Annotation manuelle d'occurrences de candidats termes et écrit scientifique

Evelyne Jacquey, Laurence Kister, Simon Méoni, Sabine Barreaux, Camille Noûs

Annotation manuelle d'occurrences de candidats termes dans des articles scientifiques

Évelyne Jacquey, Laurence Kister, Simon Méoni, Sabine Barreaux, Camille Noûs

UMR ATILF CNRS Université de Lorraine

RÉSUMÉ. *Cet article compare deux campagnes d'annotation successives visant l'identification manuelle des occurrences de candidats termes qui relèvent effectivement de la discipline scientifique de l'article considéré. Les deux campagnes se distinguent par leurs objectifs. La première visait l'enrichissement de terminologies existantes. La seconde avait l'objectif de mesurer la difficulté de la tâche d'annotation en sciences humaines et sociales (SHS) par rapport aux sciences dites exactes. Les corpus produits ne permettant pas de comparer les deux campagnes directement, nous exploitons ces corpus comme corpus d'apprentissage dans une tâche test qui consiste à automatiser l'annotation manuelle. L'objectif est de savoir si le corpus de la seconde campagne permet d'augmenter les performances de la tâche test par rapport à celui de la première campagne.*

ABSTRACT. *This paper compares two successive annotation campaigns aimed at manually identifying the occurrences of candidate terms that actually fall within the scientific domain of the annotated document. The two campaigns are distinguished by their objectives. The first aimed at the enrichment of existing terminological resources. The second had the objective of measuring the difficulty of the annotation task in the human and social sciences compared to the so-called hard sciences. A direct comparison between both campaigns is not possible on the basis of the produced corpora. To do this, we use these corpora as learning corpus in the context of a test task. The role of this task is to automatise the manual annotation. The goal is to determine if the second corpus is of better quality than the first one with regards to the test task performances.*

MOTS-CLÉS : *Terminologie textuelle, Annotation manuelle, Langue de spécialité.*

KEYWORDS: *Textual Terminology, Manual Annotation, Domain specific Languages.*

1re soumission à *TAL 61-1*, le 02/12/2019

1. Introduction

Sur le plan des publications scientifiques, la situation actuelle se caractérise par l'augmentation constante de la quantité et de la diversité des articles scientifiques publiés, et leur mise à disposition sur diverses plateformes comme HAL, ArXiv, ISTEX. Cette situation rend plus difficile l'objectif d'appréhender dans leur ensemble les publications qui s'intéressent à une problématique donnée tant sur le plan de leur quantité que sur le plan de la masse et de la diversité des connaissances expertes nécessaires. Pour faire face à cette situation, l'extraction des vocabulaires de spécialité est un des éléments pouvant contribuer à un accès efficace au contenu des publications scientifiques. Cette extraction est prise en charge par des outils automatiques, les extracteurs terminologiques. Ceux-ci s'appuient sur des critères statistiques, des critères linguistiques ou font interagir approche linguistique et approche statistique (Cabré Castelli *et al.*, 2001 ; Jacquemin, 2003 ; Daille *et al.*, 2004). Quelle que soit leur dominante méthodologique, l'intervention d'experts du domaine de spécialité est nécessaire pour évaluer *in fine* la qualité des terminologies produites. Ces interventions expertes sont d'une part coûteuses et peu reproductibles (Vàsquez et Oliver, 2018), et d'autre part, ne peuvent pas suivre l'évolution lexicale et phraséologique des domaines scientifiques au rythme où les articles scientifiques sont publiés et mis à disposition. Il apparaît donc nécessaire de mettre sur pied une procédure plus souple sous la forme de corpus annotés au sein desquels le vocabulaire de spécialité est clairement identifié. Par ailleurs, comme l'a souligné Nazarenko *et al.* (2009), de tels corpus représentent une méthode alternative pour l'évaluation des outils d'extraction terminologique.
À l'intérieur du vaste domaine de l'accès au contenu de documents à partir de leur vocabulaire spécifique, les travaux que nous présentons ici visent à décrire et à analyser deux étapes méthodologiques mises en œuvre successivement pour la constitution de corpus au sein desquels le vocabulaire de spécialité est identifié.

Pour réaliser de tels corpus, les documents sont passés à un extracteur terminologique afin de limiter le recours aux experts *ex nihilo*. L'extracteur terminologique utilisé produit des listes de candidats termes, éléments du vocabulaire de spécialité, dont les occurrences sont détectées au sein des documents sur la base de critères morpho-syntaxiques uniquement et non grâce à une sélection d'ordre sémantique. Les occurrences de candidats termes apparaissent ainsi ambiguës entre une acception relevant de la langue de spécialité (par la suite *occurrence disciplinaire*) et d'autres acceptions n'en relevant pas (langue générale, vocabulaire savant, lexique scientifique, etc.). Dans l'exemple (1) ci-dessous, issu d'un article de chimie, les occurrences des candidats *espèce* et *réaction* sont faciles à identifier comme relevant de la langue de spécialité.

(1) Au cours de la nitration des noyaux aromatiques par le N 2 0 4 en milieu aprotonique, la présence d'un nitrate métallique tel que Zn(N0 3) 2 , Cu(N0 3) 2 ou U0 2 (N0 3) 2 , génère l'*espèce*$_{[+disciplinaire]}$ catalytique NO + selon la *réaction*$_{[+disciplinaire]}$ N 2 0 4 + M(N0 3) 2 <=t NO + + M(N0 3).

L'identification des occurrences disciplinaires n'est cependant pas toujours aussi évidente. Un cas de polysémie assez simple est celui de *type*. Dans le domaine de la linguistique de corpus, *type* s'oppose à *token* par exemple dans une expression comme *token-type ratio*. On trouvera de nombreuses autres occurrences de *type* dans beaucoup de contextes et de disciplines avec des expressions comme *des résultats de plusieurs types* ou encore l'expression *de type [...]*. Un exemple plus complexe est celui du candidat *contexte* ci-dessous, issu d'un article de linguistique. Dans le premier exemple (2.a), il peut être considéré comme ayant une acception disciplinaire ou comme étant une précision du sens général de l'occurrence du même candidat dans le second exemple (2.b).

(2.a) La morpholexicalité, la structuration syntaxique, l'effet paronomastique et le $contexte_{[+disciplinaire]}$ participent, conjointement ou non, à (re)former l'expression figée et à en actualiser le sens.

(2.b) Le second touche à des développements de recherche récents sur l'interaction exolingue dans une diversité de $contextes_{[-disciplinaire]}$ à partir de 1980.

Enfin, les occurrences de candidats termes peuvent être ambiguës entre plusieurs disciplines. Le candidat *patient*, par exemple, désigne deux concepts différents selon que l'on est dans le domaine de la linguistique ou dans celui de la psychologie[1]. L'Homme (2004) et Rastier (2005) précisent que si un terme, en tant qu'étiquette de concept, est non ambigu à l'intérieur d'une terminologie et pour une application donnée, ses réalisations linguistiques en contexte peuvent l'être. Lorsqu'un article scientifique s'intéresse à une problématique connexe entre la linguistique et la psychologie, l'ambiguïté en contexte des occurrences du candidat *patient*, qui est d'ordre disciplinaire, viendra s'ajouter à l'ambiguïté entre vocabulaire de spécialité (occurrences disciplinaires) et tous les autres vocabulaires apparaissant dans l'article.

Produire des corpus annotés en vocabulaire de spécialité suppose alors une annotation manuelle des occurrences de candidats termes afin de filter celles qui sont jugées comme relevant de la langue de spécialité du document par les annotateurs. Cette tâche d'annotation revient donc à une tâche de désambiguïsation lexicale simplifiée dans la mesure où il n'y a que deux choix considérés : l'occurrence annotée relève d'un emploi disciplinaire ou non. Dans l'inventaire des ressources textuelles annotées[2] réalisé par le groupe de travail n°8 (*GT8*) « Annotation du plus haut niveau » du consortium « Corpus, Langues et Interactions » *(CORLI)*, aucun corpus de référence annoté en occurrences disciplinaires de candidats termes n'est recensé pour le français. Cependant, dans le domaine de la désambiguïsation lexicale, deux ressources diffusées par ELRA ont été recensées par le GT8 : le corpus ROMANSEVAL (Véronis, 2001) et le cor-

1. On peut ainsi comparer l'occurrence de *patient* qui apparaît dans un article de linguistique étudiant la langue de spécialité du monde médical « *C'est surtout dans le discours médecin / patient qu'est présente une grande variabilité dans la langue* », avec l'acception courante en linguistique que l'on trouvera dans des expressions comme « *le patient de ce prédicat* ».

2. Inventaire du GT8 (co-animé par Amalia Todirascu et Agnès Tutin) : `https://listes.cru.fr/wiki/corpus-ecrits/public/groupe-8`. [Page consultée le 10 octobre 2018]

pus ESTER2 (Galliano *et al.*, 2009). Le corpus ESTER2 servira de *baseline* pour les travaux qui seront présentés ici dans la mesure où c'est l'expérience la plus similaire.

