Approches à base de fréquences pour la simplification lexicale

Anne-Laure Ligozat[1,2] Cyril Grouin[1,3]
Anne Garcia-Fernandez[4] Delphine Bernhard[5]
(1) LIMSI–CNRS, Orsay (2) ENSIIE, Évry (3) INSERM U872 Eq 20 & UPMC, Paris
(4) LAS, CNRS/EHESS/Collège de France, Paris (5) LiLPa, Université de Strasbourg, Strasbourg

RÉSUMÉ

La simplification lexicale consiste à remplacer des mots ou des phrases par leur équivalent plus simple. Dans cet article, nous présentons trois modèles de simplification lexicale, fondés sur différents critères qui font qu'un mot est plus simple à lire et à comprendre qu'un autre. Nous avons testé différentes tailles de contextes autour du mot étudié : absence de contexte avec un modèle fondé sur des fréquences de termes dans un corpus d'anglais simplifié ; quelques mots de contexte au moyen de probabilités à base de n-grammes issus de données du web ; et le contexte étendu avec un modèle fondé sur les fréquences de cooccurrences.

ABSTRACT

Studying frequency-based approaches to process lexical simplification

Lexical simplification aims at replacing words or phrases by simpler equivalents. In this paper, we present three models for lexical simplification, focusing on the criteria that make one word simpler to read and understand than another. We tested different contexts of the considered word : no context, with a model based on word frequencies in a simplified English corpus ; a few words context, with n-grams probabilites on Web data, and an extended context, with a model based on co-occurrence frequencies.

MOTS-CLÉS : simplification lexicale, fréquence lexicale, modèle de langue.

KEYWORDS: lexical simplification, lexical frequency, language model.

1 Introduction

La simplification textuelle consiste à rendre les textes plus faciles à lire, par exemple pour des enfants ou des locuteurs non natifs. Des documents de tout type peuvent ainsi être rendus accessibles à différents publics ; dans notre travail, nous considérerons un public de locuteurs non natifs de l'anglais et des documents de domaine général.

Deux sous-tâches sont généralement distinguées dans la simplification textuelle automatique, bien qu'elles ne soient pas totalement déconnectées : la simplification syntaxique et la simplification lexicale. Nous nous intéressons plus particulièrement à la problématique de la simplification lexicale. Ce type de simplification consiste à remplacer des mots ou des phrases par des équivalents plus simples. Afin de procéder à de telles substitutions, il importe d'abord d'identifier des mots équivalents qui correspondent au contexte, puis de choisir le mot le plus simple. Dans le cadre de nos travaux sur la simplification, nous nous sommes intéressés à la problématique de la simplification lexicale, et plus particulièrement à l'évaluation de mots équivalents en contexte,

en fonction de leur degré de simplicité. Dans cet article, nous présentons les expériences supplémentaires que nous avons menées à partir des systèmes que nous avons créés lors de notre participation à cette campagne (Ligozat *et al.*, 2012). Nous avons défini trois types de critères fondés sur les fréquences des mots à simplifier et de leurs substituts : les critères sur le mot lui-même, des critères reposant sur les contextes locaux, et des critères sur les contextes thématiques. Ce dernier type de critère constitue une expérience nouvelle par rapport à notre participation d'origine à SemEval 2012.

La simplification lexicale est proche de plusieurs tâches. Sa première étape consiste à choisir les substituts possibles d'un mot donné et requiert une désambiguïsation sémantique au niveau du mot et une recherche de paraphrases. La seconde étape considère tous les substituts ou paraphrases possibles, et vise à ordonner ces éléments en fonction de leur niveau de simplicité. La simplification peut également être considérée comme une tâche de traduction entre une langue standard et une version simplifiée de cette langue ; nous notons que dans les traductions habituelles, il est difficile de produire des corpus totalement parallèles.

2 État de l'art

Alors que la simplification syntaxique a fait l'objet d'un grand nombre de travaux (Siddharthan, 2006; Woodsend et Lapata, 2011; Watanabe *et al.*, 2009), la simplification lexicale a comparativement été moins traitée.

Les premiers travaux sur la simplification lexicale ont consisté à remplacer des mots par des synonymes plus communs issus de WordNet ou d'autres dictionnaires (Devlin, 1999; Carroll *et al.*, 1999; Lal et Rüger, 2002). La complexité lexicale est généralement estimée en termes de (i) longueur du mot *(nombre de caractères)* ou nombre de syllabes, ou (ii) de fréquence du mot, fondée sur une analyse en corpus ou une base de données, telle que la base de données psycholinguistique MRC (Lal et Rüger, 2002). Drndarević et Saggion (2012) ont montré que la fréquence des mots et leur longueur en nombre de caractères ou de syllabes étaient des indicateurs utiles de complexité lexicale à partir d'un corpus parallèle espagnol.

