

## Construire un corpus monolingue annoté comparable Expérience à partir d'un corpus annoté morpho-syntaxiquement

**Résumé.** Motivé par la problématique de construction automatique d'un corpus annoté morpho-syntaxiquement distinct d'un corpus source, nous proposons une définition générale et opérationnelle de la relation de la comparabilité entre des corpus monolingues annotés. Cette définition se veut indépendante du domaine applicatif. Nous proposons une mesure de la relation de comparabilité et une procédure de construction d'un corpus comparable. Enfin nous étudions la possibilité d'utiliser la mesure de la perplexité définie dans la théorie de l'information comme moyen de prioriser les phrases à sélectionner pour construire un corpus comparable. Nous montrons que cette mesure joue un rôle mais qu'elle n'est pas suffisante.

**Abstract.** This work is motivated by the will of creating a new part-of-speech annotated corpus in French from an existing one. In this context, we propose a general and operational definition of the comparability relation between annotated monolingual corpora. We propose a comparability measure and a procedure to build semi-automatically a comparable corpus from a source one. We study the use of the perplexity (information theory motivated measure) as a way to rank the sentences to select for building a comparable corpus. We show that the measure can play a role but that it is not sufficient.

**Mots-clés :** Corpus comparable, Corpus monolingue, Corpus annoté, Mesure de la comparabilité, Construction de corpus comparable, Analyse morpho-syntaxique.

**Keywords:** Comparable corpus, Monolingual corpus, Annotated corpus, Measuring comparability, Building comparable corpus, Part-of-Speech tagging.

## 1 Introduction

La question de construction de corpus comparables est souvent abordée dans le contexte applicatif d'extraction terminologique multilingue<sup>1</sup> (Bo et al., 2011). Cette question a aussi son importance dans un contexte monolingue où l'on peut vouloir construire un corpus avec des propriétés linguistiques similaires à un corpus source. En effet, il n'est pas rare pour un chercheur de souhaiter diffuser des corpus et des résultats d'analyse associés et ce, afin de permettre à ses pairs de vérifier ou de poursuivre ses expériences (Nielsen, 2011). Néanmoins, en pratique ce type d'ambition se trouve souvent compromis pour des questions de droit visant la protection de données personnelles ou le respect d'une licence d'exploitation associée aux données utilisées.

Dans ce travail, nous nous situons dans un contexte monolingue avec comme contraintes de ne pas pouvoir diffuser tout ou partie d'un corpus source mais d'avoir à disposition un corpus, que nous appellerons *corpus relais*, de taille importante et ne présentant pas les restrictions d'exploitation du corpus source. Dans ce cadre, nous explorons la possibilité de construire un corpus comparable au corpus source à partir d'extraits du corpus relais.

Nous postulons que la détermination de la *relation de comparabilité* entre deux corpus est fonction du traitement d'analyse (e.g. segmentation en mots, étiquetage morpho-syntaxique, reconnaissance des entités nommées...) à laquelle on destine ces corpus. Plus précisément nous qualifierons de *comparables*<sup>2</sup> des corpus à partir desquels on peut construire des modélisations<sup>3</sup> pour un traitement donné, qui produisent des performances équivalentes lorsqu'elles sont évaluées sur un corpus tiers.

1. <http://comparable.limsi.fr/bucc2013>

2. Le qualificatif de «comparable» est généralement utilisé pour décrire un corpus dont les composantes (des sous-corpus) sont comparables. En ce qui nous concerne, nous utiliserons le terme «corpus comparable» pour désigner directement ces composantes.

3. Nous interprétons le terme «modélisation» au sens large aussi bien pour désigner des règles construites manuellement que pour désigner un modèle probabiliste.

Dans cet article, nous nous intéressons au problème de comparabilité en termes d’annotation morpho-syntaxiques. Nous supposons à disposition un corpus source annoté morpho-syntaxiquement. Nous nous interrogeons sur la possibilité de construire automatiquement un corpus comparable distinct à partir duquel on puisse entraîner un étiqueteur morpho-syntaxique dont la performance ne sera statistiquement pas différente de celle d’un étiqueteur construit à partir du corpus source.