Dans la lignée de travaux antérieurs décrivant la production et l'exploitation de corpus annotés en vocabulaire de spécialité (Zesch et Gurevych, 2009 ; Nazar, 2016 ; Bougouin *et al.*, 2016 ; Vàsquez et Oliver, 2018), cet article décrit et analyse deux campagnes d'annotation successives avec l'objectif général de savoir si la seconde campagne a permis de produire un corpus de référence de meilleure qualité que celui produit par la première campagne.

Pour les deux campagnes, les documents annotés sont extraits d'un corpus d'articles scientifiques français en libre accès pour la communauté scientifique appelé par la suite *corpus-source*. Bien que les données proviennent de la même source (le corpus de la seconde campagne est un sous-ensemble du corpus de la première campagne), les deux campagnes se différencient en fonction de leur finalité globale. La première campagne, dite *campagne* C_1, visait l'enrichissement de terminologies existantes dans trois disciplines de SHS (archéologie, linguistique et psychologie). La seconde campagne, dite *campagne* C_2, avait un double objectif : (1) comparer deux disciplines de SHS (archéologie et linguistique) avec une discipline de sciences dites exactes (chimie) et (2) mettre à l'épreuve un protocole d'annotation fondé sur les bonnes pratiques (Bonneau-Maynard *et al.*, 2005 ; Fort, 2016).
Elles se différencient aussi sur le plan de leur protocole. La campagne C_1 n'a pas fait l'objet d'une annotation multiple étant donné la quantité d'occurrences à désambiguïser (un peu plus de 145 000 occurrences présentes dans 70 articles intégraux). La campagne C_2, d'un volume beaucoup plus réduit (un peu plus de 2 500 occurrences de candidats termes dans 89 résumés), a bénéficié d'une annotation multiple et de la mise en œuvre des bonnes pratiques en matière d'annotation manuelle et de constitution de corpus de référence.
La dernière différence entre les deux campagnes concerne les disciplines. La campagne C_1 porte sur trois disciplines SHS : l'archéologie, la linguistique et la psychologie. La campagne C_2 porte sur deux disciplines SHS (archéologie et linguistique) et une discipline de sciences exactes pour comparaison (la chimie).

Pour comparer ces deux campagnes, il n'est pas possible d'utiliser les mesures courantes d'accord inter-annotateurs puisque les deux campagnes n'ont pas été réalisées suivant le même protocole. La qualité de l'annotation produite est mesurée par une évaluation dite par la tâche. La tâche servant de cadre d'évaluation consiste à automatiser l'identification manuelle des occurrences disciplinaires en utilisant les deux corpus annotés comme corpus d'apprentissage.

Cet article s'organise de la manière suivante : la section (2) est consacrée à la description des données annotées et des protocoles d'annotation, la section (3) détaille l'évaluation de la campagne C_2 et la section (4) compare les résultats de l'évaluation par la tâche des deux campagnes.

2. Données et protocoles d'annotation

2.1. *Pré-annotation*

Les données du corpus-source sont encodées selon un même format XML inspiré des recommandations internationales de la TEI[3], TBX (Lommel *et al.*, 2014 ; Melby, 2015) et *Stand-Off* (Ide et Romary, 2006 ; Romary, 2014). Dans ce format (figure 1), le texte est un flux linéaire de *tokens*, chacun doté d'un identifiant unique (les « t_i » dans l'élément <text>). Ces identifiants permettent de positionner de manière fiable tout enrichissement du texte sous la forme d'annotations déportées (les « #t_i », valeurs de l'attribut @target). Le corpus-source enrichi comporte deux types d'annotations : (1) la segmentation lexicale et l'étiquetage morphosyntaxique (<ns :standOff type='wordForms'>) et (2) l'identification des occurrences de candidats termes (<ns :standOff type='candidatsTermes'>).

```
<teiHeader> < !-- Métadonnées du document -->
<ns:standOff type="wordForms"> < !-- première couche d'annotation, POS tagging -->
    <ns:listAnnotation> < !-- élément englobant des annotations -->
        <span target="#t1">
        @lemma = du ; @pos = PRP:det
        </span>
        <span target="#t2">
        @lemma = fouille ; @pos = NOM
        </span>
        ...
<ns:standOff type="candidatsTermes"> < !-- deuxième couche d'annotation, Occurrences de candidats termes -->
    <ns:listAnnotation> < !-- élément englobant des annotations -->
        <span target="#t2" corresp="#TS1.4-entry-727968" ana="#disciplinaire">
        @inflexionsWord= fouilles
        </span>
        ...
 <text> < !-- texte tokenisé où chaque token est associé à un identifiant unique -->
     <div type="abstract" xml:lang="fr">
     <p>
     <w xml:id="t1">Des</w> <w xml:id="t2">fouilles</w> <w xml:id="t3">récentes</w>
     <w xml:id="t4">sur</w> <w xml:id="t5">deux</w> <w
     xml:id="t6">gisements</w> ....
```

FIGURE 1. *Illustration du format XML des articles traités*

La segmentation lexicale et l'étiquetage morphosyntaxique sont réalisés avec TreeTagger (Schmid, 1994), nativement intégré à la plateforme multilingue TTC-TermSuite (Rocheteau et Daille, 2011) qui se charge aussi de l'extraction des candidats termes et de la détection de leurs occurrences. Cet extracteur utilise les variations linguistiques courantes des candidats termes (orthographiques, flexionnelles, syntaxiques) ainsi que des critères statistiques, en particulier la fréquence relative d'un candidat et sa spécificité (*termhood*) par rapport à un corpus de langue générale (pour le français, 82 millions de tokens issus du journal Le Monde, (Daille *et al.*, 2016)). À l'issue du calcul de la liste de tous les candidats potentiels, les formes variantes des candidats sont regroupées autour de ceux qui sont considérés comme les meilleurs pivots. Au sein des choix possibles dans TermSuite, nous avons fixé trois paramètres : la

3. `http://www.tei-c.org/Guidelines/` [Pages consultées le 14 novembre 2019]

catégorie des candidats (noms ou adjectifs), un seuil de fréquence absolue (minimum de 5 occurrences) et un seuil de taille de liste (les 10 000 premiers candidats dans la liste triée par spécificité décroissante)[4]. L'enrichissement automatique des données textuelles est réalisé par une chaîne de traitement dédiée qui intègre ces outils afin de produire le corpus-source enrichi.
Dans les deux campagnes, les annotateurs sont intervenus sur des données pré-annotées et n'ont pas eu à identifier les annotables, l'*unitizing* au sens de Krippendorff (1995).

2.2. *Campagne* $\mathbf{C}_1$

2.2.1. *Corpus, protocole et environnement*

La première campagne a porté sur le texte intégral de documents de trois disciplines de SHS (11 articles en archéologie, 29 en linguistique et 30 en psychologie). Le corpus comporte 568 646 tokens pour 145 445 occurrences de candidats termes identifiées automatiquement puis annotées manuellement. Le protocole de la campagne C_1 subdivise la prise de décision au travers de deux étapes successives recevant une réponse binaire.
Pour la première étape (validité linguistique), les annotateurs sont des linguistes expérimentés qui se sont appuyés sur leur compétence linguistique en français et sur un guide d'annotation accessible en ligne dont les recommandations principales sont : (1) dans les candidats de type groupe nominal étendu, le candidat maximal et les candidats inclus de plus petite taille et autonomes sont conservés ; (2) les candidats correspondant à des noms ou à des adjectifs substantivés sont conservés (*palatale, occlusive*) ; (3) les candidats, simples ou complexes, correspondant à des verbes, des conjonctions et tout adjectif isolé, sont rejetés.
Ainsi dans *Hommes du Paléolithique supérieur*, six candidats sont détectés selon le découpage suivant : $[_{C1}\ [_{C2}\ [_{C3}$*Hommes*$_{C3}]$ *du* $[_{C4}\ [_{C5}$ *Paléolithique*$_{C5}]\ _{C2}]$ $[_{C6}$*supérieur*$_{C6}]\ _{C4}]\ _{C1}]$. Parmi ces candidats, $C1$, $C3$, $C4$ et $C5$ sont jugés valides tandis que $C2$ et $C6$ sont jugés invalides.
La seconde étape d'annotation (appartenance disciplinaire) est réalisée sur le sous-ensemble d'occurrences jugées linguistiquement correctes. Les annotateurs, en tant que spécialistes des domaines traités, se sont appuyés sur leur expertise et sur les recommandations du guide d'annotation. Les occurrences des candidats sont jugées disciplinaires si elles désignent des concepts du domaine scientifique traité (*définition circulaire, dictionnaire, schwa, variation graphique* en linguistique), des concepts d'un autre domaine suffisamment intégrés dans le domaine scientifique traité (*apprentissage, communication, récit* en linguistique) ou des concepts propres au discours du domaine (*construction, forme, usage* pour les monolexicaux en linguistique ou

4. Le silence lors de la détection des candidats termes n'a pas été évalué car aucun test de ce type n'a été réalisé avant le lancement de la campagne C_1. Pour la campagne C_2, une partie des erreurs de segmentation du pré-traitement réalisé par TermSuite a été résolue grâce à l'externalisation de ce pré-traitement.

polarité négative et positive, terme affectif français, syllabe initiale pour les complexes). Les occurrences de candidats sont jugées non disciplinaires si elles dénotent des concepts relevant d'autres domaines scientifiques qui ne sont pas suffisamment intégrés dans le domaine scientifique traité (*carte conceptuelle, élève* en linguistique) ou si ces occurrences désignent des notions relevant du discours scientifique en général ou de la langue générale (*application, hypothèse, processus* en linguistique).

