

# Une archive numérique de comptes rendus de conseils d'administration universitaire (1986-2018) annotée pour l'analyse de la représentation des prises de parole

Timothée Premat<sup>1,2</sup>, Cyrielle Montrichard<sup>1</sup>, Frédérique Sitri<sup>1\*</sup>, Virginie Lethier<sup>3</sup>

<sup>1</sup>Université Paris-Est Créteil, *Céditec*, 61 avenue du Général de Gaulle, 94010 Créteil, France

<sup>2</sup>Université de Neuchâtel, *chaire de dialectologie*, Pierre-à-Mazel n°7, 2000 Neuchâtel, Suisse

<sup>3</sup>Université Marie & Louis Pasteur, *ELLIADD*, 30-32 rue Mégevand, 25030 Besançon, France

**Résumé.** Cet article documente la constitution et l'annotation d'une vaste archive de comptes rendus de conseils d'administration de l'université Paris Nanterre (1986-2018), réalisée dans le cadre du projet ArchivU, projet qui vise à interroger les évolutions de l'institution universitaire sous l'angle de l'analyse du discours. Après avoir présenté les caractéristiques de l'archive numérique mise à disposition de la communauté scientifique sur la plateforme Ortolang, nous expliquons les enjeux de l'annotation de la représentation des tours de parole des membres du conseil, représentation constitutive d'un genre ayant pour fonction non seulement d'enregistrer les décisions prises mais aussi de représenter les dires tenus lors de la réunion et les prises de position des uns et des autres. Nous présentons ensuite la méthodologie d'annotation semi-automatique des tours de parole des membres du conseil et nous en évaluons les résultats. Enfin, nous exposons quelques exemples d'exploitation des possibilités offertes par l'annotation des tours de parole.

## 1 Introduction

Le projet ArchivU [1] vise à étudier les transformations de l'Université française de 1971 à nos jours à partir de l'analyse de l'évolution de deux genres de textes, le compte rendu (CR) de conseil d'université (avant 1986)<sup>i</sup> ou de conseil d'administration (CA, depuis 1986) et le rapport de laboratoire, genres étroitement liés à l'institution qui les produit, et dont on fait l'hypothèse qu'ils participent à construire une image de cette institution. Nous nous intéressons, dans le présent article, uniquement au corpus de comptes rendus de conseils d'administration de l'université de Nanterre.

Pour explorer ces données selon les méthodes de l'analyse du discours outillée par l'informatique et la statistique [2] et en garantissant la pérennité et l'interopérabilité, les archives étudiées dans le cadre d'ArchivU ont fait l'objet d'un processus de numérisation, de structuration et d'annotation fine selon la norme XML-TEI P5<sup>ii</sup>, puis ont été déposées sur la

---

\* Auteur correspondant : [frederique.sitri@u-pec.fr](mailto:frederique.sitri@u-pec.fr)

plateforme de partage Ortolang, de sorte à les rendre accessibles et réutilisables par la communauté scientifique. Au total, ce sont 351 comptes rendus, produits entre 1986 et 2018, et représentant 2 848 375 occurrences, qui ont été traités. Dans une optique de mise à disposition des données de la recherche (principes FAIR), nous souhaitons que les données de l'archive numérique déposée (ou « base » [3]) puissent être extraites, sélectionnées et réorganisées en nouveaux corpus d'étude, construits en fonction d'objectifs scientifiques différents des nôtres<sup>iii</sup>.

À cet effet, nous nous proposons dans cet article de documenter les opérations de philologie numérique [4] qui ont présidé à sa constitution et à son annotation. Cette dernière répond à deux objectifs principaux : d'une part, préserver les informations relatives à la structure physique des textes ; d'autre part, enrichir les informations liées aux propriétés structurelles du genre, telles que la représentation des prises de parole des membres du conseil. Nous présenterons tout d'abord les caractéristiques de l'archive numérique mise à disposition de la communauté scientifique sur la plateforme Ortolang, avant d'exposer les enjeux de l'annotation de la représentation des tours de parole dans le genre du compte rendu et de présenter la méthodologie semi-automatique choisie. Ensuite, nous évaluerons la qualité de l'annotation semi-automatique des tours de parole effectuée à partir d'un corpus échantillon. Enfin, nous présenterons quelques exemples d'exploitation de cette annotation en relation avec nos objectifs de recherche.

## 2 Présentation de l'archive numérique ArchivU

### 2.1 Des archives à l'archive numérique

Documents administratifs destinés à être archivés [5] dans le cadre d'une politique de « transparence » caractéristique des institutions publiques, les CR de CA de l'université Paris Nanterre sont inscrits sur différents supports d'origine en fonction de leur date de publication. Jusqu'en 1988, ils sont tapés à la machine sur des feuilles (libres ou collées dans des cahiers). Certains de ces documents contiennent des traces d'annotations manuscrites qui témoignent d'une chaîne d'écriture collective [6]. Le passage au traitement de texte, en 1988, n'engendre pas de changement de support de conservation et de diffusion. Ce n'est en effet qu'à partir de 2008 que les CR sont conservés et diffusés sur l'intranet sous forme de PDF.

Les CR de la période 1971-2008, préalablement numérisés, ont été océrisés par lots, tout comme les PDF de la période 2008-2018. Les données ont été converties automatiquement au format XML et annotées selon les recommandations TEI-P5, pour être ensuite soumises à un processus de vérification manuelle de la fidélité de l'état de texte numérique à celui du support d'origine, impliquant :

- de contrôler la graphie des formes lexicales reconnues automatiquement et de rétablir, lorsque nécessaire, la graphie fautive observable sur le support d'origine ;
- de documenter la structuration physique du document (paginations, annotations, etc.) ;
- d'annoter la structure logique participant à l'organisation du texte et à la construction du sens (paragraphes, tableaux, listes, sections et sous-sections, titres et sous-titres, etc.).

