

Université de Nantes

Reconnaissance des langues traitées dans
les articles de TALN
&
Applications à la mesure de la diversité
linguistique

WIBAUX Robin & COILLEROT Carol

Travail d'Étude et de Recherche

UFR Sciences & Techniques
Département informatique

18 juin 2018

Table des matières

1	Introduction	1
2	La diversité des langues dans le monde	2
2.1	Un baromètre des langues	2
2.2	La dotation des langues	4
2.3	Niveaux de langues, politiques linguistiques	6
3	Méthodologie	8
3.1	Constitution du corpus	8
3.2	Mesurer un baromètre du TALN	8
3.3	Limites de la méthode	9
3.4	Langage utilisé	10
4	Mise en place des données	11
4.1	Constitution du corpus brut	11
4.2	Constitution de la liste des langues	12
5	Prétraitement du corpus brut	14
5.1	Les outils de conversion des documents pdf	16
5.2	Nettoyage des données	17
5.2.1	Problèmes lors des conversions	18
5.2.2	Élagage des sources de bruit	20
6	Lectures et Mesures	22
6.1	Présentation des programmes d'analyse	22
6.2	Comparaison des résultats manuels avec les résultats du programme	24
6.3	Analyses entre corpus	26
6.3.1	Résultat du Fréquencier comparatif sur les corpus ACL et LREC	26
6.3.2	Détails du résultat du Fréquencier comparatif sur un échantillon de 18 articles	29
6.3.3	Bilan du processus d'élagage	31
6.4	Analyse diachronique	32
6.5	Étude sur les noms de corpus	34
7	Améliorations & Discussion	38
7.1	Pistes d'améliorations	38
7.1.1	Amélioration des prétraitements	38

7.1.2	Le problème de déductions implicites	39
7.1.2.1	Travaux bilingues	39
7.1.2.2	Familles de langues	40
7.2	Conclusion	41
A	Tableau d'évaluation du niveau informatique d'une langue	42
B	Liste de 8436 langues	44
C	Scripts bash	46
C.1	links_ACL.sh	46
C.2	dl_ACL.sh	48
D	Glossaire	49
	Bibliographie	50

1. Introduction

Ce travail de recherche a pour objectif de déterminer dans quelle mesure les travaux en Traitement Automatique des Langues Naturelles (TALN) sont consacrées à certaines langues plutôt qu'à d'autres.

Tout d'abord, nous introduirons le contexte et les enjeux de la diversité linguistique dans le monde, à l'origine des motivations de nos recherches.

Nous décrirons ensuite le choix de notre méthodologie, ses fondements et ses failles.

Puis nous détaillerons la préparation des données, de leur mise en place aux processus de prétraitements.

Alors, nous présenterons nos différentes mesures, leurs objectifs, les résultats et les conclusions à en tirer.

Enfin, nous discuterons des pistes possibles pour poursuivre ce travail.

2. La diversité des langues dans le monde

2.1 Un baromètre des langues

Environ 7000 langues sont dénombrées aujourd’hui dans le monde, dont 386 qui sont parlées par plus d’un million de personnes en tant que langue maternelle^[1]. Ainsi 5% des langues sont la langue maternelle de 95% de la population mondiale, et inversement 95% des langues ne sont la langue maternelle que de 5% du monde.

Le linguiste Louis-Jean Calvet a présenté en 2010 un baromètre des langues ^[2] – réactualisé en 2012 – pour proposer une mesure la diversité des langues. Ce baromètre mesure l’usage d’une langue selon une dizaine de critères parmi lesquels le nombre de locuteurs, le nombre d’utilisations en tant que langue officielle, en tant que langue source/cible d’une traduction ou encore le nombre de prix internationaux de littérature. Deux paramètres sont directement liés à l’informatisation d’une langue : le nombre de pages **wikipédia** et le **taux de pénétration internet**.

Nous verrons ces deux indicateurs en détail, ainsi que les classements qui en résultent. Pour commencer, observons d’abord le critère du nombre de locuteurs. Il est fondé sur les données fournies par le site ethnologue¹, et ne considère que les locuteurs dont la langue est maternelle : un anglais qui parle le français comme seconde langue n’est pas comptabilisé pour cette langue et ce critère.

1. <https://www.ethnologue.com/>

Les cinq premières langues en nombre de locuteurs selon le baromètre de L.J. Calvet (2012) :

1. mandarin (845 033 031 locuteurs)
2. anglais (326 985 909 locuteurs)
3. espagnol (327 380 862 locuteurs)
4. bengali (180 624 200 locuteurs)
5. hindi (180 469 200 locuteurs)

Wikipédia est la plus grande encyclopédie participative et libre. L'hypothèse selon laquelle une page écrite dans une langue serait le fait d'un locuteur natif, est invalidée par les observations : le wiki en bambara a été initié par le néerlandais Kasper Souren. Cependant cet indicateur reste intéressant car la quantité (environ deux millions d'articles écrits en français) permet d'éviter ce biais. Les données utilisées par le baromètre proviennent directement de wikipédia².

Les cinq premières langues en nombre de pages Wikipedia selon le baromètre de L.J. Calvet (2012) :

1. anglais (3 850 426 articles)
2. allemand (1 301 944 articles)
3. français (1 166 939 articles)
4. italien (860 818 articles)
5. polonais (837 989 articles)

Le score d'accès à internet représente le pourcentage de la population d'un pays, ou d'une région, qui utilise internet. Ce score est tiré du site InternetWorldStats³, pour lequel un utilisateur d'internet est toute personne en capacité d'utiliser internet. C'est à dire qu'il a un accès disponible à internet et les connaissances minimales pour l'utiliser. Le **taux de pénétration internet**, quant à lui, calcule pour une langue sa proportion de locuteurs dans chaque pays et multiplie celle-ci par le score d'accès à internet de ces pays. Ainsi, pour la langue somali, 65% des locuteurs se trouvent en Somalie, 30% en Éthiopie, 3% au Kenya et 2% à Djibouti, le calcul est le suivant :

$$Internet_{somali} = 0.65 \cdot Internet_{Somalie} + 0.3 \cdot Internet_{Ethiopie} + 0.03 \cdot Internet_{Kenya} + 0.02 \cdot Internet_{Djibouti}$$

Remarquons que le total n'atteint pas 100% : en effet si le nombre de locuteurs d'une langue situés dans un pays est faible (par exemple inférieur à 1%) alors il est jugé négligeable et n'est pas pris en compte. Les données utilisées ont été relevées en 2011.

2. [https://meta.wikimedia.org/wiki/List_of_Wikipedias#Grand.Total](https://meta.wikimedia.org/wiki/List_of_Wikipedias#Grand_Total)

3. <https://www.internetworldstats.com/stats.htm>

Les six premières langues selon leur taux de pénétration internet, baromètre de L.J. Calvet, 2012 :

1. norvégien (94,40% des locuteurs)
2. suédois (92,14% des locuteurs)
3. groninois (88,30% des locuteurs)
4. danois (85,88% des locuteurs)
5. néerlandais (85,42% des locuteurs)
6. finnois (85,20% des locuteurs)

À l'exception du groninois, toutes les langues sont du nord de l'Europe, probablement parce que l'accès à internet y est largement répandu. Quant au groninois, elle est la seule de ces six langues à être parlée par moins d'un million de locuteurs (elle compte 592 mille locuteurs selon le baromètre de 2012 de Calvet). De nombreuses langues peu parlées figurent en tête du classement.

2.2 La dotation des langues

Les résultats observés sur le baromètre des langues montrent que les langues ne sont pas représentées sur internet proportionnellement au nombre de leurs locuteurs : par exemple, le mandarin, langue parlée par la plus grande population, ne se positionne qu'en treizième position selon le nombre de pages wikipédia et en cent-huitième place selon le taux de pénétration internet. Se pose alors la question suivante : pourquoi certaines langues sont-elles moins représentées que d'autres ? Cela peut être dû à des habitudes de vies différentes des locuteurs de ces langues, ou à des situations économiques, culturelles et sociales différentes. Une autre possibilité encore serait le manque d'outils informatiques pour les locuteurs de ces langues.

Vincent Berment s'est intéressé au problème des outils informatiques disponibles pour les différentes langues. Les outils informatiques en question sont ceux qui permettent à un locuteur de s'exprimer sur ordinateur dans sa langue : traitement de texte, communication orale, dictionnaires et traductions, *etc.* Dans le cadre de sa thèse[3], V. Berment a alors proposé d'estimer la qualité de ces outils informatiques selon la langue, introduisant la notion de **dotation informatique** d'une langue. Plus une langue donnée dispose d'outils informatique permettant son expression, mieux elle est considérée comme dotée.

Berment a ainsi proposé cette estimation sous la forme d'un tableau :

	Services / ressources	Criticité C_k (0 à 10)	Note N_k (/20)	Note pondérée ($C_k N_k$)
Traitement du texte				
	Saisie simple			
	Visualisation / impression			
	Recherche et remplacement			
	Sélection du texte ²			
	Tri lexicographique			
	Correction orthographique			
	Correction grammaticale			
	Correction stylistique			
Traitement de l'oral				
	Synthèse vocale			
	Reconnaissance de la parole			
Traduction				
	Traduction automatisée			
ROC				
	Reconnaissance optique de caractères			
Ressources				
	Dictionnaire bilingue			
	Dictionnaire d'usage			
Total		ΣC_k		$\Sigma C_k N_k$
Moyenne (/20)				$\Sigma C_k N_k / \Sigma C_k$

Figure 1 : Tableau d'évaluation du niveau d'informatisation d'une langue

Chaque ligne du tableau correspond à une ressource. Pour Berment, toute fonctionnalité ou tout service disponible pour l'utilisation d'une langue est une ressource : saisie simple, visualisation, recherche et remplacement, dictionnaire bilingue...

Chaque ressource est notée selon sa criticité, c'est-à-dire son importance dans l'utilisation de l'outil informatique au quotidien, mais aussi selon sa disponibilité : la ressource existe-t-elle ? Et si oui est-elle bien implémentée ou de qualité ? Par exemple, la saisie simple dans la langue maternelle d'un utilisateur obtiendrait une note de criticité de 10, et sa note d'implémentation dépendrait de la possibilité par exemple de bien noter les accents ou autres spécificités. À partir de ces deux notes et pour chaque ressource, une note pondérée est calculée ; la moyenne des notes pondérées est l'estimation du niveau d'informatisation de la langue.

Deux tableaux remplis sont proposés en annexe (chapitre A).

Berment définit alors une langue peu dotée notée π comme une langue dont la moyenne est inférieure à 10. Entre 10 et 13.99 une langue est considérée comme moyennement dotée et notée μ ; enfin une langue bien dotée est notée τ , quand son score est au-dessus de 14.

2.3 Niveaux de langues, politiques linguistiques

Les technologies informatiques sont de plus en plus utilisées dans les communications quotidiennes (mobiles multifonctions, ordinateurs personnels) et officielles (administration), devenant un facteur de création de richesse, d'évolution sociale et de développement humain. Dans le but de mieux comprendre les enjeux posés par l'utilisation des langues, il faut définir quelques notions : les différents **niveaux de langues** et les **politiques linguistiques**.

Nous distinguons quatre niveaux de langues :

- Une langue internationale est une langue officielle des Nation Unies (anglais, français, espagnol, arabe, russe, chinois, portugais)
- Une langue officielle est la langue qui s'impose aux services officiels de l'État
- Une langue nationale est considérée comme choisie par un pays, une nation ou même une communauté et dont la définition est variable selon les pays. Une langue nationale peut aussi être langue officielle.
- Une langue locale est une langue parlée non reconnue comme langue nationale

Ainsi la Guinée[4] compte sept langues nationales : le kpellé, le kissi, le bassari, le malinké, le loma, le poular, le soussou et le konianké. Cependant pour des motifs politiques (il s'agit de ne pas mettre en valeur une langue nationale par rapport aux autres), c'est la langue française, alors considérée comme neutre qui y est devenue la langue officielle. Enfin, une dizaine d'autres langues sont parlées, ce sont des langues locales.

À travers cet exemple se dessinent des dynamiques entre langues : le français est la langue officielle de la Guinée alors que seulement 15% à 25% de ses habitants la parlent. L'établissement du français en tant que langue officielle impose son utilisation, d'abord d'un point de vue administratif, via notamment la loi, les actes de mariages ou de décès. Mais aussi d'un point de vue social, car la langue officielle est bien souvent la langue de l'ascension sociale.

Une langue peut ainsi prendre le pas sur une autre. En conséquence, les États définissent des politiques linguistiques, c'est-à-dire l'ensemble des choix conscients effectués dans le domaine des rapports entre langue et vie nationale. Les États ne sont pas seuls à définir de telles politiques. L'UNESCO dans une déclaration de 2001⁴ vise à défendre et sauvegarder les quelques 7000 langues et fait des propositions notamment au niveau de l'informatisation des langues.

La mesure de cette informatisation permet alors de voir les dynamiques à l'oeuvre. La possibilité d'utiliser sa langue sur Internet va déterminer le degré d'implication de

4. <http://unesdoc.unesco.org/images/0012/001246/124687f.pdf#page=78>

chacun dans les sociétés du savoir émergentes. Le domaine du TALN est au coeur des dynamiques autorisant l'utilisation, basique ou de plus en plus poussée, d'une langue en informatique.