Les textes sont annotés dans un environnement HTML dédié que nous ne détaillerons pas faute de place. Les éléments importants de cette interface sont que les occurrences à annoter sont présentées à l'annotateur dans leur ordre linéaire d'apparition dans le texte sans qu'il y ait mémorisation et propagation des choix antérieurs de l'annotateur. Chaque occurrence à annoter est donc considérée isolément.

En ce qui concerne le déroulement de la campagne, la sélection des occurrences linguistiquement correctes a été effectuée par deux linguistes expérimentés pour l'ensemble des 70 fichiers, 35 fichiers par annotateur. La sélection des occurrences disciplinaires a été réalisée par un ou deux ingénieurs documentalistes spécialistes des disciplines traitées ainsi que par les deux annotateurs linguistes. Six annotateurs ont été mobilisés en tout : trois en linguistique, deux en psychologie et un en archéologie. Pour les disciplines traitées par plusieurs annotateurs experts, les fichiers ont été répartis entre eux.
Le temps d'annotation n'a pas été mesuré : chaque annotateur était autonome et travaillait à son rythme après appropriation de l'environnement et du guide d'annotation. Il est enfin à préciser que les auteurs de l'article ont participé aux deux étapes d'annotation dans cette campagne.

2.2.2. *Bilan de la campagne* C_1

La campagne d'annotation a concerné 17 523 candidats pour 145 445 occurrences (voir tableau 1). Au sein des candidats, les plus fréquents sont *site* en archéologie (274 occurrences), *langue* en linguistique (514 occurrences) et *scolaire* en psychologie (832 occurrences).

	Corpus		Archéologie		Linguistique		Psychologie	
Articles	70		11		29		30	
Tokens	568 646		88 441		220 118		260 087	
	100 %		*15,55 %*		*38,71 %*		*45,74 %*	
	Types	Occ.	Types	Occ.	Types	Occ.	Types	Occ.
Candidats	17 523	145 445	4 194	23 857	5 819	34 986	7 510	86 602
Valides ling.	12 756	103 431	2 884	14 258	4 441	23 092	5 431	66 081
% / Candidats	*72,80 %*	*71,11 %*	*68,76 %*	*59,76 %*	*76,32 %*	*66,00 %*	*72,32 %*	*76,30 %*
Disciplinaires	9 171	71 493	2 696	13 561	3 303	15 662	1 907	42 270
% / Valides ling.	*71,90 %*	*69,12 %*	*93,48 %*	*95,11 %*	*74,38 %*	*67,82 %*	*58,41 %*	*63,97 %*

TABLEAU 1. *Effectifs de la campagne* C_1

Le taux de validation linguistique moyen est de 71,11 % des occurrences (ligne « Valides ling. » dans le tableau 1), 59,76 % en archéologie, 66,00 % en linguistique et 76,30 % en psychologie. Autrement dit, un peu moins de 30 % des occurrences sont rejetées en moyenne dans cette campagne du fait d'incorrections linguistiques.
Le taux de validation disciplinaire moyen est de 69,12 % des occurrences (ligne « Disciplinaires » dans le tableau 1) avec l'archéologie qui se distingue nettement des deux autres disciplines sur ce point : 95,11 % en archéologie, 67,82 % en linguistique et 63,97 % en psychologie.
Un résumé de ces observations est que l'archéologie compte moins d'occurrences linguistiquement valides mais que parmi celles-ci, davantage d'entre elles sont jugées disciplinaires. Nous reviendrons plus en détail sur la composition des candidats termes de cette campagne et de ceux de la campagne C_2 dans la section (4.1) consacrée à leur comparaison.

En termes de qualité des annotations, bien que cette première campagne n'ait pu faire l'objet d'un calcul inter-annotateur, l'utilisation des données produites dans des travaux postérieurs a fait apparaître un manque de cohérence entre des choix d'annotation dans des contextes similaires et des annotations oubliées. Ces constats, bien que n'ayant pas pu être mesurés de manière systématique, nous ont conduits à analyser les conditions d'annotation de cette première campagne et nous ont ensuite amenés à définir une seconde campagne dont les différences majeures se situent sur plusieurs plans : (1) la quantité de données à annoter, (2) la clarté des consignes d'annotation, (3) la mise en œuvre des bonnes pratiques en la matière et (4) le rejet d'un environnement *ad hoc* au profit d'environnements connus et utilisés dans la communauté scientifique.

2.3. *Campagne* C_2

2.3.1. *Corpus*

La campagne C_2 porte sur 89 résumés d'articles (11 368 tokens) au sein desquels 2 511 occurrences de candidats termes (1 506 candidats) sont pré-annotées (voir tableau 2, p. 9).

Les candidats les plus fréquents sont *funéraire* et *sépulture* en archéologie (13 occurrences), et *article* en linguistique (25 occurrences). Le taux de validité linguistique est nettement plus élevé dans cette campagne : 93,03 % des occurrences à comparer avec les 71,11 % des occurrences en C_1 (tableau 1). Cette différence s'explique principalement par la modification des recommandations d'annotation et par la correction d'erreurs de segmentation et d'étiquetage via l'externalisation de TreeTagger dans notre utilisation de la plateforme TTC-TermSuite (voir section 2.3.2 ci-après).
Parmi les occurrences linguistiquement valides, le taux de validation disciplinaire est en baisse toutes disciplines confondues : 61,30 % des occurrences en C_2 à comparer avec les 69,12 % des occurrences en C_1. Au niveau des disciplines communes, cette baisse apparaît très fortement pour l'archéologie (validation disciplinaire à 95,11 %

	Corpus		Archéologie		Linguistique		Chimie	
Résumés		89		30		30		29
Tokens		11 368		4 454		3 588		3 326
				39,18 %		*31,56 %*		*29,26 %*
	Types	Occ.	Types	Occ.	Types	Occ.	Types	Occ.
Candidats	1 506	2 511	425	582	452	778	629	1 151
Valides ling.	1 380	2 336	393	541	409	719	578	1 076
% / Candidats	*91,63%*	*93,03%*	*92,47%*	*92,96%*	*90,49%*	*92,42%*	*91,89%*	*93,48%*
Disciplinaires	850	1 432	240	361	236	391	374	680
% / Valides ling.	*61,59%*	*61,30%*	*61,07%*	*66,73%*	*57,70%*	*54,38%*	*64,71%*	*63,20%*

TABLEAU 2. *Effectifs de la campagne* C_2

des occurrences en C_1 et à 66,73 % en C_2) et apparaît nettement en linguistique (validation disciplinaire à 67,82 % des occurrences en C_1 et à 54,38 % en C_2).

2.3.2. *Protocole d'annotation de la campagne* C_2

L'annotation est réalisée en une seule étape au moyen d'un jeu de trois étiquettes.

Valeur	Signification
« 0 »	occurrences linguistiquement incorrectes
« 1 »	occurrences linguistiquement correctes qui n'ont pas un sens disciplinaire dans le texte à annoter
« 2 »	occurrences linguistiquement correctes dont le sens est disciplinaire

Concernant les recommandations d'annotation, la principale différence entre les deux campagnes se situe au niveau des critères de validité linguistique : la recommandation du rejet des adjectifs isolés a été supprimée quand ces éléments étaient bien formés. De plus, les conditions de rejet ont été affinées : rejet systématique d'occurrences en langue étrangère sauf si le candidat correspond à un terme du domaine dans cette langue, rejet des découpages syntaxiques incorrects dus à des erreurs d'étiquetage restantes (*résonance de* dans le candidat plus étendu *résonance de spin*) ou d'éléments appartenant à une unité phraséologique (*objet* dans l'expression *faire l'objet de*).

Les critères de choix entre les valeurs « 1 » (linguistiquement correct mais non disciplinaire) et « 2 » (disciplinaire) s'appuient sur la consultation de ressources terminologiques de référence car, dans cette seconde campagne, les annotateurs ne sont pas forcément des experts des domaines scientifiques traités. L'ordre de consultation des ressources est le suivant en tenant compte des difficultés qu'elles posent ou des lacunes que ces ressources comportent par rapport à l'utilisation qui en est faite dans cette campagne : (1) les ressources terminologiques de référence fournies par l'un des participants de la campagne (absence de définitions pour les termes) ; (2) le Grand

Dictionnaire Terminologique (*GDT*)[5] (couverture incomplète des disciplines de la campagne et variations diatopiques possibles du français québécois) ; (3) la base de données multilingue InterActive Terminology for Europe (*IATE*)[6] (orientation vers la traduction dans des domaines non aisément compatibles avec les disciplines de la campagne). Lorsque l'occurrence à annoter correspond à un terme dans l'une des ressources terminologiques de référence, l'annotateur doit vérifier si le sens de l'occurrence est conforme à celui du terme correspondant tel qu'il peut l'inférer à partir de la ressource terminologique consultée. Si aucun terme ne correspond à l'occurrence à annoter quelle que soit la ressource terminologique de référence consultée, l'annotateur est autorisé à consulter Wikipedia et à faire une recherche au moyen de Google bien que ces deux solutions doivent rester exceptionnelles car le guide n'indique pas de méthodologie contrôlée.

