Évaluer et améliorer une ressource distributionnelle : protocole d'annotation de liens sémantiques en contexte

Clémentine Adam* — **Cécile Fabre*** — **Philippe Muller****

* *CLLE-ERSS, 5 allées Antonio Machado – F-31058 Toulouse*
prenom.nom@univ-tlse2.fr
** *IRIT, 118 route de Narbonne – F-31500 Toulouse*
muller@irit.fr

RÉSUMÉ. *L'application de méthodes d'analyse distributionnelle pour calculer des liens de proximité sémantique entre les mots est devenue courante en TAL. Toutefois, il reste encore beaucoup à faire pour mieux comprendre la nature de la proximité sémantique qui est calculée par ces méthodes. Cet article est consacré à la question de l'évaluation d'une ressource distributionnelle, et de son amélioration ; en effet, nous envisageons la mise en place d'une procédure d'évaluation comme une première étape vers la caractérisation de la ressource et vers son ajustement, c'est-à-dire la réduction du bruit en faveur de paires de voisins distributionnels exhibant une relation sémantique pertinente. Nous proposons un protocole d'annotation en contexte des voisins distributionnels, qui nous permet de constituer un ensemble fiable de données de référence (couples de voisins jugés pertinents ou non par les annotateurs). Les données produites sont analysées, puis exploitées pour entraîner un système de catégorisation automatique des liens de voisinage distributionnel, qui prend en compte une large gamme d'indices et permet un filtrage efficace de la ressource considérée.*

ABSTRACT. *Using distributional analysis methods to compute semantic proximity links between words has become commonplace in NLP. This paper focuses on the issues of evaluating a distributional resource. We consider that setting up an evaluation procedure is a first step towards the characterization of the resource, and a way to improve its overall quality. We then propose a new protocol for in-text annotation of distributional neighbors, which is used to build a reliable reference data set. The data generated are analyzed and used to guide the automatic categorization of distributional links.*

MOTS-CLÉS : *Analyse distributionnelle, ressources lexicales, évaluation.*

KEYWORDS: *Distributional analysis, lexical resources, evaluation.*

TAL. Volume 54 – n° 1/2013, pages 71 à 97

1. Introduction

L'application de méthodes d'analyse distributionnelle pour calculer des liens de proximité sémantique entre les mots est devenue courante en TAL. Les procédures de calcul de la similarité distributionnelle sont désormais bien définies, des expérimentations ont été menées dans différents contextes applicatifs et sur plusieurs langues, et des synthèses récentes (Baroni et Lenci, 2010 ; Turney *et al.*, 2010 ; Clark, à paraître) ont permis de dresser un panorama cohérent du champ de la sémantique distributionnelle, en précisant les paramètres qui varient selon les méthodes – principalement : la taille et la nature des contextes considérés, les mesures de similarité employées, les méthodes mises en œuvre pour optimiser le calcul. Si ces aspects méthodologiques sont maintenant clarifiés, il reste encore beaucoup à faire pour mieux comprendre la nature de la proximité sémantique qui est calculée par ce biais. Comme le dit Sahlgren (2006, p. 57), l'hypothèse distributionnelle repose sur des fondements très peu contraints sur le plan sémantique : « *It states that differences of meaning correlate with differences of distribution, but it neither specifies what kind of distributional information we should look for, nor what kind of meaning differences it mediates*. » Dans cet article, nous proposons une démarche d'évaluation et de filtrage d'une ressource distributionnelle qui vise à progresser dans la compréhension et la maîtrise de l'information sémantique qu'elle contient.

En effet, la très grande variété des relations sémantiques qui sont mises au jour par des techniques d'analyse distributionnelle automatique a été souvent soulignée. Celles-ci détectent tout le spectre de relations lexicales couramment répertoriées, mais mettent également au jour des relations moins bien spécifiées et qualifiées à défaut de lâches ou d'associatives (van der Plas, 2008). Cette diversité est parfois vue comme un atout, car elle donne accès à des relations non systématiques qui peuvent être pertinentes pour certaines applications (Morris et Hirst, 2004). Néanmoins, les moyens de la maîtriser sont pour l'instant mal définis, dans la mesure où la plupart des travaux se sont concentrés sur l'acquisition de relations lexicales canoniques (en premier lieu, la synonymie) pour évaluer la qualité de la ressource distributionnelle. Nous cherchons ici à considérer la proximité sémantique détectée par le calcul distributionnel dans sa plus large extension.

La question du corpus, qui devrait être centrale s'agissant de techniques d'acquisition sémantique, est généralement évacuée, avec une focalisation exclusive sur la taille des données. Il est frappant de voir réapparaître le slogan du « gros, c'est beau » (Habert, 2000), que la linguistique de corpus avait cherché à nuancer en faveur d'une spécification précise de la nature des corpus employés : « *The size of the corpus is an important factor in the quality of the extracted relations, with the general message that more data is better data*. » (Clark, à paraître, p. 17). Il s'agit donc généralement de constituer un corpus le plus volumineux possible, par exemple issu du Web (Curran, 2004 ; Agirre *et al.*, 2009) ou du regroupement de collections de textes différentes (Baroni et Lenci, 2010), atteignant à chaque fois plusieurs milliards de mots pour l'anglais. Ces auteurs font alors l'hypothèse qu'un modèle sémantique général de la langue est ainsi accessible. Dans ce contexte, nous avons au contraire choisi d'ana-

lyser un corpus issu d'une source unique, Wikipédia, de taille beaucoup plus modeste (200 millions de mots). Ce choix est bien sûr guidé par des contraintes matérielles de disponibilité de très grands corpus en français. Mais il est aussi lié au souhait de mieux contrôler la source des relations sémantiques qui sont construites : bien qu'il ne s'agisse pas d'un corpus spécialisé (les textes sont homogènes du point de vue du genre, mais hétérogènes du point de vue des domaines couverts), le fait de disposer de données issues d'une source unique facilite le retour aux contextes et la compréhension des relations sémantiques mises au jour. La contrepartie de ce choix est potentiellement la qualité moindre des rapprochements effectués, puisque le calcul statistique s'appuie sur des contextes moins nombreux ; par conséquent, il rend nécessaire une phase de filtrage de la ressource obtenue pour compenser les effets d'un corpus peu volumineux. Le critère de proximité distributionnelle produit en effet des relations de qualité très variable. Les unités peuvent être rapprochées selon des dimensions très ténues. À titre d'exemple, certains noms peuvent être considérés comme voisins distributionnels parce qu'ils partagent une série d'adjectifs de nationalité. C'est le cas des mots *compositeur* et *drapeau*, qui partagent trente-neuf contextes de ce type dans notre corpus, et sont donc rapprochés avec un score de similarité assez élevé. La relation est valide sur le plan distributionnel, mais elle ne se traduit pas en termes de proximité sémantique.

L'objectif du travail présenté dans cet article a été (1) de mettre en place une évaluation de notre ressource distributionnelle qui prenne en considération des relations de proximité sémantique au sens large, mais n'éclipse pas le problème du bruit généré par le calcul distributionnel, et (2) de nous appuyer sur cette évaluation pour trouver des moyens de réduire ce bruit, c'est-à-dire d'exclure des paires de voisins distributionnels telles que *compositeur*/*drapeau* en faveur de paires présentant une relation sémantique pertinente. Nous mettons donc l'accent non pas sur l'utilisation d'une ressource distributionnelle (*cf.* (Adam, 2012) pour ce volet), mais sur la possibilité de l'ajuster en amont pour en améliorer la pertinence. Dans la section 2, nous discutons des méthodes classiques pour l'évaluation de ressources lexicales, et justifions l'intérêt de notre approche, qui s'écarte des options existantes pour mettre en place une tâche d'annotation contextuelle des liens sémantiques. Dans la section 3, nous décrivons le protocole d'annotation mis en place. Dans la section 4, nous analysons les données obtenues au terme de l'annotation, ce qui nous permet de mettre en correspondance la proportion de liens pertinents et la mesure de similarité distributionnelle. Dans la section 5, nous proposons une large gamme d'indices pour le filtrage en amont de la base distributionnelle. Nous montrons qu'un classifieur automatique exploitant ces indices variés et les données annotées permet un filtrage efficace de la ressource distributionnelle considérée.