Ont également été distingués le corps de texte à proprement parler (<body>) et le texte préliminaire (<front>) dont relève par exemple, dans le genre du CR, la liste des présents. La balise <back>, recommandée pour l'annotation des annexes, n'a pas été utilisée car ces dernières, regroupant des documents variés (motions, tableaux de données chiffrées, etc.), n'ont pas été intégrées à l'archive numérique conformément à l'hypothèse d'homogénéité générique sur laquelle est construit le projet.

Pour les CR antérieurs à 1988, la qualité de l'océrisation, particulièrement dégradée en raison du caractère artisanal de la composition du texte et de la présence d'annotations manuscrites, a impliqué un important travail de correction des graphies et des marques de mise en forme indûment « reconnues » par le logiciel d'OCR<sup>iv</sup>, notamment en ce qui concerne les tableaux, souvent mal identifiés. À ce jour, les données de la période 1971-1984, toujours en cours de vérification, n'ont pas encore été intégrées à l'archive numérique.

## 2.2 Enjeux d'une annotation adaptée aux particularités génériques du compte rendu

Nous situant dans le cadre de la représentation du discours autre (RDA), par lequel Authier-Revuz [7] redéfinit le champ du discours rapporté, nous analysons le compte rendu de réunion comme un genre « de la RDA », c'est-à-dire un genre pour lequel la relation de représentation se situe entre le texte écrit et la réunion qu'il représente [8]. De ce fait, même en l'absence de marque explicite, le texte est interprété comme représentant les dires tenus pendant la réunion. Si tout compte rendu peut « tenir lieu » de la réunion qu'il représente, le CR de CA, une fois validé, devient procès-verbal et acquiert une valeur juridique de « substitut » [9] : il « vaut pour » la réunion qu'il représente. Le CA étant un organe législatif, les décisions validées par vote en son sein s'appliquent à la communauté universitaire. Mais le CA est également une instance politique, dont les membres élus représentent – avec des pondérations différentes selon les époques – les différentes catégories constituant la communauté universitaire, de sorte que les prises de parole doivent aussi s'entendre en relation avec cette dimension de représentation ou de porte-parolat. Étant donné l'importance de cette instance, la rédaction des CR de CA est assurée par un personnel dédié et est relativement normalisée par rapport aux comptes rendus d'autres instances universitaires. À l'université de Nanterre, cette rédaction est assurée par un prestataire de service externe à l'université à partir de 2013 ; sur l'ensemble de la période, le mode de RDA dominant est le discours indirect (DI).

Par conséquent, annoter la structure logique des CR de CA implique de prendre en compte la structuration du texte par l'alternance de la représentation des tours de parole. Or, si elles constituent un cadre de référence général pour la description et le partage des données textuelles, les recommandations TEI ne comportent pas de balise correspondant au cas de figure du compte rendu de réunion. La TEI présente en revanche l'avantage de sa souplesse et de son adaptabilité aux données qu'il s'agit de décrire. Ainsi, nous avons décidé d'utiliser la balise <u>, empruntée au module dédié à l'oral, conformément au choix opéré pour la description et l'annotation de la RDA (au discours direct) du corpus ParlaMint [10].

## 3 L'annotation des tours de parole : méthode et évaluation

### 3.1 Méthode

Nous avons développé un outil de repérage et d'annotation de la représentation des tours de parole (désormais « tours de parole ») qui s'appuie sur un ensemble de patrons lexico-syntaxiques identifiés comme marquant l'ouverture d'un tour, à partir d'un corpus échantillon couvrant toute la période. Ces patrons, enrichis par des indices typographiques (saut de ligne, majuscules, mise en gras, etc.), ont constitué les règles à partir desquelles a été développé un programme Python, repérant, dans nos données préalablement étiquetées d'un point de vue morphosyntaxique, l'ouverture d'une séquence au discours direct (*M*|*M*<sup>me</sup> + : + « ... »)<sup>v</sup>, au discours indirect (*[M*|*M*<sup>me</sup> + nom propre] ou [*Article défini* + *Président*|*Conseil*] + verbe introducteur de parole) ou encore correspondant à une

modalisation en discours second (*Selon* +  $M|M^{me}$  + *nom propre*). Des phases de tests menées sur un ensemble élargi de données ont permis d'enrichir notre système de règles afin de l'adapter aux variations des modalités d'ouverture d'un tour observées en diachronie.

Très concrètement, lorsque le programme rencontre un patron lexico-syntaxique introduisant un tour de parole, il ouvre une balise <u> et renseigne l'identité du locuteur dont la parole est représentée (attribut *NormalizedName*) en fonction du nom qui précède le verbe. Le programme attribue ensuite un statut de locuteur (attribut *label<sup>vi</sup>*) en fonction de la table des présents, organisée par collèges. Enfin, il attribue un genre au locuteur représenté (attribut *gender*) à partir du genre grammatical du titre, abrégé ou non, qui précède le nom (M., M<sup>me</sup>, M<sup>lle</sup>, etc.). Le programme ne ferme l'environnement <u> que lorsqu'il rencontre de nouveau une construction introduisant un nouveau tour de parole, ou lorsqu'il rencontre un élément structural (titre, fin du document, figure, etc.) ou un énoncé rituel dont le patron a été intégré au programme (p. ex. *La séance est interrompue|levée*).

Nous avons ensuite mené une campagne de révision visant à améliorer le taux d'identification des statuts des locuteurs (*label*). Notre programme se basant sur l'identification du locuteur pour lui attribuer un statut, les nombreux cas où le locuteur était mal identifié engendraient une non-identification de son statut. Or, la correspondance entre le nom en plein texte (source du *NormalizedName*) et le nom dans la table des présents (source du *label*) est loin d'être systématique, notamment en raison de coquilles dans la source. Pour certains CR de la décennie 2000, l'absence d'une table des présents empêchait l'assignation d'un statut aux locuteurs représentés. Notons d'ailleurs que les tables des présents ne sont pas toujours exhaustives ; c'est donc à partir d'un travail de recherches historiques et/ou intertextuelles dans les CR que nous avons rétabli le statut du locuteur. Nous avons également corrigé de nombreux cas où un tour de parole était attribué à une organisation (p. ex. un syndicat au lieu du représentant syndical qui s'exprime, avec des formules comme *L'UNEF s'oppose à*) ou à un locuteur (*Monsieur Jospin souligne*) dont les propos étaient rapportés par un membre du conseil (en l'occurrence, le président, qui relate les propos de M. Jospin lors d'une réunion).