Des approches plus récentes se sont intéressées à l'acquisition de simplifications lexicales. Les travaux de Yatskar *et al.* (2010) ont porté sur l'obtention de simplifications lexicales (« *collaborate* » → « *work together* ») à partir des révisions des pages Wikipedia rédigées en anglais simplifié [1]. Les auteurs dérivent ainsi des probabilités de simplification au moyen d'un modèle fondé sur les méta-données d'édition de chaque page. Les 100 plus importantes paires extraites par ces modèles constituent des simplifications avec une précision élevée (86 % sur le meilleur modèle), ce qui représente un point de départ intéressant pour l'acquisition de simplification lexicale. Précisons que ce modèle ne tient cependant pas compte du contexte.

Biran *et al.* (2011) s'appuient sur des paires de substitution apprises à partir du corpus de la Wikipedia en anglais et en anglais simplifié, en fonction de la similarité des contextes des mots, de leur fréquence et de leur longueur. Ces paires sont ensuite utilisées pour simplifier certains mots d'une phrase, en tenant compte de la similarité entre la phrase et les contextes des mots considérés.

1. L'encyclopédie collaborative en ligne Wikipedia propose, pour certains articles, une version en anglais simplifié appelé « Simple English » à destination des locuteurs non natifs de l'anglais.

Woodsend et Lapata (2011) ont implémenté une approche de simplification fondée sur une grammaire quasi synchrone, qui apprend des réécritures de simplification à partir de phrases source/cible extraites des pages Wikipedia rédigées en anglais et en anglais simplifié. Ce modèle intègre également des substitutions lexicales, avec pour objectif le remplacement d'un mot en fonction de son contexte syntaxique. L'acquisition de substituts lexicaux reste cependant limitée aux termes présents en corpus, ce qui réduit l'intérêt d'un tel modèle pour une tâche de simplification lexicale.

Dans ce travail, nous envisageons d'étudier la simplification lexicale en elle-même, en nous attachant à identifier les critères qui font qu'un mot est plus simple à lire et à comprendre qu'un autre mot. Notre approche repose principalement sur les modèles à base de n-grammes, tels que les modèles décrits par Jauhar et Specia (2012). Nous avons cependant essayé d'affiner ces modèles en tenant compte des différents contextes d'apparition du mot étudié. Notre travail repose sur le cadre expérimental fourni par la tâche de simplification lexicale proposée par la campagne d'évaluation SemEval 2012 (Specia *et al.*, 2012).

3 Critères de simplification d'un élément lexical

Nous nous proposons donc d'étudier la simplification lexicale sous l'angle de la caractérisation du caractère simple d'éléments lexicaux en contexte. L'étude de la littérature et du corpus de la campagne SemEval 2012 nous a permis de dégager plusieurs critères pour choisir un élément lexical dans un contexte donné (voir par exemple François et Fairon (2012); Jauhar et Specia (2012)) :

– des critères concernant l'élément lui-même, principalement issus des mesures de lisibilité de textes : taille de l'élément en nombre de caractères ou de syllabes, fréquence de l'élément en corpus, présence de cet élément dans des listes de mots simples, caractéristiques psycholinguistiques de l'élément (comme par exemple caractère concret, âge d'acquisition ou autres provenant de la MRC Psycholinguistic Database)...
– le contexte local de l'élément, et notamment dans le cas de l'appartenance à une collocation. Dans la phrase « *Put granola bars in bowl* . », le contexte local « *granola* » nous permet d'identifier le substitut « *bar* » comme meilleur choix possible ;
– le contexte plus général de l'élément, notamment son contexte thématique. Ainsi, dans la phrase « *The film shows Afghan mercenaries to be involved with the separatists , suggesting that the present struggle in Kashmir has been hijacked by foreign extremists , who are shown discussing the loss of Bangladesh in the 1971 war , providing it as a justification for their present acts of revenge* . », il est nécessaire de prendre en compte tout le contexte du mot cible « *film* » pour identifier le substitut « *documentary* » comme meilleur choix par rapport aux autres substituts possibles « *film, movie, picture* ».

Nous émettons l'hypothèse que l'utilisation d'un contexte plus important permet de mieux tenir compte des spécificités sémantiques des substituts et de l'environnement linguistique dans lequel ces substituts évoluent.

4 Corpus

Dans le cadre de ce travail, nous avons poursuivi les expériences que nous avions menées lors de notre participation à la tâche de simplification lexicale de l'anglais proposée par la campagne SemEval 2012 [2]. À ce titre, nous avons appliqué nos méthodes et effectué de nouvelles expériences en nous appuyant sur les corpus de la campagne.

4.1 Présentation

Dans le cadre de cette tâche, deux corpus ont été fournis. Le corpus d'apprentissage contient 300 instances tandis que le corpus de test, utilisé pour l'évaluation, se compose de 1 710 instances. Le corpus d'apprentissage s'accompagne des annotations de référence pour permettre le développement des systèmes.