2. Évaluation d'une ressource distributionnelle

L'opposition définie par (Galliers et Jones, 1993) entre les méthodes d'évaluation extrinsèques et intrinsèques (Curran, 2004 ; Poibeau et Messiant, 2008) s'applique au

cas de l'évaluation de ressources lexicales construites de manière distributionnelle. Les méthodes extrinsèques, ou évaluation par la tâche, consistent à évaluer la ressource du point de vue de sa fonction : on apprécie sa capacité à améliorer le système dans lequel elle a été implémentée. Dans le cas de l'analyse distributionnelle, ce type d'évaluation s'est fondé sur différentes tâches de TAL comme la recherche d'information (van der Plas, 2008) ou la désambiguisation (Weeds et Weir, 2005) ; nous avons nous-mêmes fait appel à la tâche de segmentation thématique pour mesurer l'apport de relations sémantiques plus variées que la simple répétition ou la synonymie (Adam *et al.*, 2010). D'autres travaux ont confronté les résultats du calcul distributionnel au jugement humain dans des tâches de détection de synonymes ou de jugement de similarité (Pado et Lapata, 2007 ; Baroni et Lenci, 2010). Baroni et Lenci (2011) font état des limites de ce type d'approche : en particulier, les données utilisées ont été conçues pour d'autres objectifs que celui de l'évaluation de modèles distributionnels, elles amènent à privilégier un type de relation parmi celles que l'analyse distributionnelle peut détecter (et ne fournissent donc qu'une indication partielle de la qualité de la ressource), et leur définition pose divers problèmes, comme celui de la nature des distracteurs utilisés (mots reliés par une relation sémantique, ou choisis aléatoirement). Les méthodes intrinsèques évaluent, quant à elles, la ressource du point de vue de son objectif propre, en utilisant des ressources manuellement construites qui servent alors de *gold standard* – dictionnaires de synonymes ou thesaurus (Weeds, 2003 ; Bordag, 2008 ; Anguiano *et al.*, 2011). Les limites d'une évaluation intrinsèque de ce type sont évidentes : tout d'abord, l'analyse distributionnelle met au jour des liens de proximité sémantique qui excèdent le périmètre des relations décrites dans les ressources lexicales disponibles ; celles-ci ne permettent donc pas d'évaluer la qualité globale des résultats mais seulement d'estimer la proportion de relations lexicales classiques qu'ils comportent. Par ailleurs, le propre des techniques d'acquisition sémantique est de repérer des liens de similarité sémantique construits dans le corpus, qui ne sont pas nécessairement identifiés en langue par les concepteurs de ressources lexicales généralistes, et qui offrent ainsi la possibilité d'en étendre la couverture. En utilisant des ressources existantes comme étalon, on réduit donc la portée de la base distributionnelle que l'on veut évaluer.

Notre objectif a été de concevoir une démarche d'évaluation intrinsèque de notre ressource distributionnelle qui permette de valider les liens de proximité sémantique dans leur plus large extension, c'est-à-dire sans se conformer au modèle réducteur des relations lexicales classiques. D'autres relations, exhibant un certain degré de proximité sémantique, doivent pouvoir être considérées, comme par exemple les relations de type actanciel (*auteur*/*publier*, *auteur*/*publication*) ou de type associatif (*acteur*/*cinéma*) qui sont largement mises au jour par la méthode distributionnelle. Par ailleurs, l'évaluation doit permettre de tenir compte du caractère inductif de la procédure de rapprochement, et ne pas exclure comme non pertinentes des relations de similarité qui opèrent dans le corpus considéré.

Nous montrons dans cet article qu'une procédure visant à exercer le jugement de similarité en contexte et non plus hors contexte permet à la fois de ne pas restreindre ce jugement à des cas de figure répertoriés *a priori*, de tenir compte des liens de

proximité sémantique qui sont construits dans le discours, et de garantir néanmoins la fiabilité du jugement de similarité. Nous montrons par ailleurs qu'elle permet de fournir des données d'évaluation plus étendues, parce que collectées plus facilement et de manière plus fiable que les tests de jugement de similarité existants.

3. Proposition d'un protocole d'annotation des voisins distributionnels

Cette section est consacrée au protocole d'annotation que nous avons mis en place pour ajuster notre ressource distributionnelle. Dans un premier temps, nous présentons les données que nous avons utilisées (section 3.1) : corpus mobilisé et ressource résultant de l'analyse distributionnelle de ce corpus. Dans un deuxième temps, nous décrivons une première expérience montrant que l'annotation de paires de voisins remises en contexte est effectivement une tâche plus facile pour les annotateurs que celle consistant à porter un jugement sur des paires de voisins présentées hors contexte (section 3.2). Enfin, nous décrivons notre protocole concernant les décisions prises sur les modalités d'annotation, l'interface développée et les consignes aux annotateurs (section 3.3).

3.1. *Données mobilisées*

Nous disposons d'une ressource distributionnelle en français, les *Voisins de Wikipédia*, calculée à partir d'un corpus de 200 millions de mots tirés de Wikipédia. Elle est construite selon les principes décrits par Bourigault (2002) à partir d'un modèle structuré (Baroni et Lenci, 2010), c'est-à-dire exploitant des contextes syntaxiques et non de simples cooccurrences. Ces contextes syntaxiques ont été calculés par l'analyseur Syntex, et le calcul distributionnel utilise une des mesures définies pour cette tâche, celle de Lin (1998).

Nous résumons les étapes de l'analyse distributionnelle opérée à partir d'une phrase d'exemple : « Le Kilimandjaro éveille l'intérêt des explorateurs ».

Dans un premier temps, l'ensemble des triplets <*gouverneur*, *relation*, *dépendant*> sont extraits. Dans notre exemple, il s'agira des triplets :
<*éveiller*, *obj*, *intérêt*>
<*éveiller*, *suj*, *Kilimandjaro*>
<*intérêt*, *de*, *explorateur*>

Ces triplets sont ensuite ramenés à des couples <*prédicat*, *argument*> en accolant la relation au gouverneur. Ainsi la différence entre les gouverneurs et les dépendants est conservée, puisque les prédicats (définis comme un gouverneur + une relation) ne seront rapprochés qu'avec d'autres prédicats et les arguments (dépendants syntaxiques) seront rapprochés avec d'autres arguments. Dans notre exemple, on obtient les couples <*prédicat*, *argument*> suivants :
<*éveiller_obj*, *intérêt*>
<*éveiller_suj*, *Kilimandjaro*>

<*intérêt_de*, *explorateur*>
On peut constater qu'un même mot (ici *intérêt*) peut être prédicat dans certains couples et argument dans d'autres.

Enfin, la similarité des distributions de l'ensemble des prédicats et de l'ensemble des arguments extraits est calculée en utilisant le score de Lin.

Le prédicat *éveiller_obj* a par exemple une forte similarité distributionnelle avec le prédicat *exciter_obj*, grâce à des contextes partagés tels que *curiosité*, *convoitise*, *appétit* ou encore *imagination*. L'argument *intérêt* est quant à lui rapproché de l'argument *importance* grâce à des contextes partagés tels que *juger_sans*, *attacher_obj* ou *se mesurer_suj*. On parle de *voisins distributionnels* (ou *voisins*) pour désigner les paires de prédicats ou d'arguments rapprochés — *éveiller_obj*/*exciter_obj* ou *intérêt*/*importance*.

Le score de Lin donne des valeurs comprises entre 0 et 1. La ressource distributionnelle les *Voisins de Wikipédia* recense tous les couples de mots (prédicats ou arguments) dont le score de Lin est supérieur ou égal à 0,1, c'est-à-dire 1 383 774 couples de voisins. La répartition des scores de Lin est illustrée par le graphique 1 : on peut constater que le nombre de couples décroît rapidement avec l'augmentation du score de proximité distributionnelle : 97 % des voisins sont dotés d'un score de Lin compris entre 0,1 et 0,29.

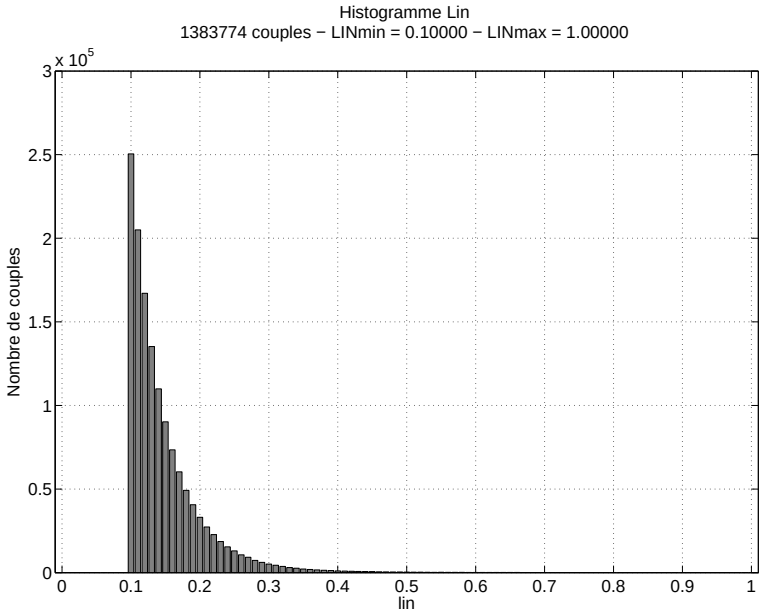

Figure 1. *Histogramme du score de Lin des* Voisins de Wikipédia

Pour faciliter l'exploitation des *Voisins de Wikipédia*, nous avons simplifié les prédicats en fusionnant les ensembles de voisins des différents prédicats portant sur la même unité lexicale ; par exemple, nous avons considéré comme voisins d'*intérêt* les voisins des prédicats *intérêt_de*, *intérêt_pour* et *intérêt_mod*[1].

Nous avons utilisé la ressource décrite pour différentes tâches – segmentation thématique, détection de relations de discours de diverses natures (Adam, 2012).

3.2. *Première expérience : validation de l'approche*

Nous avons opté pour une annotation en contexte des liens de voisinage distributionnel, en faisant les hypothèses suivantes : (1) la tâche est plus naturelle car elle s'appuie sur des exemples illustrant l'emploi des formes présentées, (2) le contexte permet de lever les ambiguïtés liées à la polysémie, et (3) la ressource distributionnelle étant le reflet des liens de proximité sémantique qui se construisent dans le corpus, une évaluation qui ne tient pas compte de la source risque d'être biaisée.