Lors de cette vague de correction, et notamment à l'occasion des nombreux retours au texte qu'elle a impliqués, nous avons également pu corriger manuellement des positions de balises, des identifiants de locuteurs et leur genre, ou encore fusionner certaines catégories dont la dénomination est fluctuante en diachronie (p. ex. ATOS-ITA, ATOS, IATOS et BIATSS).

Cette campagne de révision a amélioré sensiblement le taux de reconnaissance des informations relatives au statut du locuteur, dans la mesure où le nombre de mots correspondant à un tour de parole sans statut de locuteur est passé de 30 % à 10 % (pourcentages donnés sur l'ensemble des mots du corpus). Nous avons également pu introduire une distinction, dans le post-traitement, entre les cas où un locuteur est identifié mais n'a pas de statut (*label="other"*, 7 % du corpus) et les cas où le locuteur n'est même pas identifiable (*label="unknown"*, 3 % du corpus). Cette phase de post-traitement a aussi impliqué la fusion de tours de parole successifs correspondant au même locuteur : l'algorithme évoqué plus haut ferme et ouvre en effet un environnement <u> à chaque fois qu'un patron d'ouverture de tour de parole revient, même avec un même locuteur. L'état de corpus issu de cette phase de post-traitement est celui qui est distribué sur Ortolang. Nous avons évalué la qualité de l'annotation selon une méthodologie détaillée ci-dessous.

### 3.2 Évaluation

Pour mesurer la qualité de l'annotation semi-automatique, nous avons mis en place une procédure d'évaluation qui consiste à vérifier manuellement chaque tour de parole, par la

mise en parallèle du document TEI et du facsimilé original, et l'observation de la présence de la balise, de sa position et du locuteur identifié. Trois cas ont été considérés :

- annotation inexacte : un écart existe entre ce que les marques formelles indiquent en termes de tours de parole et l'annotation XML-TEI. L'extrait (1) ci-dessous montre un passage d'un tour de parole de M<sup>me</sup> Duthu qui a été traité par l'algorithme comme étant partie intégrante du tour de parole du président.

(1) <p gender="M" NormalizedName="president" label="presidential\_office">  
 <p><hi rend="bold"> Le <roleName>Président</roleName></hi> ajoute qu'un certain nombre d'entreprises ou banques ont été contactées pour un éventuel partenariat. </p>  
 <p><hi rend="bold"><persName>Madame DUTHU</persName></hi> fait remarquer que les formations d'enseignement supérieur sont de la compétence de la région et non du département.</p>  
 </u>

(Extrait du document XML-TEI du CR de CA du 27 janvier 1992)

- annotation acceptable : il n'y a pas suffisamment de marques formelles permettant de remettre en cause l'annotation. La séquence « Cette proposition est acceptée » dans l'exemple (2) ci-après montre un cas où un tour de parole est ouvert et attribué à un locuteur non identifié. Au vu de notre connaissance du genre et de la situation d'énonciation – indiquant une demande de report d'un vote formulée par le président au conseil –, on pourrait considérer ici qu'il aurait été plus juste d'attribuer cette séquence au conseil. Cependant, en l'absence d'indices formels, nous considérons que l'annotation opérée n'est pas inexacte et est donc acceptable.

(2) <p gender="M" NormalizedName="audeoud" label="presidential\_office">  
 <p><hi rend="bold"><persName>Monsieur AUDEOUD</persName></hi> lui répond qu'il va les poser après.</p>  
 <p>Il propose aux membres du Conseil d'Administration de remettre le vote sur le PV du <date when="2008-12-17"> 17 décembre 2008 </date> à la prochaine séance.</p>  
 </u>  
 <u gender="unknown" NormalizedName="unknown" label="unknown">  
 <p>Cette proposition est acceptée.</p>  
 </u>

(Extrait du document XML-TEI du CR de CA du 28 janvier 2008)

- annotation valide : l'annotation correspond aux marques formelles présentes dans le texte, comme c'est le cas pour l'extrait (3) par exemple.

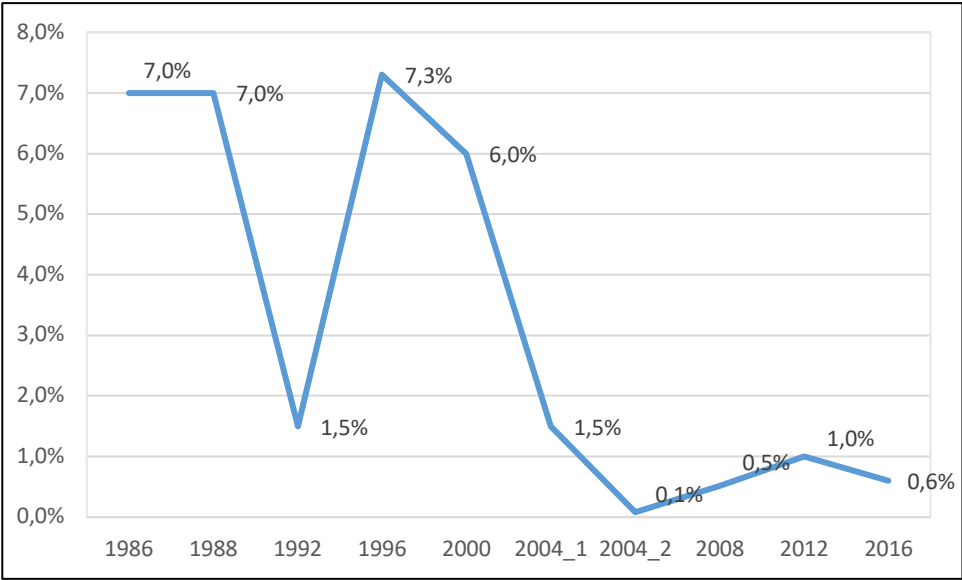
(3) <p gender="M" NormalizedName="laccassagne" label="college\_etudiant">  
 <p><hi rend="bold"><persName>M. LACCASSAGNE</persName></hi> pense que ce débat, traditionnel au moment des élections, est inutile. Ces personnes sont irresponsables.</p>  
 </u>  
 <u gender="M" NormalizedName="massengo" label="college\_etudiant">  
 <p><hi rend="bold"><persName> M. MASSENGO</persName></hi> condamne fermement les excès commis dans la nuit du 16 janvier, qui sont dus à des personnes extérieures à l'Université. Sans excuser ce qui s'est passé le lendemain, ces excès l'expliquent en partie.</p>  
 </u>  
 <u gender="F" NormalizedName="duthu" label="college\_b">  
 <p><hi rend="bold"><persName> Mme DUTHU</persName></hi> déclare que rien ne justifie ce genre de comportement.</p>  
 </u>