Ce travail poursuit les travaux de (Hernandez & Boudin, 2013) qui ont montré qu’un étiqueteur entraîné sur un corpus annoté automatiquement pouvait obtenir des performances statistiquement non différentes d’un étiqueteur entraîné sur un corpus annoté validé manuellement, et ce pourvu que la quantité de données d’entraînement fût suffisante. En d’autres termes, les auteurs ont montré que sous certaines conditions un corpus avec une annotation manuellement validée et un corpus automatiquement annoté pouvaient être comparables. Nous nous posons ici la question de savoir si les observations tenues par (Hernandez & Boudin, 2013) reposent seulement sur la quantité des données d’entraînement ou bien si une sélection éclairée des phrases d’entraînement pourrait conduire au même résultat.

Dans la section 2, nous proposons une définition générale et opérationnelle de la notion de comparabilité. Nous l’accompagnons d’une proposition d’une procédure générique visant la construction d’un corpus comparable. A la section 3, nous instancions notre réflexion sur la problématique de construction d’un corpus comparable annoté morpho-syntaxiquement en mettant notamment en avant la possibilité d’utiliser la mesure de perplexité (théorie de l’information) comme moyen d’ordonner les phrases à sélectionner pour constituer un corpus comparable. Après avoir décrit notre cadre expérimental à la section 4, nous rapportons à la section 5 les résultats d’expériences de construction menées à partir de différents corpus relais. Enfin nous discutons notre approche par rapport à l’existant 6.

## 2 Corpus comparables monolingues

Nous posons qu’un corpus peut être vu comme un ensemble d’unités textuelles et que ses unités illustrent des phénomènes linguistiques de différentes natures (lexicales, syntaxiques, sémantiques, stylistiques, discursives...).

Un traitement d’analyse sera sensible à un sous-ensemble défini de ces phénomènes. Le résultat du traitement sera fonction d’une part d’une modélisation de ceux-ci construite a priori par l’observation d’un corpus d’entraînement et d’autre part, de l’observation de la distribution de ces phénomènes dans le corpus nouvellement analysé.

Le traitement d’analyse par le type d’information qu’il requiert en entrée contraint la nature des unités textuelles manipulées. Suivant le traitement l’unité considérée sera la phrase, le document...

Soient trois corpus à disposition : un *corpus source*  $S$  pour lequel on veut construire un corpus comparable, un *corpus relais*  $R$  dont on souhaite extraire le corpus comparable  $\tilde{R}$  et un *corpus de référence*  $Q$  sur lequel on peut évaluer et comparer des modélisations construites à partir du corpus source et d’extraits du corpus relais. Soit  $t$  un traitement d’analyse sensible à un ensemble de phénomènes noté  $F_t$ . Soient  $o(F_t, \tilde{R})$  et  $o(F_t, S)$ , les observations que l’on peut capturer selon une même procédure des phénomènes discriminants pour la tâche  $t$  respectivement dans les corpus  $\tilde{R}$  et  $S$ . Soient  $m(o(F_t, \tilde{R}))$  et  $m(o(F_t, S))$ , des modélisations construites selon une même procédure et qui capturent la distribution des phénomènes discriminants pour la tâche  $t$  respectivement sur les corpus  $\tilde{R}$  et  $S$ . Nous noterons  $card_{\tilde{R}}(x)$  et  $card_S(x)$  le nombre d’occurrences du phénomène linguistique  $x$  observés respectivement dans les corpus  $\tilde{R}$  et  $S$ .

