Annotation linguistique de corpus : vers l'exhaustivité par la convivialité

Jean-Marie Viprey, Virginie Léthier

UMR Bases, Corpus, Langages et Maison des Sciences de l'Homme de Franche-Comté
EA 3187 Archives, Textes, Sciences des Textes, Université de Franche-Comté, Besançon

jean-marie.viprey@univ-fcomte.fr / virginie.lethier@yahoo.fr

Abstract

It is a long time now that linguistic annotations (e.g. lemmatisation and others) have not been contradictory to preservation of the surfacing graphic forms. The matter is no longer "Should we lemmatise?" but rather, "How to enrich textual resources with lexical, morphological and syntactical information?" Some textometrical operations following the urn principles can be safely led with automatical annotation and its processing of residual ambiguities based on probabilist printings. Yet, TAD (Textual Analysis of Discourses) - among others disciplines resorting to textual materiality, demands a cautious and optimal control of the choices made and registered. This is made necessary given the need to go back to the text by a statistically assisted digitized exploration. In the *numerical scriptorium* perspective, DiaTag is an annotation environment alternating between automatic and dialogue stages, whose ergonomics are improved for the skilled operator facing textual bases of billions of words. The experiment described here was carried out on a significant sample of written French press. The aim was to start from the paper material and to process it in order to enable a full-text exploitation and a skilled research. The experiment shows that tagging cases unworkable for robots under human control become realistic, allowing us to reach deep and elaborated levels of the linguistic structure.

Résumé

L'annotation linguistique (lemmatisation et autres) n'est depuis longtemps plus contradictoire avec la préservation des formes graphiques de surface. Le problème n'est plus « devons-nous lemmatiser ? » mais « comment enrichir les ressources textuelles d'informations lexicales, morphologiques et syntaxiques ? ». Or, même si certaines opérations textométriques (reposant sur le schéma d'urne) se satisfont de l'annotation automatique où l'ambiguïté résiduelle est traitée selon des scripts, l'Analyse Textuelle des Discours (parmi d'autres carrefours disciplinaires recourant aux matérialités textuelles) exige un contrôle raisonné et optimal des choix effectués et enregistrés. Cela est nécessaire au regard des nécessités d'un « retour au texte » ou plus exactement de l'exploration numérisée et assistée statistiquement. Dans l'optique du *scriptorium numérique*, DiaTag est un environnement d'annotation entièrement constitué d'alternances automates/dialogues, où ces derniers sont ergonomisés pour l'opérateur expert confronté à des bases de millions de mots. L'expérience décrite ici, menée sur un gros échantillon de presse française saisie à la source papier, qu'il faut amener à l'exploitation en mode plein-texte et recherche experte, montre qu'il devient réaliste d'annoter sous contrôle humain l'ensemble des cas rebelles aux automates tout en allant jusqu'à des niveaux relativement profonds et élaborés de la structuration linguistique.

Mots-clés : analyse textuelle des discours, lemmatisation, annotation, logométrie, TEI, TreeTagger, Cordial, NooJ, DiaTag.

1. Introduction

Au fil des JADT successives, la question de la *lemmatisation* a presque disparu des discussions de notre communauté. Aux JADT 2006 à Besançon, seules 3 communications portaient cette notion parmi leurs mots-clés. L'une (Tomasetto *et al.*) pour montrer son faible

rendement dans les travaux sur corpus visant des résultats rapides (tout en réaffirmant son utilité pour certaines recherches) ; la seconde (Heitz) à titre très accessoire, pour la réduire au statut de l'une des opérations de pré-traitement à intégrer dans un panel de cycles en vue de modélisation. Seule la troisième (Mayaffre) thématise réellement la *lemmatisation*, pour l'essentiel dans sa conclusion, tandis que l'article porte plutôt sur l'exploitation des autres aspects de la reconnaissance linguistique (annotation grammaticale) ; Mayaffre souligne que la *lemmatisation* est fréquemment sous-exploitée dans les approches statistiques, notamment dans la perspective logométrique ; il soutient par ailleurs que la *lemmatisation* est désormais *rapide* et *fiable*.

Il est établi (Brunet 2000) que la lemmatisation ne modifie pas substantiellement les résultats des analyses statistiques classiques (schéma d'urne). Il reste à démontrer qu'il en va de même pour des analyses apparues plus récemment, notamment celles portant sur la cooccurrence et plus globalement sur les structures du vocabulaire, si l'on entend par vocabulaire la constitution lexicale des textes et si l'on prend au sérieux les avancées de la Linguistique Textuelle (LT) et de l'Analyse Textuelle des Discours (ATD).

Cette démonstration n'est pas l'objet du présent article. Nous souhaitons préalablement interroger la *rapidité* et la *fiabilité* de la lemmatisation à ce jour, en les rapportant, comme il se doit, à des conditions et à des objectifs de recherche bien spécifiés. L'Analyse Textuelle des Discours est l'un de ces cadres qui intéressent au plus haut chef notre domaine JADT, puisqu'en même temps elle est certes l'un de ses débouchés, l'une de ses applications majeures, mais aussi l'un de ses terrains d'expérimentation les plus exigeants.