(Extrait du document XML-TEI du CR de CA du 29 février 1996)

Seuls les cas d'inexactitude ont été pris en compte pour évaluer l'annotation des tours de parole. Le contrôle manuel a été effectué sur dix comptes rendus – pour un total de 141 pages,

73 614 mots et 1 047 balises de tours de parole – sélectionnés à intervalles réguliers sur les années 1986, 1988, 1992, 1996, 2000, 2004, 2008, 2012 et 2016. Les annotations inexactes constituent des erreurs d’annotation.

Pour comptabiliser le taux d’erreur, deux méthodes de calcul ont été mises en place, avec une lecture en double aveugle des données afin de prévenir les oublis et l’erreur humaine. Cependant, nous n’avons pas mobilisé les outils de calcul d’accord inter-annotateur de type Kappa de Cohen dans la mesure où il ne s’agissait pas d’interpréter les données mais de vérifier la justesse de l’annotation, selon les principes exposés ci-dessus. La première méthode, dont les résultats sont exposés dans la figure 1, consiste à calculer le nombre d’occurrences qui sont intégrées à des tours de parole correspondant à une annotation inexacte. Cette méthode de décompte permet d’obtenir un pourcentage d’occurrences erronées par rapport au nombre d’occurrences total de chaque CR.



**Fig. 1.** Pourcentage d’occurrences inexactement annotées par CR sur les années traitées

En moyenne, le taux d’occurrences inexactement annotées sur l’échantillon traité est de 2,2 %, résultat qui semble satisfaisant en l’état. En appliquant l’intervalle de confiance de Wilson (adapté aux petits échantillons et aux données de type binaire comme c’est le cas ici : annotation exacte *versus* inexacte), les résultats obtenus montrent que l’on peut s’attendre à un taux d’erreur compris entre 2,1 % et 2,3 % sur l’ensemble des données. Ce taux d’erreur semble particulièrement acceptable si on le compare aux différents travaux concernant d’autres types d’annotation : par exemple, on trouve pour les annotateurs morphosyntaxiques des taux de fiabilité aux alentours de 97 % dans le cas de l’annotation en partie du discours [11]. Par ailleurs, on peut considérer que ce taux ne viendra pas perturber les résultats statistiques globaux obtenus sur le corpus.

Notons encore que le pourcentage d’erreurs diminue en diachronie, passant de 6 % sur la période 1986-2000 à 0,6 % entre 2004-2016. Cette évolution est intéressante dans la mesure où elle permet d’observer que l’algorithme, bien que conçu à partir d’un échantillonnage aléatoire en diachronie, semble plus adapté aux années les plus récentes. En effet, on note une évolution de la pratique du CR vers un patron standardisé du DI garantissant l’attribution quasi-systématique des dire. Ainsi, lors de l’étape d’exploitation, il s’agira de tenir compte

du taux d’erreur plus important sur la période 1986-2000 afin d’éviter de surinterpréter les résultats qui seraient peu significatifs.

La seconde méthode vise à quantifier le nombre d’erreurs en fréquence brute selon le type d’erreur. Nous nous sommes appuyés sur une typologie inspirée des travaux visant à qualifier les erreurs dans les textes générés automatiquement [12], plus proche de nos besoins de classification que d’autres typologies issues du même domaine [13] ou liées à la traduction [14]. La catégorisation consiste à évaluer si l’inexactitude tient à l’omission d’une balise, l’addition d’une balise non essentielle, le remplacement d’un locuteur par un autre, ou la position des balises (ouvertes ou fermées au mauvais endroit). Dans un second temps, une fois l’erreur source observée, il s’agissait de voir si elle en entraînait d’autres, que nous désignons comme secondaires. Par exemple, pour l’erreur d’annotation observée dans l’extrait (1), l’omission du tour de parole de M<sup>me</sup> Duthu – erreur primaire – conduit à inclure le tour de parole de M<sup>me</sup> Duthu dans celui du président – erreur secondaire.

Le tableau ci-après rend compte du nombre total d’erreurs repérées dans l’échantillon et de leur type.

**Tableau 1.** Type et nombre d’erreurs primaires et secondaires

| Type d’erreur primaire                           | Nombre d’erreurs primaires | Nombre d’erreurs secondaires associées |
|--------------------------------------------------|----------------------------|----------------------------------------|
| Addition (la balise ne devait pas être présente) | 3                          | 1                                      |
| Omission (une balise a été oubliée)              | 9                          | 1                                      |
| Erreur dans la position des balises              | 21                         | 27                                     |
| Erreur de locuteur dans le contenu de la balise  | 4                          | 0                                      |

Le tableau permet de relever que le problème principal réside davantage dans la position des balises que dans le repérage du début et de la fin du tour. Tout en étant un point fort de la méthode d’annotation (qui omet peu de tours de parole et identifie correctement le locuteur associé), les erreurs liées à la position des balises entraînent un nombre d’erreurs secondaires plus important (1,3 erreur secondaire pour une erreur primaire). Plus précisément, un tri de ces cas montre que plus de 90 % d’entre eux concernent une mauvaise position de la balise de fin de tour de parole. Les problèmes liés à la position des balises sont pour l’essentiel dus à la difficulté qu’il peut y avoir à identifier la clôture d’une intervention dans les cas où elle ne correspond pas à un nouveau tour. Ainsi, alors que les indices de début d’une prise de parole sont généralement faciles à identifier de façon automatique, ce n’est pas le cas des indices de fin, sauf quand il y a une nouvelle prise de parole.