L'exemple (1) présente un cas dans lequel le contexte guide fortement le jugement porté sur un couple de voisins : le contexte instaure ici une relation – qui est loin d'aller de soi « en langue » – entre les mots *insecte* et *racine*, qui désignent tous deux un aliment du hérisson, et sont donc en relation de co-hyponymie.

(1) Bien que faisant partie des insectivores, les hérissons sont quasiment omnivores. Ils se nourrissent d'**insectes**, [...] de **racines**, de melons et de courges. (Wikipédia, article « Hérisson »)

Au vu de ce premier exemple, on pourrait objecter qu'il sera toujours possible d'interpréter le lien entre deux mots présentés dans un contexte réduit comme une relation sémantique pertinente, mais ce n'est pas le cas. L'exemple (2) présente des liens qui, selon nous, devraient être rejetés. Dans cet exemple, *espace* est voisin avec les mots *animal* et *majorité*[2] ; toutefois, il nous paraît extrêmement difficile d'établir une relation sémantique pour ces deux paires. Nous verrons dans ce qui suit que les annotateurs s'accordent effectivement à rejeter de nombreuses paires.

(2) Les impalas sont des **animaux** diurnes ; ils passent donc la **majorité** de la nuit à se reposer et à ruminer et se déplacent le jour afin de trouver de nouveaux **espaces** nourriciers. (Wikipédia, article « Impala »)

1. Lorsque deux des prédicats ont le même voisin, les scores de Lin sont moyennés.

2. *Espace* est rapproché d'*animal via* des contextes partagés tels que les adjectifs *domestique*, *protégé* et *sauvage*, et de *majorité via* des contextes partagés tels que *disposer de*, *requérir_obj*, et *changement_de*.

Nous présentons dans cette section un prétest, une première expérience d'annotation à petite échelle, ayant pour objectif de montrer qu'il est effectivement plus pertinent d'annoter des liens en contexte que hors contexte. Pour atteindre cet objectif, notre stratégie consiste à mettre en place deux annotations de couples extraits des *Voisins de Wikipédia* ; lors de la première annotation, les couples sont présentés en dehors de tout contexte ; lors de la seconde, les couples sont présentés à l'intérieur d'un paragraphe au sein duquel ils cooccurrent. Ces deux phases d'annotation sont effectuées par trois mêmes annotateurs, ce qui nous permet alors de comparer les accords interannotateurs obtenus pour l'annotation dite *hors contexte* et l'annotation dite *en contexte*. Les trois annotateurs sont des linguistes ; deux d'entre eux (par la suite « Ann1 » et « Ann3 ») connaissent la ressource évaluée et son mode de construction ; le dernier (« Ann2 ») n'est pas expert du domaine. Nous détaillons ci-dessous la mise en place de cette expérience.

3.2.1. *Sélection des couples à annoter*

Pour chaque annotation, cent couples ont été sélectionnés. Les contraintes posées pour la sélection de ces cent couples étaient les suivantes :

– pour l'annotation *hors contexte*, les couples candidats devaient avoir un score de Lin supérieur à $0,2$; $14,1$ % des couples de voisins répondent à cette contrainte ;

– pour l'annotation *en contexte*, la seule contrainte était que les couples cooccurrent au moins une fois dans un même paragraphe du corpus ayant servi à construire la base de voisins, afin de pouvoir les présenter au sein de ce paragraphe à l'annotateur. Si un couple apparaît dans plusieurs paragraphes, le contexte d'apparition présenté à l'annotateur est sélectionné aléatoirement.

Dans un cas comme dans l'autre, les cent couples finalement présentés aux annotateurs ont été piochés aléatoirement parmi tous les candidats possibles.

3.2.2. *Consigne aux annotateurs*

Afin de ne pas biaiser l'expérience, la consigne donnée était la même dans les deux cas : « Les deux mots proposés présentent-ils un lien de proximité sémantique ? En d'autres termes, existe-t-il une relation sémantique entre eux, qu'elle soit classique (synonymie, antonymie, hyperonymie, co-hyponymie, méronymie, co-méronymie) ou non classique (la relation peut être glosée mais n'appartient pas aux relations précédemment citées) ? »

3.2.3. *Résultats*

Les tableaux 1 et 2 présentent les matrices de confusion obtenues par chaque paire d'annotateurs pour les annotations *hors contexte* (1) et *en contexte* (2). La classe 1 correspond aux couples jugés pertinents, et la classe 0 aux couples jugés non pertinents.

On constate un net déséquilibre (non contrôlable puisqu'il ne pouvait émerger qu'après annotation) entre les deux ensembles de données à annoter : pour les don-

		Ann2			Ann3					Ann3		
		1	0	Tot.	1	0	Tot.			1	0	Tot.
Ann1	1	28	16	44	20	24	44	Ann2	1	20	15	35
	0	7	49	56	6	50	56		0	6	59	65
	Tot.	35	65	100	26	74	100		Tot.	26	74	100

Tableau 1. *Matrices de confusion des annotations hors contexte*

		Ann2			Ann3					Ann3		
		1	0	Tot.	1	0	Tot.			1	0	Tot.
Ann1	1	80	7	87	81	6	87	Ann2	1	78	4	82
	0	2	11	13	2	11	13		0	5	13	18
	Tot.	82	18	100	83	17	100		Tot.	83	17	100

Tableau 2. *Matrices de confusion des annotations en contexte*

nées présentées hors contexte, la classe majoritaire est, selon tous les annotateurs, la classe 0 (liens non pertinents), tandis que les liens présentés en contexte sont jugés très majoritairement pertinents (classe 1). Ce déséquilibre peut être attribué à deux facteurs :

– la différence entre les contraintes qui ont conditionné la sélection des couples à annoter dans les deux versants de l'expérience : la contrainte de cooccurrence au sein d'un même paragraphe sélectionnant manifestement plus de couples pertinents que la contrainte de score de Lin supérieur à 0,2 ;

– la différence entre les deux types de tâches : l'annotation en contexte amène à considérer des liens de proximité sémantique qui opèrent dans le texte et sur lesquels on n'a pas d'intuition *a priori*, ce qui pourrait expliquer des jugements moins sévères de la part des annotateurs.

D'un point de vue méthodologique, ce déséquilibre ne nuit pas à la comparaison entre les deux annotations, dans la mesure où l'on ne cherche pas à interpréter les taux d'accord. Dans le tableau 3, nous fournissons, en plus des taux d'accord, les scores obtenus en appliquant le coefficient Kappa de Cohen (1960), qui tient compte de l'accord aléatoire et permet donc la comparaison.

Notre hypothèse selon laquelle l'annotation de liens en contexte est une tâche plus naturelle et donnant prise à moins d'ambiguïtés est largement confirmée. La comparaison des coefficients Kappa fait ressortir une différence moyenne de 0,22 point. Selon l'échelle d'interprétation du Kappa proposée par Landis (1977), l'accord *hors contexte* est *faible* à *modéré*, tandis que le score obtenu *en contexte* correspond à un accord *fort*. Cela nous permet d'envisager l'utilisation d'annotations posées en contexte comme référence pour le paramétrage et l'évaluation d'une ressource distributionnelle. De plus,

Annotateurs	Hors contexte		En contexte	
	Taux d'accord	Kappa	Taux d'accord	Kappa
Ann1+Ann2	77 %	0,52	91 %	0,66
Ann1+Ann3	70 %	0,36	92 %	0,69
Ann2+Ann3	79 %	0,50	92 %	0,69
Moyenne	75, 3 %	0,46	91, 7 %	0,68

Tableau 3. *Accords interannotateurs selon le coefficient Kappa, en contexte* vs *hors contexte*

la qualité de l'accord observé pour l'annotation *en contexte* est pour nous extrêmement encourageante. En effet, un bon accord prouve la validité de la tâche demandée aux annotateurs.

Afin d'évaluer plus précisément l'impact du contexte sur les jugements posés, il serait intéressant de réitérer cette expérience en soumettant des couples de mots identiques à des annotateurs différents, en faisant varier les conditions d'apparition (hors contexte et dans des contextes variés et non pas un contexte unique). Cette extension, qui demande de faire appel à un nombre plus important d'annotateurs, fait partie de nos perspectives mais n'était pas nécessaire dans le cadre de la validation de notre approche.

3.3. *Protocole d'annotation des couples de voisins*

Dans cette section nous présentons les décisions que nous avons prises concernant l'annotation (3.3.1), le sous-corpus que nous avons mobilisé pour cette annotation (3.3.2) et l'interface que nous avons développée (3.3.3).

3.3.1. *Décisions prises sur l'annotation*

Pour définir les modalités d'annotation, nous avons dû prendre un certain nombre de décisions concernant la nature des objets à annoter, le type de jugement demandé à l'annotateur et l'information qui lui est présentée.

3.3.1.1. Quels objets annote-t-on ?

L'importance du contexte est manifeste dans la décision de l'annotateur ; se pose alors la question de savoir si un même couple de voisins pourrait permettre de mettre au jour un lien lexical pertinent dans certains contextes, et non dans d'autres. Il est donc difficilement envisageable de proposer d'étendre la validité d'une annotation posée dans un contexte donné à l'ensemble des réalisations possibles d'un couple de voisins.

Faut-il dès lors faire appel au jugement d'un annotateur pour chaque lien de voisinage au sein d'un texte ? Cette solution reviendrait à demander quatre jugements

différents sur la paire *atteindre*/*mètre* dans l'exemple (3), (liens particuliers « atteignent »/« mètre », « atteignent »/« mètres », « atteindre »/« mètre » et « atteindre »/« mètres »). On le voit, cette solution extrêmement fastidieuse peut difficilement être mise en pratique.