L'ATD est l'Analyse du Discours (AD) en tant qu'elle surmonte ce que Rastier, Sarfaty, Adam pointent comme son *déficit philologique*, son évitement de la matérialité textuelle. Comme l'AD en général, elle est interdisciplinaire sur le terrain d'ensemble des sciences humaines ; elle vise à enrichir et infléchir les pratiques scientifiques en les historicisant sous l'espèce des discours humains en général, et plus particulièrement encore à les articuler entre elles (sociologie des sciences, par exemple, ou terminologie) ; elle apporte des correctifs non négligeables à la linguistique elle-même, par le biais notamment de l'approche en corpus et de la réflexion sur cette notion. L'ATD a pour programme de réintégrer le *texte* dans toutes ses dimensions aux objets de l'AD ; elle a donc partie liée d'un côté avec la LT, de l'autre avec la techno-logie, et plus largement avec la *philologie numérique*.

Dans cette optique, la *lemmatisation* ne peut plus être considérée comme un simple pré-traitement de données soumises ensuite à une analyse dont seuls les résultats importent. Comme toutes les autres opérations d'annotation, et comme les résultats statistiques eux-mêmes, elle devient un *vecteur* du retour au texte, de son exploration systématique assistée. C'est pourquoi nous supposons que l'ATD a besoin d'une *autre* lemmatisation que l'Analyse de Données Textuelles au sens strict (ADT). A savoir une lemmatisation contrôlée, explicitée, justifiée, et si possible par le chercheur lui-même.

L'application de logiciels entièrement automatisés (même s'ils sont paramétrables) est un recours précieux, *rapide* et *fiable* dans l'optique définie par Tomasetto ou Heitz. Dans la perspective de l'ATD, elle laisse de côté des attentes fondamentales. Ce n'est pas une simple question de perfectionnement. Certes, le résidu semble se restreindre sous l'effet des améliorations et raffinements successifs. Mais outre qu'il n'est pas compressible à l'infini (loin de là), l'essentiel est que les attentes expertes croissent aussi vite que les dits raffinements. Dans les seuls domaines (limités au français moderne et contemporain) [a] de l'ambiguïté graphique des formes simples entre plusieurs lemmes, entre plusieurs valeurs

(genre, nombre, personne, tiroir verbal) [b] de la reconnaissance des lexies composées qu'un dictionnaire ne saurait assurer à lui seul (*au cœur de*) [c] de l'identification des formes verbales auxiliées discontinues, qui sont aujourd'hui considérés comme relevant du *b-a ba* de l'annotation linguistique, les logiciels couramment appliqués (*Tree Tagger*, *Cordial*) échouent quand ils ne renoncent pas de prime abord.

Le pôle *Archive Bases Corpus* (ABC) de la MSH de Franche-Comté a entrepris, en partenariat avec la Bibliothèque d'Etudes de Besançon, la numérisation d'un fonds de presse régionale du XIX^{e} siècle en mode texte et en vue de son exploitation experte. Au-delà des objectifs de conservation et de communication, qui se satisfont en règle générale du mode image, il s'agit d'explorer les voies de la *nouvelle archive*, si peu et si peu systématiquement frayées, et avec un retard presque militant, par les tutelles des bibliothèques patrimoniales.

Nous laisserons ici de côté la discussion quant à savoir si, sur cette voie, on doit ou non s'en tenir à l'océrisation, laissant ou non aux usagers actuels et futurs le soin de cette première interprétation qu'est nécessairement l'annotation linguistique (à commencer par la segmentation). Notons seulement que cette discussion ne sera réellement possible que si certains montrent concrètement ce que peut être l'annotation linguistique poussée d'une vaste base documentaire textuelle.

Nous pouvons laisser cela de côté dans la mesure où nous avons pris le risque d'entreprendre, sur cette base en voie de constitution, une recherche en ATD centrée sur une thèse de 3^{e} cycle. Cette recherche requiert d'ores et déjà un établissement linguistique aussi poussé qu'il sera possible et c'est le domaine d'application que nous soumettons à la présente problématique.

2. Présentation de DiaTag

Dans le cadre d'*ABC* et de ses laboratoires[1], nous avons depuis 1998 entrepris de développer un environnement spécifiquement dédié à l'annotation linguistique experte dans son ensemble (sans négliger pour autant l'apport de modules spécifiques à des environnements déjà constitués[2]). Cet environnement s'intitule *DiaTag* (pour *dialogic tagging*) et il est étroitement articulé à l'environnement *Astartex* dédié à l'exploration assistée elle-même et aux calculs statistiques. Il est, dans cette phase, strictement dédié au français moderne[3].

Le principe général de DiaTag[4] est de faire alterner des phases automatiques et des phases de dialogue où les situations non univoques sont soumises à une décision humaine, de la manière la plus ergonomique possible. Un autre principe majeur est de permettre l'enrichissement et l'édition progressifs des ressources, notamment des dictionnaires « livrés » avec le système, dans le cours même des opérations d'annotation. Un objectif à moyen terme est de constituer des réseaux d'utilisateurs maintenant collectivement les ressources (et les diversifiant)[5].

---

[1] Archives, Textes, Sciences des Textes (ATST), EA 3187 et Laboratoire de Sémio-Linguistique, Didactique, Informatique (LASELDI), EA 2181. Université de Franche-Comté.

[2] C'est le cas en particulier pour NooJ, développé autour de Max Silberztein au LASELDI.

[3] Des applications ont néanmoins été réalisées, au coup par coup, pour l'espagnol et l'anglais.

[4] DiaTag est conçu pour recevoir, en principe, tout texte et ensemble de textes au format .txt ou .html. Il comporte un module « passerelle » destiné à mettre les ressources dans un format propriétaire, convertible en fin de course en .xml normalisé TEI. Il applique aux ressources un balisage propriétaire minimal Astartex destiné à marquer éventuellement des partitionnements à l'intérieur du même fichier de texte.