Un dernier point sur les recommandations données aux annotateurs concerne les travaux de L'Homme (2005) qui fournissent une synthèse des règles à suivre pour identifier des occurrences de termes au sein d'une terminologie candidate ou en corpus. Elles n'ont pas été utilisées ici car seules des occurrences disciplinaires devaient être identifiées, et non des occurrences terminologiques. Ce choix est tout d'abord motivé par les difficultés constatées lors de la campagne C_1 même lorsque l'annotation était réalisée par des experts du domaine scientifique traité. Ensuite, l'arbitrage entre « disciplinaire » et « terminologique » est fortement contraint par l'exploitation envisagée pour l'annotation. Dans le cas de la campagne C_2, l'objectif n'est pas d'enrichir des terminologies existantes mais de tester une méthodologie pour la constitution de corpus de référence annotés manuellement. L'arbitrage entre « disciplinaire » et « terminologique » pourrait effectivement intervenir dans le cadre d'une troisième campagne fondée sur les mêmes objectifs que la campagne C_1.

2.3.3. *Gestion de la campagne* C_2

La campagne avait pour objectif secondaire de tester trois environnements d'annotation mis à la disposition de la communauté scientifique : BRAT (Stenetorp *et al.*, 2012), GATE (Cunningham *et al.*, 2013) et GLOZZ (Widlöcher et Mathet, 2012)[7]. Les 89 résumés d'articles ont été répartis en trois ensembles de dix documents par discipline dans chacun des trois environnements d'annotation excepté pour la chimie où l'un des résumés a été supprimé en raison d'erreurs de segmentation constatées après le démarrage de la campagne.

La campagne C_2 a été gérée selon les recommandations de Fort (2016). Une pré-campagne a eu lieu avec les futurs arbitres de la campagne effective en vue de mettre sur pied les différents guides (guide d'annotation, guide d'utilisation des différents environnements et des différentes ressources externes). Cette pré-campagne a permis

5. `http://www.granddictionnaire.com/` [Page consultée le 14 novembre 2019]

6. `https://iate.europa.eu/home` [Page consultée le 14 novembre 2019]

7. BRAT a été choisi du fait de son utilisation en ligne et de sa prise en main très facile, GATE pour son intégration aisée avec des chaînes de traitement TAL et GLOZZ pour son ergonomie et ses nombreuses utilisations dans des campagnes d'annotation complexes.

d'estimer à une demie-heure le temps de traitement d'un résumé et donc de planifier le recrutement des annotateurs.
À l'issue de la pré-campagne, les données ont été réparties en quatre phases pour chacune des trois disciplines. Les objectifs de la première phase étant de vérifier la bonne compréhension des recommandations d'annotation et l'utilisation adéquate des environnements d'annotation, elle n'a concerné que trois textes par discipline soit un texte par environnement. Les trois autres phases, vues comme des phases de production, ont porté sur neuf textes par discipline, trois par environnement, sauf pour la phase 3 en chimie où deux textes seulement ont été traités sous GLOZZ du fait d'erreurs de segmentation signalées ci-dessus. Entre chaque phase, un temps d'arrêt a été prévu afin de réaliser plusieurs tâches :

– arbitrage collégial des annotations réalisées ;

– corrections des recommandations d'annotation, mise à jour de l'annotation de référence antérieure et interactions avec les annotateurs (explications et discussion autour des confusions fréquentes, etc.) ;

– calcul des accords inter-annotateurs.

L'arbitrage est préparé par des scripts créés pour la campagne qui fusionnent les annotations réalisées dans un tableau. Chaque cas de désaccord entre les annotateurs est accompagné du contexte de l'occurrence. Lors de la première phase de la campagne, les arbitres ont procédé à une vérification de la totalité des annotations réalisées. Dans les trois phases suivantes, seuls les cas de désaccord ont été analysés en détail mais une vérification rapide de toutes les annotations a été faite.
L'environnement d'annotation et de gestion de campagne WebAnno (Yimam *et al.*, 2013) a été testé lors de la pré-campagne. Il a été rejeté pour des raisons d'ergonomie et de recours beaucoup trop massif au clic, tant pour les annotateurs que les arbitres.

3. Évaluation de la campagne C_2

La campagne C_2 s'étant déroulée dans un cadre d'annotation multiple, la qualité intrinsèque des annotations produites est calculée selon des mesures d'accord inter-annotateurs. À la suite de (Artstein et Poesio, 2008 ; Fort, 2016 ; Mathet et Widlöcher, 2016) notamment, nous utilisons un Kappa de Cohen pour les disciplines annotées par deux annotateurs (Cohen, 1960) (chimie et linguistique), et un Kappa de Fleiss lorsqu'il y a trois annotateurs (Davies et Fleiss, 1982) (archéologie). Les mesures de Kappa sont une manière de normaliser l'accord observé entre les différents annotateurs en fonction d'un accord attendu. Mathet et Widlöcher (2016) parlent de « correction par la chance ». L'idée générale est de calculer l'accord attendu à partir de la distribution observée des annotations. La différence entre les Kappa de Cohen et de Fleiss repose sur le fait qu'avec le Kappa de Cohen, la distribution des annotations par annotateur est prise en compte. Ce n'est pas le cas avec le Kappa de Fleiss où les distributions des annotations de tous les annotateurs sont fusionnées.

Cependant, le Kappa de Cohen ne peut être utilisé que pour deux annotateurs.

Des Kappa ont été calculés pour chaque phase de l'annotation (phases 1 à 4) et pour chaque discipline. Pour interpréter les scores obtenus, nous utilisons ceux de l'annotation en entités nommées de la campagne ESTER2 comme *baseline*, car ESTER2 semble plus proche de l'annotation menée dans la campagne C_2 que ne l'est la désambiguïsation lexicale réalisée dans la campagne ROMANSEVAL à partir des sens lexicographiques du dictionnaire Le Larousse. Le Kappa de Cohen pour le premier niveau d'annotation d'ESTER2 (identification des entités nommées et catégorisation selon sept catégories de haut niveau) est de 0,71 selon Fort (2012). Selon Artstein et Poesio (2008), un Kappa de 0,8 caractérise une bonne qualité des données annotées et un Kappa situé entre 0,6 et 0,8 traduit une qualité moyenne des données annotées.
Afin de comparer la campagne C_2 avec ESTER2, les accords inter-annotateurs par phase sont exploités de deux manières : le Kappa en phase 1 est pris en compte de manière isolée car cette phase est une étape d'appropriation, les Kappa des phases 2 à 4 sont regroupés par un calcul de moyenne.

3.1. *Comparaison des disciplines*

Les diagrammes par discipline pour les SHS montrent tout d'abord que l'accord moyen des phases 2 à 4 en archéologie est quasiment identique à celui d'ESTER2 contrairement à celui obtenu en linguistique, qui est légèrement en deçà (figure 2, p. 13). En archéologie (barre en gris clair dans la série « Moyenne des phases 2 à 4 »), la moyenne est de 0,72 pour les phases 2 à 4. Par ailleurs, cet accord moyen connaît une augmentation de 0,4 points par rapport à l'accord en phase 1 (0,32, barre en gris clair dans la série « Phase 1 »). On analyse les métriques reproduites pour la linguistique de la même manière (barres en gris moyen) : la moyenne des taux d'accord est de 0,67 pour les phases 2 à 4 et cette moyenne est relativement stable par rapport à l'accord en phase 1 (0,58) puisque cette moyenne augmente seulement de 0,09 points.

La chimie (barres en gris foncé), testée à titre de comparaison, affiche des valeurs d'accord nettement supérieures : moyenne de 0,92 pour les phases 2 à 4 et une augmentation forte au fur et à mesure des phases (+0,50 points) entre la phase 1 et les suivantes).
Les différences constatées entre SHS et chimie d'une part, et entre les deux disciplines de SHS d'autre part, pourraient s'expliquer par le caractère plus ou moins disjoint du lexique de la discipline par rapport à celui de la langue générale. Le lexique de la chimie semble nettement plus disjoint que celui de l'archéologie ou celui de la linguistique comme le montrent les exemples fournis dans l'introduction.