3.3 Discussion

Si l’on se penche sur l’origine probable des erreurs d’annotation en tours de parole, on peut relever plusieurs cas de figure.

Un certain nombre de cas sont explicables par une défaillance de l’annotation morpho-syntaxique. Ainsi dans l’exemple (1) ci-dessus il est probable que la locution « fait remarquer » n’ait pas été identifiée car « fait » n’aurait pas été annoté comme un verbe. À l’inverse, dans l’extrait suivant, le début d’un tour a été repéré alors que le verbe était au passé composé, sans doute parce que l’auxiliaire a été analysé comme un présent.

(4) `<head n="title_lvl_2">`  
`<hi rend="underline"> Les I.U.T.</hi> :`  
`</head>`  
`<u gender="M" NormalizedName="piejus" label="presidential_office">`  
`<p><hi rend="bold"><persName> Monsieur PIEJUS</persName></hi>a été informé,`  
`par un conseiller du <roleName>Recteur</roleName>, de l'accord passé entre l'Etat,`  
`représenté par le Préfet de Région, et le Conseil Régional pour l'implantation de`  
`départements d'IUT en Ile de France. </p>`  
`</u>`

(Extrait du document XML-TEI du CR de CA du 27 janvier 1992)

Par ailleurs, le marquage par un tour au passif est une source d'erreur. Dans l'exemple (5), le tour de parole de M<sup>me</sup> Le Pellec devrait s'arrêter avant la représentation des opérations de vote, qui constitue un autre tour de parole, attribuable à un autre locuteur (le président ou son représentant pour l'appel au vote et le conseil pour le vote lui-même). Le marquage par deux tours au passif, « soumis au vote » et « sont approuvés » – dont un en apposition –, n'a pas été identifié :

(5) `<u gender="F" NormalizedName="lepellec" label="college_b">`  
`<p><hi rend="bold"><persName>Mme LE PELLEC</persName></hi> précise que les`  
`financements des services Médecine préventive et Handicap concernent : </p>`  
`<list rend="bulleted">`  
`<item>pour la Médecine préventive : des actions dans le domaine de la contraception,`  
`des journées d'action et de mobilisation contre le VIH, un groupe de travail sur la prise`  
`de parole, des formations aux premiers secours, </item>`  
`<item> pour le Service Handicaps : il s'agit d'actions pour les non voyants et au titre`  
`de la Langue des Signes (matériels spécifiques, petites fournitures...)</item>`  
`</list>`  
`<p><hi rend="bold"> Les financements proposés par la Commission Sociale`  
`d'Etablissement, soumis au vote, sont approuvés à l'unanimité des membres présents ou`  
`représentés moins </hi> <hi rend="bold"> 1 abstention (Cf annexe 2). </hi> </p>`  
`</u>`

(Extrait du document XML-TEI du CR de CA du 26 janvier 2004)

Le repérage des formes de modalisation en assertion seconde peut également produire des erreurs. Ainsi, dans l'exemple suivant, la séquence introduite par « d'après » ne marque-t-elle pas le début d'un tour mais réfère à un dire tenu en dehors du conseil :

(6) `<u gender="F" NormalizedName="moutel" label="consulting_members">`  
`<p>D'après les informations données par <persName>Madame MOUTEL</persName>,`  
`les produits financiers seront en légère baisse et les produits divers augmenteront`  
`légèrement.</p>`  
`<p>En ce qui concerne les projets de budget des UFR, il y aura nécessairement des`  
`arbitrages qui seront discutés en Commission Budgétaire et arbitrés par le Conseil`  
`d'Administration.</p>`  
`</u>`

(Extrait du document XML-TEI du CR de CA du 29 janvier 1992)

D'une façon générale, il semble d'ailleurs que « d'après » soit peu employé pour introduire un tour de parole.

Par ailleurs, un nombre non négligeable de séquences doivent être interprétées, en raison des spécificités génériques du compte rendu (voir ci-dessus), comme représentant un tour de parole, alors qu'elles ne sont pas explicitement introduites. Nous avons décidé de ne pas considérer comme inexacte l'absence d'annotation sur ces séquences même en présence d'un titre comme « informations du président » indiquant une source possible, considérant qu'il s'agissait d'un indice et non d'une marque univoque.

De même, le cas de figure suivant apparaît trop compliqué pour donner lieu à une annotation automatique. C'est la prise en compte de la table des présents qui nous indique que le participe détaché en tête représente le tour de parole de M<sup>me</sup> Persyn.



(7) <u gender="M" NormalizedName="imberty" label="presidential\_office">  
 <p>Répondant à une question de <hi rend="bold"> <persName>Mme  
 PERSYN</persName> </hi>, <hi rend="bold">M. IMBERTY</hi> précise que 18  
 "vrais" postes seront créés à Nanterre en septembre 1996. L'Université a demandé plus  
 de catégories A et B (B surtout) que de catégorie C, mais c'est le Ministère qui décidera  
 de leur répartition en catégories. Ce seront des postes de titulaires accessibles par  
 concours ou mutation. </p>  
 </u>

(Extrait du document XML-TEI du CR de CA du 29 janvier 1996)

Il arrive également qu'une annotation soit juste alors qu'elle semble contrevenir aux instructions données à l'algorithme. C'est le cas dans l'exemple suivant, où la balise fermée avant l'intertitre est rouverte juste après.