3. Méthodologie

3.1 Constitution du corpus

Dans notre travail de recherche, nous nous intéressons aux travaux qui sont réalisés en TALN. Dans le but de constituer un corpus, nous avons fouillé dans les conférences publiées par l'Association pour la Linguistique Informatique (ou Association for Computational Linguistic, ACL). Cette association présente chaque année des conférences sur le TALN. Les articles publiés lors de ces conférences sont hébergés sur le site internet de l'ACL¹, qui regroupe ainsi plus de 44 000 articles de recherches, organisés selon leurs conférences et années de publication.

Nous nous intéressons à deux des conférences : l'ACL et le LREC (International Conference of Language Resources and Evaluation), regroupant à eux deux près de 9000 articles. L'ACL est une référence mondiale dans le domaine du TAL. Quant au LREC, s'intéressant particulièrement à l'évaluation et aux ressources des langues, nous supposons pouvoir y trouver une plus grande quantité de travail sur les différentes langues du monde. Ce sont deux conférences internationales, avec des articles écrits en anglais. Nous travaillerons donc sur un corpus anglais.

Nous pouvons aussi penser aux conférences d'EACL, la section européenne d'ACL, mais craignons alors une sur-représentation des langues européennes, due au contexte géographique des conférences.

3.2 Mesurer un baromètre du TALN

Ce travail de recherche vise à fournir un baromètre des langues dans le domaine du TALN.

1. <https://aclweb.org/anthology/>

Nous souhaitons donc savoir, pour chaque langue, la proportion d'articles de recherche qui s'y intéresse. Nous définissons l'**effectif** d'une langue comme étant le nombre d'articles traitant de cette langue.

Pour réaliser nos mesures, nous faisons l'hypothèse qu'un article relatant un travail de recherche sur une langue mentionne celle-ci au moins une fois. Il s'agit par exemple de chercher le nombre d'articles contenant le mot "French" pour estimer le nombre d'articles traitant de la langue française. Nous définissons la **fréquence** d'une langue comme le nombre d'articles mentionnant une langue dans le corpus exploré. Le nombre de fois qu'un même article mentionne la langue n'a pas d'importance ; nous regardons **si** la langue est mentionnée, **ou non**.

Intuitivement, nous nous attendons par exemple à ce qu'il y ait bien plus d'articles de recherches travaillant sur l'anglais que d'articles portant sur n'importe laquelle des langues africaines.

3.3 Limites de la méthode

Rappelons notre hypothèse de départ :

Un article relatant un travail de recherche sur une langue mentionne celle-ci au moins une fois.

Premièrement, nous sommes conscients des cas de figure où cette hypothèse ne s'applique pas ou peu. Prenons par exemple un article écrit en anglais et présentant une étude fondée sur l'anglais : les auteurs peuvent ne pas préciser qu'ils travaillent sur l'anglais. Autre cas de figure, il se peut qu'un article mentionne, non pas des noms de langues, mais des noms de corpus particuliers. Le corpus nommé Verbmobil, par exemple, est un corpus pouvant contenir des phrases anglaises, allemandes ou japonaises.

Dans tous ces cas de figure, il s'agit de devoir trouver la langue d'étude de manière plus ou moins implicite, ce qui est impossible avec notre méthode. Il s'agit donc de potentielles sources de **faux-négatif**, c'est-à-dire d'articles que notre méthode ne permettra pas de reconnaître comme travaillant sur une langue. C'est ce qui différencie l'**effectif** de la **fréquence**.

Deuxièmement, remarquons que notre hypothèse de départ est très peu réciproque : ce n'est pas parce qu'un article mentionne une langue qu'il présente un travail de recherche sur cette langue. Un nom de langue peut être mentionné pour de nombreuses raisons. Il peut s'agir d'un homonyme ("A German student eating french fries" : deux langues reconnues à tort), d'une langue mentionnée en comparaison avec la langue traitée par l'article, ou encore d'une langue figurant dans le nom d'une parution citée.

Tous ces exemples sont des **faux-positifs**, c'est-à-dire des articles qui, selon notre hypothèse, seront identifiés comme traitant de recherches sur une langue alors qu'il n'en est rien. Différencions, pour une langue donnée, la fréquence **brute** de la fréquence **utile**. La **fréquence brute** d'une langue est le nombre d'articles la mentionnant – elle ou un homonyme –, tandis que sa **fréquence utile** est le nombre d'articles la mentionnant **et** la traitant. Le problème de faux-positif est alors ce qui différencie la **fréquence brute** de la **fréquence utile**.

Notre méthode est donc sujette aux faux-positifs et faux-négatifs. S'il est possible de réduire le taux de faux-positifs (nous verrons comment dans le chapitre 2 consacré aux prétraitements), le problème des faux-négatifs est plus compliqué à résoudre. Nous pourrions envisager d'autres moyens de chiffrer l'intérêt porté aux différentes langues. Une autre estimation complémentaire serait de connaître les montants des budgets de recherche en taln pour chaque recherche ou le nombre de recherches, ou les cursus universitaires...

3.4 Langage utilisé

Nous travaillons principalement en Perl. Le langage Perl permet d'utiliser les expressions régulières afin de fouiller un corpus.

Bien qu'idéalement, nous souhaiterions obtenir l'effectif des langues, c'est-à-dire le nombre d'articles travaillant sur chaque langue, nos scripts Perl ne peuvent que leur mesurer une fréquence, c'est-à-dire le nombre d'article les mentionnant. Nous pouvons, au mieux, réduire les sources de bruits – faux-positifs – pour nous rapprocher de la fréquence utile des langues, à savoir le nombre d'articles mentionnant une langue **et** travaillant dessus.

Nous nommons **fréquence observée** la mesure (en nombre d'articles) que nous obtenons pour une langue, et que nous nous efforçons de rapprocher de la fréquence utile.

4. Mise en place des données

4.1 Constitution du corpus brut

Pour rappel, le corpus brut est constitué des articles scientifiques issus des conférences de l'ACL et du LREC. Ces articles sont hébergés sur le site de l'ACL : aclweb.org/anthology/. Les conférences de l'ACL comptent près de 5000 articles, et celles du LREC en comptent 4000, ce qui totalise 9000 documents.

Nous avons automatisé le téléchargement des documents via des scripts bash. Les documents d'ACL se téléchargent en exécutant le script *links_ACL.sh*, puis le script *dl_ACL.sh*. Les documents de LREC se téléchargent via deux autres scripts nommés et fonctionnant similairement. Ces scripts tirent profit de l'organisation du site internet, dont l'ensemble des adresses internet respecte un formatage bien défini : tous les articles des conférences ACL sont accessibles via un lien de la forme : aclweb.org/anthology/P/P**/, où ** correspond à l'année de publication et où le 'P' représente la conférence ACL. Pour les conférences LREC, il suffit de remplacer 'P' par 'L' : aclweb.org/anthology/L/L**/.

Le script *links_ACL.sh* crée un dossier *ACL/* contenant des sous-dossiers numérotés selon les années de publication des articles (préfixé d'un 'P', pour la conférence ACL). Il ne prend aucun paramètre. A l'issue de son exécution, chaque sous-dossier contient un fichier nommé *links* où figurent l'adresse internet de chaque document à télécharger.

```
./pdfs/ACL/P07/link :  
https://aclweb.org/anthology/P/P07/P07-1000.pdf  
https://aclweb.org/anthology/P/P07/P07-1001.pdf  
https://aclweb.org/anthology/P/P07/P07-1002.pdf  
https://aclweb.org/anthology/P/P07/P07-1003.pdf  
...
```

Note : Les fichiers numérotés 1000, 2000, 3000... (les multiples de 1000) ne sont pas des publications scientifiques mais des sommaires des articles qui suivent.

Le script *dlACL.sh* doit être exécuté à la suite du précédent, dans le même répertoire. Il ouvre chaque fichier *links* et télécharge les documents ciblés par les adresses internet. A l'issue du téléchargement, les documents sont présents dans le sous-dossier correspondant à leur année de publication et dans le dossier correspondant à leur conférence – en l'occurrence ici, la conférence ACL –.

Cet ensemble de fichiers pdf constitue le corpus brut.

4.2 Constitution de la liste des langues

Notre méthodologie consiste à compter le nombre d'articles dans lesquels le nom d'une langue apparaît, pour déterminer la fréquence de cette langue. Nous l'appliquons via l'utilisation de scripts que nous avons écrit en Perl.

Notre travail faisant la suite des travaux de C. Enguehard et M. Mangeot[5], nous avons repris la liste de langues qu'ils avaient composée. Initiée sur la base des cent langues les plus parlées dans le monde selon le baromètre des langues de Louis-Jean Calvet¹, cette liste a été complétée à l'aide d'un concordancier. Ce programme recherche dans le corpus toutes les mentions du mot "*language*" et affiche la phrase (ou le groupe de mots) où apparaît ce mot. En étudiant les résultats, il a été possible de trouver manuellement des langues non présentes dans la liste initiale. La liste que nous avons récupérée présentait ainsi 289 noms de langues. Nous l'avons retouchée par la suite pour éliminer ou renommer des langues qui génèrent trop de faux-positifs comme Mod (que nous avons supprimé) et Sign (renommé "Sign language").

Une autre approche possible est de récupérer une liste de langues sur Internet, pouvant comporter idéalement tous les noms des langues existantes – soit près de 7000 –. Il est possible de trouver de telles listes, qui sont principalement l'œuvre du Summer Institute of Linguistics. Le site du World Atlas of Language Structures² propose une liste avec 2680 noms de langues. La liste fournie par le site Wiktionary³ en comporte quant à elle 8046 mais pose plusieurs problèmes.

Tout d'abord, le nombre de langues est étrange puisqu'il y a plus de langues dans leur liste que de langues parlées dans le monde, or elle ne comporte pas les langues mortes. Ensuite, les informations renseignées sont parfois non fondées, avec le latin comme seul système d'écriture pour le Bambara. Enfin des langues comme l'Akar sont comptabilisées alors même qu'il n'existe aucune page *wiktionary* écrite en cette langue.

1. <http://wikilf.culture.fr/barometre2012/>

2. <http://wals.info/>

3. https://en.wiktionary.org/wiki/Wiktionary:_List_of_languages

Nous avons finalement opté pour la liste disponible sur le site glottolog⁴ qui dénombre 8436 langues. Le fait que cette liste recense un nombre de langue supérieur au nombre de langues parlées dans le monde s'explique par la présence des langues mortes comme le grec ancien, l'hébreu ancien, etc... Comme pour la première liste, nous avons éliminé de nombreux noms de langues, notamment les noms sur deux lettres, et les noms correspondant à un nom courant en anglais : The, As, Even...

Opter pour une liste exhaustive présente plusieurs inconvénients. Restreindre la taille de la liste et la constituer soi-même permet de mieux connaître les données sur lesquelles on travaille et de garder un meilleur contrôle sur les recherches. La liste exhaustive peut non seulement renfermer du bruit, comme les "The", "As", "Even" que nous évoquions précédemment, mais également des erreurs cachées ou des ambiguïtés. Il existe par exemple des langues ayant plusieurs noms alternatifs.

L'ensemble des mesures que nous présentons dans le chapitre 6 ont été effectués sur la liste restreinte des langues. Nous avons aussi choisi d'effectuer des expérimentations⁵ fondées sur la liste exhaustive afin de comparer les résultats. Nous pourrions déterminer les apports et les inconvénients d'une liste complète par rapport à une liste réduite.

4. <http://glottolog.org/glottolog/language>

5. De brefs résultats de ces expérimentations sont mis en annexe, chapitre B

5. Prétraitement du corpus brut

L'objectif premier du prétraitement est d'obtenir un corpus textuel à partir du corpus brut constitué de documents *pdf*. Il est possible de convertir les documents *pdf* au format *txt*, au prix de la perte de la mise en page du document, ou dans un format permettant de récupérer cette mise en page, comme le *html* ou le *xml*.

Le second objectif est de réduire certaines sources de bruit. Comme nous l'avons vu en introduction, un bruit correspond à la mention d'une langue (ou d'un homonyme) dans un article sans que cette langue ne soit le support de travail ou le sujet d'étude de l'article. Si notre programme considère l'article comme mentionnant la langue à cause de ce bruit, il s'agit alors d'un faux-positif. Par exemple, un nom de langue présent dans une citation est considéré comme du bruit, de même que la mention d'une langue en référence à d'autres travaux externes à l'article.

Nous avons pu identifier quelques sources récurrentes de bruits dans les articles. Les articles scientifiques sont tous – ou presque – structurés de manière très similaires, comme suit : le titre – avec les noms des auteurs –, un résumé – l'*abstract* –, une introduction, le corps de l'article et enfin la bibliographie. La liste n'est pas exhaustive mais contient toutes les parties qui nous intéressent. En effet, les deux parties qui sont régulièrement sources de faux-positifs sont l'introduction et la bibliographie.

La bibliographie d'un article peut contenir des noms d'articles ou de livres comportant des noms de langues (ex : "*Automatic identification of word translations from unrelated **english** and **german** corpora.*", Reinhard Rapp, 1999). Or, ces mentions de langues ne sont jamais révélateurs du sujet de l'article.

L'introduction peut être elle-même composée de plusieurs sous-parties, telles que les "travaux en lien", les "travaux précédents" ou la contribution des auteurs à ces travaux précédents. Ainsi, toutes ces parties peuvent mentionner des langues en lien avec des travaux autres que ceux présentés par l'article, devenant ainsi des sources de bruit.