[5] Le modèle d'un tel réseau a été construit notamment autour d'Intex, puis NooJ, par Maurice Gross et Max Silberztein.

2.1. Segmentations

La première opération est la segmentation en « mots » et plus généralement en *unités textuelles atomiques UTA*. Ce découpage linéaire est presque entièrement automatisé pour le français, car il existe très peu de cas où une liste de séparateurs et des listes d'exceptions puissent être prises en défaut. Si l'on excepte les textes très dégradés (pour lesquels une vérification préalable s'impose), ou très peu normalisés (pour lesquels la reconnaissance linguistique sera en tout état de cause difficile et peu rentable), on ne rencontre guère que des séquences graphiques comme *rendez-vous* ou *d'Alembert* pour devoir être désambiguïsées manuellement. Les valeurs des apostrophes et des tirets[6] peuvent être différenciées. Les anomalies (supposées ou réelles), comme la succession d'un point et d'un caractère de bas de casse, lorsque le mot qui précède n'est pas une abréviation recensée, sont enregistrées dans un index permettant à l'utilisateur qui le souhaite d'opérer une vérification et le cas échéant une correction[7]. Pour cette phase, les ressources linguistiques se limitent à la liste des formes couramment élidées en français et à celle des formes post-clitiques soudées par un tiret ; les ressources lexicales, à deux listes d'abréviations et de sigles courants.

La segmentation en phrases est certes beaucoup plus délicate, et DiaTag ne l'opère pas. En aval, Astartex propose en effet une définition des limites de phrase qui reste grossière et n'est utilisée que pour des calculs statistiques supportant l'approximation (relevé de contextes). Elle n'est pas utilisée pour la navigation experte ni pour l'affichage des concordances. La segmentation en paragraphes est traitée en amont[8].

A ce stade, le fichier texte est déjà utilisable par Astartex aux fins de son exploration en corpus limitée bien sûr aux formes graphiques. Chaque UTA « mot » est représentée par un conteneur du format <*M k*=valeur *f*=valeur > ou *k* est l'attribut *position du premier caractère du mot* et f la *forme graphique exacte* (en Ascii). Chaque ponctuation par un conteneur introduit en <*PC*. Les blancs typographiques seuls sont maintenus hors des conteneurs, et tous les conteneurs de forme <contenu> sont conservés. Ils seront assimilés par tous les traitements ultérieurs à une zone « espace » combinant les blancs et les conteneurs, entre deux mots ou ponctuations. Ils pourront bien sûr être lus et interprétés si besoin. Ils seront donc restitués en sortie à tout moment.

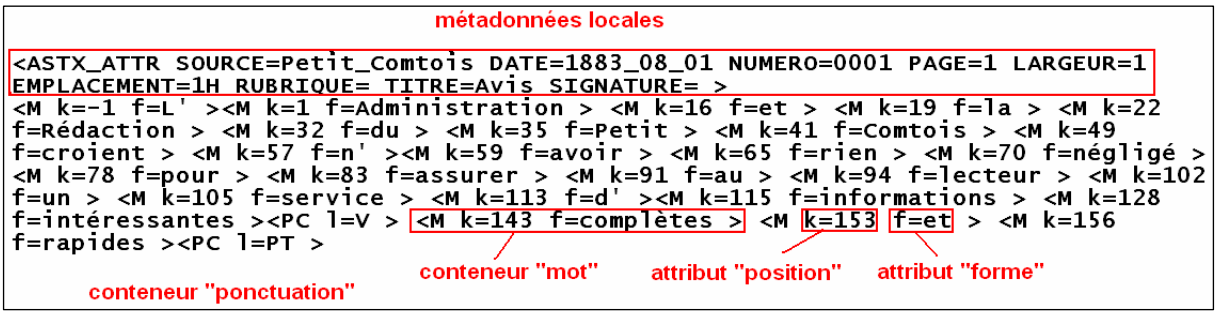

Fig. 1 : Exemple de codage après la passe de segmentation.

[6] Ainsi que des points et virgules dans des expressions numériques, notamment.

[7] Le cadre DiaTag permet de distinguer les corrections restituant l'état de la source (en cas d'erreur d'océrisation notamment), de celles qui au contraire s'en écartent (coquilles, etc).

[8] Lors de l'acquisition des données, on distinguera les balises   et la balise
, cette dernière destinée notamment, si le fichier n'est pas aux normes TEI, à marquer les vers, les versets, les sauts de lignes à l'intérieur d'une même réplique, etc.

2.2. Identification des séquences graphiques non reconnues

Une fois la segmentation effectuée, et éventuellement révisée, l'opérateur est invité à activer le recensement des formes, afin notamment de faire la liste de celles qui ne sont pas reconnues par les dictionnaires DiaTag. Un vocabulaire fréquentiel est constitué (consultable dans un tableur), ainsi qu'une liste indexée des formes non reconnues.

DiaTag propose alors de traiter cette liste dans une interface spécialisée où chaque item sera présenté en concordance. Sur cette interface, on peut préparer l'intégration d'une forme inconnue soit au dictionnaire du système, soit à un dictionnaire du corpus[9], et au-delà de cette seule forme, de l'ensemble des flexions correspondant à son lemme. On peut aussi noter l'appartenance de cette forme à une séquence composée, qui peut être un nom propre, ou indexer une forme en vue de sa vérification dans la source et de son éventuelle correction.