Le corpus se compose de textes courts issus de documents récupérés sur internet, dans lesquels un mot cible a été choisi, et pour lequel plusieurs substituts possibles doivent être ordonnés. Dans l'exemple suivant, le mot « *outdoor* » est la cible à traiter et tous les autres mots du texte constituent le contexte de ce mot cible.

```
<instance id="270">
<context>With the growing demand for these fine garden furnishings , they found it
necessary to dedicate a portion of their business to <head>outdoor</head> living
and patio furnishings .</context>
</instance>
```

Pour cette cible, les substituts proposés sont les suivants : *{alfresco, outside, open-air, outdoor}*. Les informations disponibles sur la constitution de la référence nous permettent de savoir que ces substituts ont été ordonnés par des locuteurs non natifs de l'anglais (respectivement 4 et 5 annotateurs pour les corpus d'apprentissage et de test) selon leur degré de simplicité décroissant. Nous n'avons cependant pas connaissance d'un guide d'annotation auquel se référer. L'objectif de la tâche consiste donc à ordonner ces différents substituts en fonction de leur degré de simplicité.

La séquence de référence associée à ces substituts est la suivante : *(outdoor, open-air, {outside, alfresco})*, où « *outdoor* » est considéré comme le substitut le plus simple, tandis que « *outside* » et « *alfresco* » sont considérés comme les substituts les plus complexes à égalité.

4.2 Statistiques

Nous donnons ci-après quelques éléments statistiques calculés sur les corpus d'apprentissage et de test, afin de représenter la difficulté de la tâche.

Nombre de tokens dans chaque contexte. Dans un premier temps, nous avons étudié le nombre de tokens dans chaque contexte, un token étant considéré comme une chaîne de caractères entre deux espaces. Alors que les contextes les plus longs se retrouvent dans le corpus de test, nous avons relevé que les contextes sont, en moyenne, plus courts dans le corpus de test

2. http://www.cs.york.ac.uk/semeval-2012/task1/

que dans le corpus d'apprentissage, avec un nombre moyen de 27,6 tokens dans le test contre 28,9 dans l'apprentissage (tableau 1, gauche). Les contextes les plus courts se composent de 5 tokens dans les deux corpus. Ainsi, le contexte « *Well , perhaps not .* » se rapporte au mot cible « *well* » dans le corpus d'apprentissage alors que le contexte « *The spin's are flat .* » se rapporte au mot cible « *flat* » dans le corpus de test. Cela signifie qu'il existe des informations contextuelles pour pratiquement tous les mots cibles, et que ce contexte peut être utilisé pour choisir les substituts qui conviennent le mieux à ce contexte.

Corpus	**Nombre de tokens**			**Nombre de substituts**		
	Min	Max	Moy	Min	Max	Moy
Apprentissage	5	76	28,9	2	9	4,8
Test	5	92	27,6	1	10	5,0

TABLE 1 – Nombre minimum, maximum et moyen de tokens par contexte (gauche) et nombre minimum, maximum et moyen de substituts par instance (droite)

Nombre de substituts par contexte. Nous avons également calculé le nombre de substituts proposés pour chaque cible dans chaque instance à traiter. Il y a, en moyenne, cinq substituts proposés par instance dans les deux corpus (tableau 1, droite). Chaque instance se compose ainsi de plusieurs substituts à ordonner.

Fréquence d'utilisation des substituts en corpus. Un point intéressant concerne le nombre de fois que chaque substitut est proposé dans chacun des corpus. La majorité des ensembles proposés de substituts se composent de substituts proposés une seule fois. Nous remarquons cependant qu'il y a davantage de substituts proposés une seule fois dans le corpus d'apprentissage que dans le corpus de test. Nous reportons sur le graphique 1 le pourcentage d'utilisation des substituts, classés par nombre d'occurrences décroissant, proposés respectivement dans les corpus d'apprentissage (en rouge) et de test (en bleu).

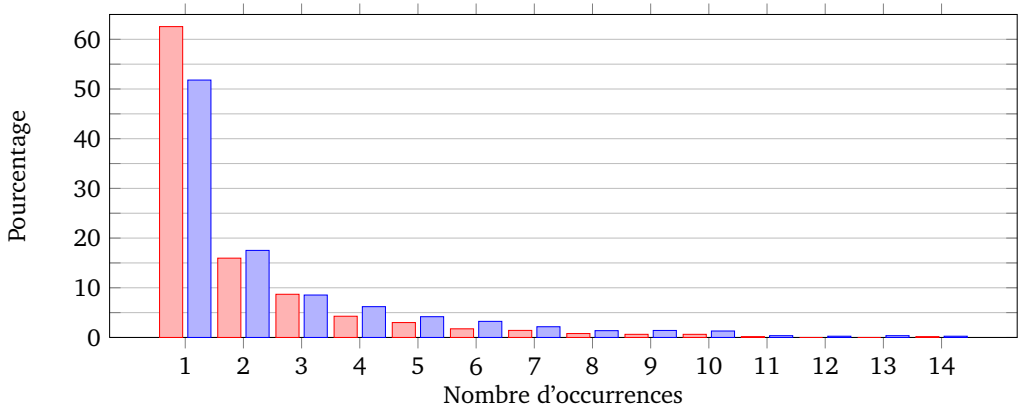

FIGURE 1 – Pourcentage d'utilisation de chaque substitut sur le corpus d'apprentissage (rouge) et de test (bleu), classé par nombre d'occurrences décroissant

Si la majorité des substituts est proposée une seule fois (62,6 % des substituts du corpus d'apprentissage et 51,8 % dans le corpus de test ne sont présentés qu'une seule fois), certains apparaissent néanmoins un nombre élevé de fois (les substituts présentés deux fois constituent 16,0 % et 17,5 % du nombre total de substituts). En matière de présentation maximum, le substitut « *unpleasant* » est proposé jusqu'à 14 fois dans le corpus d'apprentissage (sur un total de 633 substituts) alors que « *consequently* » est proposé 26 fois dans le corpus de test (sur un total de 2 774 substituts).