(3) Redressés, les gorilles atteignent une taille de 1,75 mètre, mais ils sont en fait un peu plus grands car ils ont les genoux fléchis. L'envergure des bras dépasse la longueur du corpus et peut atteindre 2,75 mètres. (Wikipédia, article « Gorille »)

Nous avons considéré le texte comme le palier susceptible de garantir une certaine stabilité des jugements pour un couple donné. Chaque couple de voisins est donc annoté une seule fois pour un texte donné. Nous suivons en cela un postulat bien connu en désambiguïsation lexicale, selon lequel toutes les occurrences d'un mot dans un même texte correspondent au même sens (« *One sense per discourse* », (Gale *et al.*, 1992)). Nous supposons que, de la même manière que le sens des mots pris individuellement, la nature de la relation qui unit ces mots reste stable dans un texte donné. Il s'agit ici d'une hypothèse assez forte. Pour s'assurer de sa plausibilité, il faut être attentif à l'accord entre annotateurs émergeant de ce dispositif d'annotation, qui doit rester comparable à celui du prétest (section 3.2).

Ainsi, le jugement de l'annotateur n'est sollicité qu'une seule fois pour la paire *atteindre*/*mètre* dans le texte ci-dessus. Par contre, un même couple de voisins peut être annoté plusieurs fois s'il apparaît dans deux textes différents, comme c'est le cas ci-dessous avec les exemples (4) (précédemment commenté dans la section 3.2) et (5). Ici, nous pensons que le couple *insecte*/*racine* est pertinent dans le texte « Hérisson » (car les insectes et les racines font tous deux partie des aliments du hérisson), mais plus difficilement dans le texte « Annélation ».

(4) Bien que faisant partie des insectivores, les hérissons sont quasiment omnivores. Ils se nourrissent d'**insectes**, [...] de **racines**, de melons et de courges. (Wikipédia, article « Hérisson »)

(5) Le mot annélation peut aussi décrire l'écorçage total d'un arbre, d'une branche, d'une **racine** ou d'une tige par un animal (**insecte**, rongeur, grand ou petit herbivore). (Wikipédia, article « Annélation »)

3.3.1.2. Quel type de jugement doit poser l'annotateur ?

De la même manière que dans le prétest, nous avons choisi de demander à l'annotateur un jugement binaire. Ce jugement, comme nous l'avons signalé, peut être porté en même temps sur plusieurs liens dans le cas où les deux voisins considérés ont plusieurs occurrences dans le texte présenté à l'annotateur.

3.3.1.3. Que montre-t-on à l'annotateur ?

Pour cette phase d'annotation, nous avons choisi de présenter des textes entiers aux annotateurs, pour les raisons suivantes :

– dans la mesure où le jugement de l'annotateur doit être valide pour la totalité d'un texte donné, il était logique qu'il puisse visualiser ce texte dans son ensemble. Toutes les occurrences des deux voisins dont la relation est évaluée sont mises en surbrillance ;

– même lorsqu'un couple n'a qu'une seule réalisation, il peut s'agir d'un lien connectant deux items très distants, d'où la nécessité d'afficher l'ensemble du texte.

3.3.2. *Corpus*

Pour cette phase d'annotation, nous avons constitué un sous-corpus de quarante-deux textes sélectionnés de manière arbitraire dans Wikipédia. Nous avons repéré automatiquement dans chaque texte du corpus toutes les paires de voisins (recensées par les *Voisins de Wikipédia*) qu'il contient. Aucun filtrage des voisins (par exemple sur la base du score de Lin[3] ou de la distance entre deux items dans le texte) n'a été appliqué, dans la mesure où nous souhaitons nous appuyer sur l'annotation effectuée notamment pour évaluer la validité de telles stratégies de filtrage. Le tableau 4 indique les caractéristiques essentielles du corpus.

Nombre de textes	42
Nombre de mots	15 527
Nombre de liens	3 476 785
Nombre de couples	801 196

Tableau 4. *Caractéristiques du corpus d'annotation*

Le nombre de couples de voisins indiqué correspond non pas au nombre de couples différents dans tout le corpus, mais à l'addition du nombre de couples de chaque texte ; ce sont donc bien ces objets que nous souhaitons annoter, comme nous l'avons développé dans la section précédente. Le but de l'annotation n'est évidemment pas de fournir un jugement sur tous ces couples, et la taille importante du corpus constitué s'explique par le souci d'assurer une variété suffisante des annotations.

3.3.3. *Interface développée et modalités d'annotation*

Nous avons développé en PhP/MySQL et JavaScript une interface d'annotation dédiée au protocole décrit. Quand un utilisateur accède à l'interface, un texte choisi aléatoirement parmi les quarante-deux textes du corpus est affiché. Dans ce texte, un voisin (qu'on appellera mot-cible) est choisi au hasard, et tous ses voisins apparaissent surlignés dans le texte. Le rôle de l'annotateur est alors de juger de la pertinence de

3. Rappelons que pour que deux mots soient considérés comme voisins, celui-ci doit être supérieur à 0,1.

tous les liens dans lesquels est impliqué le mot-cible. L'annotation se fait uniquement en cliquant sur les voisins du mot-cible : au premier clic, le voisin apparaît en rouge (pour signifier qu'il est jugé non pertinent), et au second clic, il apparaît en vert (pour signifier qu'il est jugé pertinent).

La figure 2 montre une annotation en cours sur un extrait du texte « Impala ». Dans cet extrait, le mot-cible est *corne*. Le couple *corne/patte* a été jugé pertinent dans ce contexte. Le couple *corne/herbe* a été rejeté. Les couples *corne/queue* et *corne/oreille* doivent encore être annotés.

Impala

Description physique
Les impalas font partie de la famille des antilopes, ils ressemblent aux kobs, gazelles et aux cerfs, ils mesurent généralement de 1,10 m à 1,50 m de longueur. Les mâles mesurent de 85 à 95 cm à l'épaule, les femelles mesurent de 75 à 85 cm à l'épaule. Les mâles ont une masse plus élevée selon les individus ; de 45 à 75 kg (60 en moyenne) contre 35 à 55 kg (40 en moyenne) pour les femelles. Le mâle, comme la femelle, est d'un brun rougeâtre sur le dos et beige sur les côtés. Le ventre de l'impala de même que ses lèvres et sa queue sont blancs. Il faut aussi mentionner leurs lignes noires uniques à chaque individu au bout des oreilles, sur le dos de la queue et sur le front. Ces lignes noires sont très utiles aux impalas puisque ce sont des signes qui leur permettent de se reconnaître entre eux. Ils possèdent aussi des glandes sécrétant des odeurs sur les pattes arrières et sur le front. Ces odeurs permettent également aux individus de se reconnaître entre eux. Il a également des coussinets noirs situés, à l'arrière de ses pattes. Les impalas mâles et femelles ont une morphologie différente. En effet, on peut facilement distinguer un mâle par ses cornes en forme de S qui mesurent de 40 à 90 cm de long.

Habitat
Les impalas vivent dans les savanes où l'herbe (courte ou moyenne) abonde. Bien qu'ils apprécient la proximité d'une source d'eau, celle-ci n'est généralement pas essentielle aux impalas puisqu'ils peuvent se satisfaire de l'eau contenue dans l'herbe qu'ils consomment. Leur environnement est relativement peu accidenté et n'est composé que d'herbes, de buissons ainsi que de quelques arbres.
[...]

Figure 2. *Interface d'annotation : exemple avec le mot-cible* corne*, dans le texte « Impala »*

4. Analyse des données constituées et évaluation de la ressource distributionnelle

Dans cette section, nous décrivons les données résultant de l'annotation mise en place et présentée dans la section 3 :

– nous proposons tout d'abord un bilan quantitatif de l'annotation : nombre de couples annotés, proportion de couples jugés pertinents et qualité de l'accord interannotateur (section 4.1) ;

– nous examinons ensuite le lien entre la pertinence d'un couple de voisins et son score de Lin (section 4.2).

4.1. *Bilan de l'annotation*

Deux annotateurs experts ont participé à l'annotation. Leur accord a été évalué sur cent vingt couples répartis dans les différents textes du corpus. Le tableau 5 donne la matrice de confusion des deux annotateurs sur ces cent vingt couples. Le tableau 6 indique les taux d'accord et coefficient Kappa correspondants.

	Pert.	**Non pert.**	**Tot.**
Pert.	35	6	41
Non pert.	5	74	79
Tot.	40	80	120

Tableau 5. *Matrice de confusion*

Taux d'accord	Coefficient Kappa
90,8 %	0,80

Tableau 6. *Taux d'accord et coefficient Kappa*

Le coefficient Kappa de $0,80$ indique un accord interannotateur encore meilleur que lors du prétest. Ce meilleur score peut notamment s'expliquer par :

– la quantité supérieure d'informations présentées à l'annotateur : un contexte plus large (tout le texte) et plus d'occurrences des deux voisins impliqués dans le couple à annoter ;

– le plus grand entraînement des annotateurs, qui sont les mêmes que lors du prétest.

Ce très bon accord interannotateur montre que le compromis décrit en 3.3.1 (sur la nature des objets annotés) n'a pas affecté la qualité des jugements posés par les annotateurs. La présentation simultanée de (parfois) plusieurs occurrences, pour un même couple de voisins, a manifestement été ressentie par les annotateurs non pas comme une source de confusion, mais comme une indication supplémentaire.