### 2.1 Propriétés d’un corpus en relation de comparabilité avec un autre corpus

Nous énonçons qu’un corpus  $\tilde{R}$  est dit comparable à un corpus  $S$  (i.e.  $\tilde{R} \mathcal{C} S$ ) si l’on peut observer les propriétés suivantes :

**Propriété**  $S$  et  $\tilde{R}$  ne sont pas constitués des mêmes unités textuelles.

$$S \neq \tilde{R} \Leftrightarrow \forall u, \text{ unité textuelle, } (u \in S \not\Rightarrow u \in \tilde{R}) \wedge (u \in \tilde{R} \not\Rightarrow u \in S) \quad (1)$$

**Propriété** Les phénomènes linguistiques discriminants pour la tâche  $t$  observés dans les unités de  $S$  sont aussi observés dans celles de  $\tilde{R}$ ,

$$o(F_t, S) \subseteq o(F_t, \tilde{R}) \Leftrightarrow \forall x, (x \in o(F_t, S) \Rightarrow x \in o(F_t, \tilde{R})) \quad (2)$$

**Propriété** ... et ce, dans les mêmes proportions.

$$\forall x \in o(F_t, S), \forall y \in o(F_t, \tilde{R}), (x = y \Rightarrow \text{card}_S(x) = \lambda * \text{card}_{\tilde{R}}(y)) \text{ avec } \lambda > 0 \quad (3)$$

## 2.2 Mesure de la relation de comparabilité

En pratique, il n'est pas toujours simple de mesurer directement ces propriétés. La principale raison vient du fait qu'il n'est pas toujours aisé d'appréhender précisément un phénomène linguistique discriminant pour une tâche. En effet, un phénomène peut être porté par une ou plusieurs expressions textuelles distinctes. Celles-ci ne sont pas toujours simples à délimiter et peuvent participer à l'expression de plusieurs phénomènes dans un texte. C'est pour cette raison que nous proposons de comparer des corpus à travers la comparaison des résultats qu'obtiennent des systèmes entraînés respectivement sur chacun d'eux. Les systèmes définissent, de fait, un type d'information en entrée qui leur permettent d'observer les phénomènes opportuns.

La comparaison de systèmes est possible à travers un *test statistique de significativité* qui va permettre de mesurer la significativité statistique des différences entre deux ensembles de scores. Ce type de test retourne une probabilité *pvalue* que l'on discute par rapport à des seuils pré-établis. Suivant les contextes applicatifs et les communautés scientifiques, les seuils  $\alpha$  considérés sont 0,01, 0,05 et 0,1.

Une valeur de probabilité s'interprète ainsi comme suit. Si elle est inférieure à un seuil de 0,05 par exemple, alors on peut affirmer avec moins de 5% de risques de se tromper (c'est-à-dire avec un niveau de confiance de 95%) que les scores sont significativement différents. De la même manière, si ils sont supérieurs à un seuil de 0,05, alors on peut affirmer avec plus de 5% de chances de ne pas se tromper que les scores ne sont pas significativement différents.

Pour obtenir les ensembles de scores, on découpe notre corpus de référence en  $n$  partitions et on évalue les systèmes sur chacune de ces partitions pour obtenir un score par partition et par système.

Ainsi si les corpus  $\tilde{R}$  et  $S$  sont comparables du point de vue du traitement  $t$  alors les modélisations  $m(o(F_t, \tilde{R}))$  et  $m(o(F_t, S))$ , mises en oeuvre dans le traitement  $t$  pour analyser les  $n$  partitions  $Q_i$  d'un corpus tiers, donneront des résultats  $res_i$  ne présentant pas de différences statistiquement significatives.

$$pvalue(res_i(t(m(o(F_t, \tilde{R})), Q_i)), res_i(t(m(o(F_t, S)), Q_i))) > \alpha \quad (4)$$

## 2.3 Procédure de construction d'un corpus comparable

On souhaite construire un corpus  $\tilde{R}$  à partir d'extraits d'un corpus  $R$  qui permette d'entraîner un système pour un traitement  $t$  dont les performances sont comparables à celles obtenues avec un système entraîné sur un corpus source  $S$ .

Donné un corpus relais très grand, une possibilité est d'explorer toutes les combinaisons d'unités textuelles (e.g. les phrases) possibles jusqu'à en trouver une qui vérifie notre souhait. Cette approche présente de nombreux inconvénients. D'abord elle est coûteuse en temps (du fait du nombre de corpus candidat à tester et du temps de construction des modélisations), ensuite la procédure ne permet d'identifier le meilleur des candidats qu'après les avoir tous évalués.