DiaTag génère sur ce mode deux fichiers dits d'*apports*, éditables en mode texte, constitués pour chaque item du lemme à intégrer et d'un code flexionnel. Il comporte un *fléchisseur*, qui génère à son tour une liste d'éléments désormais présentés au format *DiaTag* : forme, lemme, valeur flexionnelle. Cette liste, éditable, peut alors être intégrée à l'un des deux dictionnaires par une fonction DiaTag.

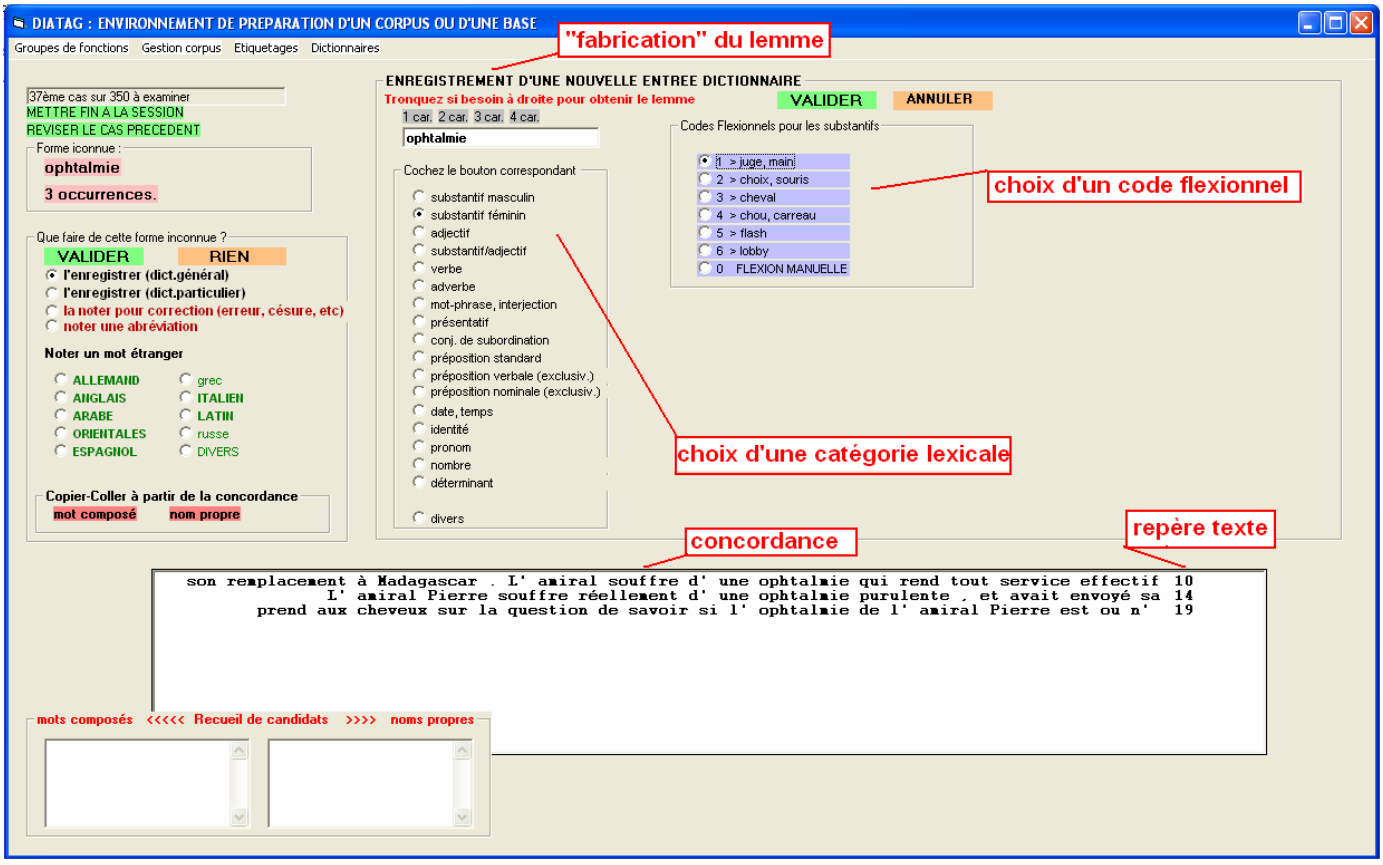

Fig. 2 : Interface de saisie d'une forme inconnue.

Si une ligne de l'apport concerne une forme graphique déjà recensée, deux cas se présentent : soit l'interprétation lexico-morphologique est déjà là, et l'apport est rejeté ; dans le cas

[9] Si cette acquisition risque de fausser le dictionnaire avec une idiosyncrasie, ou si l'on veut temporiser. Le dictionnaire de corpus sera utilisé au même titre que le dictionnaire du système pour annoter les ressources.

contraire, DiaTag ajoute cette interprétation et l'item, s'il n'était pas déjà ambigu, le devient[10].

Les formes marquées comme relevant de séquences composées donnent lieu au même processus interactif, où l'opérateur crée une liste d'apports pour le dictionnaire des mots composés du système et/ou pour un dictionnaire des noms propres spécifiques au corpus[11].

Il faut noter que cette passe est tout-à-fait utile pour la « correction » des textes. *Le Petit comtois* par exemple a été numérisé en bibliothèque, puis océrisé en laboratoire. De nombreuses coquilles ont été marquées lors de l'océrisation (selon le principe de la conservation de la graphie du document, même fautive, et de son signalement), mais pas toutes ; par ailleurs, les erreurs d'océrisation ont été rectifiées, mais partiellement. La plupart de ces deux classes d'omissions sont trahies par l'application d'un dictionnaire et peuvent dès lors être corrigées.