La rapidité de l'annotation constitue un autre aspect important. Notre protocole et l'interface que nous avons développée ont été partiellement conditionnés par l'objectif de mettre en place une annotation rapide, pouvant être réitérée facilement sur d'autres données. Cet objectif a été atteint : un annotateur entraîné travaillant 20 minutes sans changer de texte traite environ 850 couples de voisins.

Au total, 9 885 couples ont été annotés en plusieurs séances courtes réparties sur quelques jours. Environ 11 % de ces couples présentent, selon les annotateurs, un lien de proximité sémantique (*cf.* figure 3). Cette proportion peut paraître faible de prime abord ; mais il faut rappeler qu'elle est le résultat de la projection de tous les couples de voisins rapprochés par l'analyse distributionnelle du corpus, sans aucun seuillage posé sur leur score de Lin, donc présentant parfois une proximité distributionnelle très faible. Pour réellement évaluer la ressource, il faut mettre en rapport les jugements portés par les annotateurs avec le score de similarité distributionnelle.

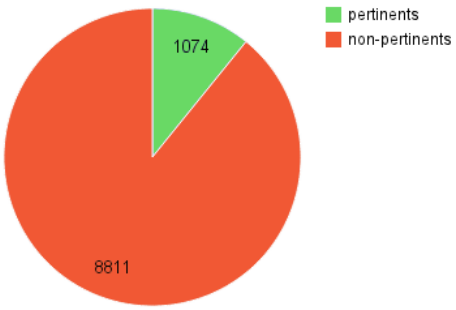

Figure 3. *Jugements des annotateurs sur les couples de voisins annotés*

4.2. *Pertinence des liens de voisinage et score de Lin*

Les données produites constituent une référence qui peut être exploitée pour explorer de nombreux aspects du traitement distributionnel. Le plus important, dans l'optique d'une évaluation de la ressource, est la corrélation entre le score de proximité distributionnelle (ici le score de Lin) et la pertinence des couples de voisins selon les annotateurs.

Le calcul de la corrélation de Pearson montre que le score de Lin est effectivement corrélé à la pertinence des couples de voisins ($r = 0,159^{***}$). À partir des données annotées, nous pouvons observer l'effet d'un seuil sur le score de Lin. La construction d'une ressource distributionnelle implique un tel seuil, qui permet de ne retenir que les couples de mots présentant une similarité distributionnelle supposée suffisante. La figure 4 permet de visualiser le nombre de couples retenus par le calcul distributionnel, pertinents ou non, en fonction du seuil posé sur le Lin. On peut observer que ce nombre décroît très rapidement ; parmi eux, le nombre de couples pertinents décroît moins rapidement, ce qui signifie une augmentation de la précision de la ressource.

La figure 5 montre l'évolution des courbes de précision et de rappel, définis de la manière suivante :

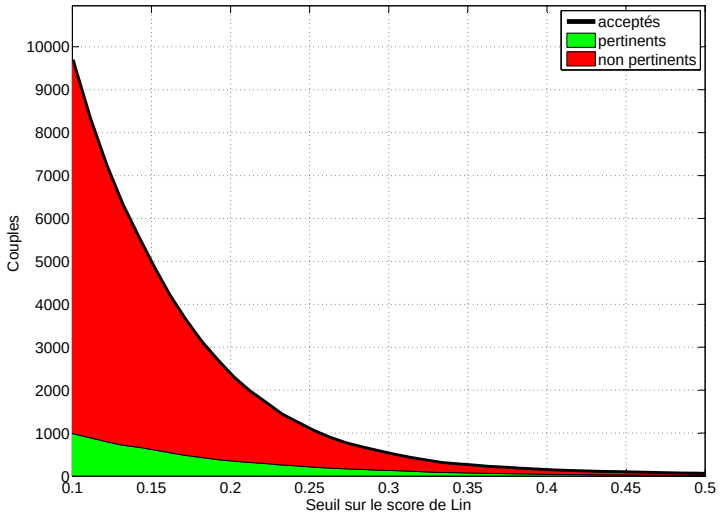

Figure 4. *Nombre de liens retenus et proportion de liens pertinents en fonction du seuil sur le score de Lin*

– la précision est la proportion de couples pertinents parmi tous les couples retenus ;

– le rappel est la proportion de couples retenus parmi tous les couples pertinents.

Cette courbe peut être utilisée pour guider le choix d'un seuil, selon que l'on veuille favoriser la précision (accepter moins de couples, mais de meilleure qualité) ou le rappel (viser une forte couverture mais en acceptant beaucoup de bruit). Si l'on ne veut favoriser ni la précision ni le rappel, un seuil sur le point de croisement des courbes est approprié : avec un seuil à environ 0,24 sur le score de Lin, on obtient un rappel et une précision légèrement inférieurs à 25 %.

Ainsi, si le score de Lin est bien corrélé à la pertinence des liens de voisinage, la qualité du filtrage obtenu en posant un score sur le score de Lin laisse à désirer. Nous pensons qu'un meilleur filtrage est possible en prenant en compte des indices plus variés. La section suivante est consacrée à l'exploitation des données constituées pour l'ajustement de la ressource.

5. Exploitation des données constituées pour l'amélioration de la ressource distributionnelle

Dans cette section, nous discutons de l'exploitation des données constituées pour l'ajustement de la ressource distributionnelle (la réduction du bruit en faveur de paires

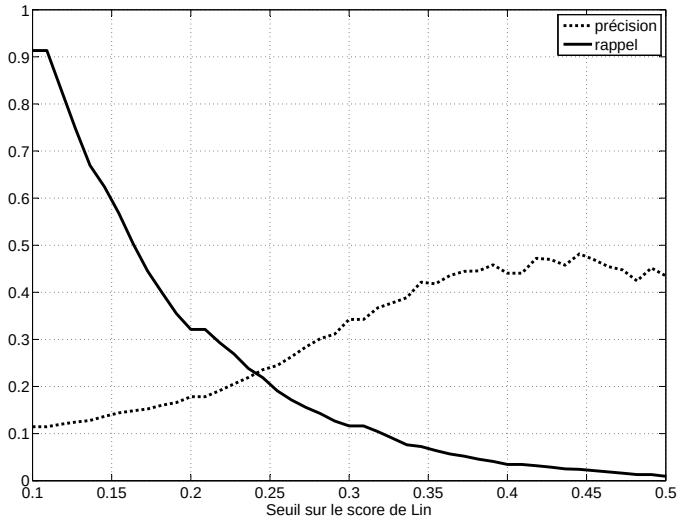

Figure 5. *Précision et rappel en fonction du seuil sur le score de Lin*

de voisins distributionnels exhibant une relation sémantique pertinente). Le prolongement naturel de la démarche menée jusqu'ici consiste en effet à mettre en place un dispositif d'apprentissage automatique supervisé s'appuyant sur les données annotées pour mener une classification automatique des couples de voisins. Nous avons donc isolé un ensemble de traits qui nous semblaient utiles pour prédire la pertinence d'un lien en contexte, calculé leurs valeurs sur l'ensemble des instances précédemment annotées, et appliqué la méthodologie classique consistant à entraîner un modèle inductif sur une partie de ces données, et à évaluer la correction du modèle sur les données annotées restantes. Dans la section 5.1 nous présentons les traits que nous proposons d'exploiter dans ce cadre, et dans la section 5.2, nous montrons les résultats obtenus.

5.1. *Définition d'indices pour le filtrage des voisins*

Au terme de l'annotation de liens en contexte décrite dans la section 3, nous disposons d'une masse relativement importante de données constituées de couples de voisins, et de jugements linguistiques posés sur les réalisations de ces couples dans des textes particuliers. Ces jugements sont assez fiables si l'on se réfère à l'accord interannotateur ; on peut donc envisager d'exploiter ces données pour mettre en place une catégorisation automatique supervisée des couples de voisins.

Dans cette section, nous listons les indices que nous avons définis pour le filtrage des liens de voisinage projetés dans les textes. Nous avons classé les indices définis

selon leur origine : sont-ils calculés à partir du corpus ? Résultent-ils de la construction de la base distributionnelle ? Ou enfin, émergent-ils après la projection des voisins distributionnels dans le texte ?

5.1.1. *Indices émanant du corpus*

Pour un couple *voisin*$_a$/*voisin*$_b$ donné, nous avons calculé, à partir de Wikipédia, deux informations principales.

La première concerne les fréquences des deux voisins dans le corpus. Nous appelons $freq_a$ la fréquence de *voisin*$_a$ et $freq_b$ la fréquence de *voisin*$_b$. À partir de ces deux fréquences, trois indices sont définis :

– $freq_{min}$ a pour valeur la plus faible fréquence parmi $freq_a$ et $freq_b$;

– $freq_{max}$ a pour valeur la plus forte fréquence parmi $freq_a$ et $freq_b$;

– $freq_{\times}$ combine les deux fréquences en un seul attribut ; il est défini comme le log du produit de $freq_a$ et $freq_b$

La deuxième information concerne l'association syntagmatique de *voisin*$_a$ et *voisin*$_b$. Bien que les voisins distributionnels soient rapprochés selon des critères paradigmatiques, ils présentent souvent également une certaine affinité d'ordre syntagmatique. L'indice que nous utilisons pour en rendre compte est une information mutuelle (im) (Church et Hanks, 1990) fondée sur la cooccurrence des deux membres du couple dans une fenêtre d'un paragraphe au sein de Wikipédia. On peut noter que le critère du paragraphe est une contrainte relativement faible ; on espère ainsi une meilleure couverture de l'indice (qui avec cette contrainte peut être calculé pour 70 % des couples). On a par ailleurs constaté lors du prétest que cette contrainte de cooccurrence au sein d'un même paragraphe constitue un très bon indice de la pertinence d'un couple de voisins (*cf.* section 3.2).