Pour réduire les sources de bruit, nous supprimons de chaque article son début (jusqu'à la fin de l'introduction) et sa fin (depuis le début de la bibliographie). Cette tâche nécessite de récupérer la mise en page des documents : s'il est aisé de trouver où commence la partie introduction (il s'agit la plupart du temps d'une ligne ne comptant que le mot "Introduction"), il reste à trouver la fin de cette partie. Nous pouvons pour cela rechercher le début de la partie suivante, à savoir le prochain titre, après introduction. C'est précisément ce dernier point qui requiert d'avoir la mise en page. Nous avons donc choisi de travailler sur des documents convertis en XML.

P07-1033.pdf (corpus brut) :

Frustratingly Easy Domain Adaptation

Hal Daumé III
School of Computing
University of Utah
Salt Lake City, Utah 84112
me@hal13.name

Abstract

We describe an approach to domain adaptation that is appropriate exactly in the case when one has enough "target" data to do slightly better than just using only "source" data. Our approach is incredibly simple, easy to implement as a preprocessing step (10 lines of Perl!) and outperforms state-of-the-art approaches on a range of datasets. Moreover, it is trivially extended to a multi-domain adaptation problem, where one has data from a variety of different domains.

1 Introduction

The task of domain adaptation is to develop learning algorithms that can be easily ported from one domain to another—say, from newswire to biomedical documents. This problem is particularly interesting in NLP because we are often in the situation that we have a large collection of labeled data in one "source" domain (say, newswire) but truly desire a model that performs well in a second "target" domain. The approach we present in this paper is based on the idea of transforming the domain adaptation learning problem into a standard supervised learning problem to which any standard algorithm may be applied (e.g., maxent, SVMs, etc.). Our transformation is incredibly simple: we augment the feature space of both the source and target data and use the result as input to a standard learning algorithm. There are roughly two varieties of the domain adaptation problem that have been addressed in the literature: the fully supervised case and the semi-

supervised case. The fully supervised case models the following scenario. We have access to a large, annotated corpus of data from a source domain. In addition, we spend a little money to annotate a small corpus in the target domain. We want to leverage both annotated datasets to obtain a model that performs well on the target domain. The semi-supervised case is similar, but instead of having a small annotated target corpus, we have a large but *unannotated* target corpus. In this paper, we focus exclusively on the fully supervised case.

One particularly nice property of our approach is that it is incredibly easy to implement: the Appendix provides a 10 line, 194 character Perl script for performing the complete transformation (available at <http://hal13.name/easyadapt.pl.gz>). In addition to this simplicity, our algorithm performs as well as (or, in some cases, better than) current state of the art techniques.

2 Problem Formalization and Prior Work

To facilitate discussion, we first introduce some notation. Denote by X the input space (typically either a real vector or a binary vector), and by Y the output space. We will write D^S to denote the distribution over source examples and D^T to denote the distribution over target examples. We assume access to a samples $D^S \sim D^S$ of source examples from the source domain, and samples $D^T \sim D^T$ of target examples from the target domain. We will assume that D^S is a collection of N examples and D^T is a collection of M examples (where, typically, $N \gg M$). Our goal is to learn a function $h : X \rightarrow Y$ with low expected loss with respect to the target domain.

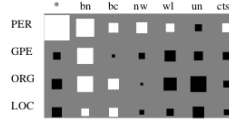


Figure 5: Hinton diagram for membership on a list of names at current position.

news and it is not quite so good for people in usenet. The first is easily explained: in broadcast news, it is very common to refer to countries and organizations by the name of their respective leaders. This is essentially a metonymy issue, but as the data is annotated, these are marked by their true referent. For usenet, it is because the list of names comes from news data, but usenet names are more diverse.

In general, the weights depict for these features make some intuitive sense (in as much as weights for any learned algorithm make intuitive sense). It is particularly interesting to note that while there are some regularities to the patterns in the five diagrams, it is definitely *not* the case that there are, e.g., two domains that behave identically across all features. This supports the hypothesis that the reason our algorithm works so well on this data is because the domains are actually quite well separated.

5 Discussion

In this paper we have described an *incredibly* simple approach to domain adaptation that—under a common and easy-to-verify condition—outperforms previous approaches. While it is somewhat frustrating that something so simple does so well, it is perhaps not surprising. By augmenting the feature space, we are essentially forcing the learning algorithm to do the adaptation for us. Good supervised learning algorithms have been developed over decades, and so we are essentially just leveraging all that previous work. Our hope is that this approach is so simple that it can be used for many more real-world tasks than we have presented here with little effort. Finally, it is very interesting to note that using our method, shallow parsing error rate on the

CoNLL section of the treebank improves from 5.35 to 5.11. While this improvement is small, it is real, and may carry over to full parsing. The most important avenue of future work is to develop a formal framework under which we can analyze this (and other supervised domain adaptation models) theoretically. Currently our results only state that this augmentation procedure doesn't make the learning harder—we would like to know that it actually makes it easier. An additional future direction is to explore the kernelization interpretation further: why should we use 2 as the "similarity" between domains—we could introduce a hyperparameter α that indicates the similarity between domains and could be tuned via cross-validation.

Acknowledgments. We thank the three anonymous reviewers, as well as Ryan McDonald and John Blitzer for very helpful comments and insights.

References

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. 2006. Analysis of representations for domain adaptation. In *Advances in Neural Information Processing Systems (NIPS)*.

John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. Domain adaptation with structural correspondence learning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Barcelona, Spain.

Ciprian Chelba and Alex Acero. 2004. Adaptation of maximum entropy classifier: Little data can help a lot. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Barcelona, Spain.

Hal Daumé III and Daniel Marcu. 2006. Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence Research*, 26.

Hal Daumé III, John Langford, and Daniel Marcu. 2007. Search-based structured prediction. *Machine Learning Journal (submitted)*.

Hal Daumé III. 2004. Notes on CG and LM-BFGS optimization of logistic regression. Paper available at http://pub.hal13.name/daume04cg-bfgs_implementation available at <http://hal13.name/daume/>, August.

Christopher Manning. 2006. Doing named entity recognition? Don't optimize for F₁. Post on the NLPers Blog, 25 August. <http://nlpers.blogspot.com/2006/08/doing-named-entity-recognition-dont.html>.

256

263

Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, pages 256–263, Prague, Czech Republic, June 2007. ©2007 Association for Computational Linguistics

Figure 2 : En rouge, ce que nous retirons de l'article.

L'étude de la mise en page des documents xml offre également d'autres opportunités de prétraitement, tels que l'ajout de certaines annotations ou la séparation – sans suppression – des titres, auteurs ou origines géographiques des articles. Cependant, nous ne traitons pas ces possibilités dans notre travail.

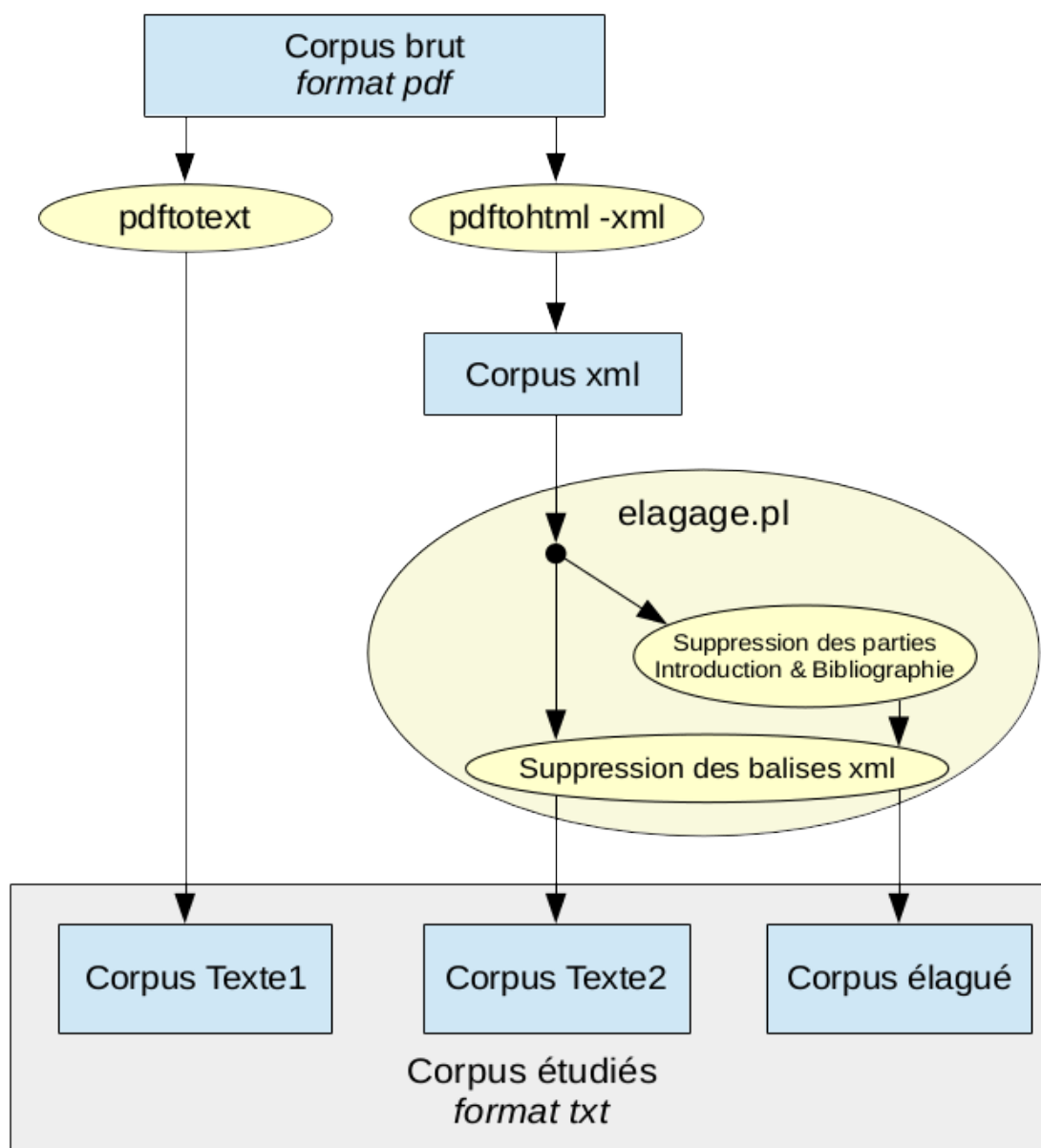


Figure 3 : Étapes de prétraitements du corpus brut :

5.1 Les outils de conversion des documents pdf

Travaillant sur GNU/Linux, nous utilisons pour convertir le corpus brut les outils fournis par *poppler-utils*, des outils utilisant la librairie Poppler¹. Ces outils sont utilisés via des lignes de commandes. Voici les deux que nous utilisons :

```
pdftotext doc.pdf
```

```
pdftohtml -xml -i -s doc.pdf # -i : ignore les images
```

```
# -s : ne produit qu'un seul fichier
```

1. <https://poppler.freedesktop.org/>

Le premier, *pdftotext*, produit un fichier nommé *doc.txt*. Le second, *pdftohtml* (avec l'option *-xml*), produit un fichier *doc.xml*.

Il est possible de supprimer toutes les balises d'un document xml, produisant ainsi l'équivalent d'un document *txt*. Le processus de prétraitement (que nous décrivons en 5.2.2) supprime les balises.

Nous avons alors comparé un document converti en *txt* (corpus **Texte1**) avec un document converti en xml duquel nous avons supprimé les balises xml (corpus **Texte2**). Dans les deux cas, nous avons pris soin de remplacer tous les retours à la ligne par des espaces pour réduire les différences observées. Nous observons des différences qui ont pour origine des particularités de mise en page du document *pdf* : lorsqu'un mot est indicé, le convertisseur en *txt* note le mot attaché à son indice (passé en taille normale), tandis que le convertisseur en *xml* place l'indice à la ligne, le séparant du mot.

Il peut arriver que le mot indicé soit un nom de langue ; notre programme reconnaîtra donc cette langue dans le corpus **Texte2** mais pas dans le corpus **Texte1**.