Catégories morpho-syntaxiques des substituts. Chaque mot cible relève d'une catégorie morpho-syntaxique parmi quatre catégories possibles. Nous avons étudié la répartition des mots cibles en fonction de leur catégorie d'appartenance (tableau 2).

Catégorie	Apprentissage	Test
Adjectif	26,5 %	27,5 %
Adverbe	14,7 %	17,5 %
Nom	23,5 %	29,2 %
Verbe	23,5 %	25,7 %
Adjectif ou Nom	5,9 %	—
Nom ou Verbe	5,9 %	—

TABLE 2 – Pourcentage de mots cibles appartenant à chaque catégorie morpho-syntaxique

Nous remarquons que la répartition des mots cibles dans chacune des catégories morpho-syntaxique est similaire entre les deux corpus. Cependant, le corpus d'apprentissage se compose de mots cibles ambigus dans la mesure où certains de ces mots peuvent relever de deux catégories potentielles (adjectif ou nom, nom ou verbe). Cette ambiguïté disparait dans le corpus de test.

4.3 Expériences de base

Trois expériences de base *(baselines)* ont été fournies par les organisateurs en accompagnement du corpus d'apprentissage :
- La première relève d'un simple ordonnancement au hasard des substituts de chaque ensemble proposé ;
- La seconde conserve la liste de substituts proposés dans l'ordre dans lequel elle est fournie ;
- La troisième (appelée « fréquence simple ») repose sur l'utilisation des fréquences des termes présents dans le corpus Google Web 1T.

Ces expériences de base permettent, d'une part de fixer le seuil minimum à atteindre, et d'autre part de présenter de premières approches simples pour résoudre la problématique soulevée.

5 Méthodes

Nous avons implémenté trois modèles distincts qui correspondent à différentes tailles de contextes que nous avons envisagés autour des mots cibles : (i) pas de contexte, (ii) quelques mots, et

(*iii*) le contexte entier. L'idée sous-jacente de ces expériences concerne le fait qu'un substitut peut être préféré à un autre parce qu'il est plus fréquent *(le contexte n'est donc pas nécessaire)*, parce qu'il appartient à une expression *(auquel cas, le contexte composé de quelques mots se révèle utile)*, ou bien, parce qu'il est le plus adapté sur le plan sémantique *(l'ensemble du contexte est alors utilisé)*.

5.1 Modèle fondé sur les fréquences des termes

Notre premier modèle ne prend pas en compte le contexte d'utilisation des mots et repose sur les fréquences des substituts trouvées dans un corpus rédigé en anglais simplifié, la *Simple English Wikipedia* (SEW). Notre hypothèse de travail repose sur le fait que les mots les plus fréquemment employés dans ce corpus seront préférés par les locuteurs non natifs de l'anglais. Ce public correspond au profil des annotateurs utilisés pour la tâche de simplification lexicale. La SEW a déjà été utilisée dans des travaux portant sur la simplification automatique de textes (Yatskar *et al.*, 2010). D'autre part, puisque les corpus SemEval sont constitués de données issues d'internet, nous estimons qu'ils sont proches des textes de Wikipedia du point de vue linguistique.

Dans un premier temps, nous avons converti la SEW au format texte à partir de l'archive du 27 février 2012 dont nous avons extrait le contenu textuel grâce à l'outil wikipedia2text [3]. Le fichier texte final contient approximativement 10 millions de mots.

Nous avons ensuite extrait des n-grammes de mots, en variant la taille des n-grammes de 1 à 3 mots, ce qui est suffisant pour la plupart des substituts. Le corpus d'apprentissage contient seulement deux substituts composés de quatre mots, ce qui constitue la taille la plus importante. Nous avons néanmoins constaté que le corpus de test comprend des substituts pouvant aller jusqu'à sept mots, tel que « *cause your outer work to be more* » ou « *stop at the side of the road* », qui seront de toute façon moins fréquents que des mots plus courts. Nous avons ensuite calculé des fréquences de n-grammes depuis ce corpus grâce au module Perl Text-NSP [4] et le script associé count.pl qui produit la liste des n-grammes d'un document avec leurs fréquences. Nous renseignons dans le tableau 3 du nombre de n-grammes produits en fonction de la taille des n-grammes.

taille des n-grammes	1	2	3	1 à 3
nombre de n-grammes	301 718	2 517 394	6 680 906	9 500 018

TABLE 3 – Nombre de n-grammes distincts extraits de Wikipedia, version anglais simplifié

Certains des n-grammes ne sont pas valides et résultent d'erreurs lors de l'extraction du texte des pages Wikipedia : « 27|ufc 1 » correspond ainsi à une syntaxe du wiki. Puisqu'il est impossible de trouver ce type de n-gramme en corpus, nous n'avons pas cherché à nettoyer nos listes.