2.3. Annotation des formes composées

La reconnaissance et le marquage des formes composées sont considérés comme prioritaires par rappport à ceux des formes simples (ils permettent notamment de lever un nombre conséquent d'ambiguïtés graphiques). Néanmoins, ils restent facultatifs ; rien n'interdit de passer directement aux formes simples, notamment si l'on récuse cette notion ou si l'on veut lui appliquer une autre procédure simplifiée[12].

Lors d'une passe propre, DiaTag applique deux dictionnaires : l'un comportant les séquences qui sont toujours des composés (soit parce que l'un des constituants n'a aucune autre existence autonome, comme *parce que*, soit par jugement linguistique du concepteur du dictionnaire[13], comme par défaut dans DiaTag *à côté de*) ; s'il rencontre une telle séqence, DiaTag crée un conteneur <*MC k*=valeur *f*=valeur *l*=valeur *x*= > où l introduit le lemme et x la valeur flexionnelle si elle est pertinente[14] ; il supprime les conteneurs correspondant aux constituants[15].

L'autre dictionnaire comporte les séquences pouvant ou non être des composés (*bien que*, *au cœur de*, etc). DiaTag crée un conteneur <*MK k*=valeur *f*=valeur *l*=valeur *x*= >, tout en maintenant les conteneurs correspondant aux constituants. En même temps, ces marquages sont indexés en vue de la phase de levée conviviale des ambiguïtés ainsi repérées.

Cette levée s'opère en dialogue avec l'utilisateur. Toutes les occurrences sont présentées en concordance, type par type, en commençant bien sûr par les plus longs (en nombre de

[10] Nous laissons ici de côté la description technique du format du dictionnaire.

[11] Les noms propres sont traités dans les dictionnaires courants, soit comme formes simples, soit comme formes composées. Le dictionnaire des noms propres du corpus ne contient que des noms propres.

[12] Par exemple, un traitement préalable à la segmentation, au moyen d'underscores au lieu d'espaces, l'underscore étant considéré tant par Diatag que par Astartex comme un caractère interne s'il est précédé et suivi immédiatement d'une lettre ou d'un chiffre quelconque.

[13] Rien n'interdit à un utilisateur d'éditer le dictionnaire, voire d'en constituer un de toutes pièces, pourvu qu'il respecte le format prescrit.

[14] Le lemme inclut l'indication de catégorie même si celle-ci n'est pas ambiguë ; le format est du type AMOUR_SM, TENDRE_V, TENDRE_J, LIRE_V, LIRE_SF etc. La valeur flexionnelle est vide pour les catégories non fléchies.

[15] L'information graphique est retenue dans l'attribut f= du composé, où les blancs sont remplacés par des underscores.

constituants) de manière à éviter la présentation de candidats inclus dans des séquences à reconnaître d'abord (exemple : *jusqu'à la fin de / à la fin*). La concordance, triable selon le contexte droit ou gauche, est dotée de cases à cocher qui enregistrent la décison, validée au moment du passage au cas suivant. Un historique des décisions est créé, une fonction retour permet de réviser les 5 derniers cas, et une session peut être interrompue. Pour notre corpus, DiaTag a repéré (à partir de son dictionnaire, encore très lacunaire pour des composés de médiocre fréquence) 3 575 cas différents à traiter pour près de 80 000 occurrences (contre environ 73 000 occurrences de composés jugés non ambigus).

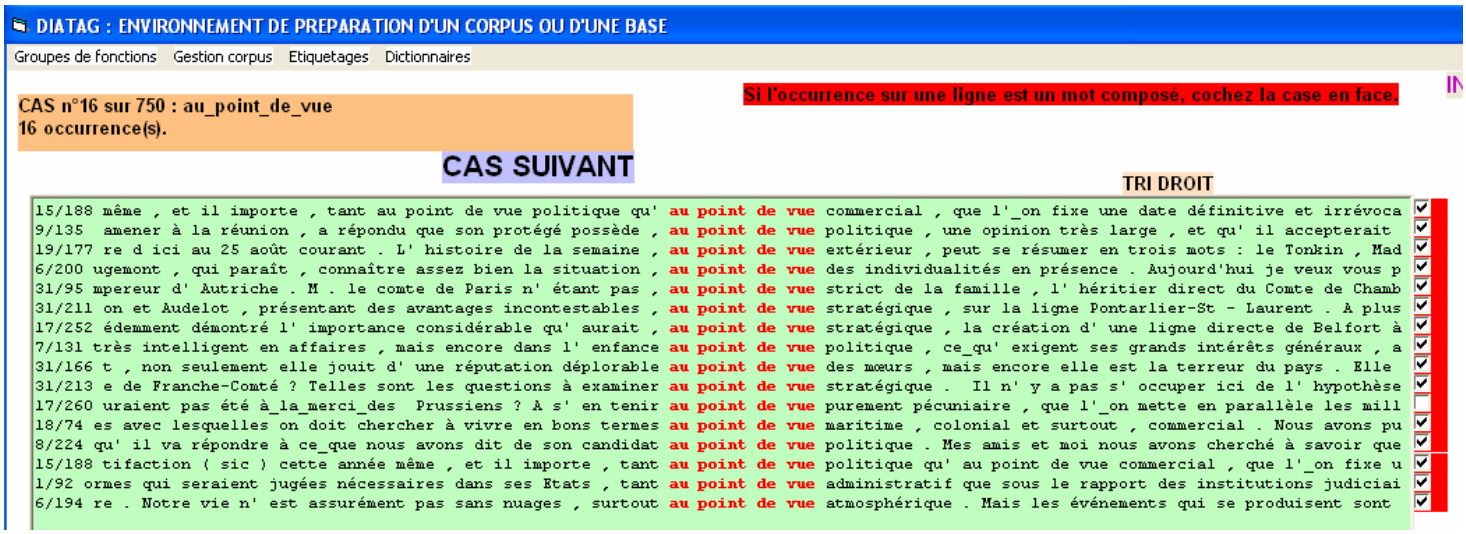

Fig. 3 : Interface de décision « mots composés ».