Ces indices sont résumés dans le tableau 7.

Indice	Description	Valeurs possibles
$freq_{\min}$	$\min(freq_a, freq_b)$	$freq_{\min} \in \mathbb{N}^*$
$freq_{\max}$	$\max(freq_a, freq_b)$	$freq_{\max} \in \mathbb{N}^*$
$freq_{\times}$	$\log(freq_a \times freq_b)$	$freq_{\times} \in \mathbb{N}^*$
im	$im = \log \frac{P(a,b)}{P(a) \cdot P(b)}$	$im \in \mathbb{R}$

Tableau 7. *Indices émanant du corpus*

5.1.2. *Indices émanant de la base distributionnelle*

Utiliser pour filtrer les voisins des informations résultant de leur construction est la solution la plus évidente. Plus particulièrement, c'est précisément la vocation du score de Lin que d'évaluer la qualité d'un rapprochement distributionnel. On a toutefois eu l'occasion de constater la possible insuffisance de ce score ; comme souligné en 4.2, nous pensons que de meilleures stratégies de filtrage sont possibles. Dès lors

que l'on considère que le score « final » de proximité n'est pas à lui seul suffisant, toutes les informations sur les couples de voisins peuvent alors servir d'indices, et c'est la confrontation avec les données annotées qui devra faire émerger les indices (ou configurations d'indices) les plus saillants.

Outre le score de Lin (lin), nous avons donc défini d'autres indices « quantitatifs » qui émergent de la base de voisins distributionnels :

– la productivité des voisins : $prod_a$ est défini comme le nombre de voisins de *voisin*$_a$ dans les *Voisins de Wikipédia*, et $prod_b$ comme le nombre de voisins de *voisin*$_b$ dans cette même base. Tout comme la fréquence dans la section précédente, la productivité des voisins est déclinée en trois indices : $prod_{min}$, $prod_{max}$ et $prod_{\times}$. Nous pensons que les voisins trop productifs donnent lieu à beaucoup de bruit ;

– les rangs des voisins (qui découlent directement du lin) : $rang_{a-b}$ est défini comme le rang de *voisin*$_a$ parmi tous les voisins de *voisin*$_b$ triés par score de Lin ; $rang_{b-a}$ est défini réciproquement. Là encore, trois indices sont considérés : $rang_{min}$, $rang_{max}$ et $rang_{\times}$.

Nous avons également défini deux indices catégoriels plus « linguistiques » :

– le premier ($cats$) concerne les catégories morphosyntaxiques des deux voisins considérés : s'agit-il de noms, de verbes, d'adjectifs ? Lors de la phase d'annotation, il nous a semblé que certaines catégories donnaient plus souvent lieu à des liens pertinents que d'autres : par exemple, les couples NN (nom-nom) nous paraissaient plus souvent pertinents que les couples VV (verbe-verbe) ; nous avons souhaité vérifier cette intuition ;

– le second ($predarg$) concerne la distinction prédicat/argument, une des spécificités des *Voisins de Wikipédia* : le couple de voisins considéré résulte-t-il d'un rapprochement entre prédicats, ou entre arguments ? Il s'agit là d'une information disponible dans la base de voisins, que nous ajoutons donc à la liste des indices, mais sur laquelle nous n'avons aucune intuition *a priori*.

Le tableau 8 résume l'ensemble des indices « distributionnels » utilisés.

Indice	Description	Valeurs possibles
lin	score de Lin	$0 \leq lin \leq 1$
$rang_{\min}$	$\min(rang_{a-b}, rang_{b-a})$	$rang_{\min} \in \mathbb{N}^*$
$rang_{\max}$	$\max(rang_{a-b}, rang_{b-a})$	$rang_{\max} \in \mathbb{N}^*$
$rang_{\times}$	$\log(rang_{a-b} \times rang_{b-a})$	$rang_{\times} \in \mathbb{N}^*$
$prod_{\min}$	$\min(prod_a, prod_b)$	$prod_{\min} \in \mathbb{N}^*$
$prod_{\max}$	$\max(prod_a, prod_b)$	$prod_{\max} \in \mathbb{N}^*$
$prod_{\times}$	$\log(prod_a \times prod_b)$	$prod_{\times} \in \mathbb{N}^*$
$cats$	catégories des voisins	AA,AN,AV,NN,NV,VV
$predarg$	prédicats ou arguments ?	pred,arg

Tableau 8. *Indices émanant de la base distributionnelle*

Ainsi, nous envisageons plusieurs indices liés à la base distributionnelle et facilement accessibles. Nous ne faisons en revanche pas varier de paramètres régissant la construction de la base (par exemple, les seuils posés en aval du calcul distributionnel : fréquence minimale des mots pris en compte, nombre de contextes, etc.) pour examiner leur impact.

5.1.3. *Indices émanant du texte après projection*

Si l'on considère qu'un couple de voisins peut donner lieu à des réalisations pertinentes (dans certains textes) et à d'autres qui ne le soient pas (dans d'autres textes), il est dès lors nécessaire de développer des indices dépendant du texte pour filtrer ces couples. Nous pensons que de tels indices peuvent compléter efficacement les indices fondés sur les propriétés distributionnelles des paires de voisins.

Les indices que nous avons définis visent à appréhender l'« importance » des items considérés au sein du texte, leurs configurations (plus particulièrement *via* la notion de distance entre les deux voisins), ou encore certaines propriétés du sous-graphe des voisins impliqués dans le texte.

5.1.3.1. Des indices mesurant l'importance d'un lemme au sein du texte

La première information prise en compte concerne les fréquences des items au sein du texte, déclinées en trois indices : $freqtxt_{\min}$, $freqtxt_{\max}$ et $freqtxt_{\times}$.

Le *tf·idf* (Salton *et al.*, 1975), qui évalue l'importance d'un mot dans un document en se fondant sur sa fréquence dans ce document et sa distribution dans une collection de documents, a donné lieu à des adaptations pour identifier des mots importants localement, dans une certaine zone de texte, relativement à la totalité du texte. Ainsi, Dias *et al.* (2007) proposent un *tf·isf* (*term frequency · inverse sentence frequency*), utilisé dans le cadre de la tâche de segmentation thématique. Nous avons calculé pour chaque voisin un score fondé sur sa fréquence dans le ou les paragraphes où il apparaît et sa distribution dans le texte ; l'indice *tf·ipf* est défini comme le produit des scores calculés pour $voisin_a$ et $voisin_b$.

5.1.3.2. Des indices mesurant la distance entre les deux voisins

Lors de la phase d'annotation, il a été observé que les voisins apparaissant plus proches l'un de l'autre dans le texte semblaient plus souvent entretenir un lien jugé pertinent que les voisins plus éloignés. Nous avons défini différents indices pour rendre compte de cette intuition :

– des indices comptant cette distance dans le texte en nombre de mots. Dans la mesure où chaque voisin d'un couple peut avoir plusieurs occurrences dans le texte, nous calculons :

- une distance minimale (ppd pour « plus petite distance ») définie comme le nombre de mots séparant les deux occurrences les plus proches de $voisin_a$ et $voisin_b$,

- une distance maximale (pgd pour « plus grande distance ») définie comme le nombre de mots séparant les deux occurrences les plus éloignées de *voisin*$_a$ et *voisin*$_b$,

- une distance moyenne (md) ;

– des indices booléens indiquant si oui ou non *voisin*$_a$ et *voisin*$_b$ sont, au moins une fois dans le texte, coprésents au sein de la même phrase ($copr_{ph}$) ou du même paragraphe ($copr_{para}$).

Ces évaluations de distance ou de proximité n'épuisent pas la notion de « configuration » des deux voisins ; on aurait pu imaginer, par exemple, des indices rendant compte de l'ordre d'apparition de *voisin*$_a$ et *voisin*$_b$.

5.1.3.3. Des indices liés au sous-graphe de voisinage associé au texte

Les voisins impliqués dans un texte forment un graphe, qui est un sous-graphe de l'ensemble des voisins distributionnels. À partir de ce sous-graphe, nous relevons les informations suivantes pour chaque couple de lemmes considéré :

– la productivité dans le texte de chacun des lemmes reliés, c'est-à-dire le nombre de couples différents dont ils font partie pour le texte considéré, (indices $prodtxt_{\min}$, $prodtxt_{\max}$), et leur produit $prodtxt_{\times}$; en termes de graphes, il s'agit du degré du lemme considéré ;

– l'appartenance du couple à une composante connexe (un sous-graphe connexe maximal) qui ne soit pas la composante connexe principale (présentant le plus grand nombre de nœuds) du texte (indice booléen cc). Nous avons observé que les liens extraits de cette manière paraissaient souvent pertinents, probablement parce qu'ils portent sur un champ lexical plus circonscrit. La figure 6 montre des exemples de composantes connexes de moins de cinq sommets dans un extrait de l'article « Gorille » (comme le groupe *pelage*, *dos*, *fourrure*).