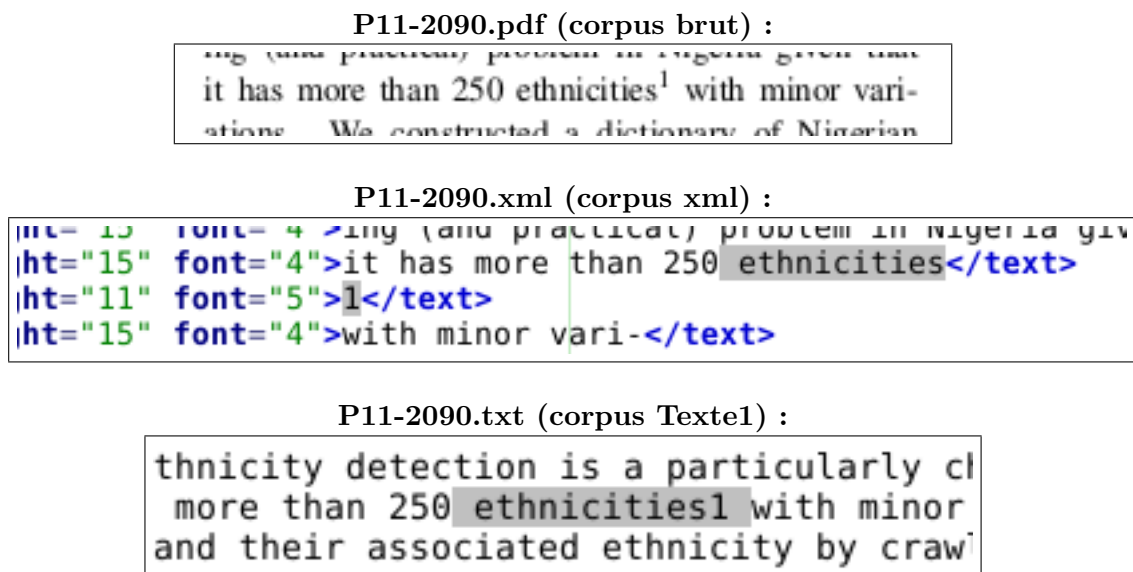


Figure 4 : Comparaison du traitement d'une même phrase par deux convertisseurs différents.

5.2 Nettoyage des données

Le nettoyage des données s'applique au corpus xml. Il procède en deux étapes : la mise à l'écart des fichiers défectueux (que nous décrivons 5.2.1) puis l'élagage des sources de bruit – Introduction & Bibliographie –.

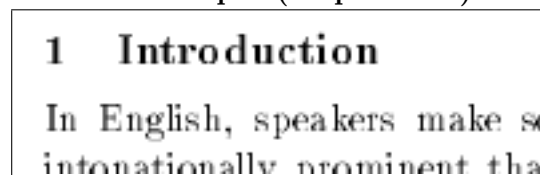
5.2.1 Problèmes lors des conversions

Commençons par décrire plus en détail les fichiers obtenus par le convertisseur en xml, composant le corpus **xml**.

Le fichier issu de la conversion d'un *pdf* en *xml* contient une ligne par mise en page spécifique. Chaque ligne concerne la mise en page d'un groupe de mots occupant un bloc. Le bloc possède une position verticale et horizontale, une largeur et une hauteur. Le formatage du groupe de mots comprend la police d'écriture utilisée, la taille de la police et sa couleur.

Intuitivement, une ligne du document *pdf* devrait former un seul bloc. Mais lorsqu'un mot de cette ligne possède une mise en page spéciale (un mot mis en valeur, un indice, etc), la ligne peut être découpée en trois blocs. Dans les faits, il arrive que chaque ligne ne contienne que deux ou trois mots, parfois même un seul mot découpé en plusieurs lignes :

P00-1030.pdf (corpus brut) :



P00-1030.xml (corpus xml) :



Figure 5 : Exemple de mauvaise conversion d'un fichier pdf en format xml.

Dans cet exemple, les mots *Introduction* et *speakers* ont été découpés. Le second est même découpé au sein d'une même ligne par un espace. Ce découpage des mots rend impossible leur reconnaissance par nos programmes de recherche. En l'occurrence, le découpage du titre *Introduction* rend impossible l'application de nos prétraitements sur ce fichier. Mais ce découpage des mots a aussi pu atteindre des noms de langues, les rendant eux-aussi illisibles.

Concernant également les conversions en *txt*, le document *pdf* d'origine peut être illisible pour le convertisseur, qui ne produira alors que des caractères spéciaux. Ces problèmes surviennent le plus souvent sur les documents *pdf* les plus anciens. Nous pouvons supposer que la récupération du texte de ces documents nécessite un logiciel de reconnaissance visuel, mais que l'aspect altéré de ces documents empêche le bon déroulement de ces reconnaissances.

P84-1074.pdf (corpus brut) :

... (f-structures, Warren 80) and functional structures
(f-structures) are generated during the parsing
process.

I INTRODUCTION

The fundamental purposes of syntactic
analysis are to check the grammaticality and to
clarify the mapping between semantic structures

P84-1074.txt (corpus Texte1) :

... (f-structures, Warren 80) and functional structures
(f-structures) are generated durlnK the parsing
process.

I

INTRODUCTIONr~

The
fundamental
purposes
of
syntactic
analysis are to check the Eramnatlcallty and to
clarify the mapping between semantic structures

*Figure 6 : Document datant de 1984 présentant
des problèmes de conversion.*

Pour mettre à l'écart les fichiers *xml* défectueux, nous avons écrit le script *NettoyageCorpus.pl*. Ce script Perl détecte les fichiers défectueux et ajoute leur nom à une liste nommée *badFiles*. Ces fichiers sont reconnus en comptant le nombre de mots présents dans chaque ligne (plus exactement, compris entre chaque balise *text*) : les fichiers ayant une moyenne inférieure à 3.5 (seuil trouvé par expérimentation) sont identifiés comme défectueux.

Ces fichiers sont retirés du corpus par l'ajout de l'extension *.back*, pour qu'ils ne soient plus considérés comme des fichiers xml à prétraiter. Nous aurions pu chercher à "réparer" ces fichiers en détectant les mots abîmés et en les "reconstruisant". Nous expliquerons notre choix après la présentation des résultats de ce script.

Lors de l'exécution du script sur les documents d'ACL, une trentaine de fichiers mal convertis ont été détectés parmi les documents datant du XXI^e siècle. La quasi totalité des documents d'avant l'an 2000 ont été mal convertis.

Corpus (format)	Conférence ACL	Conférence LREC
Corpus brut (pdf)	5530 articles	4133 articles
Corpus Texte1 (txt)	5525 articles	4133 articles
Corpus xml (xml)	5525 articles dont 1259 avant l'an 2000	4133 articles
Corpus xml (xml illisible)	1291 articles dont 1258 avant l'an 2000	30 articles
Corpus Élagué (txt, issu de l'xml)	3932 articles	3335 articles
Total exploitable	71% du corpus brut	80% du corpus brut

À l'issu du nettoyage, nous perdons donc la quasi-totalité des documents d'avant 2000. À l'image de l'exemple de la figure 6, ces documents sont bien trop abîmés pour être restaurés. Les conversions en xml de ces documents produisaient même des fichiers vides de texte.

Quant aux documents plus récents (à partir de l'an 2000), seuls 30 d'entre eux sont défectueux. Tenter de les récupérer n'en valait donc pas la peine.

5.2.2 Élagage des sources de bruit

L'élagage des sources de bruit correspond ici à la suppression des parties Introduction et Bibliographie. Pour procéder à l'élagage, nous avons écrit le script Perl *elagage.pl*.

Ce script cherche dans chaque fichier xml la ligne correspondant au titre de l'introduction. Dans près de 95% des cas, il s'agit d'une ligne ne contenant que le mot *Introduction*. Lorsque la ligne est trouvée, son *font* (l'identifiant de sa mise en page) est récupéré, car la même mise en page (et donc le même *font*) est normalement utilisé

pour tous les titres du document. La prochaine ligne utilisant ce *font* correspond donc au prochain titre et indique la fin de l'introduction. Tout ce qui se trouve avant est supprimé.

Ce choix de supprimer également tout ce qui se trouve avant l'introduction peut être discutable, car cela comprend également le titre de l'article et son résumé. Nous avons fait ce choix pour des questions de simplicité, mais il peut être intéressant de récupérer ces parties.

La partie Bibliographie est recherchée de la même manière ; plusieurs mots clés sont utilisés : *bibliography*, *reference(s)*, *sources*... Tout ce qui se trouve après le titre de cette partie est supprimé.

À la suite de cet élagage, les balises xml sont supprimées, de même que les retours à la ligne (remplacés par un espace) et les césures (un mot coupé en fin de ligne est alors recollé). Un nouveau fichier est finalement créé, nommé avec le suffixe *_elague.txt* (*doc.xml* devient *doc_elague.txt*). L'adresse de ce nouveau fichier est ajoutée au fichier *textes_adresses.txt*.

```
./textes_adresses.txt :
./pdfs/ACL/P07/P07-1112_elague.txt
./pdfs/ACL/P07/P07-2049_elague.txt
./pdfs/ACL/P07/P07-1050_elague.txt
./pdfs/ACL/P07/P07-1062_elague.txt
...
```

L'ensemble de ces fichiers constitue ainsi le corpus **Elagué**.

En plus des fichiers privés de leurs introductions et bibliographies, *elagage.pl* crée un deuxième fichier par document xml. Ce fichier conserve son introduction et sa bibliographie, mais subit le même nettoyage des balises xml, retours à la ligne et coupures en fin de ligne. Il est alors nommé avec le suffixe *_entier.txt* (*doc.xml* devient *doc_entier.txt*). L'ensemble des fichiers créés constitue le corpus **Texte2**. Du fait du convertisseur utilisé, ce corpus a de légères différences avec le corpus **Texte1**, différences que nous avons vu précédemment.

À l'issue du prétraitement, nous avons donc trois corpus :

- Le corpus **Texte1**, issu directement de la conversion de *pdf* à *txt*,
- Le corpus **Texte2**, initialement convertis en xml et resté entiers,
- Le corpus **Elagué**, initialement converti en xml et débarrassé des introductions et bibliographies.

6. Lectures et Mesures

Nous avons réalisé plusieurs programmes Perl pour effectuer nos mesures et études. Dans l'ensemble, il s'agit toujours de compter le nombre d'articles qui mentionnent chaque langue d'une liste, pour tenter de mesurer la fréquence utile de chacune de ces langues.

Nous présenterons aussi deux mesures alternatives ; la première consiste à comparer les fréquences obtenues sur deux corpus distincts (mais issus du même corpus brut), et la seconde à subdiviser un corpus en plusieurs périodes pour comparer les fréquences sur ces périodes.

6.1 Présentation des programmes d'analyse

Fréquencier :

Ce programme mesure pour chaque langue d'une liste sa fréquence dans un corpus donné, c'est-à-dire le nombre d'articles dans lesquels apparaît la langue.

Afar	2
Afrikaans	5
Akan	1
Albanian	4
Amahric	0
Amazigh	0
Amharic	1
Arabic	245

Figure 7 : Les huit premières lignes des mesures du Fréquencier sur le corpus Élagué composé de 3932 documents.

Fréquencier diachronique :

Ce programme fournit les mêmes mesures que le Fréquencier, mais subdivisées en un nombre donné de périodes (au sein d'un corpus, les documents sont organisés par années). L'utilisateur choisit une durée en années et le programme fournit la fréquence

de chaque langue pour chaque période.

Dans l'exemple qui suit, un score est donné en pourcentage de la fréquence sur la somme totale des fréquences, pour chaque période.

Language	2000 – 2004	2005 – 2009	2010 – 2014	2015 – 2017
Afar	0.0 %	0.0 %	0.0 %	0.1 %
Afrikaans	0.0 %	0.2 %	0.1 %	0.1 %
Akan	0.0 %	0.0 %	0.0 %	0.0 %
Albanian	0.0 %	0.1 %	0.001	0.0 %
Amahric	0.0 %	0.0 %	0.0 %	0.0 %
Amazigh	0.0 %	0.0 %	0.0 %	0.0 %
Amharic	0.0 %	0.0 %	0.1 %	0.0 %
Arabic	2.0 %	4.6 %	4.0 %	2.6 %
...
Total	100.0 %	100.0 %	100.0 %	100.0 %

Figure 8 : Les huit premières lignes des mesures du Fréquencier diachronique, sur une période de 5 ans.

Fréquencier comparatif :

Ce programme mesure les fréquences pour deux corpus distincts donnés et compare les résultats. Les deux corpus doivent provenir du même corpus brut et être composés des mêmes articles (dans des versions issues de conversions différentes). Le programme indique alors, pour chaque article, les langues présentes uniquement dans l'une ou l'autre des deux versions.

Langue	Élagué	Texte2	
Afar	2	2	./pdfs/ACL/P07/P07-2055 Langue : Chinese Present dans Élagué Langue : German Absent dans Élagué
Afrikaans	5	8	
Akan	1	2	
Albanian	4	6	./pdfs/ACL/P07/P07-2015 Langue : French Absent dans Élagué
Amahric	0	0	
Amazigh	0	0	
Amharic	1	2	./pdfs/ACL/P07/P07-1052 Langue : Dutch Absent dans Élagué
Arabic	245	303	

*Figure 9 : À gauche, les huit premières lignes des mesures ;
À droite, des comparaisons détaillées.*

6.2 Comparaison des résultats manuels avec les résultats du programme

Nous avons vérifié la fiabilité du Fréquencier en comparant ses résultats sur un échantillon analysé à la main.