Si l'occurrence est validée, <MK est remplacé par <MC et les conteneurs des constituants sont supprimés. Si elle est rejetée, le conteneur <MK est supprimé[16]. Si une occurrence validée « contenait » à son tour un composé plus court, celui-ci est neutralisé, même s'il est un <MC.

Il est à noter que cette passe peut être recommencée à partir du dictionnaire ou à partir d'une liste restreinte. Les annotations <MC sont par principe réversibles. Rappelons aussi que les noms propres composés sont marqués à ce stade s'ils ont été recensés.

2.4. Annotation automatique des formes simples

L'annotation lexico-morphologique des formes simples comporte, elle aussi, une phase automatique, ou plus exactement deux. Tout d'abord, DiaTag applique son dictionnaire à toutes les unités <M > restantes. Il en résulte l'addition, dans chaque conteneur, d'un attribut l= et d'un attribut x=, sauf pour les formes non recensées (si l'identification des séquences inconnues et/ou la correction de la ressource n'ont pas été poussées à leur terme). L'attribut x= est vide si le lemme n'est pas fléchissable. L'attribut l= est un lemme à part entière (lemme + code de catégorie) si l'interprétation est univoque, mais il peut aussi avoir la forme ***valeur1/valeur2/valeurN*** s'il y a N interprétations possibles, non départagées à ce stade. C'est la phase d'application des scripts contextuels qui lève certaines de ces ambiguïtés. Le plus grand nombre possible d'indices contextuels sont employés pour résoudre chaque type

[16] En réalité, les conteneurs sont d'abord « vidés », afin de conserver l'indexation mot par mot en service jusqu'à la fin de l'opération globale. C'est seulement en toute fin que les conteneurs vides sont supprimés.

d'homographies, et ces indices sont recrutés sur la base de la certitude[17] (et non d'une quelconque probabilité).

Dans l'exemple de la fig. 4, *avoir* a été désambiguïsé par la précession du préverbal *ne*, et *négligé* par celle du pronom *rien*.

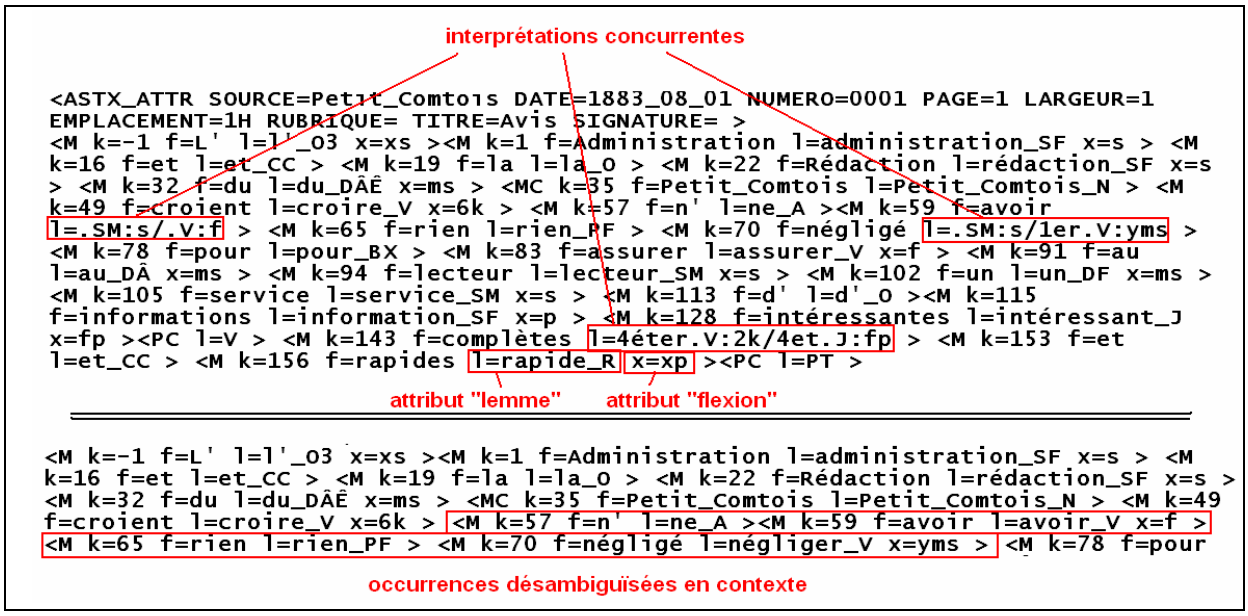

Fig. 4 : Application du dictionnaire, puis des scripts contextuels.