Cependant, les particularités de la langue de spécialité ne sont pas la seule piste d'explication. Une autre piste semble être le profil des annotateurs. On observe, en effet, que le taux d'accord inter-annotateurs décroît en même temps qu'augmente le

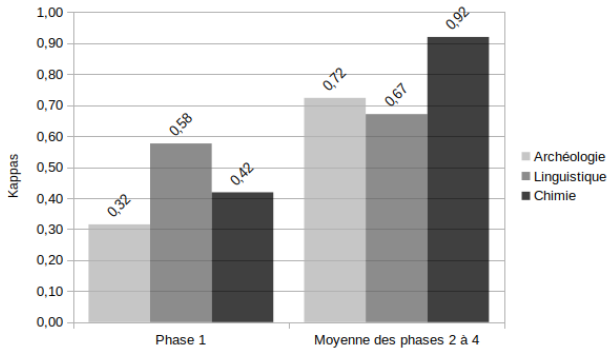

FIGURE 2. *Accords inter-annotateurs par discipline*

nombre d'annotateurs experts du domaine scientifique. La linguistique est annotée par deux linguistes exclusivement, l'archéologie par une archéologue et deux étudiants en terminologie et traduction que nous assimilerons à des terminologues dans la suite de notre propos. La chimie est exclusivement annotée par les terminologues qui ne sont pas des experts du domaine au sens de Fort (2015). Cette observation nous incite à faire l'hypothèse que la formation initiale des annotateurs experts du domaine scientifique des résumés fait obstacle à l'utilisation des recommandations fournies dans le guide d'annotation.
Nous avons testé cette hypothèse sur l'archéologie qui est la seule discipline comportant les deux types d'annotateurs, experts et non-experts du domaine. Pour comparer les accords obtenus entre experts et non-experts, nous avons calculé l'accord entre annotateurs et l'annotation issue de l'arbitrage. La figure (3, p. 14) montre une tendance à la diminution du taux d'accord avec l'annotateur archéologue (l'accord en gris foncé passe de 0,80 en phase 1 à 0,73 pour la moyenne des phases 2 à 4) et une augmentation du taux d'accord avec les annotateurs non experts (l'accord en gris clair passe de 0,36 en phase 1 à 0,76 pour la moyenne des phases 2 à 4).

L'hypothèse que les annotateurs terminologues deviennent progressivement experts de la tâche d'annotation est renforcée par l'évolution des taux d'accord entre les arbitres avant l'étape du consensus permettant d'arrêter une annotation de référence pour chaque résumé. La figure (4, p. 14) montre une hausse du taux d'accord dans les trois disciplines entre le taux en phase 1 d'une part et la moyenne de l'accord sur les phases 2 à 4 d'autre part.
Les arbitres linguistes ont traité les trois disciplines dans leur intégralité et ils ont été épaulés par une arbitre spécialiste de la chimie pour cette discipline. Comme les deux arbitres linguistes ont traité la totalité du corpus de la campagne C_2, on peut légitimement considérer que leur expertise dans la tâche d'annotation s'est accrue au fur et à mesure de la campagne. L'influence de l'expertise dans la tâche semble donc prendre le pas sur l'expertise disciplinaire. Ainsi, si un grand nombre de données avaient été annotées par les annotateurs spécialisés en linguistique et en archéologie, il est pro-

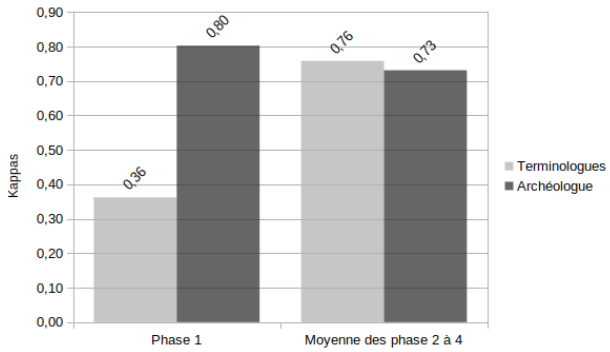

FIGURE 3. *Accord à la référence : terminologues et archéologue*

bable que l'évolution de leurs taux d'accord aurait été similaire à celle qu'on constate avec les arbitres.

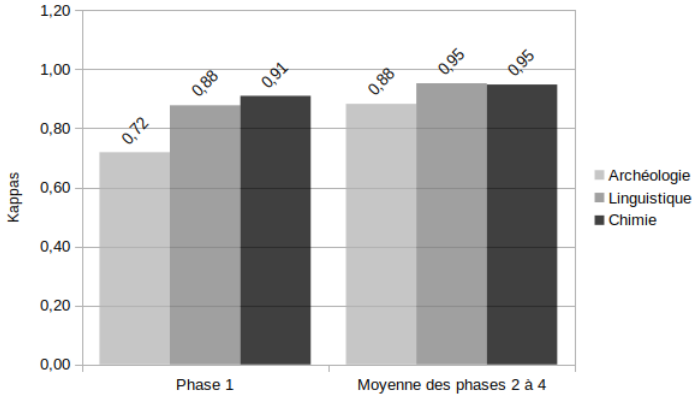

FIGURE 4. *Taux d'accord entre arbitres par discipline*

4. Comparaison des deux campagnes

À ce stade, on peut se demander quelle est la plus-value de la seconde campagne par rapport à la première, en particulier parce que son étendue est beaucoup plus réduite que celle de la première campagne. La question générale est de savoir si le fait de privilégier les bonnes pratiques dans la réalisation d'une campagne d'annotation est plus important que de privilégier la quantité de données. Pour apporter des éléments de réponse, nous avons procédé à deux types de comparaison. Dans la section (4.1), nous comparons ce qui peut l'être directement, à savoir les distributions de candidats termes et leurs occurrences. Pour aller au-delà, étant donné que les deux campagnes

n'ont pas suivi le même protocole (annotation non croisée pour la première, annotation multiple pour la seconde), nous avons utilisé une tâche externe nécessitant un apprentissage sur corpus avec l'objectif d'être ainsi en mesure de comparer la qualité des annotations produites dans chaque campagne (section 4.2).

4.1. *Examen des occurrences jugées disciplinaires*

Les effectifs des deux campagnes (tableau 1, p. 7 et tableau 2, p. 9) montrent que la campagne C_2 est plus sélective au niveau de la validation disciplinaire. En archéologie, 95,11 % des occurrences linguistiquement correctes sont jugées disciplinaires en C_1 contre 66,73 % en C_2. En linguistique, 67,82 % des occurrences linguistiquement correctes sont jugées disciplinaires en C_1 contre 54,38 % en C_2. Au-delà de ces nombres globaux, le tableau (3) fournit des informations plus détaillées sur les occurrences jugées disciplinaires dans chacune des campagnes.

	C_1		C_2	
	Archéologie	Linguistique	Archéologie	Linguistique
% / occurrences disciplinaires				
C.T.S.	76,07 % (10 316)	79,70 % (12 482)	56,79 % (205)	79,54 % (311)
C.T.C.	23,93 % (3 245)	20,30 % (3 180)	43,21 % (156)	20,46 % (80)
Taux de répétition moyen des C.T.				
C.T.S.	7,07 (+/-17,18)	6,22 (+/-22,27)	1,80 (+/-1,95)	1,86 (+/-1,74)
C.T.C.	2,63 (+/-4,87)	2,45 (+/-3,52)	1,24 (+/-0,65)	1,14 (+/-0,43)
Taille moyenne des C.T.C.	4,00 (+/-1,58)	4,67 (+/-2,16)	2,46 (+/-0,59)	2,34 (+/-0,54)

TABLEAU 3. *Distribution des occurrences disciplinaires dans les deux campagnes*

Pour la linguistique, la répartition des occurrences disciplinaires entre les candidats termes simples (occurrences monolexicales, *C.T.S.* dans le tableau 3) et complexes (occurrences composées de plusieurs lexèmes, *C.T.C.* dans le tableau 3) est similaire pour les deux campagnes : 80 % environ pour les C.T.S. *versus* 20 % pour les C.T.C. Pour l'archéologie, la répartition est proche de celle de la linguistique en C_1 mais en C_2, la répartition entre les deux catégories de candidats s'approche davantage de 50 % : en C_2, un peu plus de 56 % pour les C.T.S. *versus* un peu plus de 43 % pour les C.T.C.

La seconde caractéristique quantitative que nous avons calculée est celle du taux de répétition moyen pour les candidats simples *versus* les candidats complexes. Nous donnons la moyenne pondérée ainsi que l'écart-type entre parenthèses. De manière non suprenante, on constate que les candidats termes simples ont un taux de répétition plus important par rapport aux candidats complexes (par exemple, taux de répétition de 7,07 pour les C.T.S. en archéologie en C_1 contre 2,63 pour les C.T.C.) et que le taux de répétition en C_1 est plus important qu'en C_2, tous types de candidats confondus.

Ceci s'explique par le fait que le corpus de C_1 est composé d'articles intégraux alors que le corpus de C_2 est composé de résumés.

Enfin, toujours lié à la différence de nature dans les documents qui composent les corpus des deux campagnes, la taille moyenne des candidats complexes est plus élevée dans la première campagne que dans la seconde. La taille moyenne des candidats complexes est de 4,00 en archéologie et de 4,67 en linguistique dans la campagne C_1. Elle est de 2,46 en archéologie et de 2,34 en linguistique dans la campagne C_2.

Ces éléments chiffrés, pour cohérents qu'ils soient avec la nature des documents annotés dans les deux campagnes, ne fournissent pas d'explication au taux de validation disciplinaire qui est moindre en C_2. Ce taux de validation moins important n'est par ailleurs pas une information permettant de mesurer la qualité des annotations produites dans chacune des campagnes, ni de la comparer entre les deux campagnes.
Pour avoir une vision plus objective et plus synthétique de la qualité des annotations produites dans les deux campagnes, nous avons mis en œuvre leur comparaison via une tâche externe.

