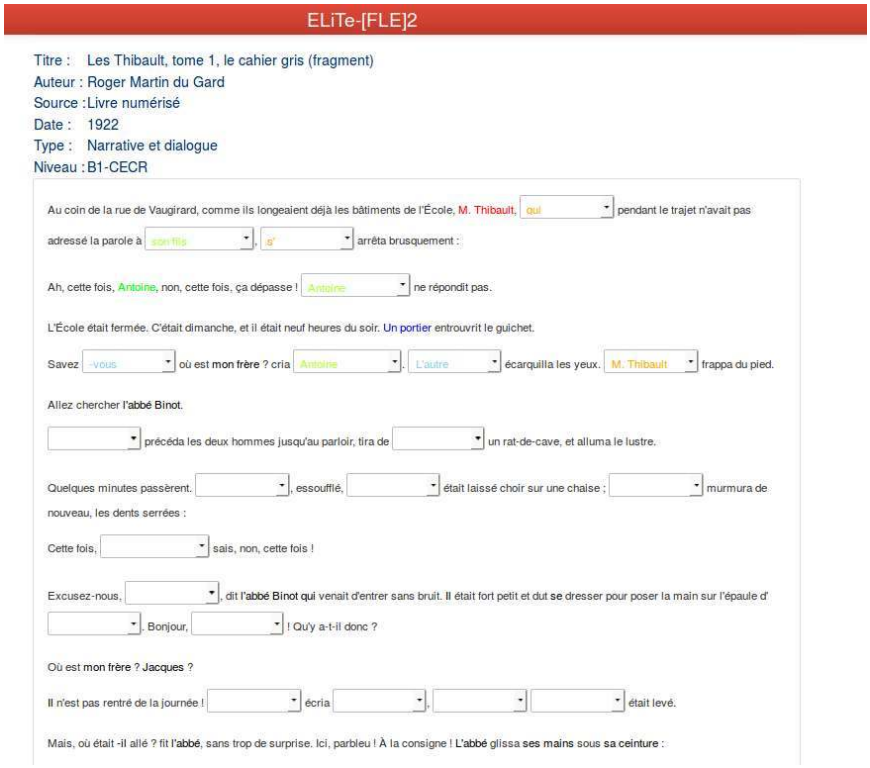

Figure 12: Excercice de systématisation 1

55 Une fois que les étudiants ont fini de faire l'exercice de systématisation 1, ils passent à une nouvelle phase. Dans un autre exercice (figure 13, Molina Mejia & Antoniadis, 2015) ils doivent choisir l'élément qui convient par rapport aux chaînes de coréférence. Les étudiants devront inférer, une nouvelle fois, le type d'élément selon le contexte et le co-texte dans lequel celui-ci se trouve, en tenant compte des référents principaux. À chaque groupe de coréférents correspond une couleur en particulier. L'idée est de leur faire travailler ces notions en regardant le contexte droit et le contexte gauche dans lequel se trouvent les éléments manquants dans le texte.

56 Pour chaque espace à remplir il n'y aura que les éléments qui appartiennent à la chaîne de coréférence qui a été annotée.

57 Une dernière activité de systématisation permet aux étudiants en formation de mieux saisir les notions d'anaphores nominales et pronominales. L'image ci-dessous (figure 14, Molina Mejia & Antoniadis, 2015) nous permet de voir de quelle manière les étudiants peuvent diviser les éléments en deux types selon l'appartenance aux coréférents nominaux (GN ou N), et aux coréférents pronominaux (Pron).

4. Analyses et résultats par rapport au corpus

58 Afin de mieux cibler nos activités, nous avons décidé d'adapter les annotations et l'étiquetage de notre corpus en fonction des notions issues de la linguistique textuelle. De ce fait, une typologie des notions liées à la linguistique textuelle a été établie. Nous partons, par exemple, non de la notion de phrase (très commune dans les grammaires d'ordre structuraliste et autres), mais de la notion de proposition, plus axée dans le sens de la linguistique textuelle (Adam, 2011) et de la linguistique énonciative (Benveniste,

1970). Ceci nous a permis d'alléger notre corpus le rendant plus dynamique, car au lieu des phrases (dont la notion et la délimitation sont parfois assez difficiles à cerner), ou des *chunks* (pour lesquels certains linguistes n'arrivent pas encore à se mettre d'accord sur leur valeur dans le domaine de la linguistique[14]), nous avons préféré d'utiliser la notion de proposition telle qu'elle est analysée par Adam (2011), et de la manière dont elle est décrite dans la grammaire méthodique du français (Riegel *et al.* 2009). En effet, la proposition (ou proposition-énoncée) est une unité minimale porteuse de sens, car il s'agit de : « une unité de base, effectivement réalisée et produite par un acte d'énonciation, donc comme un *énoncé minimal* » (Adam, 2011).