2.5. Annotation interactive des ambiguïtés résiduelles

Il faut bien noter ici qu'il y a deux sortes d'échecs à l'annotation automatique (qu'elle concerne ou non spécifiquement les formes simples). L'environnement peut renoncer à attribuer une interprétation parce que l'ambiguïté n'entre pas dans un schema de résolution jugé sûr à 100% (exemple ultime, *La petite brise la glace*). Mais un autre environnement choisira une interprétation, la plus probable selon des critères souvent fort divers et fort opaques ; en ce cas, ne sera considéré comme échec qu'un choix erroné ; le taux d'échec en ce cas est beaucoup plus bas que dans le modèle précédent, puisque les choix justes en sont décomptés, même s'il est difficile, voire impossible de savoir s'ils ont été atteints par une voie plus ou moins fondée en raison. Ce qui est sûr, c'est que si nous voulons vérifier les choix effectués, nous ne pourrons le faire qu'en révisant tous les cas ; par définition, les choix erronés, voire les choix très risqués mais justes, etc, ne seront pas marqués.

Dans une optique philologique, on souhaitera visualiser l'ensemble des cas où l'environnement n'est pas fiable à 100%.

Le corpus du *Petit Comtois* renvoie les données suivantes :

[17] Certitude fondée bien sûr sur une norme grammaticale du français écrit moderne, qui peut être prise en défaut par l'écriture créative ou par l'incertitude éditoriale, notamment. Exemple : précession obligatoire de l'élément conjugué du verbe par la séquence ordonnée des clitiques verbaux.

	Sans passe "composés"	% a	% b	Après passe "composés"	% a	% b
Nb de formes simples	4201796			4034607		
Non reconnues	7729	0,18%		7728	0,19%	
Ambiguës	549130	13,07%		531421	13,17%	
Identifiées par le contexte	275759		50,22%	265760		50,01%
Simplifiées par le contexte	77466		14,11%	81129		15,27%
Résiduelles	195905	4,66%	35,68%	184532	4,57%	34,72%

Fig. 5 : Statistiques de la reconnaissance des formes simples.

La levée des ambiguïtés résiduelles s'effectue dans le même cadre ergonomique que celle des mots composés. Au lieu d'une seule case à cocher, chaque ligne de concordance comporte autant de cases de couleurs différentes qu'il reste d'interprétations pour cette série de cas. Les mêmes procédures d'enregistrement des résultats sont mis en œuvre ; le lemme et l'information flexionnelle sont stockées progressivement dans les conteneurs de « mots ».

2.6. Annotation des formes verbales auxiliées

Dans une perspective textométrique (*logométrique* Mayaffre 2002) il est essentiel de pouvoir reconnaître et recenser les formes verbales auxiliées, à commencer par les passés composés et les passifs canoniques. Le passé composé constitue l'un des marqueurs privilégiés de certaines oppositions génériques fondamentales, ainsi d'ailleurs que l'aspect accompli que certaines de ses occurrences partagent avec le paradigme du passé composé (plus que parfait, passé antérieur, etc). Le passif, quant à lui, est très important pour esquisser une approche de l'organisation thématique/rhématique du texte.

Le simple recensement des participes passés, dans la mesure déjà où il serait juste (ce qui suppose de donner toute sa rigueur à la phase précédemment décrite), ne donne qu'une idée très approximative et surtout amalgamante de ces deux réalités distinctes.

Le passif ne se résume en aucun cas aux formes canoniques *être*+part.passé. On voudra aussi reconnaître les formes *se faire*+inf., les participes passés sans auxiliaire à valeur passive, etc.

La modalité en français s'exprime en grande partie grâce à l'importante série des dits « semi-auxiliaires » (*pouvoir, devoir*…) sans parler des nombreux cas où il est conséquent de distinguer les emplois modaux des emplois pléniers (*aller, -se-voir, venir-de-* etc).

DiaTag propose (sans l'imposer en première intention) un appareil de marquage de tous ces faits, qui repose sur le même principe de dialogue ergonomisé avec l'opérateur expert. On peut, de même, distinguer les emplois des principaux verbes français entre « verbes supports » et verbes pleins (*prendre, mettre*, etc).

2.7. Interopérabilité, réversibilité, authentification

Dans tous les cas, l'opérateur expert sera amené à prendre des décisions, dont chacune pourra être sujette à caution. Deux types de biais peuvent se présenter.

Les décisions peuvent tout d'abord être plus ou moins fautives. Pour aborder l'une quelconque des phases de l'annotation DiaTag, il est nécessaire de détenir un point de vue informé sur la/les notion-s en jeu. Qu'est-ce, par exemple, qu'un *mot composé* ? Qu'est-ce éventuellement qu'un *verbe support* ? DiaTag contient, par ses dictionnaires, ses listes et ses scripts, des choix implicites quant à ces notions, choix qui sont eux aussi critiquables et que nous cherchons du moins à expliciter du mieux que nous pouvons, dans le manuel

d'utilisation. Mais rien n'empêche l'opérateur de se référer à d'autres interprétations grammaticales, pourvu qu'il (se) les explicite au mieux et qu'il sache organiser ses critères. Par ailleurs, il faut noter que DiaTag insère un item de *header* qui identifie l'/les auteur-s de l'annotation, la date du travail et tout autre paramètre que l'opérateur voudra enregistrer.

Un second biais, qui est déjà effleuré implicitement ci-dessus, concerne toutes les opérations d'annotation qui ne sont pas explicitement prévues. Il est certain qu'une fois la ressource annotée, les faits marqués choisis par DiaTag prennent une importance et un relief potentiels, dans l'hypothèse d'une exploration assistée, au détriment de ce qui est laissé dans l'ombre.