4.2. *Évaluation comparée par la tâche*

La campagne C_1 n'a pas été évaluée par des mesures d'accords inter-annotateurs puisqu'elle n'a pas été réalisée dans un contexte d'annotation multiple. La campagne C_2, plus réduite, plus ciblée et fondée sur un protocole d'annotation inspiré des bonnes pratiques, a été évaluée de cette manière. De ce fait, les deux campagnes ne sont pas directement comparables. Pour réaliser cette comparaison et ainsi évaluer la qualité du protocole d'annotation de la campagne C_2 par rapport à celui de la campagne C_1, nous mesurons les performances d'une même application qui utilise un apprentissage sur corpus. La comparaison entre les deux campagnes peut alors se faire en utilisant les corpus produits par chacune des campagnes. En cela, nous appliquons la stratégie courante utilisée dans les campagnes d'évaluation mais en changeant d'objectif. L'enjeu n'est pas de comparer des systèmes mais de comparer les corpus d'apprentissage et de déduire une évaluation de leur qualité en tant que corpus de référence ainsi qu'une évaluation des protocoles d'annotation sous-jacents.
L'application utilisée ici consiste à automatiser l'identification des occurrences disciplinaires de candidats termes. La procédure est décrite dans la figure (5, p. 17).

Le système d'annotation automatique décrit à l'aide de la procédure ci-dessus a été présenté dans le cadre de conférences centrées sur le traitement automatique ou sur les statistiques textuelles. Il s'inscrit dans le champ des méthodes de désambiguïsation lexicale de type probabiliste supervisée. Ce système calcule des profils caractéristiques des deux acceptions entre lesquelles il doit décider : l'acception disciplinaire qui a reçu l'étiquette « 2 » lors de l'annotation manuelle et l'acception non disciplinaire qui a reçu les étiquettes « 0 » ou « 1 ». Plus précisement, ce système

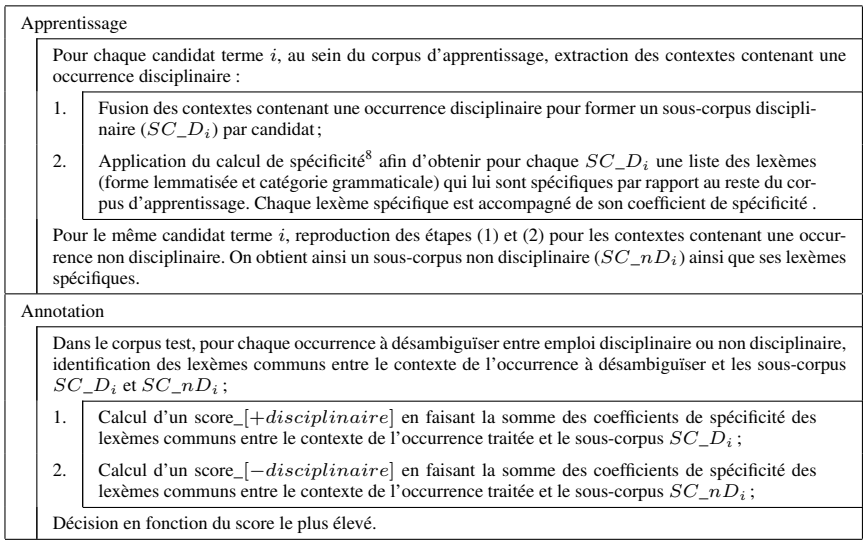

FIGURE 5. *Procédure du système d'annotation automatique*

s'inspire des algorithmes de désambiguïsation sémantique de type Lesk (1986) à la différence que dans notre cas, au lieu d'utiliser des dictionnaires, nous utilisons le corpus d'apprentissage pour calculer les profils de chaque acception. C'est pourquoi, lorsque le candidat à annoter est absent du corpus d'apprentissage, aucun modèle n'est appris et le système ne prend pas de décision.

Préalablement à l'application de la tâche utilisée, nous créons un échantillon-test qui correspond à l'intersection des corpus produits lors des deux campagnes afin de réaliser les tests sur des données comparables (tableau 4).

Disciplines	Archéologie et linguistique
Type de documents	Résumés de C_2 qui correspondent aux résumés des articles intégraux de C_1
Nombre de documents	Alignement sur le plus petit nombre de document de chaque corpus : 11 documents en archéologie et 30 documents en linguistique
Choix des annotations	Candidats et occurrences communs entre les deux corpus

TABLEAU 4. *Critères de constitution de l'échantillon commun entre* C_1 *et* C_2

L'échantillon extrait en fonction de ces critères est très restreint par rapport au nombre de candidats dans chacune des campagnes (tableau 5).

Sur le plan qualitatif, on observe une différence du nombre d'occurrences jugées disciplinaires au sein du même échantillon extrait des corpus annotés dans chacune des

	Archéologie		Linguistique		Échantillon	
	Types	Occurrences	Types	Occurrences	Types	Occurrences
Candidats	8	19	47	157	55	176
Disciplinaires C_1	5	10	21	62	26	72
Disciplinaires C_2	6	15	27	87	33	102

TABLEAU 5. *Distribution des candidats termes de l'échantillon-test*

campagnes. En archéologie (nombres d'occurrences pour les lignes « Disciplinaires C_1« *versus* « Disciplinaires C_2 » dans le tableau 5), la différence de 5 occurrences vient entièrement du changement des recommandations d'annotation pour la validité linguistique entre les deux campagnes. Dans la campagne C_2, la recommandation de rejeter tout élément isolé, et notamment les adjectifs, a été supprimée. De ce fait, lors de la campagne C_2, 5 occurrences supplémentaires, correspondant à 2 candidats adjectivaux (*archéologique* et *magdalénien*), ont été conservées en tant qu'occurrences linguistiquement correctes puis en tant qu'occurrences disciplinaires. En linguistique (mêmes lignes dans le tableau 5 mais pour la linguistique), il y a une différence de 25 occurrences entre C_1 et C_2. Parmi celles-ci, 18 occurrences correspondant à 8 candidats adjectivaux, sont conservées en C_2. Les 7 occurrences disciplinaires restantes viennent de choix différents réalisés par les arbitres de C_2 et par les annotateurs de C_1.

Sur le plan quantitatif, étant donné la taille de l'échantillon, et à l'intérieur de celui-ci, le très petit nombre d'occurrences disciplinaires pour chaque campagne (en particulier en archéologie), nous avons utilisé le test du χ^2 pour déterminer si cet échantillon était représentatif des campagnes[9]. Pour calculer les valeurs de χ^2, nous avons comparé les distributions des candidats (nombre d'occurrences de candidats / nombre d'occurrences disciplinaires parmi elles) de l'échantillon avec celui de chaque campagne et nous avons reproduit ces comparaisons pour chaque discipline. Les valeurs obtenues pour chaque comparaison sont données dans le tableau (6).

L'interprétation des valeurs obtenues reproduit le raisonnement de Laurencelle (2012, p. 175). On observe que toutes les valeurs obtenues sont inférieures à la valeur théorique de χ^2, qui est de 3,86 avec une probabilité d'erreur de 0,05 et un degré de liberté égal à 1. On en déduit que l'échantillon est représentatif de chacune des campagnes, et à l'intérieur de celles-ci, de chacune des disciplines concernées. Dit autrement, nous déduisons des valeurs de χ^2 ci-dessus que la différence de distribution entre l'échantillon et chaque campagne n'est pas significative.

9. Nous remercions les relecteurs des versions précédentes de cet article pour leur vigilance sur ce point.

	Occurrences de candidats		Occurrences disciplinaires		
	Campagne	Échantillon	Campagne	Échantillon	χ^2
C_1	58 843	176	29 223	72	1,74
Archéologie_C_1	23 857	19	13 561	10	0,000 014
Linguistique_C_1	34 986	157	15 662	62	0,58
C_2	1 360	176	752	102	0,08
Archéologie_C_2	582	19	361	15	0,26
Linguistique_C_2	778	157	391	87	0,35

TABLEAU 6. *Valeurs de χ^2 pour l'échantillon et les campagnes d'annotation*

Suite à la sélection de cet échantillon, l'évaluation des performances du système qui est appliquée suit l'approche de Daille *et al.* (2016). Il s'agit de mettre en œuvre une évaluation endogène par lots pour éviter une nouvelle phase de validation manuelle des résultats du système. Le principe de cette évaluation consiste à subdiviser les données testées en deux sous-corpus : 90 % du corpus initial est assimilé à un corpus d'apprentissage et 10 % à un corpus de validation. Dans le cas de l'archéologie, cette subdivision est approchée en prenant dix résumés comme corpus d'apprentissage et un résumé comme corpus de validation. En linguistique, le corpus d'apprentissage comporte vingt-sept résumés et le corpus de validation, trois résumés. Afin que toutes les occurrences de candidats termes de l'ensemble des fichiers puissent être évaluées, la méthode consiste à créer autant de lots qu'il y a de subdivisions possibles, soit onze lots en archéologie et dix en linguistique. Les résultats pour chaque lot sont ensuite fusionnés pour chaque candidat afin d'obtenir un F-score par candidat sur l'ensemble des lots.