(8) <head n="title\_lvl\_1">ELECTIONS D'ETUDIANTS A DIFFERENTES  
 COMMISSIONS.</head>  
 <u gender="F" NormalizedName="segalen" label="presidential\_office">  
 <p><hi rend="bold"> <persName>Mme SEGALEN</persName> </hi> invite les  
 membres du CA à se référer au document qui leur a été adressé à ce titre.</p>  
 </u>  
 <head n="title\_lvl\_2"><hi rend="bold">Bureau du Conseil d'administration.</hi>  
 </head>  
 <u gender="F" NormalizedName="segalen" label="presidential\_office">  
 <p>Il convient de procéder à l'élection d'un étudiant titulaire et d'un  
 <roleName>suppléant</roleName>. L'élection a lieu par l'ensemble du Conseil, à la  
 majorité des suffrages exprimés.</p>  
 <p><hi rend="bold"> <persName>Mme SEGALEN</persName> </hi> précise que le  
 Bureau du CA se réunit chaque mois ; il est essentiellement chargé d'établir l'ordre du  
 jour du CA et d'examiner les conventions.</p>  
 <p>Elle procède à l'appel à candidature.</p>[...]  
 </u>

(Extrait du document XML-TEI du CR de CA du 26 janvier 2004)

De fait, la représentation des prises de parole, surtout quand elles sont un peu longues ou techniques, peut être structurée par des titres ou des intertitres. Cependant, étant donné que le titre constitue une instruction de fermeture de balise, il est difficile de comprendre pourquoi il y a réouverture juste après, alors que le patron indiquant le début d'un tour survient quelques lignes plus bas (« Mme Segalen précise que »).

## 4 Exploitation de l'annotation : quelques premiers résultats

### 4.1 Mesure de la répartition des tours de parole dans le corpus

L'annotation des tours de parole permet de mesurer le nombre de mots attribués aux différents groupes statutaires identifiés par l'attribut *label* : 25 % des mots du corpus sont attribués à la présidence (*presidential\_office*) et 24 % aux invités (*consulting\_members*). Les autres collèges se partagent moins de la moitié du volume textuel, avec 11 % pour le collège A, 8 % pour le collège B, 6 % pour les étudiants et 4 % pour les BIATSS. Deux types sont très peu représentés : le conseil, abstraction correspondant à des énoncés tels que « le conseil considère que... », et les membres extérieurs (personnalités extérieures à l'université mais disposant d'un droit de vote, tel le représentant de l'établissement public d'aménagement de la défense). La distribution de la somme de mots attribuée à chaque statut de locuteur est représentée en figure 2 ; sur cette figure, la catégorie « *background* » correspond aux mots qui ne sont pas reconnus comme appartenant à un tour de parole.

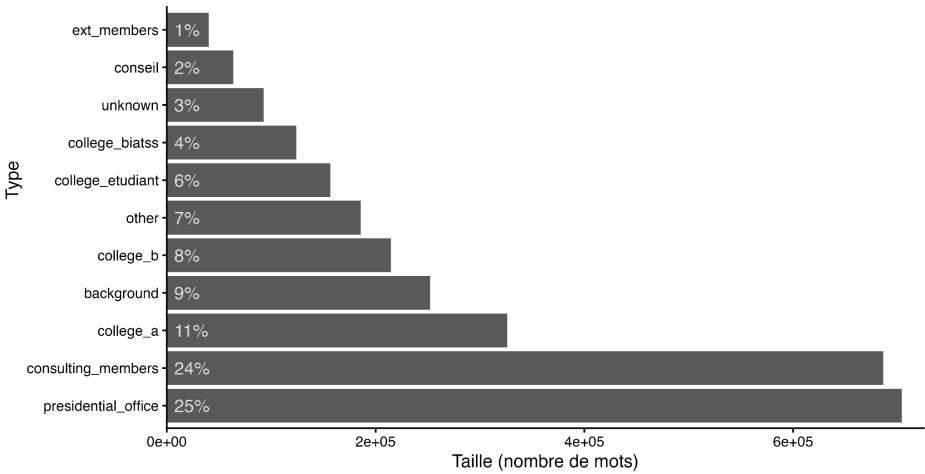


Fig. 2. Fréquence relative du nombre de mots selon le statut

Cette distribution sur l’ensemble du corpus n’est néanmoins pas stable : on note de fortes variations entre les CR, en lien notamment avec les thématiques traitées lors du CA. Le graphique en figure 3 représente la variation diachronique de cette distribution des tours de parole en fonction des statuts. Sur ce graphique, nous avons éliminé le volume textuel non reconnu comme un tour de parole (*background*) ou dont le statut du locuteur n’est pas reconnu (*other* et *unknown*) ainsi que les catégories les plus rares.

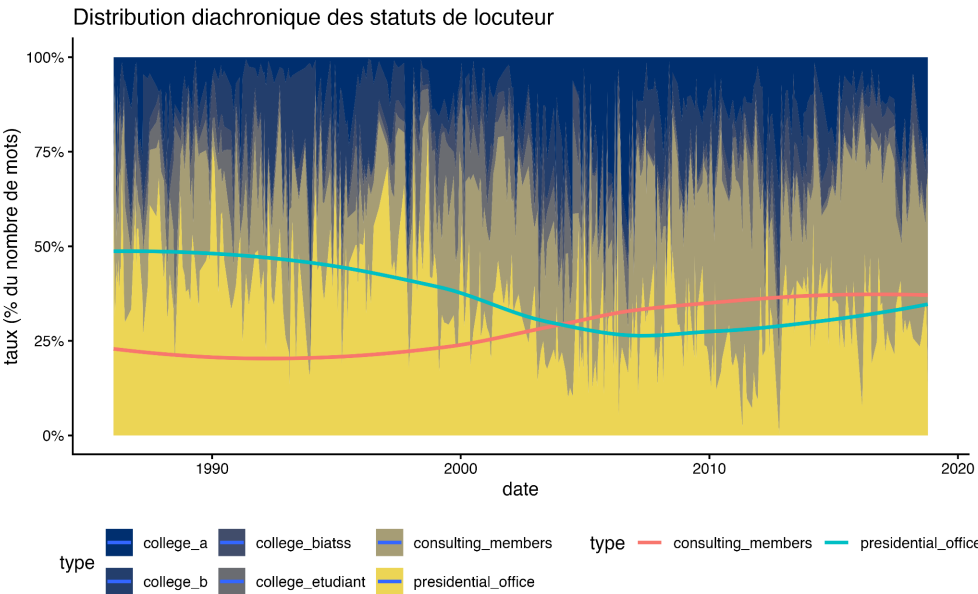


Fig. 3. Évolution diachronique de la taille des tours de parole

Au-delà de la variation très forte entre chaque CR, une évolution principale se dessine : jusqu’au milieu de la décennie 2000, le poids relatif des occurrences attribuées à la présidence (*presidential\_office*) décroît, tandis que celui des personnalités invitées (*consulting\_members*) augmente, comme le représentent les courbes de régression locale en figure 3. Si cette tendance peut s’analyser comme une montée en puissance de la logique