Fichier	Analyse manuelle	Analyse Fréquencier	Commentaires
P00-1009	<i>English, German, Japanese</i>	-	L'article ne mentionne aucune langue, que les noms des corpus utilisés ; les corpus mentionnés concernent les trois langues que nous avons alors déduit.
P00-1025	Arabic, Malay, Indonesian	-	Le fichier xml étant illisible, l'analyse par le Fréquencier était impossible.
P00-1052	English	English, <i>Italian</i>	La langue <i>Italian</i> est mentionnée en note de bas de page. Il s'agit donc d'un Faux positif
P01-1012	Dutch	Dutch	-
P01-1045	German, English	German, English, <i>Mod</i>	<i>Mod</i> est un diminutif de <i>Modulo</i> , donc Faux positif .
P02-1016	-	-	L'article ne mentionne ni langue ni nom de corpus ; nous avons été incapable d'en déduire une langue support spécifique.
P03-2019	Chinese, Czech, English, French, German, Italian, Japanese, Polish, Slavic, Spanish	Chinese, Czech, English, French, German, Italian, Japanese, Polish, Slavic, <i>Sign</i> , Spanish	Le terme <i>Sign</i> est utilisé dans une formule mathématique : Faux positif

Fichier	Analyse manuelle	Analyse Fréquentier	Commentaires
P04-1077	-	-	Tout comme pour l'article P02-1016, aucune langue support n'a été déduite.
P05-3015	Arabic, English	Arabic, English	-
P06-4018	-	-	L'article concerne une étude sur l'enseignement et non sur le TALN.
P07-1031	English	-	La langue English était présente dans le résumé qui a été supprimé lors de l'élagage : Faux négatif
P08-4008	English	-	La langue English a été supprimé lors de l'élagage : Faux négatif
P09-2044	English	-	La langue English n'était pas mentionnée dans l'article. Nous l'avons déduit du corpus cité par l'article.

La liste de langue utilisée lors de ces tests contenait encore les langues Mod et Sign, que nous avons par la suite supprimé pour l'un et renommé pour l'autre.

Il est remarquable que nous soyons tombés sur deux articles dont l'élagage a produit des faux-négatifs.

Ces tests ont mis en évidence l'importance de la déduction implicite, à priori impossible par l'usage des Fréquentiers. C'est à l'occasion de ces tests que nous est venue l'idée de nous intéresser à la recherche des noms de corpus dans le but d'en déduire les langues utilisées. Nous sommes en effet tombés sur un article où seul le nom du corpus pouvait nous permettre de déduire les langues étudiées. Nous présentons nos recherches sur les noms de corpus dans la partie 6.5 : *Étude sur les noms de corpus*.

6.3 Analyses entre corpus

6.3.1 Résultat du Fréquencier comparatif sur les corpus ACL et LREC

Nous avons comparé les résultats de notre Fréquencier sur le corpus Élagué et sur le corpus Texte2. Pour rappel, le corpus élagué correspond au corpus entier duquel les débuts (jusqu'à introduction inclus) et fins d'articles (à partir de la bibliographie) ont été retirées. Cet élagage est censé nous rapprocher de la fréquence utile en réduisant le nombre de faux-positifs, c'est-à-dire en supprimant des mentions de langues qui ne sont pas le sujet de l'article en question.

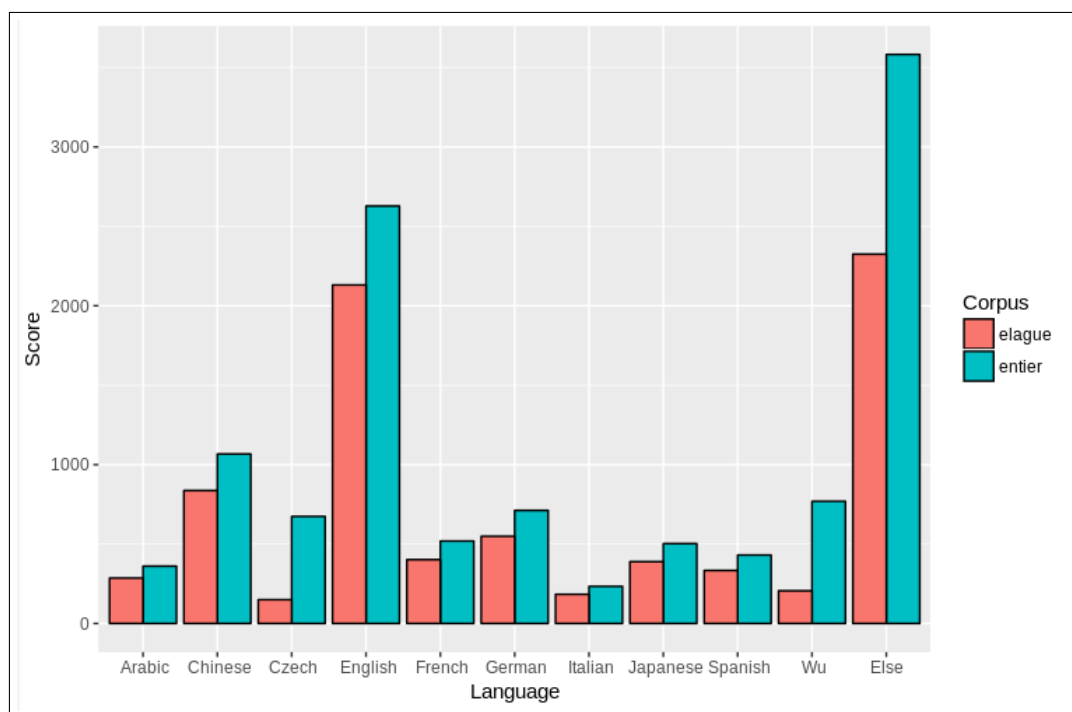


Figure 10 : Les dix langues les plus présentes dans le corpus ACL (en nombre d'articles) + la somme des autres langues.

La langue tchèque est, dans cette liste, la langue souffrant le plus de l'élagage. Il s'avère qu'en 2007, les conférences étaient hébergées en République Tchèque. Or, le nom du pays hébergeant les conférences est indiqué au pied de la première page d'un article. Ainsi, dans les articles de 2007, la langue "Czech", homonyme avec le nom du pays "Czech Republic", est systématiquement détectée dans le corpus entier **Texte2**. Étant donné que cette indication ne se trouve qu'à la première page, elle a donc toujours été supprimée lors de l'élagage et n'était plus présente dans les articles du corpus **Élagué**.

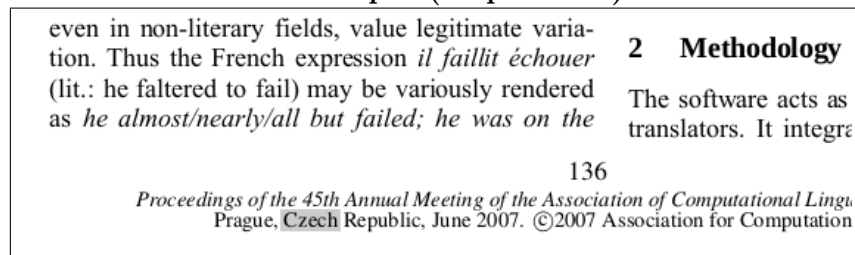
P07-1018.pdf (corpus brut) :

Figure 11 : L'indication du pays d'accueil ne se trouve qu'à la première page des articles.

À priori, la langue Wu semble avoir elle aussi beaucoup souffert de l'élagage, mais des lectures d'articles nous ont indiqué que "Wu" est aussi un nom de famille apparaissant relativement souvent en référence.

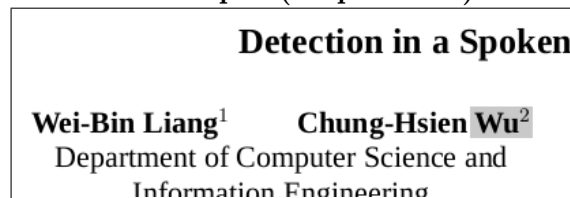
11-2106.pdf (corpus brut) :

Figure 12 : Entête d'un article scientifique avec noms des auteurs.

Nous avons vérifié à l'aide d'un concordancier dans quel contexte est mentionné le mot "Wu" dans l'ensemble du corpus Élagué. Il s'avère que, même après élagage, "Wu" correspond toujours à un nom d'auteur cité pour ses travaux. Nous avons donc décidé de supprimer la langue Wu de notre liste des langues pour la suite des mesures (à partir de l'analyse diachronique, partie 6.4).

Notons que le Wu est la deuxième langue chinoise la plus parlée (77 millions de locuteurs, pour reprendre les chiffres du baromètres de 2012 de L.J. Calvet).

La catégorie "Autres" a aussi été bien diminué lors de l'élagage : près d'un tiers de ces langues étaient mentionnées en introduction ou en référence.

Observons maintenant les résultats obtenus sur le corpus LREC.

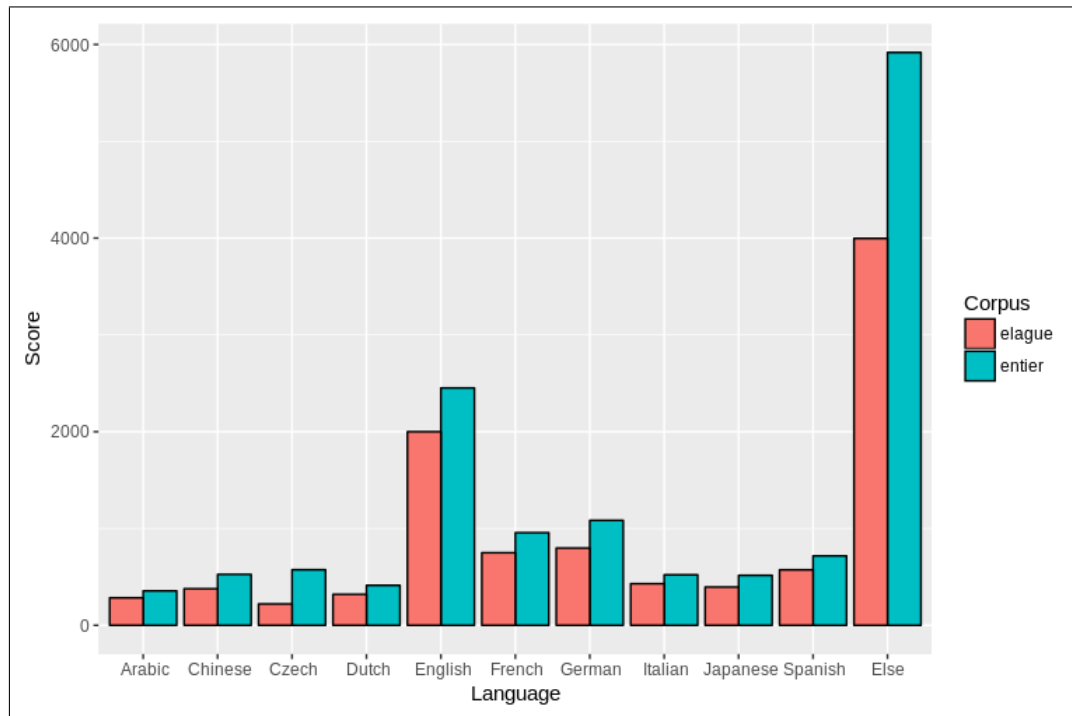


Figure 13 : Les dix langues les plus présentes dans le corpus LREC (en nombre d'articles) + la somme des autres langues.

Concernant le corpus LREC, neuf des dix langues les plus présentes le sont aussi dans le corpus ACL élagué.

Le corpus LREC contient globalement plus d'occurrences de langues que le corpus ACL : la catégorie "autre" contient par exemple 4000 occurrences (6000 avec les introductions et références) contre 2300 occurrences pour le corpus ACL (3600 avec les introductions et références).

Une fois de plus, le tchèque est parmi les dix premières langues, la langue la plus citée en introduction et bibliographie (proportionnellement au nombre total de mentions). Toutefois, les conférences LREC se déroulent tous les deux ans et n'ont pas eu lieu en 2007. Nous supposons malgré cela que de nombreuses références à des articles venant de République Tchèque ont été élaguées.

Enfin, dans un corpus comme dans l'autre, l'anglais est toujours la langue prédominante. Dans le corpus ACL élagué, elle apparaît presque autant de fois que la totalité des langues "autres". Dans le corpus LREC, la somme des dix premières langues dépasse celle de toutes les autres.

Ces deux graphiques montrent que le processus d'élagage a eu un réel impact sur les fréquences observées des langues. Nous n'avons pas effectué de comparaison entre le retrait des débuts d'articles et des fins d'articles. Un tel test pourrait être intéressant.