Sur les corpus de la tâche SemEval et pour une instance donnée, nous avons ordonné les substituts proposés par fréquence d'apparition décroissante dans la SEW. Ainsi, sur l'ensemble de substituts *{intelligent, bright, clever, smart}*, les fréquences calculées sur la SEW sont respectivement de

3. Voir http://www.polishmywriting.com/download/wikipedia2text_rsm_mods.tgz et http://blog.afterthedeadline.com/2009/12/04/generating-a-plain-text-corpus-from-wikipedia

4. http://search.cpan.org/~tpederse/Text-NSP-1.25/lib/Text/NSP.pm

(206, 475, 141, 201) ; notre classement final sera donc *{bright, intelligent, smart, clever}*.

Sur cette base de travail, nous avons réalisé plusieurs expériences. Nous avons utilisé la version texte brut de la SEW, ainsi que la version lemmatisée, puisque les substituts proposés sont des lemmes. Nous avons réalisé cette étape de lemmatisation grâce au TreeTagger [5] (Schmid, 1994) que nous avons appliqué sur l'ensemble du corpus, avant d'effectuer les décomptes de n-grammes.

D'autre part, puisque les bigrammes et trigrammes augmentent le volume des données, nous avons cherché à mesurer leur influence sur les résultats produits. En se fondant sur les unigrammes uniquement, 158 substituts du corpus d'apprentissage sont absents des annotations de référence ; ce nombre se réduit à 105 en ajoutant les bigrammes et à 91 lorsque l'on ajoute les trigrammes. Deux substituts se composent donc de quatre mots et 89 substituts sont absents de notre corpus SEW. Les n-grammes manquants (en utilisant des uni-, bi- et tri-grammes) semblent cependant très peu fréquents, tels que « *undomesticated* » ou « *telling untruths* ».

5.2 Probabilités des termes en contexte

Notre deuxième modèle repose sur les modèles de langue, méthode utilisée par les organisateurs dans leur expérience de base sur les fréquences simples. Alors que les organisateurs ont utilisé les n-grammes de Google [6] pour ordonner les substituts par fréquence d'utilisation décroissante, nous avons utilisé les n-grammes du service Microsoft Web en retenant le même principe de tri par fréquence décroissante. Nous avons également ajouté les contextes à chaque substitut. Notre approche repose sur les n-grammes proposés par le service Microsoft Web [7], via la librairie Python [8], pour obtenir la probabilité de regroupement d'unités textuelles. Parmi les différents modèles de n-grammes disponibles, nous avons utilisé le modèle *bing-body/apr10/*.

Nous avons ainsi étudié une unité textuelle composée de l'élément lexical et d'une fenêtre contextuelle reposant sur les quatre tokens encadrant l'élément lexical de part et d'autre. Ainsi, sur l'exemple ci-dessous, nous avons testé la portion d'origine « *He brings an incredibly rich and diverse background that* » et les versions dans lesquelles le mot cible est remplacé par un substitut, telles que « *He brings an incredibly lush and diverse background that* ».

```
<instance id="118">
<context>He brings an incredibly <head>rich</head> and diverse background
that includes everything from executive coaching , learning &amp; development
and management consulting , to senior operations roles , mixed with a masters in
organizational development.</context>
</instance>
```

L'une des faiblesses de ce modèle est qu'il ne prend en compte qu'un contexte local, alors que des mots plus éloignés du contexte pourraient également être utiles au choix. Pour tester cette hypothèse, nous avons mis en œuvre un troisième modèle, qui utilise le texte entier comme contexte.

5. http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/
6. Le corpus Google Web 1T utilisé dans l'expérience de base des organisateurs n'est pas disponible gratuitement.
7. http://research.microsoft.com/en-us/collaboration/focus/cs/web-ngram.aspx
8. http://web-ngram.research.microsoft.com/info/MicrosoftNgram-1.02.zip

5.3 Comparaison des contextes et co-occurrents

Afin de tester la phrase entière comme contexte, nous avons utilisé deux ressources de co-occurrences : Wortschatz (Quasthoff *et al.*, 2006) d'une part, et une liste de co-occurrents que nous avons construite depuis la SEW d'autre part.

Wortschatz se compose de listes de co-occurrents provenant de plusieurs corpus tels que des corpus d'informations ou Wikipedia [9]. Dans un premier temps, nous avons utilisé l'un des corpus disponible pour l'anglais, composé d'articles Wikipedia potentiellement proches du corpus de SemEval, de manière à produire des listes de co-occurrents avec leur fréquence en corpus. Ces co-occurrences ne sont pas dirigées, c'est-à-dire que les contextes droit et gauche ne sont pas distingués.

Nous avons également construit une deuxième ressource à partir de la *Simple English Wikipedia*, en retenant tous les mots qui co-occurrent avec un substitut dans la même phrase. Nous avons néanmoins limité les co-occurrences testées à certaines catégories des parties du discours telles que les noms, les noms propres, les verbes, etc.