gestionnaire, il convient de noter que l'évolution semble s'inverser après 2013 : le volume de la parole présidentielle rejoint celui des membres invités, tandis que le poids relatif des occurrences attribuées au collège B et au collège étudiant diminue, comme le suggère l'écrasement des deuxièmes et quatrièmes strates en partant du haut du graphique.

4.2 Vers l'analyse des variations d'emploi des verbes introducteurs de représentation du dire en fonction des statuts

L'annotation des tours de parole permet également d'observer la variation des formes de RDA en fonction du statut du locuteur dont la parole est représentée. En guise d'illustration, nous nous intéresserons, dans une perspective d'analyse du discours ouverte aux rapports de pouvoir, à la façon dont les verbes introducteurs de discours indirect tendent à varier en fonction du statut du locuteur représenté (désormais : l).

Après avoir partitionné notre corpus en fonction du statut (« label ») des locuteurs (collège A, collège B, BIATSS, étudiant, membres extérieurs, etc.), nous avons établi une matrice lexicale des verbes (sous forme de lemmes) suivis par « que ». Nous avons ensuite soumis la matrice obtenue à l'analyse des spécificités [15], afin d'identifier les verbes témoignant d'un excédent ou d'un déficit d'emploi significatif pour un statut donné par rapport aux autres statuts.

Si l'on considère les verbes les plus significativement suremployés par le collège étudiant, on remarque que ces derniers partagent d'être des verbes d'affect, renseignant sur l'attitude émotionnelle dysphorique du locuteur source par rapport au contenu représenté : « craindre », « regretter », « déplorer ». On notera avec attention que ces mêmes verbes présentent également un profil excédentaire dans les représentations de la parole des locuteurs des collèges BIATSS et B. En revanche, ils attestent d'un sous-emploi très net dans les partitions qui représentent les invités (*consulting\_members*) et l'équipe présidentielle (*presidential\_office*).

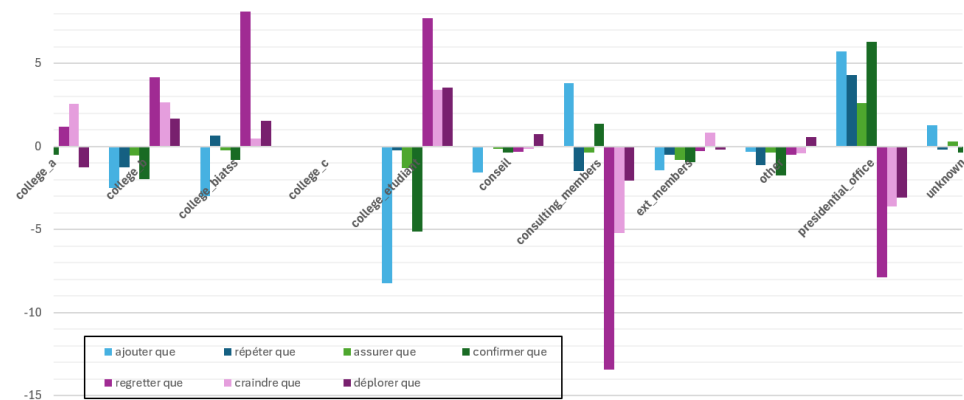


Fig. 4. Ventilation d'une sélection de verbes introducteurs de discours indirect (lemmes) en fonction du statut

Concernant la représentation des dires des locuteurs de l'équipe présidentielle, celle-ci se singularise par le suremploi de verbes porteurs de temporalité (« ajouter », « répéter ») ou des verbes renseignant l'attitude du locuteur vis-à-vis de son dire (« confirmer », « assurer »). L'extrait ci-dessous illustre ainsi la dissymétrie des formes de représentation du discours rapporté selon le statut des locuteurs. Introduit par un verbe d'affect dysphorique, le dire d'une membre du collège B est cadré dans un registre émotionnel : sa position est représentée comme étant de l'ordre du ressenti. À l'inverse, la parole des membres de l'équipe

présidentielle est représentée à l'aide de verbes introducteurs qui participent à construire l'homogénéité d'une parole présidentielle s'exprimant à l'unisson et dans la continuité de principes partagés et stabilisés.

(9) **Mme PERRIN-CHENOUR** [college b] regrette que la communication entre les étudiants des différentes UFR soit rendue difficile par la coupure des bâtiments. Il lui paraît nécessaire – si cette coupure des bâtiments de A à E est inévitable – de remettre sur pied le projet de Maison des étudiants.

**M. LEGRAND** [presidential office] remarque que ce projet n'est pas abandonné mais qu'il ne figure plus en priorité du fait des réactions des organisations étudiantes. Il ajoute que l'équipe présidentielle est sensible à la nécessité d'aménager des lieux d'animation dans cette université.

**M. DENEUX** [presidential\_office] ajoute que le projet d'aménagement du terrain de sport pour 2001-2002, le montre bien. Cela s'inscrit dans la volonté de favoriser une communication collective qui ne peut résulter seulement d'une circulation intérieure dans les bâtiments A à E. (Extrait du CR du CA du 24 septembre 2001, le gras est de la source, nous soulignons)

## 5 Conclusion

Documentant les choix méthodologiques et techniques opérés pour la constitution d'une archive de comptes rendus destinée à être réutilisée par la communauté scientifique, cette contribution montre comment une annotation de la représentation des tours de parole, fondée sur le langage XML et la norme TEI-P5, permet différentes exploitations en analyse du discours.