6.3.2 Détails du résultat du Fréquencier comparatif sur un échantillon de 18 articles

Pour en revenir au corpus ACL, voici un tableau détaillant la comparaison entre la version entière et la version élaguée pour un échantillon de dix-huit articles d'ACL pris au hasard :

Fichier	Détectées dans Corpus Entier	Détectées dans Corpus Élagué	Remarques
P01-1055	French, Italian, Greek, Japanese	French, Italian, Greek	Japanese : langue présente en référence
P06-4013	Czech	-	Czech : Langue présente en référence
P09-1029	English	English	-
P10-1159	Czech, English	-	Czech, English : Langues présentes en référence
P11-2053	English, German	English, German	-
P11-2106	Chinese, Mandarin, Wu	Chinese, Mandarin	Wu : Nom d'un des auteurs de l'article, cité avant l'introduction
P11-3007	English, Portuguese, Brazilian, Thai	English, Portuguese, Brazilian	Thai : langue citée en référence
P12-1101	Chinese, English, Japanese, German, French	Chinese, English, Japanese, German, French	-
P12-3003	English	English	-
P13-2029	-	-	-
P13-4025	English, French, Ido	English, French	Ido : nom en référence
P15-1054	Chinese, English, Wu	-	Chinese : Présent dans la partie "Related Work", une sous-partie de l'introduction ; English : Présent dans l'Introduction, dans un schéma ; Wu : nom en référence

Fichier	Détectées dans Corpus Entier	Détectées dans Corpus Élagué	Remarques
P15-2044	Chinese, Spanish, English, Russian, German, French, Marathi, Romanian, Czech, Albanian, Latin, Lithuanian, Maori, Persian, Swahili, Hebrew, Afrikaans, Croatian, Irish, Serbian	Chinese, Spanish, English, Russian, German, French, Marathi, Romanian, Czech, Albanian, Latin, Lithuanian, Maori, Persian, Swahili, Hebrew, Afrikaans	Croatian, Irish, Serbian : langues présentes dans la partie "Our Contribution", une sous partie de l'Introduction. Possiblement des faux-négatif.
P15-2105	Wu, Xiang	-	Wu, Xiang : noms d'auteur en référence
P16-1033	Chinese, English, Japanese, Yue, Irish, Swedish, Wu	Chinese, English, Japanese	Yue : nom de l'auteur de l'article (faux positif) ; Irish, Swedish : langues mentionnées en références ; Wu : nom de l'auteur de l'article
P16-1076	English	English	-
P17-1165	Kurdish, Wu	-	Kurdish : mentionné en Introduction (ne fait pas référence à la langue kurde) ; Wu : Nom d'auteur en référence
P17-4006	English, German, Ido	English, German	Ido : Nom d'auteur en référence

La plupart des noms de langues éliminées lors de l'élagage sont présentes dans la bibliographie de l'article. Nous pouvons considérer que tout ce qui vient de la bibliographie est source de bruit car dénué de tout contexte propre à l'article. En revanche, nous pouvons légitimement nous poser la question de la pertinence de l'élagage du début de l'article (titre, auteurs, résumé et introduction).

Deux des articles de notre échantillons contiennent, parmi les noms de langues, le nom d'un des auteurs de l'article, affiché alors avant l'introduction.

Nous retrouvons dans nos échantillons les noms d'auteurs pris par le Fréquencier pour des noms de langues : **Wu**, **Xiang**, **Yue**. Il peut même s'agir du nom d'un

des auteurs d'un article, comme pour l'article P11-2106 (illustré dans la figure X vue précédemment). Tout comme pour le **Wu**, nous avons à la suite de ces mesures supprimé le **Xiang** et le **Yue** de notre liste des langues.

L'introduction elle-même peut contenir plusieurs sous-parties comme, dans notre échantillon, les "travaux en lien" avec l'article ou "[leur] contribution" aux précédents travaux que les auteurs présentes. La question est alors de savoir, premièrement, si les auteurs traitent dans l'introduction du sujet de leur article et des langues concernées et, secondement, si les langues citées (que nous cherchons à identifier) sont à nouveau citées dans la suite de l'article. Quand une langue est support ou sujet d'un article, et lorsqu'elle est citée dans l'introduction, il est rare qu'elle ne paraisse pas dans la suite de l'article. Mais l'un des article de l'échantillon (l'article P15-2044) présente possiblement ce cas-là :

"**Our contribution** (...) Additionally, we make the POS tagging models for 100 languages publicly available and extend the mappings in Petrov et al. (2011) for six new languages (Hindi, *Croatian*, *Icelandic*, Norwegian, Persian, and *Serbian*)"

Dans cet extrait, élagué lors du prétraitement, trois langues (que nous avons mises en italique) sont citées et ne sont plus présentes dans la suite de l'article. Cet article cite de nombreuses langues et indique même qu'une centaine de langues sont concernées par leurs recherches. Il y a alors plus de chances pour que, d'une, des langues concernées ne soient pas du tout citées dans l'article et que, de deux, certaines ne soient citées que dans l'introduction. Dans ce second cas, l'élagage peut alors être source de faux-négatifs. Mais il s'agit pour nous d'un cas particulier, loin d'être une généralité.

6.3.3 Bilan du processus d'élagage

Nous ne voyons pas de contre-indication à l'élagage des fins d'articles. À partir de la bibliographie ne peuvent se trouver que des références à des auteurs et articles, ou éventuellement des glossaires et annexes. Le bilan de cet élagage est donc pour nous entièrement positif.

Discutons maintenant de l'élagage des débuts d'articles.

En premier lieu, les langues peu sujettes au bruit, telles que l'anglais, le chinois, le français ou l'arabe, ont été relativement peu affectées par l'élagage. Rappelons cependant que des **faux-négatifs** ont été constatés dans le tableau ci-dessus.

L'élagage des débuts d'articles comprend les parties suivantes : le titre, les auteurs, le résumé, l'introduction (pouvant comprendre plusieurs sous-parties) et le pied de la

première page.

L'élagage du nom des auteurs et du pied de la première page est selon nous efficace. Ces parties ne peuvent contenir de nom de langues, si ce n'est des homonymes (noms d'auteur ou de pays) qui sont donc par définition de potentiels faux-positifs.

Le titre de l'article n'aurait pas besoin d'être élagué. Un titre devant être court et concis, trouver des faux-positif nous semble invraisemblable. Il peut même contenir des informations utiles, notamment des noms de langues, alors assurément sujets de l'article.

Le résumé de début d'article semble lui aussi être d'avantage source de vrai-positif que de faux-positif, pour des raisons similaires au titre – doit être court et concis –, bien que le résumé soit plus développé que le titre.

La partie introductive, enfin, est la plus délicate. Le fait est qu'elle peut être constituée de plusieurs sous-parties, telles que les "travaux en liens" ou les "précédents travaux" (à priori sources de bruit), mais aussi "nos contributions" ou, même, une forme de résumé (donc à priori sources de vrai-positifs).

Comme nous l'avons indiqué précédemment, nous pensons qu'il est peu commun qu'une langue qui soit le sujet d'un article ne soit pas mentionné par celui-ci **dans son corps**. Dés lors, nous privilégions la suppression des sources de faux-positifs au maintien des sources de vrai-positifs.

Nous considérons donc en l'état que l'élagage des débuts d'article est globalement positif, même s'il mériterait d'être affiné pour traiter différemment les titres et résumés. Bien que nécessitant plus de travail, un tri pourrait aussi se faire au sein de l'introduction.

En conclusion, l'élagage des fins d'articles est donc pour nous entièrement bénéfique, de même que l'élagage des débuts d'articles, mais dans une moindre mesure car il est potentiellement source de faux-négatifs.

Un affinage de l'élagage des début d'articles pourrait être fait, de même qu'une comparaison entre les fréquences de deux corpus débarrassés de leur bibliographie, l'un avec ses débuts d'articles et l'autre sans.

6.4 Analyse diachronique

Nous avons réalisé une analyse diachronique sur le corpus ACL Élagué, avec des périodes de cinq ans. Le graphique que nous présentons ci-dessous comporte les dix langues les plus mentionnées sur la totalité du corpus Élagué, ainsi que la somme des fréquences de toutes les langues restantes (**Else(154)**, correspondant aux 154 langues mentionnées au moins une fois dans le corpus). Notons que l'unité correspond au rapport de la fréquence d'une langue sur la somme de toutes les fréquence – ce qui permet de

normaliser les valeurs, malgré la variation du nombre d'articles par périodes –, et que l'échelle du graphique est logarithmique.

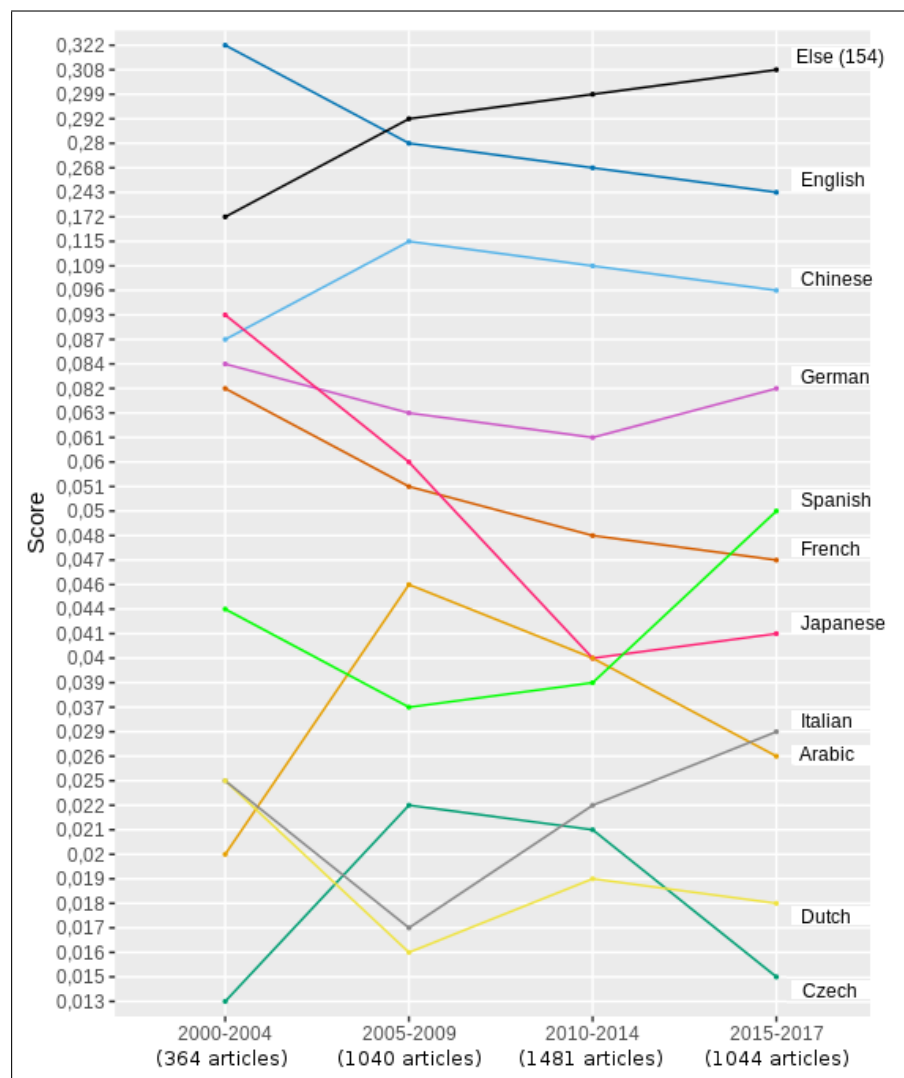


Figure 14 : Analyse diachronique sur corpus ACL Élargé avec période de 5 ans.

Étudions en premier lieu l'évolution du nombre d'articles par période. Pour rappel, dans la partie sur les problèmes de conversion (partie 5.2.2), nous avons pu voir que sur les 5530 articles de l'ACL, 1259 dataient d'avant l'an 2000 (contre 4266 datant du XIXe siècle), soit près d'un quart. Sachant que les plus vieux articles datent de 1979, nous sommes passés de 1259 articles en 21 ans, à 4266 en 17 ans. La production d'articles scientifiques par années a donc plus que triplé durant notre siècle.

Sur la figure ci-dessus, nous pouvons voir que nous sommes passés de 364 articles sur les cinq premières années à 1040 sur les cinq suivantes, puis 1481 sur la troisième période. De la première période à la deuxième, la quantité d'articles a triplé. Quant à la dernière période (ne comptant que trois ans), nous n'observons qu'une légère hausse par

rapport à la période d'avant (nous passons d'environ 296 articles/ans à 348 articles/ans).

Cette hausse de production en littérature scientifique est peut être à l'origine de la hausse des langues "autres". Les langues "autres" (*Else(154)*) correspond aux 154 langues ayant été mentionnées dans au moins un article, mais ne figurant pas dans les dix langues les plus mentionnées. En l'occurrence, cette catégorie a presque doublé d'importance entre la première et la dernière période.

Voyons maintenant comment lire ce graphique : le score a pour but de classer les langues entre elles. Une langue peut baisser de score même si elle gagne en fréquence ; il suffit pour cela que toutes les autres langues aient d'avantage gagné en fréquence qu'elle.

Comme attendu, la langue en tête est l'anglais. Bien que l'anglais semble chuter d'un quart entre la première et la dernière période, l'anglais est en fait mentionné par environ 50% des articles quelle que soit la période. C'est en réalité l'importance relative de l'anglais sur les autres langues qui diminue avec le temps. Cela démontre une possible hausse d'intérêt envers toutes les autres langues (comme le montre aussi la catégorie "*Else*").

Le tchèque double de score à la deuxième période et revient à son taux initial sur la dernière période. Nous serions tentés de faire un lien avec l'année 2007 (située dans la deuxième période) où les conférences avaient eut lieu en République Tchèque. Cependant, à la lecture d'une analyse diachronique sur des périodes d'un an, nous avons constaté que le tchèque a gagné en fréquence avant 2007 (le tchèque est passé de 3 articles le mentionnant en 2004, à 6 articles en 2005 puis 10 articles en 2006, avant de retomber à 8 articles en 2007).

En conclusion, l'analyse diachronique montre une fois de plus que les langues les plus mentionnées viennent de pays riches ou alors sont des langues internationales. Mais elle permet aussi de voir des dynamiques et tendance, que ce soit avec l'évolution du nombre d'articles par période ou des fréquences de certaines langues, voire de l'ensemble des langues.

6.5 Étude sur les noms de corpus

Dans la section 6.2 (Comparaison des résultats manuels avec les résultats du programme), nous avons, pour certains articles, trouvé la ou les langues supports sans qu'elles ne soient explicitement mentionnées. Nous avons trouvé ces langues de manière implicite.