Pour chacune de ces deux ressources, nous avons considéré que la fréquence des co-occurrents formait un vecteur pour un substitut particulier, et avons calculé le produit scalaire avec les mots du contexte. Par exemple, le vocabulaire du contexte suivant est composé des termes « *(and, the, morans, have, to, ruin, Beethoven's, 6th, in, process, too)* » qui apparaissent une seule fois dans la phrase, sauf l'article « *the* » qui apparait trois fois. Leur fréquence de co-occurrence avec le substitut « *audacity* » est (0, 102, 0, 29, 0, 0, 0, 0, 3, 0, 0), le produit scalaire final est de 338.

```
<instance id="217">
<context>And the morans have the <head>gall</head> to ruin Beethoven's 6th in
the process , too .</context>
</instance>
```

Sur la base de ces listes de co-occurrents, nous avons ordonné les substituts en nous fondant sur le poids calculé, en classant les substituts par ordre décroissant.

6 Évaluation

L'évaluation officielle de la tâche de simplification lexicale repose sur une comparaison par paire des listes de rangs fournis par le système avec les rangs de référence (Specia *et al.*, 2012). Pour chaque paire de substituts, le script d'évaluation compare la position de chaque terme de la paire entre l'hypothèse et la référence en termes de position dans la hiérarchie (position identique, plus haute, plus basse). Le score final d'un jeu de substituts correspond à la moyenne des coefficients κ d'accord inter-annotateur (Formule 1) calculés sur chaque paire d'un contexte.

$$\kappa = \frac{Po - Pe}{1 - Pe} \quad (1)$$

Dans cette formule, pour un jeu de substituts donné, « Po » renvoie à la probabilité observée (le nombre d'accords divisé par le nombre total de paires) tandis que « Pe » correspond à la

9. http://corpora.informatik.uni-leipzig.de/download.html

probabilité attendue (calculée en faisant la somme des accords de position identique, plus haute, plus basse, divisée par le nombre total de paires).

Considérons la liste de substituts {A,C,B} fournie par un système et la liste de référence {A,B,C} correspondante. Sur la paire {A,B}, le terme A occupe la même position dans les deux listes de substituts ; le terme B n'occupe pas la même position mais il suit le terme A dans les deux listes. Sur cette paire, l'évaluation rapporte deux points d'accord au système : un point pour la position identique, et un point pour l'ordre identique des termes A et B dans la paire (relation « plus grand que »). Le même calcul est poursuivi sur les paires {A,C} et {B,C}.

6.1 Expériences de base

Nous indiquons dans le tableau 4 les scores calculés sur les corpus d'apprentissage et de test pour les expériences de base fournies par les organisateurs.

	Apprentissage	**Test**
Tri au hasard	0,016	—
Pas de tri	0,050	—
Fréquence simple	0,398	0,471

TABLE 4 – Résultats des expériences de base

6.2 Modèle fondé sur les fréquences des termes

Le tableau 5 résume les résultats obtenus par notre modèle fondé sur les fréquences de la SEW.

Type de n-grammes	**Lemmes**	**Apprentissage**	**Test**
Unigrammes uniquement	non	0,333	—
Uni- et bigrammes	non	0,371	—
Uni-, bi- et trigrammes	non	0,381	0,465
Uni-, bi- et trigrammes	oui	0,380	0,462
Uni-, bi- et trigrammes (Wikipedia standard)	non	0,343	—
Expérience de base (fréquence simple)		0,398	0,471
WLV-SHEF-SimpLex (meilleur système à SemEval 2012)		—	0,496

TABLE 5 – Résultats obtenus par notre système fondé sur la Simple English Wikipedia et comparaison avec d'autres expériences (Wikipedia standard, expérience de base, meilleur système à SemEval 2012)

La différence que nous observons dans les résultats entre la version lemmatisée et la version fléchie de Wikipedia s'explique de deux manières. En premier lieu, puisque les substituts proposés sont présentés sous forme lemmatisée, nous en identifions davantage dans la version lemmatisée (par exemple, le substitut « *abnormal growth* » n'est présent que sous la forme au pluriel « *abnormal*

growths » dans la version fléchie de Wikipedia). En second lieu, certains substituts font défaut dans la version lemmatisée, la plupart en raison d'erreurs du TreeTagger (par exemple, « *be scared of* » devient « *be scare of* »).

L'hypothèse selon laquelle l'utilisation de Wikipedia en anglais simplifié est plus adaptée à cette tâche que la version standard est validée, dans la mesure où nous obtenons un score plus faible en utilisant la version standard de la Wikipedia [10].

Pour l'évaluation finale, nous avons conservé le système qui a obtenu le meilleur score (0,381) sur les données d'apprentissage, en l'occurrence le système fondé sur des uni-, bi- et trigrammes non lemmatisés. Ce système a obtenu un score de 0,465 sur le corpus de test, nous classant seconds ex-æquo lors de l'évaluation SemEval.