Les figures (6), p. 20, et (7), p. 21, synthétisent les performances du système mesurées pour les occurrences des candidats de l'échantillon-test. Chaque barre verticale représente la valeur d'un F-score : les barres en gris clair représentent le F-score obtenu lorsque l'apprentissage a été réalisé avec le corpus C_1 ; les barres en gris foncé représentent le F-score obtenu avec le corpus C_2. Lorsque l'une des valeurs de F-score est égale à 0, ce qui signifie que le système s'est trompé pour toutes les occurrences du candidat, la barre verticale correspondante est absente du diagramme. Nous avons synthétisé les F-scores obtenus en fonction du type d'évolution. Les deux disciplines confondues, 5 cas de figure sont apparus : type (1a), augmentation du F-score en C_2 par rapport au F-score en C_1 ; type (1b), cas extrême de cette augmentation, c'est-à-dire évolution de 0,00 en C_1 à 1,00 en C_2 ; les types (2a) et (2b) sont les deux cas correspondant lorsque les F-score sont en baisse ; et le type (3) correspond au cas où il n'y a aucune évolution selon que l'on utilise l'un ou l'autre des corpus.

En archéologie (figure 6), les résultats obtenus sont de 3 types. Le type (1a) concerne un seul candidat, l'adjectif *magdalénien*, pour lequel le F-score passe de 0,67 avec C_1 à 1,00 avec C_2. Le type (1b) concerne 3 candidats nominaux, *biostratigraphie, paléolithique* et *sépulture*, pour lesquels le F-score passe de 0,00 avec C_1 à 1,00 avec C_2. Enfin, le type (3) concerne les 4 derniers candidats de l'archéologie (les adjectifs *archéologique* et *méthodologique*, le nom *cours* et le candidat complexe *paléolithique supérieur*) pour lesquels le F-score reste à 1,00 quel que soit le corpus d'apprentissage utilisé. Le F-score moyen pour l'ensemble des candidats passe de 0,58 avec le corpus de la campagne C_1 à 1,00 avec le corpus de la campagne C_2, soit une augmentation moyenne du F-score de 0,42.

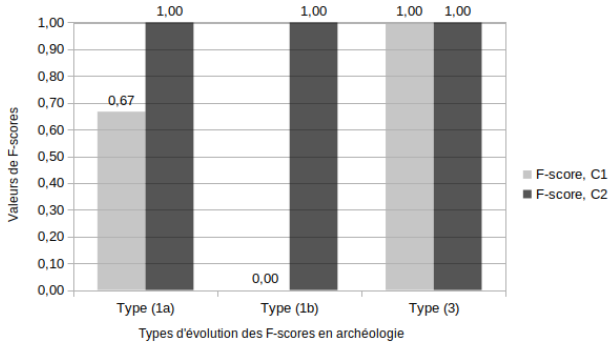

FIGURE 6. *Moyennes des F-scores obtenus par type d'évolution en archéologie*

En linguistique (figure 7), p. 21, les résultats sont moins nets et plus diversifiés. Le F-score moyen pour l'ensemble des 47 candidats (157 occurrences) passe de 0,80 avec C_1 à 0,87 avec C_2, soit une augmentation plus faible qu'en archéologie. Parmi ces 47 candidats, pour 30 d'entre eux, correspondant à 88 occurrences (56,05 % du total des occurrences de candidats en linguistique dans l'échantillon), il n'y a aucune évolution (type 3). Parmi les 30 candidats concernés, un seul reste à 0,00 (le nom *interaction*) tandis que les 29 autres restent à 1,00 (par exemple, *corpus, critère, culturel, définition, discursif, distinct, donnée* etc). Ceci explique le F-score moyen correspondant qui est de 0,97 et non de 1,00. Les 17 candidats qui connaissent une évolution se répartissent dans les 4 types d'évolution à la hausse ou à la baisse. Avec les candidats *forme, discours, langue* par exemple, 12 candidats (51 occurrences, soit 32,48 % du total des occurrences en linguistique) connaissent une amélioration du F-score de 0,40 en moyenne (type 1a). Le type (1b) concerne un seul candidat, le nom *structure*, qui apparaît à hauteur de 7 occurrences, soit 4,46 % du total des occurrences. Du côté de la dégradation des F-scores, le type (2a) concerne 2 candidats, les noms *construction* et *contexte* (6 occurrences, 3,82 % du total des occurrences), avec une baisse moyenne de 0,30. Le type (2b) enfin concerne 2 candidats, les noms *expression* et *signification* (5 occurrences, 3,18 % du total des occurrences).

En conclusion, en linguistique, les résultats sont majoritairement en augmentation avec le corpus C_2. Les cas d'augmentation représentent au total 58 occurrences

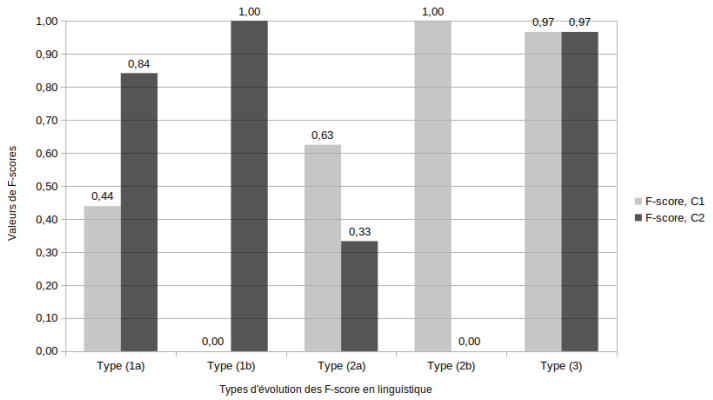

FIGURE 7. *Moyennes des F-scores obtenus par type d'évolution en linguistique*

(13 candidats), soit 36,94 % du total des 157 occurrences de candidats termes de l'échantillon-test pour la linguistique. Pour ces occurrences mieux étiquetées automatiquement, l'augmentation moyenne du F-score est de 0,44 (0,85, le F-score moyen des types (1a) et (1b) pour C_2 auquel on soustrait 0,41, le F-score moyen des types (1a) et (1b) pour C_1). Comparativement, les cas de dégradation du F-score avec C_2 représentent 11 occurrences (4 candidats), soit 7,01 % du total des 157 occurrences de candidats termes de l'échantillon-test pour la linguistique. Pour ces occurrences moins bien étiquetées automatiquement avec le corpus C_2 comme corpus d'apprentissage, la dégradation moyenne du F-score est de -0,65 (même mode de calcul que pour l'augmentation moyenne). Cependant, les variations moyennes des F-score (augmentation et baisse) rapportées aux proportions d'occurrences qu'elles concernent montrent que les cas d'augmentation sont majoritaires : +0,17 (0,44 * 36,94 %) contre -0,05 (-0,65 * 7,01 %).

Au vu des résultats obtenus pour l'archéologie et la linguistique, on peut conclure que le corpus de la campagne C_2 apparaît comme de meilleure qualité en tant que corpus d'apprentissage pour le système d'annotation automatique car il apporte une amélioration du F-score pour les deux disciplines.

5. Conclusion

Dans cet article, nous avons décrit et analysé deux campagnes d'annotation manuelle en occurrences de candidats termes. Ces deux campagnes ont été réalisées sur des corpus issus d'un même corpus-source pré-annoté : segmentation lexicale, étiquetage morphosyntaxique et identification des occurrences de candidats termes. L'annotation manuelle avait pour objectif l'identification des occurrences disciplinaires des candidats mais s'inscrivait dans des enjeux globaux différents : enrichissement

de terminologies existantes pour la campagne C_1 et comparaison entre disciplines et validation d'un protocole d'annotation conforme aux bonnes pratiques pour la campagne C_2.

Les deux campagnes portant sur des volumes de données très différents, il n'a pas été possible d'évaluer la qualité des annotations réalisées dans la campagne C_1 à l'aide de mesures d'accord inter-annotateurs. Pour la campagne C_2, en regard de la campagne ESTER2 qui nous semble la plus proche en termes de tâche d'annotation, les taux d'accord obtenus sont comparables pour l'archéologie (0,72 pour les trois dernières phases, considérées comme les phases de production contre 0,71 pour ESTER2), inférieurs en linguistique (0,67 pour les trois dernières phases) et supérieurs en chimie (0,92 pour les trois dernières phases). Outre les différences connues en matière de spécificité du vocabulaire disciplinaire entre la science dite exacte de la campagne C_2 (chimie), et les deux disciplines SHS (archéologie et linguistique), le profil des annotateurs apparaît comme une piste d'explication complémentaire. Cette piste est renforcée par l'observation de l'évolution des accords entre arbitres.