Il y a plusieurs manières de détecter la langue cible d'un article. L'une d'elles est de lire les exemples utilisés dans l'article, qui servent à décrire les étapes des démonstrations

des chercheurs et qui se font dans la langue sur laquelle les chercheurs travaillent. Le problème de détection des langues est considéré comme un problème résolu si les bonnes conditions sont remplies[6]. Entre autres, la taille des données, c'est-à-dire du texte dont il faut détecter la langue, doit être suffisante. Or, nous souhaitons trouver la langue utilisée dans les exemples d'un article, avec une taille des données bien trop faible.

D'autre part, les auteurs d'un article indiquent souvent le corpus utilisé. Un corpus est une collection de productions langagières attestées, écrites ou orales[7]. Les corpus sont l'objet de travaux en informatique[8]. Une fois un corpus constitué, les auteurs peuvent le partager, en spécifiant certaines données comme la source des données ou bien les langues. Ainsi certains corpus sont très exploités et connus : si les auteurs d'un article indiquent travailler sur le *Penn treebank*, alors nous pouvons en déduire que la langue travaillée est l'anglais.

Le **Linguistic Data Consortium** (LDC), qui rassemble notamment des universités (dont l'université de Pennsylvanie qui est à l'origine du projet) et des laboratoires, a pour but de créer et stocker des ressources langagières. À partir des corpus disponibles sur leurs sites, nous avons constitué une liste de corpus que nous avons ensuite enrichie à l'aide d'un concordancier (autour des mots *corpus* et *corpora*). Certains noms ont été supprimés car, comme pour certains noms de langues, ils génèrent beaucoup de bruit : les corpus HARD, EASY, etc. Par manque de temps nous n'avons pas pu analyser les résultats du concordancier sur l'ensemble du corpus. La liste compte finalement 438 corpus. Nous avons mesuré leurs fréquences en posant la même hypothèse que pour les langues, à savoir qu'un article mentionnant un corpus travaille sur ce corpus, et donc sur les langues concernées par ce corpus.

Les premiers résultats observables montrent la prépondérance de certains corpus déclinés pour chaque langues comme le *treebank* (cité dans 762 articles) ou le *gigaword* (cité dans 292).

Le **treebank** est un format précis dont les sources peuvent varier. Certains corpus utilisant ce format sont très connus, comme le *Penn treebank*.

Le **gigaword** est développé par le LDC, de la manière suivante : un corpus est créé par langue, et les sources sont toujours des articles de presse.

Dans la plupart des cas, les auteurs associent ces corpus à la langue utilisée (*Chinese gigaword*). Lorsqu'aucune langue ne leur est associée, nous ne pouvons donc déduire aucune langue.

Le **CoNLL** (*Computational Natural Language Learning*) est un corpus issu des conférences annuelles du même nom. Il est lui-même composé de parties d'autres corpus et celles-ci changent en fonction des années. Les langues travaillées sont toujours indiquées.

Le **brown**(261), l'**europarl** (151) ou encore le **british national corpus**(190) sont des corpus bien définis dont les langues utilisées sont connues.

Le **wall street journal**(480) est beaucoup cité mais ne semble pas être en soi un corpus, plutôt une source dont les données sont présentes dans plusieurs autres corpus. Nous pouvons en déduire la langue d'usage, il a donc été ajouté comme corpus à la liste.

Enfin les chercheurs constituent souvent eux-mêmes leurs propres corpus.

Fichier	Analyse Corpus	Analyse Fréquencier	Commentaires
P01-1025	Frankfurter Rundschau : German	German	-
P03-1016	Brown : English	Chinese, English	Le terme Brown ne correspond pas à un corpus mais au nom d'un auteur cité.
P05-1017	Brown : English, Wall Street Journal : English	English, Japanese	Les expérimentations sont réalisés sur les deux corpus cités mais les auteurs utilisent des exemples et des références japonaises.
P07-1020	Hansard : multi-language, United Nation : multi-language	Arabic, Chinese, English, Japanese	Dans cet article, il est précisé quelles langues sont associées au corpus Hansard. Le corpus Hansard correspond aux transcriptions officielles de certains débats parlementaires. Aucune langue ne peut donc lui être associée.
P15-1006	Brown : English	-	Le brown est faux-positif, il fait référence ici à une couleur.

Bien que nous n'ayons pu faire d'analyses suffisamment approfondies sur l'étude des noms de corpus, les expérimentations que nous avons réalisées jusqu'ici semblent

indiquer que l'apport des noms de corpus dans la déduction des langues étudiées risque d'être faible.

7. Améliorations & Discussion

7.1 Pistes d'améliorations

7.1.1 Amélioration des prétraitements

En conclusion de l'analyse comparative (partie 6.3.3), nous avons discuté de notre processus de prétraitement, de ses motivations, de ses effets et de ses limites. L'idée partait d'un constat : La bibliographie et l'introduction semblent être d'importantes sources de bruit. Nous pouvons à présent clarifier l'objectif que doit avoir notre prétraitement : localiser les informations que l'on cherche à garder et celles dont on souhaite se débarrasser. Nous pouvons pour cela nous aider des grandes similitudes dans les structures des articles.

Concernant les débuts d'articles, les parties à identifier sont le titre – qu'il est aussi possible de récupérer sur la page web où nous téléchargeons les documents –, les auteurs, l'*abstract* et l'introduction avec ses sous-parties. Cependant, bien que la grande majorité des articles suivent la même structure, il peut y avoir des variations. Ainsi, l'*abstract* et l'introduction peuvent avoir d'autres noms, comme *overview*.

Jusqu'ici, nos réflexions concernant le prétraitement se concentrent essentiellement sur les débuts et les fins des articles. Intéressons nous maintenant aux corps des articles, à tout ce qui se trouve après l'introduction et avant la bibliographie. Le corps d'un article peut lui aussi contenir des sources de faux-positifs, comme des langues citées dans des exemples de phrases ou dans des références à d'autres articles.

Une recherche structurelle est possible pour localiser ces sources de bruit. Les citations semblent bien codifiées, ce qui pourrait faciliter leurs détections. Par exemple, leurs positions sont généralement indiquées par les indices en fin de citations. Quant aux exemples de phrases, leurs mises en page peuvent se démarquer du reste de l'article, ce qui pourrait permettre de les détecter.

7.1.2 Le problème de déductions implicites

L'hypothèse formulée de départ suppose que chaque article mentionne toutes les langues qui y sont étudiées. Or, comme nous l'avons vu précédemment, un article peut ne pas citer toutes les langues sur lesquelles les travaux portent, ou peut se considérer multilingue.

Pour résoudre le premier problème, la non exhaustivité des langues citées, une solution serait de parvenir à identifier le groupe de langues concerné : une famille de langues, un corpus multilingue dont les langues sont renseignées... nous pourrions alors incrémenter le compteur de toutes les langues concernées, y compris de celles qui ne sont pas mentionnées.

Le second problème, la considération multilingue de l'auteur, implique en revanche de définir et de comprendre ce que l'auteur de l'article entend par multilinguisme. Il peut s'agir par exemple d'un multilinguisme intra-famille ou d'un multilinguisme porté sur d'autres critères. La définition du multilinguisme est très vague. Le dictionnaire Larousse¹ considère le multilinguisme comme un synonyme du plurilinguisme, à savoir de travailler sur *plusieurs* langues. Il en va de même pour l'encyclopédie Wikipédia². Quant à la définition de *plusieurs*, le Larousse³ et le Wiktionnaire⁴ s'entendent sur un nombre indéfini supérieur ou égal à deux. Le multilinguisme peut donc avoir de multiples interprétations selon les auteurs : dix langues, cinquante langues, "toutes les langues" (qui peut même se limiter à la famille de langues qu'étudie l'auteur)... Le travail à réaliser serait donc de créer un nouveau terme plus précis et de faire adopter son usage.

Nous proposons de poursuivre ce travail par deux axes principaux : évaluer les relations bilingues et construire une représentation en famille de langues.

7.1.2.1 Travaux bilingues

Louis-Jean Calvet a conçu un modèle gravitationnel des langues, qui repose sur l'analyse des langues en tant que couple dans une relation ordonnée ; pour Calvet les langues sont reliées entre elles par des individus bilingues. Dans ce modèle l'anglais est au centre, appelée langue hypercentrale. Autour d'elle gravitent les langues dites super-centrales : espagnol, français, russe, hindi-malais... Viennent ensuite les langues centrales (entre 100 et 200) et les langues périphériques (plus de 6000).

C'est une relation ordonnée : pour le couple français-basque par exemple, le français

1. <http://www.larousse.fr/dictionnaires/francais/multilingue/53185?q=multilingue#209884>

2. <https://fr.wikipedia.org/wiki/Multilinguisme>

3. <http://www.larousse.fr/dictionnaires/francais/plusieurs/61815?q=plusieurs#61116>

4. <https://fr.wiktionary.org/wiki/plusieurs>

est une langue supercentrale et le basque une langue périphérique. Cela signifie que le locuteur de langue basque a de très fortes chances d'avoir pour première langue le français, et qu'elle est pour lui une langue véhiculaire par rapport au basque.

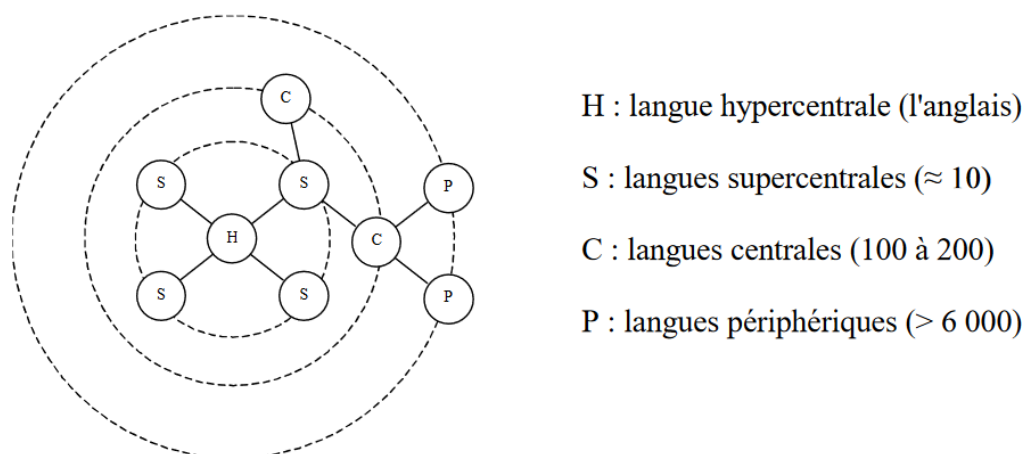


Figure 15 : Modèle gravitationnel de L-J Calvet.

Ce modèle présente en quelque sorte une image des relations entre langues : l'anglais est aujourd'hui la langue hypercentrale, la langue pivot, mais dans l'avenir une autre langue peut prendre cette place. Pour mesurer cela, nous proposons de compter les couples de langues, en cherchant les articles qui traitent des corpus parallèles, ou bien en construisant une métrique pour déterminer si une langue est assez "proche" d'une autre (par exemple en nombre de mots) et forme un couple.

7.1.2.2 Familles de langues

Dans la mesure de la diversité linguistique que nous avons réalisée, il est difficile de dégager des tendances pour plusieurs langues à cause de leur nombre élevé (que nous avons déjà restreint). Afin d'obtenir une mesure plus complète (d'avantage de langues concernées par la mesure), plus lisible et plus facile à interpréter, nous proposons de jauger selon les familles de langues. Dans ce but, il faut constituer une liste des familles de langues. Wikipédia en propose une selon la norme ISO 639-5, de 114 familles ; le site glottolog propose lui une liste au format Newick, un arbre partant de macro-familles (entre 10 et 20) pour arriver aux langues qui sont les feuilles de l'arbre. L'intérêt de cette mesure est, d'une part, d'obtenir des observations à différentes échelles et, d'autre part, d'estimer les langues pour lesquelles il y a peu de données.

7.2 Conclusion

Le problème d'identifier la langue d'étude d'un article est un problème difficile. Le meilleur moyen pour connaître la ou les langue sur lesquels portent des recherches en taln serait probablement que ces langues soient explicitement mentionnées par l'auteur. Par exemple, le site LDC indique pour chaque corpus (ou presque) la ou les langues concernées. De la même manière, les conférences ACL pourraient demander à ses participants de suivre un standard, qui comprendrait le renseignement des langues étudiées.

Notons qu'il pourrait être compliqué pour un auteur de donner de tels renseignements. Par exemple, si l'auteur a choisi une langue par défaut comme langue support (comme l'anglais ou sa propre langue d'usage), comment connaître l'éventail des langues auxquelles peuvent s'appliquer ses recherche ? À partir de quand considère-t-on qu'une langue doit être mentionnée ? Et qu'en est-il des études dites multilingues ?

Pour ce qui est de toutes nos mesures, les résultats obtenus semblent indiquer que de plus en plus de langues sont traitées par la communauté du TALN dans le cadre des conférences ACL. Les dix langues les plus travaillées sont l'anglais, le chinois, l'allemand, l'espagnol, le français, le japonais, l'italien, l'arabe, le néerlandais et le tchèque. Si en effet, dans le début des années 2000 ces langues formaient 80% des langues mentionnées en tant que cible de travail, elles n'en représentent plus que 66% entre 2015 et 2017. Toujours sur les mêmes périodes, le nombre total de langues mentionnées au moins une fois passe de 43 à 164.