59 Ensuite, nous avons annoté d'autres notions en fonction des phénomènes étudiés (*cf.* section 2.3). Par exemple :

60 Dans le cas de la coréférence textuelle, nous avons annoté les référents principaux [**type** → Nom], ensuite les coréférents [**position** → anaphores | cataphores] ; [**type** → Nom | Pron] ; [**sous-type** → GN | COD | Pers | Rel | N | Compl | Pronom | COI | NP | NC] ; [**classe** → fidèle | infidèle | déterminative | totale | partielle | intégrante | associative]. Nous avons donné à chaque référent et coréférent des identifiants qui permettent par la suite de repérer les chaînes de coréférence textuelles.

61 La DTD issue de notre corpus et des annotations ont permis de faire les activités dans lesquelles les unités textuelles peuvent être déplacées, grâce à des langages de programmation comme Java[15], ce qui permet d'isoler les unités annotées en XML et de les transformer en entités qui « bougent » et qui sont reconnues par le système. Ceci est d'extrême importance, puisque la validation des activités passe par cette étape, ce qui est l'un de points forts de notre système qui rend automatique ce processus pour n'importe quel type de texte ayant comme seule contrainte d'être conforme avec la DTD.

Conclusion et perspectives

62 Le corpus que nous avons constitué avec des textes authentiques annotés selon les notions à étudier en linguistique textuelle nous a permis dans un premier temps de mettre en œuvre un premier prototype opérationnel du système ELiTe-[FLE]². Pour l'instant nous avons développé des activités pour l'analyse des coréférents textuels, mais nous comptons dans un second temps développer d'autres activités par rapport aux autres notions annotées.

63 Une fois le système complet, avec toutes les annotations et toutes les activités mises dans la base de données, nous comptons le tester dans les cours de formation des enseignants de FLE. En effet, l'idée est d'améliorer le système en fonction de l'évaluation que les utilisateurs (enseignants-formateurs et étudiants en formation) feront dans les universités colombiennes partenaires du projet (Université d'Antioquia et Université nationale de Colombie).

BIBLIOGRAPHIE

ADAM, J.-M. (2011). *La linguistique textuelle*. Collection Linguistique Cursus, 3è édition. Armand Colin : Paris.

ASTON, G. (1995). Corpora in Language Pedagogy : Matching Theory and Practice. In Cook, G. & Seidlhofer, B. (éditeurs) : *Principle and Practice in Applied Linguistics : Studies in Honnour of H. G. Widdowson*. Chapitre 17, pp 257-270. Oxford University Press : Oxford.

ARISMENDI, F. & COLORADO, D. (2015). « La formation des enseignants de FLE en Colombie : panorama et cas de l'université d'Antioquia. » In *Dialogues et cultures : La formation initiale des enseignants de français langue étrangère*, (Revue de la FIPF), vol. 61, pp. 44-61. Bruxelles – Belgique.

BARTHE, M. & CHOVELON, B. (2009). *Le français par les textes : Volume 1*. Presses Universitaires de Grenoble : Grenoble.

BENVENISTE, E. (1970). *Problèmes de linguistique générale II*. Gallimard : Paris.

BOULTON, A. (2007). Esprit de corpus : Promouvoir l'exploitation de corpus en apprentissage des langues. In Williams, G. (éditeur) : *Actes des cinquièmes journées de la linguistique de corpus, Texte et Corpus*. Volume 3, pp 37-46. Université de Bretagne Sud : Lorient.

BOULTON, A. & TYNE, H. (2014). *Des documents authentiques aux corpus : démarches pour l'apprentissage des langues*. Langues & didactique. Didier : Paris.

CADET, L. & TELLIER, M. (2007). « Le geste pédagogique dans la formation des enseignants de LE : Réflexions à partir d'un corpus de journaux d'apprentissage. » In *Les cahiers de Théodile* N° 7, pp. 67-80.

CHANIER, T. & CIEKANSKI, M. (2010). « Utilité du partage des corpus pour l'analyse des interactions en ligne en situation d'apprentissage : un exemple d'approche méthodologique autour d'une base de corpus d'apprentissage. » In Alsic [En ligne], vol. 13 | 2010, mis en ligne le 06 décembre 2010. URL : http://alsic.revues.org/1666 ; DOI : 10.4000/alsic.1666

GIGUET, E. (1998). *Méthode pour l'analyse automatique de structures formelles sur documents multilingues*. Thèse de doctorat, Université de Caen : Caen – France.