Ces deux informations ($prodtxt$ et cc) sont liées : *via* les composantes connexes non principales, nous repérons des groupes de voisins ayant une faible productivité.

Les indices « textuels » définis sont résumés dans le tableau 9.

5.2. *Catégorisation automatique des couples de voisins*

Afin d'évaluer le pouvoir prédictif des traits choisis, nous avons mené des expériences de classification automatique à partir de l'ensemble des données annotées. Nous nous intéressons ici en fait à l'extraction des liens pertinents plus qu'à la classification pertinent/non pertinent, et nous présenterons donc principalement les résultats centrés sur la classe des liens pertinents, en précision, rappel et F-mesure (la moyenne harmonique des deux précédentes mesures). Comme nous l'avons vu plus haut, la classe non pertinente est bien plus présente que celle qui nous intéresse. Nous avons appliqué des méthodes dédiées aux problèmes mal équilibrés, pour éviter de ne prédire que des liens non pertinents.

Le gorille est après le bonobo et le chimpanzé , du point de vue génétique , l' animal le plus proche de l' humain . Cette parenté a été confirmée par les similitudes entre les chromosomes et les groupes sanguins . Notre génome ne diffère que de 2

Redressés , les gorilles atteignent une taille de 1,75 mètre , mais ils sont en fait un peu plus grands car ils ont les genoux fléchis . L' envergure des bras dépasse la longueur du corps et peut atteindre 2,75 mètres

Il existe une grande différence de masse entre les sexes : les femelles pèsent de 90 à 150 kilogrammes et les mâles jusqu' à 275. En captivité , particulièrement bien nourris , ils atteignent 350 kilogrammes .

Le pelage dépend du sexe et de l' âge . Chez les mâles les plus âgés se développe sur le dos une fourrure gris argenté , d' où leur nom de ¡¡ dos argentés ¿¿. Le pelage des gorilles de montagne est particulièrement long et soyeux .

Comme tous les anthropoïdes , les gorilles sont dépourvus de queue . Leur anatomie est puissante , le visage et les oreilles sont glabres et ils présentent des torus supra-orbitaires marqués .

Figure 6. *Composantes connexes de moins de cinq sommets ; chaque composante est indiquée par une couleur différente*

Indice	Description	Valeurs possibles
$freqtxt_{\min}$	$\min(freqtxt_a, freqtxt_b)$	$freqtxt_{\min} \in \mathbb{N}^*$
$freqtxt_{\max}$	$\max(freqtxt_a, freqtxt_b)$	$freqtxt_{\max} \in \mathbb{N}^*$
$freqtxt_{\times}$	$\log(freqtxt_a \times freqstxt_b)$	$freqstxt \in \mathbb{N}^*$
tf·ipf	$tf{\cdot}ipf(voisin_a) \cdot tf{\cdot}ipf(voisin_b)$	$0 \leq tf{\cdot}ipf \leq 1$
$copr_{ph}$	coprésence dans une même phrase	booléen
$copr_{para}$	coprésence dans un même paragraphe	booléen
ppd	+ petite distance entre $voisin_a$ et $voisin_b$	$ppd \in \mathbb{N}^*$
pgd	+ grande distance entre $voisin_a$ et $voisin_b$	$pgd \in \mathbb{N}^*$
md	distance moyenne entre $voisin_a$ et $voisin_b$	$md \in \mathbb{N}^*$
$prodtxt_{\min}$	$\min(prod_a, prod_b)$	$prod_{\min} \in \mathbb{N}^*$
$prodtxt_{\max}$	$\max(prod_a, prod_b)$	$prod_{\max} \in \mathbb{N}^*$
$prodtxt_{\times}$	$\log(prod_a \times prod_b)$	$prods \in \mathbb{N}^*$
cc	appartenance à une composante connexe	booléen

Tableau 9. *Indices émanant du texte*

Selon une méthodologie classique, nous avons effectué une évaluation par validation croisée (*10-fold cross-validation*) (Witten *et al.*, 2011, p. 152-154), en divisant les données en dix fractions, chaque fraction servant à tester un modèle entraîné sur les neuf autres. Nous avons testé plusieurs méthodes courantes de classification, et présentons les résultats obtenus avec un modèle bayésien naïf, comme étalon, et avec la

meilleure des méthodes utilisées (l'algorithme des Random Forest (Breiman, 2001))[4]. Les Random Forest sont un ensemble d'arbres de décision qui votent sur la prédiction à faire. Les différences entre les arbres sont induites par des choix aléatoires sur des sous-ensembles de traits sélectionnés pour chaque branchement de l'arbre. Nous avons aussi considéré la méthode « baseline » qui consiste à ajuster le seuil de Lin sur une partie des données pour optimiser la F-mesure résultante.

Pour compenser le déséquilibre des classes quand on cherche avant tout la classe minoritaire (ici les liens pertinents), deux grandes méthodes sont généralement appliquées : soit un rééchantillonnage des données d'entraînement pour rééquilibrer les classes et « forcer » le modèle à regarder la classe minoritaire, soit la prise en compte à l'entraînement d'un coût différent selon les classes, en pénalisant plus fortement les erreurs sur la classe que l'on vise. Une méthode classique dans le premier cas, et plus efficace qu'un simple rééchantillonnage, est la méthode Smote (Chawla *et al.*, 2002), qui synthétise de nouvelles instances par similarité avec des instances de la classe minoritaire. Un exemple classique de la deuxième stratégie, qui permet facilement d'introduire des coûts dans tout modèle d'apprentissage, est la méthode ensembliste MetaCost (Domingos, 1999). Nous avons utilisé les implémentations des diverses méthodes fournies par la bibliothèque Weka (Frank *et al.*, 2004), qui offre un cadre homogène pour l'expérimentation, même si des raffinements plus récents ont été faits depuis sur ces stratégies. Les paramètres suivants ont été choisis : pour la méthode bayésienne naïve, l'estimation de densité par noyau a été utilisée pour les traits numériques, dans la mesure où elle améliore généralement les résultats. Pour les Random Forests, nous avons utilisé une population de dix arbres, et les décisions sont prises sur un ensemble aléatoire de cinq traits à la fois. Pour le rééchantillonnage, la méthode Smote préconise de doubler le nombre d'instances de la classe minoritaire. Nous avons vérifié qu'augmenter ce nombre dégrade en effet les performances. Enfin pour l'entraînement avec coût, une option raisonnable est d'avoir un ratio inverse entre les coûts des classes minoritaires ou majoritaires et leur fréquence relative, c'est-à-dire ici de régler le coût d'une erreur sur un lien pertinent (faux négatif) à 8,5 fois celui d'une erreur sur un lien non pertinent (faux positif).

Si l'on reprend les notions de précision et de rappel telles que posées précédemment (c'est-à-dire la précision et le rappel pour la classe *pertinent*, *cf.* section 4.2), cette classification correspond à une précision de $68,1$ % et un rappel de $24,2$ %. Le F-score résultant est alors de $35,7$ %. En utilisant un modèle bayésien naïf comme indiqué ci-dessus, les scores de précision, rappel, F-mesure sont de 34,8, 51,3 et 41,5 respectivement. Si l'on se réfère à la figure 5, on voit que, pour un rappel équivalent, la précision obtenue en posant un seuil sur le score de Lin est à peine supérieure à 20 %. En réglant le meilleur seuil sur le score de Lin, la même figure indique qu'on ne peut espérer mieux qu'un F score de 25 %. Par ailleurs, si l'on classait tous les liens comme pertinents, et donc avec un rappel de 100 %, la précision de ces prédictions ne

4. Les autres méthodes classiques (*MaxEnt*, SVM...) ont abouti à des résultats légèrement inférieurs en F-mesure, même si les scores de précision et rappel peuvent montrer des variations plus importantes.

serait que de 2,6 %, et le F1 résultant de 5 %. Ces scores de bases sont améliorés avec les techniques de rééchantillonnage et d'apprentissage avec coût, et la méthode des Random Forest est celle qui bénéficie le plus de ces ajustements ; le tout est résumé tableau 10.

	Précision	Rappel	F-mesure	I.C.
Baseline (Lin ajusté)	24	24	24	
RF	**68,1**	24,2	35,7	± 3,4
NB	34,8	51,3	41,5	± 2,6
RF + rééchantillonnage	56,6	32,0	40,9	± 3,3
NB + rééchantillonnage	32,8	54,0	40,7	± 2,5
RF + coût	40,4	54,3	**46,3**	± 2,7
NB + coût	27,3	**61,5**	37,8	± 2,2

Tableau 10. *Scores de classification (en %) rapportés à la classe des liens pertinents : précision, rappel, f-mesure, intervalle de confiance sur la f-mesure. RF = Random Forest, NB = bayésien naïf.*

Ainsi, cette stratégie de filtrage s'avère payante. Elle permet d'améliorer très largement les résultats de l'analyse distributionnelle, en l'ajustant à son corpus d'origine *via* des indices prenant en compte des caractéristiques de ce corpus et des textes particuliers dans lesquels les couples de voisins apparaissent.

Pour mesurer plus précisément l'impact de ces nouveaux indices, nous avons répété la phase d'apprentissage en soustrayant à chaque fois une certaine catégorie d'indices (*cf.* section 5.1) ; les résultats sont présentés dans le tableau 11, pour la meilleure méthode (RF avec coût).