Les conditions de réalisation des deux campagnes ne permettant pas de les comparer du point de vue des taux d'accord inter-annotateurs, nous avons mis en œuvre une évaluation par la tâche à l'aide d'un système consistant à automatiser l'identification des occurrences disciplinaires de candidats termes. Lorsque le corpus annoté au cours de la seconde campagne est utilisé comme corpus d'apprentissage, les résultats augmentent globalement pour les deux disciplines communes aux deux campagnes, l'archéologie et la linguistique. L'augmentation est nette lorsqu'on ne prend en compte que les cas d'évolution positive. En archéologie, le F-score passe de 0,58 à 1,00 avec le corpus C_2 et il passe de 0,41 à 0,85 en linguistique dans les mêmes conditions. Mais cette augmentation est pondérée en linguistique par 4 candidats pour lesquels l'utilisation du corpus C_2 produit une évolution négative.

À l'issue de cette expérience, il apparaît que le protocole d'annotation de la seconde campagne peut être poursuivi et développé. Le contexte d'annotation multiple permet l'évaluation par les mesures d'accord inter-annotateurs, ceci rendant les résultats plus aisément comparables. L'amélioration constatée dans le cadre d'une évaluation par la tâche lors de l'utilisation du corpus annoté dans la campagne C_2 est sensible. Pour développer cette première expérience visant la production de corpus français annotés libres de droits, il sera cependant nécessaire de veiller à éviter le caractère fastidieux de la première campagne : étendue des données à annoter, répétition des mêmes choix un grand nombre de fois pour les occurrences d'un même candidat dans des contextes sémantiquement similaires. Pour cela, une piste possible pourrait être d'annoter des ensembles d'occurrences jugées similaires en fonction de leur contexte plutôt que d'annoter occurrence par occurrence. Une seconde piste serait d'exploiter le système d'annotation automatique afin de pré-annoter les occurrences avec l'objectif de limiter au maximum l'effort de prise de décision demandé à l'annotateur. Le dernier axe de développement possible serait d'exploiter des corpus bi-

lingues français-anglais afin de mettre en œuvre des méthodes de désambiguïsation contextuelle bilingue (Morin *et al.*, 2010).

6. Bibliographie

Artstein R., Poesio M., « Inter-coder agreement for computational linguistics », *Computational Linguistics*, vol. 34, n° 4, p. 555-596, 2008.

Bonneau-Maynard H., Rosset S., Ayache C., Kuhn A., Mostefa D., « Semantic annotation of the French Media dialog corpus », *Proceedings of the InterSpeech*, Lisbonne, Portugal, 2005.

Bougouin A., Boudin F., Daille B., « Modélisation unifiée du document et de son domaine pour une indexation par termes-clés libre et contrôlée », *TALN*, Paris, France, 2016.

Cabré Castelli M., Estopà R., Palatresi J. V., *Automatic Term Detection : A Review of Current Systems*, Recent Advances in Computational Terminology, John Benjamins Publishing Compagny, chapter 3, p. 53-87, 2001.

Cohen J., « A coefficent of agreement for nominal scales », *Educational and Psychological Measurement*, vol. 20, n° 1, p. 37-46, 1960.

Cunningham H., Tablan V., Roberts A., Bontcheva K., « Getting More Out of Biomedical Documents with GATE's Full Lifecycle Open Source Text Analytics », *PLOS Computational Biology*, 2013.

Daille B., Jacquey É., Lejeune G., Melo L. F., Toussaint Y., « Ambiguity Diagnosis for Terms in Digital Humanities », *Proceedings of Language Resources and Evaluation Conference*, Portoroz, Slovénie, Mai, 2016.

Daille B., Kageura K., Nakagawa H., Chien L.-F., *Terminology. Special Issue on Recent Trends in Computational Terminology*, vol. 10, John Benjamins Publishing Compagny, 2004.

Davies M., Fleiss J., « Measuring agreement for multinominal data », *Biometrics*, vol. 34, n° 4, p. 1047-1051, 1982.

Fort K., Les ressources annotées, un enjeu pour l'analyse de contenu : vers une méthodologie de l'annotation manuelle de corpus, PhD thesis, Université Paris- Nord - Paris XIII, 2012. 263 pages.

Fort K., « Experts ou (foule de) non-experts ? la question de l'expertise des annotateurs vue de la myriadisation (crowdsourcing) », *Journées Internationales de Linguistique de Corpus*, Orléans, France, September, 2015.

Fort K., *Collaborative Annotation for Reliable Natural Language Processing : Technical and Sociological Aspects*, Wiley-ISTE, July, 2016.

Galliano S., Gravier G., Chaubard L., « The ESTER 2 evaluation campaign for the rich transcription of French radio broadcasts », *Proceedings of the Interspeech*, Brighton, Angleterre, 2009.

Heiden S., Magué J.-P., Pincement B., « TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement », *Actes de la conférence JADT 2010 : 10th International Conference on the Statistical Analysis of Textual Data*, 2010. 12 pages.

Ide N., Romary L., « Representing linguistic corpora and their annotations », *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Gène, Italie, 2006.

Jacquemin B., French document base and text information : structuring, rephrasing, questioning, PhD thesis, Université de la Sorbonne nouvelle - Paris III, December, 2003.

Krippendorff K., « On the reliability of unitizing continuous data », *Sociological Methodology*, vol. 25, p. 47-76, 1995.

Lafon P., « Sur la variabilité de la fréquence des formes dans un corpus », *Mots*, vol. 1, p. 127-165, 1980.

Laurencelle L., « La représentativité d'un échantillon et son test par le Khi-deux », *Tutorials in Quantitative Methods for Psychology*, vol. 8, no 3, p. 173-181, 2012.

Lesk M., « Automatic sense disambiguation using MRD : how to tell a pine cone from an ice cream cone », *Proceeding of SIGDOC' 86, ACM*, New York, USA, p. 24-26, 1986.

L'Homme M.-C., *La terminologie : principes et techniques*, Presses de l'Université de Montréal, 2004.

L'Homme M.-C., « Sur la notion de « terme » », *Meta : journal des traducteurs/Meta : Translators' Journal*, vol. 50, no 4, p. 1112-1132, 2005.

Lommel A., Melby A. K., Glenn N., Hayes J., Snow T., « TBX-Min : A Simplified TBX-Based Approach to Representing Bilingual Glossaries », *Terminology and Knowledge Engineering 2014*, Berlin, Germany, p. 10 p, June, 2014.

Mathet Y., Widlöcher A., « Évaluation des annotations : ses principes et ses pièges », *Traitement Automatique des Langues*, vol. 57, no 2, p. 73-98, December, 2016.

Melby A., *TBX : A terminology exchange format for the translation and localization industry*, vol. 1 of *Handbook of Terminology*, John Benjamins Publishing Company, chapter Partie III, chapitre 6, 2015.

Morin E., Daille B., Takeuchi K., Kageura K., « Brains, not brawn : The use of 'smart' comparable corpora in bilingual terminology mining », *TSLP 7*, vol. 1, ACM DL, 2010.

Nazar R., « Distributional analysis applied to terminology acquisition », *Terminology*, vol. 22, no 2, p. 141-170, 2016.

Nazarenko A., Zargayouna H., Hamon O., Van Puymbrouck J., « Evaluation des outils terminologiques : enjeux, difficultés et propositions », *Traitement Automatique des Langues*, vol. 50, no 1 varia, p. 257-281, 2009.

Rastier F., « Pour une sémantique des textes théoriques », *Revue de sémantique et de pragmatique*, vol. 17, p. 151-180, 2005.

Rocheteau J., Daille B., « TTC TermSuite : A UIMA Application for Multilingual Terminology Extraction from Comparable Corpora », *5th International Joint Conference on Natural Language Processing (IJCNLP)*, Chiang Mai, Thailand, p. 9-12, November, 2011. System Demonstrations.

Romary L., « TBX goes TEI – Implementing a TBX basic extension for the Text Encoding Initiative guidelines », *Terminology and Knowledge Engineering 2014*, Terminology and Knowledge Engineering, TKE 2014, Berlin, Germany, June, 2014.

Schmid H., « Probabilistic Part-of-Speech Tagging Using Decision Trees », *Proceedings of International Conference on New Methods in Language Processing*, p. 154-164, 1994.

Stenetorp P., Pyysalo S., Topić G., Ohta T., Ananiadou S., Tsujii J., « BRAT : A Web-based Tool for NLP-assisted Text Annotation », *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, Association for Computational Linguistics, Stroudsburg, PA, USA, p. 102-107, 2012.

Vàsquez M., Oliver A., « Improving term candidates selection using terminological tokens », *Terminology*, vol. 24, n° 1, p. 122-147, 2018.

Véronis J., « Sense tagging : does it make sense ? », *Proceedings of the Corpus Linguistics Conference*, vol. 13, 2001.

Widlöcher A., Mathet Y., « The Glozz platform : a corpus annotation and mining tool », *in* C. Concolato, P. Schmitz (eds), *Proceedings of the ACM Symposium on Document Engineering (DocEng'12)*, Paris, France, p. 171-180, September, 2012.

Yimam S. M., Gurevych I., Eckart de Castilho R., Biemann C., « WebAnno : A Flexible, Web-based and Visually Supported System for Distributed Annotations », *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics : System Demonstrations*, Association for Computational Linguistics, Sofia, Bulgaria, p. 1-6, August, 2013.

Zesch T., Gurevych I., « Approximate Matching for Evaluation Keyphrase Extraction », *Proceedings of the International Conference RANLP*, p. 484-489, 2009.