Dans les langues pas ou peu mentionnées, nous trouvons le **javanais** (10^{ème} langue la plus parlée selon le baromètre de Calvet en 2012) ou le **bengali** (4^{ème} langue selon ce même baromètre). Le bengali est mentionné dans 24 articles d'ACL et 26 du LREC, soit moins d'un pourcent des articles du corpus Élagué. Quant au javanais, il n'est mentionné que dans deux articles (un de l'ACL, un du LREC). Ce sont alors des exemples de langues très parlées dans le monde, mais pas ou peu présentes dans les recherches en TALN, du moins concernant l'ACL et le LREC.

A. Tableau d'évaluation du niveau informatique d'une langue

Pour sa thèse, V. Berment a fait remplir son tableau par deux personnes compétentes, chacun le remplissant pour une langue différente :

*"nous avons demandé à Pierre Sein-Aye, pionnier de l'informatisation du birman, et à Michel Antelme, responsable de l'enseignement du khmer à l'INALCO, de compléter ces tableaux d'évaluation du niveau d'informatisation pour le birman et pour le khmer, en fonction de leur connaissance des logiciels et des ressources existants."*¹

Voici donc les deux tableaux remplis :

	Services / ressources	Criticité (0 à 10)	Note (/20)	Note pondérée (Criticité x Note)
Traitement du texte				
	Saisie simple	10	16	160
	Visualisation / impression	10	16	160
	Recherche et remplacement	8	0	0
	Sélection du texte	8	16	128
	Tri lexicographique	6	0	0
	Correction orthographique	4	0	0
	Correction grammaticale	4	0	0
	Correction stylistique	2	0	0
Traitement de l'oral				
	Synthèse vocale	2	0	0
	Reconnaissance de la parole	2	0	0
Traduction				
	Traduction automatisée	6	0	0
ROC				
	Reconnaissance optique de caractères	8	0	0
Ressources				
	Dictionnaire bilingue	8	0	0
	Dictionnaire d'usage	4	0	0
Total		82		448
Moyenne (/20)				448 / 82 = 5,46

Figure 16 : Tableau rempli pour le birman

1. <https://tel.archives-ouvertes.fr/tel-00006313/document> , p.21, consulté le 18 Juin 2018.

	Services / ressources	Criticité (0 à 10)	Note (/20)	Note pondérée (Criticité x Note)
Traitement du texte				
	Saisie simple	10	16	160
	Visualisation / impression	10	14	140
	Recherche et remplacement	8	12	96
	Sélection du texte	6	12	72
	Tri lexicographique	5	0	0
	Correction orthographique	2	0	0
	Correction grammaticale	0	0	0
	Correction stylistique	0	0	0
Traitement de l'oral				
	Synthèse vocale	5	0	0
	Reconnaissance de la parole	5	0	0
Traduction				
	Traduction automatisée	8	4	32
ROC				
	Reconnaissance optique de caractères	9	0	0
Ressources				
	Dictionnaire bilingue	10	4	40
	Dictionnaire d'usage	10	0	0
Total		88		540
Moyenne (/20)				540 / 88 = 6,14

Figure 17 : Tableau rempli pour le khmer

Selon les deux évaluations ci-dessus et les critères de V. Berment, le birman et le khmer sont donc deux langues peu dotées en terme de dotation informatique.

B. Liste de 8436 langues

Dans le chapitre 4 sur la mise en place des données, nous avons parlé d'expérimenter sur une liste exhaustive de langues. Nous avons effectué des mesures sur une liste de 8436 langues téléchargée sur le site glottolog¹. Voici une petite description des résultats.

530 langues ont une fréquence non nulle, contre 164 pour la liste de 284 langues. Parmi les 530 noms de langues se trouvent des mots anglais, notamment *The*, *To*, *As*, *Are*, *Were* et *Even* qui figurent dans le top 6 du résultat du fréquencier. D'autres, comme *En* ou *Ir*, ont pu être confondues avec des diminutifs de noms de langues (*En* pour *English*, *Ir* pour *Ireland*). Du reste, de nombreux noms de langues sont composés de deux à trois lettres, ce qui les rend très sensibles aux ambiguïtés.

Exploiter ces résultats nécessite donc un travail de recherches et de tri parmi les noms de langues.

Voici donc les cinquante premières lignes du résultat de notre fréquencier sur cette liste de langues. Les résultats sont présentés en pourcentage d'articles mentionnant le nom de la langue. Par exemple, 99,90% des articles mentionnent la langue *The* (un dialecte du Laos²) et 58,57% mentionnent le *Even* (une langue parlée en Sibérie³).

1. <http://glottolog.org/glottolog/language>

2. <http://glottolog.org/resource/languoid/id/thee1240>

3. <http://glottolog.org/resource/languoid/id/even1260>

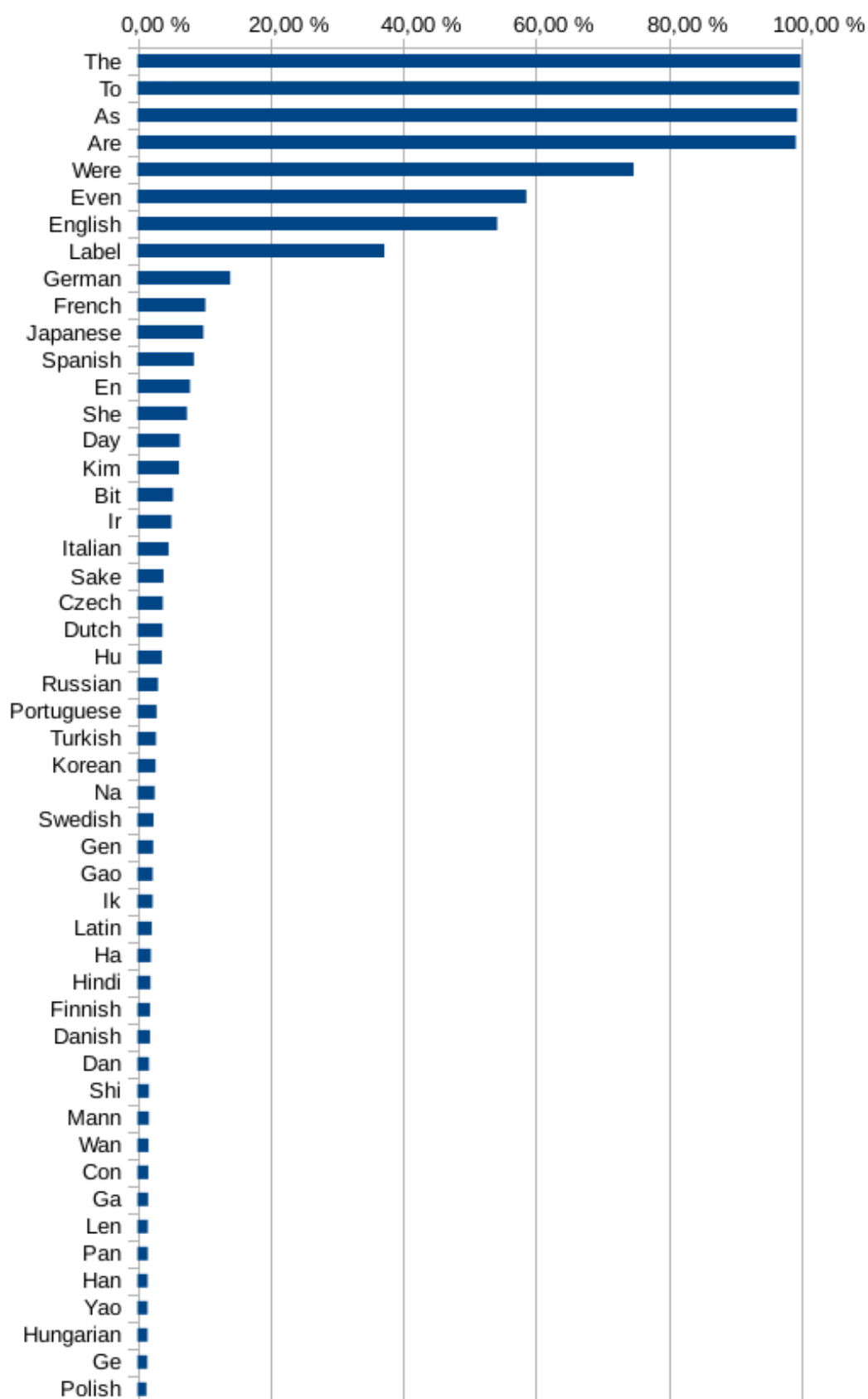


Figure 18 : cinquante premières langues selon leurs fréquences dans le corpus ACL Élagué.

C. Scripts bash

C.1 links_ACL.sh

```
#####
# Récupère les liens menant aux listes de pdf,
##
# Utilise curl pour télécharger les pages html puis grep pour rechercher
  les liens correspondant aux motifs voulus.
# Les motifs sont simples à trouver car le site source formatise
  rigoureusement ses liens.

mv ACL/ P/ #Au cas où le dossier ACL existe déjà
curl "https://aclweb.org/anthology/" > base
grep -e "P/[A-Z0-9]\{3\}/" ./base -o > result # suffixes des liens
rm base

while read line; do
    #télécharge la page contenant les pdf
    curl "https://aclweb.org/anthology/"$line > fic
    #récupère les suffixes des liens des pdfs
    grep -e "P.\{7\}.pdf" ./fic -o > ends

    #va créer un fichier de liens dans de nombreux sous-dossiers
    while read end; do
        mkdir -p $line
        cd $line
        #envoie les liens dans links :
        echo "https://aclweb.org/anthology/"$line$end >> links
        cd "../.."
    done < ends
```

```
done < result #la boucle s'exécute pour chaque ligne de ce fichier

# CAS SPÉCIAL : L'année P14 refuse de se télécharger normalement

curl "https://aclweb.org/anthology/P/P14/" > fic
grep -e "P.\{7\}.pdf" ./fic -o > ends
while read end; do
sous-dossiers
    mkdir -p ./P/P14/
    cd ./P/P14/
    echo "https://aclweb.org/anthology/P/P14/"$end >> links
    cd "../.."
done < ends

#supprime les fichiers inutiles
rm fic
rm ends
rm result

mv P/ ACL/ #renomme le dossier en ACL, la conférence concernée
```

C.2 dl_ACL.sh

```
#####
# SCRIPT À EXÉCUTER APRÈS LE SCRIPT links_ACL.sh
#
## DESCRIPTION :
# Le but ici est de télécharger tous les pdf du dossier ACL,
# dont les liens se trouvent dans des fichiers nommés link.
# wget permet de télécharger un fichier.
# wget -i récupère toutes les adresses contenus dans le fichier
# mis en paramètre.
# wget -nc ne télécharge pas le fichier s'il est déjà présent
# dans le dossier.
# wget -nc -i permet donc d'éviter de télécharger les doublons
# présents dans le fichier mis en paramètre.

cd ACL
for d in ./*/; #pour chaque dossier de ACL
do
    cd $d
    wget -nc -i links
    cd ..
done
cd ..
```

D. Glossaire

Effectif d'une langue : Nombre d'articles travaillant sur cette langue.

Fréquence (brute) d'une langue : Nombre d'articles mentionnant le nom de cette langue (ou un homonyme).

Fréquence utile d'une langue : Nombre d'articles travaillant sur cette langue et la mentionnant.

Peut être vue comme une intersection de l'Effectif et de la Fréquence brute d'une langue.

Fréquencier : Programme mesurant la Fréquence d'une ou plusieurs langues.

Fréquence mesurée : Fréquence obtenue à l'aide d'un Fréquencier ; peut être différent de la Fréquence brute car la conversion d'un article en format textuel peut rendre certains noms de langues illisibles pour le Fréquencier.

Synonyme : **Fréquence observée**.

Effectif mesuré : Effectif obtenu via une lecture manuelle (dans la mesure où on ne peut le mesurer autrement) ; peut être différent de l'Effectif réel à cause des erreurs humaines.

Synonyme : **Effectif observé**.

Bibliographie

- [1] Calvet Louis-Jean. URL https://www.youtube.com/watch?time_continue=1&v=pmSIVKtDncE.
- [2] Calvet Louis-Jean and Calvet Alain. Baromètre des langues, 2012. URL <http://wikilf.culture.fr/barometre2012/>.
- [3] Berment Vincent. *Méthodes pour informatiser les langues et les groupes de langues "peu dotées"*. PhD thesis, Université Joseph-Fourier - Grenoble I, 2004.
- [4] Calvet Louis-Jean. *La guerre des langues : Et les politiques linguistiques*. Hachette Littératures, 2005.
- [5] Mangeot Mathieu Enguehard Chantal. Favorisons la diversité linguistique en tal. 2014. URL <https://hal.archives-ouvertes.fr/hal-01096592/document>.
- [6] Cyril Goutte, Serge Léger, and Marine Carpuat. The nrc system for discriminating similar languages. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, Dubin, Ireland, August 2014.
- [7] Poudat Céline and Frédéric Landragin. *Explorer un corpus textuel*. deboek supérieur, 2017.
- [8] Haber Benoît, Nazarenko Adeline, and Salem André. *Les linguistiques de corpus*. Armand Colin, 1997.