Indices	Précision	Rappel	F-mesure
Tous	40,4	54,3	46,3
sans les traits calculés en corpus	37,4	52,8	43,8
sans les traits liés à l'analyse distrib,	36,1	49,5	41,8
sans les traits liés au contexte	36,5	54,8	43,8

Tableau 11. *Impact des différentes classes d'indices sur les meilleurs scores (en %) : plus les résultats sont faibles, plus l'impact des indices soustraits est important.*

Si les indices fournis par la base distributionnelle sont ceux qui ont le plus fort impact, on peut observer que le poids des autres catégories d'indices est également très important. Cette expérience montre que la qualité des rapprochements distributionnels dépend, en plus des paramètres bien identifiés dans la littérature (score de similarité, rangs), de certaines caractéristiques nouvelles liées à la contextualisation des paires de voisins.

6. Conclusion

Dans cet article, nous avons discuté de la nécessité de progresser dans la compréhension et la maîtrise de l'information sémantique contenue par les ressources distributionnelles ; la mise en place de procédures d'évaluation est centrale pour cet objectif. Toutefois, les méthodes classiquement utilisées pour l'évaluation de ressources lexicales présentent certaines limites lorsqu'il s'agit d'aborder une ressource distributionnelle dans un souci de l'évaluer ou de l'améliorer sans se priver de certains aspects qui font sa richesse – relations variées de proximité sémantique, relations opérant dans un corpus donné... –, car elles tendent à occulter ces aspects. Nous avons donc proposé et mis en place une nouvelle méthode d'évaluation intrinsèque, que nous avons appliquée à une ressource distributionnelle construite à partir d'un corpus de source unique et de taille modeste : les *Voisins de Wikipédia*. Cette méthode s'appuie sur l'annotation de couples de voisins remis en contexte dans leur corpus d'origine.

Nous avons tout d'abord montré qu'une annotation en contexte des liens de voisinage distributionnel constitue une démarche valide d'évaluation d'une ressource distributionnelle : elle est plus fiable qu'une annotation hors contexte, et fournit facilement des données en grand nombre. Nous avons ensuite montré comment la disponibilité de cette référence nous a permis de mettre au jour une variété d'indices qui influencent la qualité de l'analyse distributionnelle, et nous avons développé un système de catégorisation automatique des liens distributionnels qui améliore nettement la qualité de la ressource traitée. Nous avons rendu disponibles le corpus utilisé, l'ensemble des indices calculés et les annotations[5].

Cette démarche est fondée sur une conception de l'analyse distributionnelle héritée de l'hypothèse harrissienne originelle, qui consiste à considérer que les liens de proximité sémantique émergent du corpus d'analyse et sont donc le reflet des configurations sémantiques qui sont construites dans les textes du corpus. En d'autres termes, nous utilisons le corpus et les informations contextuelles (au niveau de l'annotation, puis au niveau des indices de classification) pour consolider la ressource distributionnelle, et nous ne faisons pas l'hypothèse d'une validité « en langue » des relations sémantiques qui sont détectées. La ressource est donc adaptée à des traitements sur le corpus d'origine. Cet ajustement de la ressource au corpus ne constitue pas, selon nous, une limitation, dans la mesure où les résultats présentés dans cet article montrent la validité, y compris sur un corpus de taille relativement restreinte, d'une analyse distributionnelle suivie d'une phase de filtrage, ce qui atténue la nécessité de construire une ressource généralisable.

Notre approche incite à poursuivre, en parallèle du mouvement consistant à traiter des corpus de plus en plus importants (et donc de moins en moins maîtrisés), des expériences mettant en œuvre l'analyse distributionnelle de manière contrôlée – en aval (contrôle du corpus) et en amont (examen des rapprochements calculés). Dans la continuité du travail présenté dans cet article, nous souhaitons exploiter les données

5. `http://clementine.adam.free.fr/donnees/AdamFabreMuller2013.zip`

constituées, qui donnent accès à des couples de voisins jugés pertinents et visualisables en contexte, pour caractériser plus finement le contenu sémantique de notre ressource, en mettant en place une nouvelle phase d'annotation qui conduira au typage des relations de voisinage.

7. Bibliographie

Adam C., Voisinage lexical pour l'analyse du discours, PhD thesis, Université de Toulouse, 2012.

Adam C., Muller P., Fabre C., « Une évaluation de l'impact des types de textes sur la tâche de segmentation thématique », *Actes de TALN'10*, Montréal, Canada, 2010.

Agirre E., Alfonseca E., Hall K., Kravalova J., Paşca M., Soroa A., « A study on similarity and relatedness using distributional and wordnet-based approaches », *Proceedings of Human Language Technologies : The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Association for Computational Linguistics, p. 19-27, 2009.

Anguiano E., Denis P. *et al.*, « FreDist : Automatic construction of distributional thesauri for French », *Actes de la 18e conférence sur le traitement automatique des langues naturelles*, p. 119-124, 2011. Cecile.

Baroni M., Lenci A., « Distributional memory : A general framework for corpus-based semantics », *Computational Linguistics*, vol. 36, n° 4, p. 673-721, 2010.

Baroni M., Lenci A., « How we BLESSed distributional semantic evaluation », *GEMS 2011*p. 1-10, 2011.

Bordag S., « A Comparison of Co-occurrence and Similarity Measures as Simulations of Context », *in* A. F. Gelbukh (ed.), *CICLing*, vol. 4919 of *Lecture Notes in Computer Science*, Springer, p. 52-63, 2008.

Bourigault D., « UPERY : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus », *Actes de la 9e conférence sur le Traitement Automatique de la Langue Naturelle*, Nancy, p. 75-84, 2002.

Breiman L., « Random Forests », *Machine Learning*, vol. 45, n° 1, p. 5-32, 2001.

Chawla N. V., Bowyer K. W., Hall L. O., Kegelmeyer W. P., « SMOTE : Synthetic Minority Over-sampling Technique », *J. Artif. Intell. Res. (JAIR)*, vol. 16, p. 321-357, 2002.

Church K., Hanks P., « Word association norms, mutual information, and lexicography », *Computational Linguistics*, vol. 16, n° 1, p. pp. 22-29, 1990.

Clark S., « Vector Space Models of Lexical Meaning », *in* S. Lappin, C. Fox (eds), *Handbook of Contemporary Semantics*, Wiley-Blackwell, à paraître.

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

Curran J., From distributional to semantic similarity, PhD thesis, University of Edinburgh, 2004.

Dias G., Alves E., Lopes J. G. P., « Topic segmentation algorithms for text summarization and passage retrieval : an exhaustive evaluation », *Proceedings of the 22nd national conference on Artificial intelligence - Volume 2*, AAAI'07, AAAI Press, p. 1334-1339, 2007.

Domingos P., « MetaCost : A General Method for Making Classifiers Cost-Sensitive », *in* U. M. Fayyad, S. Chaudhuri, D. Madigan (eds), *KDD*, ACM, p. 155-164, 1999.

Frank E., Hall M., , Trigg L., « WEKA 3.3 : Data Mining Software in Java », , www.cs.waikato.ac.nz/ml/weka/, 2004.

Gale W., Church K., Yarowsky D., « One Sense Per Discourse », *In Proceedings of the 4th DARPA Speech and Natural Language Workshop, New-York*, p. 233-237, 1992.

Galliers J., Jones K., *Evaluating natural language processing systems*, n° 291, University of Cambridge, Computer Laboratory, 1993.

Habert B., « Des corpus représentatifs : de quoi, pour quoi, comment », *Bilger, M., éditeur : Linguistique sur corpus. Etudes et réflexions*, n° 31, p. 11-58, 2000.

Landis J.R.and Koch G., « The measurement of observer agreement for categorical data », *Biometrics*, vol. 33 (1), p. 159-174, 1977.

Lin D., « An information-theoretic definition of similarity », *Proceedings of the 15th International Conference on Machine Learning*, Madison, p. 296-304, 1998.

Morris J., Hirst G., « Non-classical lexical semantic relations », *Proceedings of the HLT Workshop on Computational Lexical Semantics*, Boston, p. 46-51, 2004.

Pado S., Lapata M., « Dependency-based construction of semantic space models », *Computational Linguistics*, vol. 33, n° 2, p. 161-199, 2007.

Poibeau T., Messiant C., « Do we still Need Gold Standards for Evaluation ? », *Proceedings of the Language Resource and Evaluation Conference*, 2008.

Sahlgren M., The Word-Space Model : using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces., PhD thesis, Stockholm University, 2006.

Salton G., Yang C. S., Yu C. T., « A theory of term importance in automatic text analysis. », *Journal of the American Society for Information Science*, vol. 26, n° 1, p. 33-44, 1975.

Turney P., Pantel P. *et al.*, « From frequency to meaning : Vector space models of semantics », *Journal of Artificial Intelligence Research*, vol. 37, n° 1, p. 141-188, 2010.

van der Plas L., Automatic lexico-semantic acquisition for question answering, PhD thesis, Université de Groningen (Pays-bas), 2008.

Weeds J. E., Measures and Applications of Lexical Distributional Similarity, PhD thesis, University of Sussex, 2003.

Weeds J., Weir D., « Co-occurrence retrieval : A flexible framework for lexical distributional similarity », *Computational Linguistics*, vol. 31, n° 4, p. 439-475, 2005.

Witten I. H., Frank E., Hall M. A., *Data Mining : Practical Machine Learning Tools and Techniques with JAVA Implementations (Third Edition)*, Morgan Kaufmann, 2011.