GIULIANI, D. & HANNACHI, R. (2010). « Linguistique de corpus et didactique du F.L.E. Une exploitation du corpus IntUne. » In *Cahiers de praxématique* [En ligne], 54-55 | 2010, document 8, mis en ligne le 01 janvier 2013. URL : http://praxematique.revues.org/1136

HABERT, B. ; FABRE, C. & ISSAC, F. (1998). *De l'écrit au numérique : Constituer, normaliser et exploiter les corpus électroniques*. Collection Informatiques. Masson : Paris.

JOHNS, T. (1991). « From printout to handout : grammar and vocabulary teaching in the context of data-driven learning ». In : Johns T. & King P. (dir.) Classroom concordancing. English language research journal, vol. 4, p. 27-45.

KÜBLER, N. (2014). « Mettre en œuvre la linguistique de corpus à l'université : Vers une compétence utile pour l'enseignement/apprentissage des langues ? » In Recherches en didactique des langues et des cultures : Les Cahiers de l'Acedle, volume 11, numéro 1, 2014, pp 37-77.

LANDURE, C. (2011). « Data-Driven Learning : apprendre et enseigner à contre-courant. » In Mélanges CRAPEL, N° 32, pp 163-178.

LEBARBÉ, T. (2002). Hiérarchie inclusive des unités linguistiques en analyse syntaxique coopérative. Thèse de doctorat, Université de Caen : Caen – France.

MOLINA MEJIA, J. M. (2014). Conception d'un environnement informatique fondé sur la linguistique textuelle et destiné à la formation des enseignants de FLE en Colombie. In P.-A. CARON & R. CHAMPAGNAT (Éditeurs) : Actes des 5ièmes Rencontres des Jeunes Chercheurs en Environnements Informatiques pour l'Apprentissage Humain. Université de la Rochelle, 16 - 17 juin 2014, pp 23-29. La Rochelle – France.

MOLINA MEJIA, J. M. (2015). *EliTe-[FLE]² : Un environnement d'ALAO fondé sur la linguistique textuelle, pour la formation linguistique des futurs enseignants de FLE en Colombie.* Thèse de Doctorat. Soutenue le 06 Novembre 2015, 442 pp. Université Grenoble Alpes : Grenoble – France.

MOLINA MEJIA, J. M. & ANTONIADIS, G. (2014a). Toward the Constitution of a Hybrid Learning Environment for the FFL Teacher's Training in Colombian Universities Based on Text Linguistics. In G. QUIROZ & P. PATIÑO (Éditeurs) : LSP in Colombia : Advances and Challenges. Chapitre 15, Volume 175, pp 233-249. Collection Linguistic Insights. Éditions Peter Lang : Bern, Berlin, Bruxelles, Frankfurt am Main, New York, Oxford, Wien.

MOLINA MEJIA, J. M. & ANTONIADIS, G. (2014b). Conception d'un environnement informatique fondé sur la linguistique textuelle pour la formation initiale des futurs enseignants de FLE en Colombie. In Adjectif : Analyses recherches sur les TICE. Article 293, 05 Juin 2014. [En ligne] http://www.adjectif.net/spip/spip.php?article293

MOLINA MEJIA, J. M. & ANTONIADIS, G. (2014c). Conception et élaboration d'un environnement informatique fondé sur la linguistique textuelle et destiné à la formation des futurs enseignants de FLE en Colombie. In F. OLMO CAZEVIEILLE & J.-M. MANGIANTE (Éditeurs) : II Coloquio franco-español de análisis del discurso y enseñanza de lenguas para fines específicos. Lenguas, comunicación y tecnologías digitales. Universitat Politècnica de Valencia, 03 - 05 septembre 2014, pp 143-158. Colección Congresos. Editorial Universitat Politècnica de Valencia : Valencia – Espagne.

MOLINA MEJIA, J. M. & ANTONIADIS, G. (2015). « Un environnement informatique pour la formation des formateurs en FLE Colombiens, fondé sur la linguistique textuelle ». In *Alsic* [En ligne], vol. 18 | 2015, mis en ligne le 30 novembre 2015, Consulté le 06 décembre 2015. URL : http://alsic.revues.org/2843 ; DOI : 10.4000/alsic.2843.

RIEGEL, M. ; PELLAT, J.-C. & RIOUL, R. (2009). Grammaire méthodique du français. Quadrige, quatrième édition. Presses Universitaires de France : Paris.

SOUQUE, A. (2014). Modèle de vérification grammaticale automatique gauche-droite. Thèse de doctorat. Université de Grenoble : Grenoble.