L'évaluation de l'annotation semi-automatique met en évidence un taux d'erreur global faible et maîtrisé, tout en soulignant les limites structurelles liées aux phénomènes d'ambiguïté inhérents au genre. Ces résultats confirment l'intérêt d'une approche combinant automatisatation et révision humaine par retour au texte, en particulier pour des corpus diachroniques de grande ampleur. La finesse et la diversité des traits à prendre en compte pour une annotation plus fiable pourraient être explorées à l'aide des méthodes d'apprentissage profond (*deep learning*). Le présent corpus pourrait également servir de référence pour l'entraînement de tels modèles (*gold-standard*).

Sur le plan heuristique, ce travail permet d'affiner la caractérisation des modes et des formes de RDA dans le genre du compte rendu [16]. Il vise à être prolongé par l'annotation des dires représentés à l'intérieur des tours de parole, permettant d'examiner les types de locuteurs convoqués (ministère, syndicat, texte de loi), les modalités de représentations de ces dires, le tout en fonction du statut des membres du conseil et de la diachronie.

## Références

1. H. Dumoulin, C. Facq-Mellet, F. Sitri, La fabrication discursive de l'université : comptes rendus et rapports scientifiques en diachronie. Présentation, Cahiers de praxématique **78** (2022). <https://doi.org/10.4000/praxematique.8135>
2. É. Née, Méthodes et outils informatiques pour l'analyse des discours (Presses universitaires de Rennes, Rennes, 2017)
3. J.-M. Viprey, Quelle place pour les sciences des textes dans l'Analyse de Discours ? Semen **21** (2006). <http://journals.openedition.org/semen/1995>
4. F. Rastier, Arts et sciences du texte (Presses Universitaires de France, Paris, 2001)
5. F. Sitri, La genèse d'un compte rendu de réunion : dialogisme et genre de discours, Cahiers de praxématique **76** (2021). <https://doi.org/10.4000/praxematique.7299>

6. B. Fraenkel, Enquêter sur les écrits dans l'organisation, in A. Borzeix et B. Fraenkel (éds), *Langage et travail. Communication, cognition, action*, chap. 9, 231-261 (CNRS Éditions, Paris, 2001)
7. J. Authier-Revuz, *La Représentation du Discours Autre. Principes pour une description* (De Gruyter, Berlin, 2020)
8. J. Authier-Revuz, Discours « tenant lieu » d'un autre discours : un espace générique de la représentation de discours autre, *Corela HS* **41** (2024). <https://doi.org/10.4000/12yx3>
9. D. Mazzucchetti, La voix de l'Assemblée ? Le compte rendu intégral des débats parlementaires ou quand le discours fait foi, *Corela HS* **41** (2024). <https://doi.org/10.4000/12yx4>
10. T. Erjavec, M. Kopp, N. Ljubešić, et al. ParlaMint II: advancing comparable parliamentary corpora across Europe. *Lang Resources & Evaluation* **59**, 2071–2102 (2025). <https://doi.org/10.1007/s10579-024-09798-w>
11. C. Surcouf, À pas de loup dans la bergerie... La problématique du silence et du bruit dans l'étiquetage automatique du subjonctif présent en français parlé, *Corpus* **26** (2025). <https://doi.org/10.4000/1364w>
12. R. Huidrom, A. Belz, Towards a Consensus Taxonomy for Annotating Errors in Automatically Generated Text, in *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, Varna, Bulgarie, 4-6 septembre, 527-540 (INCOMA Ltd., Shoumen, 2023). <https://aclanthology.org/2023.ranlp-1.58/>
13. B. Vendeville, L. Ermakova, P. De Loor. Resource for Error Analysis in Text Simplification: New Taxonomy and Test Collection, in *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25)*, Padoue, Italie, 13-18 juillet, 3723–3732 (Association for Computing Machinery, New York, 2025). <https://doi.org/10.1145/3726302.3730304>
14. K. Al Sharou, L. Specia, A Taxonomy and Study of Critical Errors in Machine Translation. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, Ghent, Belgium, 1-3 juin, 171-180 (European Association for Machine Translation, 2022)
15. R. Lafon, Sur la variabilité de la fréquence des formes dans un corpus, *Mots* **1**, 127-165 (1980). <https://doi.org/10.3406/mots.1980.1008>
16. J. Lefebvre, F. Sitri, Le compte rendu : caractériser un genre de discours au prisme de la représentation du discours autre, *Corela HS* **41** (2024). <https://doi.org/10.4000/12yx8>

<sup>i</sup> La loi Savary qui entre autres transformations, remplace le conseil d'université par un conseil d'administration, date du 26 janvier 1984, mais à Nanterre elle ne s'applique qu'à partir de janvier 1986.

<sup>ii</sup> Ce travail a été en grande partie réalisé grâce à un appel à projet du labex « Les passés dans le présent » (ANR-11-LABX-0026-01) et grâce à un contrat post-doctoral, obtenu en réponse à un appel d'offres de l'université Paris-Est Créteil.

<sup>iii</sup> Dépôt de l'archive numérique : <https://www.ortolang.fr/market/corpora/archivu>

<sup>iv</sup> Les graphies fautives effectivement présentes sur le document d'origine ont, en revanche, été respectées et signalées à l'aide des balises XML-TEI appropriées (<sic> et <corr> au sein de <choice>).

<sup>v</sup> La barre verticale ']' signifie « ou ».

<sup>vi</sup> Les labels utilisés par le programme ont les valeurs suivantes : « collège A » = enseignants-chercheurs de rang A ; « collège B » = enseignants-chercheurs de rang B ; « collège BIATSS » = personnels administratifs et techniques ; « collège étudiants » = étudiants ; « ext members » = personnalités extérieures ; « presidential office » = président et équipe présidentielle ; « consulting members » = membres invités.