Sur ce point, on ne peut répondre que ceci : l'environnement livre un/des fichier-s dans un format de niveau au moins égal à celui qu'il a saisi initialement. S'il a reçu un fichier XML, il rend un fichier XML enrichi. S'il a reçu un fichier TXT, il rend un fichier « pseudo »-XML où toute l'annotation est protégée dans des conteneurs explicites (y compris les éléments d'annotations antérieures pourvu qu'ils aient été eux-mêmes protégés de cette façon). Ainsi, rien n'interdit de poursuivre l'enrichissement, par exemple par des marqueurs d'analyse de contenu, ou des marquages grammaticaux non prévus par DiaTag, ou encore des marqueurs de structure et de métadonnées externes. Ces éléments sont cumulatifs (sauf si l'opérateur à un moment donné décide de détruire des éléments importés) dans l'esprit de la TEI.

La réversibilité du marquage est assurée par l'enregistrement de toute l'information utile (voire dans certains cas redondante). Bien au-delà de la simple réversibilité, l'objectif est ici d'assurer l'alignement mutuel le plus complet et opérationnel possible de tous les marquages en vue de leur utilisation dans l'exploration.

3. Bilan et perspectives

L'expérience que nous avons menée sur une base de presse disponible au départ sous la seule forme d'un fonds de bibliothèque d'étude, avec une chaîne d'opérations mixte[18], tend à prouver que le marquage linguistique poussé et contrôlé est possible sur des données de grand voire de très grand volume. Si l'on tient pour admise la nécessité d'un accès plein-texte et structuré par des métadonnées, et si l'on admet que ces objectifs relèvent fondamentalement de la responsabilité des services d'archives et de documentation, la part revenant au laboratoire, au chercheur orienté vers les requêtes permises par l'annotation linguistique (et plus largement : experte) représente un surcoût de 12%[19] en temps, inférieur donc à 20 % en

[18] Le scannage s'effectue à la B.M.E. sur un matériel financé pour l'Université par la Région, dans le cadre d'un protocole précis et négocié. Les fichiers image sont d'une part aiguillés vers les archives électroniques de la Ville, où ils sont destinés à être mis en ligne au profit de lecteurs ne souhaitant consulter que les fac-similés, remplissant ainsi la mission de conservation/préservation et de communication des documents ; d'autre part ils nous sont livrés en vue de l'océrisation et de l'enrichissement destinés à l'accès en mode plein-texte qui est bien sûr à l'horizon de cette coopération conventionnelle.

[19] Tableau comptable des opérations sur *Le Petit Comtois*

	Qualification	Heures passées par million de mots	Pourcentage du temps total
numérisation-TIFF	agent bibliothèque	22	26 %
océrisation_WORD	technicien	35	41 %
vérification, balisage métadonnées	ingénieur/doctorant	15	18 %
annotation linguistique	doctorant/chercheur	13	15 %
TOTAL		85	100 %

budget. En regard, la valeur ajoutée peut être bien supérieure, et cela d'autant plus que ces opérations se font de manière systématique, communautaire, réversible, et cumulative, tous paramètres largement favorisés par l'âge numérique.

Il reste bien sûr à prévoir l'ajustement des outils d'analyse et d'exploration, face à cette disponibilité élargie de très grandes ressources. Ainsi que les réponses conceptuelles et logicielles aux nouveaux besoins, aux nouvelles requêtes qui ne manqueront pas de surgir. Il faudra aussi resserrer les liens, sur ces bases rénovées (reconnaissance de la nécessité de phases non-automatiques et du réalisme de leur mise en œuvre), avec la linguistique informatique et/ou de corpus, et envisager la généralisation de ces principes à l'ensemble des langues et états de langues.

Dans un futur proche, nous espérons en tout cas fournir aux communautés des sciences humaines une base annotée selon les lignes indiquées ci-dessus, de 250 millions de mots, pour une période s'étalant de 1883 à 1942.

Références

Adam J.-M., Heidman U. (2005). *Sciences du texte et analyse de discours : enjeux d'une interdisciplinarité.* Genève, Slatkine.

Brunet E. (2000). Qui lemmatise, dilemme attise. In *Lexicometrica* n°2.

Brunet E. (2002). Le lemme comme on l'aime. In *JADT 2002.* Saint Malo, Inria.

Habert B. (2005). *Instruments et ressources électroniques pour le Français*. Paris, Ophrys.

Habert B., Nazarenko A., Salem A. (1997). *Les linguistiques de corpus*. Paris, A.Colin.

Léthier V. (2007). Constitution d'un corpus de presse régionale du XIXe siècle : pratiques et enjeux. Communication à la *Journée des jeunes chercheurs, doctorales Jec'SIC, Association ALEC-SIC* http://www.alecsic.fr/.

Mayaffre D. (2006). Faut-il prendre en compte la composition grammaticale des textes… In *JADT 2006.* Besançon, PUFC.

Mellet S. (2003). Lemmatisation et encodage grammatical : un luxe inutile ? In *Lexicometrica*, numéro spécial.

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

Viprey J.-M. (2004). DIATAG : convivialité d'étiquetage des ambigus résiduels (communication aux journées Intex 2002, Marseille). In *Intex : pour la linguistique et le traitement automatique des langues*, éd. M.Silberztein. Besançon, PUFC.

Viprey J-M. (2005). Philologie numérique et herméneutique intégrative. In *Sciences du texte et analyse de discours : enjeux d'une interdisciplinarité,* dir. Jean-Michel Adam & Ute Heidman. Genève, Slatkine : 51-68.