6.3 Probabilités des termes en contexte

Nous avons réalisé plusieurs expériences supplémentaires, fondées sur différents modèles de n-grammes et des tailles de contexte distinctes. Les résultats les plus significatifs sont présentés dans le tableau 6.

Taille du contexte gauche	0	3	2	3	4
Taille du contexte droit	3	0	2	3	4
Score	0,362	0,358	0,365	0,358	0,370

TABLE 6 – Résultats obtenus avec le service Microsoft Web N-gram, sur le corpus d'apprentissage

Nous observons que la fenêtre de contexte composée de quatre tokens encadrant le substitut étudié est celle qui nous a permis d'obtenir les meilleurs résultats sur le corpus d'apprentissage (0,370). Avec cette configuration, nous avons obtenu un score de 0,396 sur les données de test.

6.4 Co-occurrents

Enfin, nous renseignons dans le tableau 7 des scores obtenus par notre modèle à base de co-occurrents. Sur le corpus d'apprentissage, cette méthode nous permet d'obtenir un score de 0,373 avec la meilleure configuration, celle reposant sur la ressource constituée depuis la SEW. Nous notons par ailleurs que l'ajout d'informations de parties-du-discours améliore les résultats.

Ressource	Wortschatz	Wortschatz	SEW	SEW	SEW
Paramètres	Corpus 3M	Corpus 10M	POS : NN, NP, JJ	POS + VB	Toutes les POS
Score	0,280	0,271	0,255	0,264	0,373

TABLE 7 – Scores obtenus avec le modèle de co-occurrences sur le corpus d'apprentissage

10. La Wikipedia standard étant bien plus volumineuse que la version simplifiée, nous en avons utilisé un extrait aléatoire de 375M, du même ordre de grandeur que la Wikipedia simplifiée (156M).

6.5 Évaluation sur les substituts non composés

Nous avons par ailleurs observé que nos modèles ont rencontré des difficultés à tenir compte des substituts composés de plusieurs mots. Afin de pallier cette difficulté, nous avons lancé une évaluation en ne considérant que les substituts composés d'un seul mot (tableau 8). Comme nous nous y attendions, tous les modèles voient leurs performances augmenter en ne considérant que les substituts simples (1 mot), en particulier pour le modèle fondé sur les co-occurrences.

Modèle	SEW	Web N-grams	Co-occurrences	Fréquence simple
Tous types de substituts	0,381	0,370	0,373	0,398
Substituts simples (1 mot)	0,390	0,385	0,414	0,408

TABLE 8 – Scores obtenus en considérant tous les substituts et les substituts simples sur les données d'apprentissage

7 Discussion

Malgré des performances relativement bonnes, l'une des limites du modèle fondé sur les fréquences calculées sur la SEW concerne le fait que ce modèle s'appuie uniquement sur les formes de surface des mots (ou des n-grammes), et que certaines fréquences se trouvent biaisées. Ainsi, le mot « *light* » est aussi bien un nom qu'un adjectif dans Wikipedia ; lorsque nous traitons le jeu de substituts *{light, bright, luminous, clear, well-lit}*, les fréquences des deux catégories morpho-syntaxiques du terme « *light* » sont combinées, accordant plus de poids à ce terme et permettant à ce substitut de mieux se classer. Une solution consisterait à utiliser des n-grammes annotés en parties du discours.

D'autre part, ce modèle ne tient pas compte du contexte du mot, alors que les mêmes substituts sont parfois ordonnés différemment. Dans l'exemple suivant, le mot cible « *film* » a été préféré au substitut possible « *movie* » dans les instances 16 et 19 par les annotateurs, et dans l'ordre inverse pour les instances 15 et 17.

```
<instance id="15">
<context>Film Music Literature Cyberplace - Includes <head>film</head> reviews
, message boards , chat room , and images from various films .</context>
</instance>
<instance id="16">
<context>His feature <head>film</head> debut HEROES / DE STARSTE HELTE (
1996 ) won awards at Rouen and Madrid .</context>
</instance>
<instance id="17">
<context>( Some people keep their TVs on for company. ) In Malta , news is the
main reason we turn to TV , followed by <head>films</head> , talk shows , docu-
mentaries , serials , and music , in that order .</context>
</instance>
<instance id="19">
```

<context>A fine score by George Fenton (THE CRUCIBLE) and beautiful photograhy by Roger Pratt add greatly to the effectiveness of the <head>film</head> .</context>
</instance>

Cet exemple montre que, selon le contexte du mot dans la phrase, des substituts différents peuvent être choisis.

Sur le modèle à base de co-occurrences, l'un des principaux problèmes concerne l'absence de co-occurrences dans le corpus SEW. Il ne nous a donc pas été possible d'obtenir des informations de co-occurrences pour 182 substituts du corpus d'apprentissage (sur un total de 1 452) qui n'ont donc pu être traités. L'un des moyens de pallier cette difficulté consisterait à élargir la taille de la fenêtre de recherche des co-occurrents à deux phrases par exemple, ou d'utiliser un corpus plus volumineux. Les corpus d'anglais simplifié sont cependant rares.