TEUBERT, W. (2009). La linguistique de corpus : une alternative. In Semen [En ligne], 27 | 2009, mis en ligne le 01 juin 2009, consulté le 03 mai 2014. URL : http://semen.revues.org/8914

VERGNE, J. & GIGUET, E. (1998). Regards théoriques sur le "tagging". In *Actes de TALN (Traitement Automatique des Langues Naturelles).* Paris, France, 10-12 juin 1998.

WILLIAMS, G. (2006). « La linguistique de corpus, une affaire prépositionnelle. » In Revue Texto [En ligne], URL : http://www.revue-texto.net/Parutions/Livres-E/Albi-2006/Williams.pdf

NOTES

1. Environnement informatique fondé sur la Linguistique Textuelle, pour la Formation Linguistique des futurs Enseignants de FLE.
2. *Diplôme d'Études en Langue Française*.
3. *Diplôme Approfondi en Langue Française*.
4. Le projet *InUne* est un exemple d'utilisation de corpus pour la formation des enseignants de FLE en milieu *endolingue* (Guiliani & Hannachi, 2010). Ce type de corpus prépare des étudiants en formation dont la plupart ont le français comme langue maternelle (L1).
5. Cadre Européen Commun de Référence pour les Langues.
6. Langues pour spécialistes d'autres disciplines.
7. Il s'agit d'un corpus multimodal d'apprentissage, dans le cadre du projet Mulce. http://lrl-diffusion.univ-bpclermont.fr/mulce2/index.html
8. Centre International d'Études Pédagogiques, http://www.ciep.fr/
9. Les textes retrouvés sur le site du CIEP sont libres d'utilisation dans un but pédagogique, et permettent l'entraînement aux examens DELF et DALF.
10. D'après le CECRL, les niveaux B1 et B2 correspondent à « l'utilisateur indépendant » et le niveau C1 à « l'utilisateur expérimenté ».
11. *eXtensible Markup Language* (« langage à balises extensible », en français).
12. *Document Type Definition* (« Définition du Type de Document », en français).
13. BD : Base de données.
14. La notion de *chunk* est analisée par A. Souque (2014 : 32-33). D'après cet auteur : « Pour Abney (1991) les *chunks* délimitent un groupe de mots, pour Giguet (1998) il s'agit de « syntagmes minimaux », Vergne et Giguet (1998) parlent des « syntagmes non récursifs », Lebarbé (2002) définit le *chunk* comme une unité constituée d'un mot lexical, entourée d'une constellation de mots fonctionnels (déterminants, pronoms, adverbes, etc.). » Finalement pour Souque (2014 : 33) : « Il s'agit d'une unité, que nous nommerons indifféremment *chunk* ou syntagme, définie non pas en fonction de son contenu, mais en fonction de ses frontières, principalement des mots grammaticaux [...], des marques morphologiques [...] et des ponctuations. »
15. Langage de programmation informatique orienté objet.

RÉSUMÉS

Dans cet article nous présentons des corpus qui ont été annotés afin de constituer un système d'Apprentissage des Langues Assisté par Ordinateur (ALAO) pour la formation linguistique des futurs enseignants de Français Langue Étrangère (FLE) qui se forment dans les universités colombiennes. Il s'agit des corpus qui ont été annotés suivant les procédures issues du Traitement Automatique des Langues (TAL), pour les aspects morphologiques et syntaxiques, et des procédures d'annotation manuelle afin de cibler certains phénomènes étudiés par la linguistique textuelle. Le résultat du travail d'annotation pourvoi une série d'activités qui sont prises en compte par le système informatique ELiTe-[FLE]². Ces activités cherchent l'amélioration du niveau linguistique des étudiants en formation colombiens à travers l'étude des notions linguistiques qui se trouvent contextualisées dans les textes.

In this article we present an annotated copora, created to establish a Computer Assisted Language Learning system. This system is intended to the linguistic training of future FFL (French as a Foreign Language) teachers in Colombian universities. We have a tagged corpora using Natural Language Procedures for the morphological and syntactic aspects, and manual annotations to aim at the notions studied by text linguistics. The result of these annotations provide a serie of activities that are taken into account for ELiTe-[FLE]² informatic system. Those activities look for the improvement of the linguistic level of the Colombian training students through the study of linguistic notions contextualised in the texts.

INDEX

Mots-clés : Annotation des corpus, TAL, FLE, ALAO.
Keywords : Corpus annotation, NLP, FFL, CALL.

AUTEUR

JORGE MAURICIO MOLINA MEJIA

jorge.mauricio.molina@gmail.com
Université Grenoble-Alpes, laboratoire LIDILEM