Enfin, la principale difficulté à laquelle nous avons été confrontés sur l'ensemble des modèles concerne les substituts composés de plusieurs mots, pour lesquels la comparaison des fréquences avec celles des mots simples ne s'avère guère possible.

8 Conclusion

Dans cet article, nous avons présenté trois types de critères à prendre en compte pour la tâche de simplification lexicale et mis en œuvre trois modèles fondés sur les fréquences et sur ces types de critères pour effectuer une simplification lexicale. Le premier modèle repose sur des fréquences d'utilisation de termes dans la version rédigée en anglais simplifié de la Wikipedia (SEW). Le second se fonde sur des probabilités de n-grammes fournies par le service Microsoft Web N-gram. Enfin, le dernier modèle s'appuie sur des informations de co-occurrences.

Les meilleurs scores sont obtenus avec les informations de fréquence dans la Wikipedia en anglais simplifié ; cependant, cette information seule ne suffit pas à déterminer de façon satisfaisante le substitut le plus simple. Puisque les différents modèles fournissent des caractéristiques différentes, nous considérons que la combinaison des trois modèles devrait être bénéfique. Dans cette optique, nous envisageons de tester un tel type de combinaison au moyen d'une approche d'ordonnancement à base de SVM.

Il reste bien évidemment des marges de progression, en particulier sur le traitement des substituts composés de plusieurs mots pour lesquels la mobilisation de traitements supplémentaires se révèle indispensable pour tenir compte de ces particularités.

En ce qui concerne l'application de ces méthodes au français, nous estimons que cette tâche se révèle d'autant plus difficile que sur l'anglais pour deux raisons (en plus de celles identifiées sur l'anglais) : (i) des flexions plus importantes en français qu'en anglais et (ii) de l'absence de corpus du français simplifié. Nous relevons toutefois que des travaux récents tendent à produire ce type de corpus (Brouwers *et al.*, 2012).

Références

BIRAN, O., BRODY, S. et ELHADAD, N. (2011). Putting it simply : a context-aware approach to lexical simplification. *In Proc of ACL*, pages 496–501, Portland, OR.

BROUWERS, L., BERNHARD, D., LIGOZAT, A.-L. et FRANÇOIS, T. (2012). Simplification syntaxique de phrases pour le français. *In Actes de JEP-TALN-RECITAL*, pages 211–224, Grenoble, France.

CARROLL, J., MINNEN, G., PEARCE, D., CANNING, Y., DEVLIN, S. et TAIT, J. (1999). Simplifying Text for Language-Impaired Readers. *In Proc of EACL*, pages 269–270.

DEVLIN, S. (1999). *Simplifying natural language text for aphasic readers*. Thèse de doctorat, University of Sunderland, UK.

DRNDAREVIĆ, B. et SAGGION, H. (2012). Towards automatic lexical simplification in spanish : An empirical study. *In Proc of Predicting and Improving Text Readability for target reader populations (PITR) Workshop*, pages 8–16, Montréal, Canada. NAACL-HLT.

FRANÇOIS, T. et FAIRON, C. (2012). An "AI readability" formula for french as a foreign language. *In Proc of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Jeju-do, South Korea.

JAUHAR, S. K. et SPECIA, L. (2012). UOW-SHEF : SimpLex – Lexical Simplicity Ranking based on Contextual and Psycholinguistic Features. *In *SEM*.

LAL, P. et RÜGER, S. (2002). Extract-based Summarization with Simplification. *In Proc of the Workshop on Text Summarization at DUC 2002*.

LIGOZAT, A.-L., GROUIN, C., GARCIA-FERNANDEZ, A. et BERNHARD, D. (2012). ANNLOR : A Naïve Notation-system for Lexical Outputs Ranking. *In Proc of the 6th International Workshop on Semantic Evaluation (SemEval 2012)*.

QUASTHOFF, U., RICHTER, M. et BIEMANN, C. (2006). Corpus Portal for Search in Monolingual Corpora. *In Proc of LREC*, Genoa, Italy.

SCHMID, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. *In Proc of the International Conference on New Methods in Language Processing*, Manchester, UK.

SIDDHARTHAN, A. (2006). Syntactic simplification and text cohesion. *Research on Language & Computation*, 4(1):77–109.

SPECIA, L., JAUHAR, S. K. et MIHALCEA, R. (2012). SemEval-2012 Task 1 : English Lexical Simplification. *In Proc of Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 347–355.

WATANABE, W., JUNIOR, A., UZÊDA, V., FORTES, R., PARDO, T. et ALUÍSIO, S. (2009). Facilita : reading assistance for low-literacy readers. *In Proc of ACM international conference on Design of communication*, pages 29–36. ACM.

WOODSEND, K. et LAPATA, M. (2011). Learning to simplify sentences with quasi-synchronous grammar and integer programming. *In Proc of EMNLP*.

YATSKAR, M., PANG, B., DANESCU-NICULESCU-MIZIL, C. et LEE, L. (2010). For the sake of simplicity : unsupervised extraction of lexical simplifications from Wikipedia. *In HLT'10 Human Language Technologies*, pages 365–368. ACL.