Il est nécessaire d'opter pour un moyen simple capable de sélectionner les unités de  $R$  en fonction de  $S$  mais aussi selon une certaine sensibilité à la tâche  $t$ .

On définit le problème de construction d'un corpus comparable comme un *problème d'ordonnement* des unités textuelles et de *recherche du nombre minimal* d'unités à atteindre pour constituer un ensemble qui satisfait aux mieux les propriétés énoncées à la section 2.1.

Dans un premier temps, on construit à partir du corpus source une modélisation pour la tâche  $t$  qui permet d'obtenir des scores sur les différentes partitions du corpus de référence.

Dans un second temps, on débute un mécanisme itératif qui vise la construction incrémentale du corpus comparable à l'aide d'un mécanisme de sélection prédéfini des unités textuelles du corpus relais. Chaque itération donne lieu à un corpus comparable candidat. Pour chacun d'eux, on construit une modélisation que l'on évalue sur les partitions du corpus de référence. On compare ensuite chaque modélisation avec celle construite sur le corpus source. On reproduit la procédure jusqu'à ce que l'on constate ne pas observer de différences statistiquement significatives entre les modélisations ou bien qu'il n'y ait plus d'unités à sélectionner.

Le procédé de sélection des unités textuelles du corpus relais est tel qu'à terme, il maximise la présence des phénomènes relatifs à la tâche  $t$  dans les proportions telles que celles observées dans le corpus  $S$ . En d'autres termes, ce procédé affecte un score à chaque unité qui permet de les prioriser entre elles.

Si le corpus relais n'est pas annoté, celui-ci peut l'être automatiquement en appliquant le système construit à partir du corpus source.

### 3 Construction d'un corpus comparable annoté morpho-syntaxiquement

Un *modèle de langue probabiliste* constitue une représentation d'un corpus pour laquelle la théorie de l'information offre des moyens de comparaison peu coûteux. Entre autres, elle permet d'évaluer des modèles entre eux ou bien d'évaluer un modèle sur sa capacité à «reconnaître» un texte n'ayant pas participé à sa construction.

Un modèle de langue est une fonction probabiliste  $p$  qui informe sur la probabilité d'occurrence d'une séquence de mots ( $p(W) = p(w_1, w_2, \dots, w_n)$ ) ou sur la probabilité de sortie d'un mot pour un historique de mots donné<sup>4</sup> ( $p(w_n | w_1, w_2, \dots, w_{n-1})$ ).

Nous posons l'hypothèse qu'un modèle de langue préserve les caractéristiques discriminantes pour l'entraînement d'un étiqueteur morpho-syntaxique et que, par conséquent, si des modèles de langues estimés sur des corpus différents sont de qualité comparables alors les modélisations d'étiquetage morpho-syntaxique construites de façon similaire sur ces différents corpus sont de performances équivalentes.

Un moyen pour évaluer la qualité d'un modèle de langue est de mesurer la capacité du modèle à prédire un texte inconnu. Pour ce faire on peut utiliser la mesure de l'*entropie croisée* ou celle de la *perplexité* (mesure qui découle de la première et qui est communément employé en reconnaissance de la parole).

#### 3.1 La perplexité comme critère d'ordonnement

L'entropie croisée correspond au nombre moyen de bits requis pour encoder chacun des mots du texte inconnu.

La perplexité peut être interprétée comme la capacité d'un modèle de langue à prédire le prochain mot donné son historique (les mots qui précèdent) dans un texte inconnu (n'ayant pas servi pour la construction du modèle de langue évalué). Quand la distribution est uniforme, elle peut être interprétée comme un degré de ramification et indiquer le nombre de choix possibles pour le prochain mot. On peut aussi voir ce nombre de choix comme un degré de surprise.

Si l'on pose  $q$  comme étant la distribution empirique observée sur un texte inconnu (i.e.  $q(w_i) = n_i/N$  pour le  $i^e$  mot qui apparaît  $n_i$  fois dans le texte inconnu de taille  $N$ ), alors on peut définir l'entropie croisée  $H(q, p)$  par la formule (5).

$$H(q, p) = - \sum_j q(w_j) \log_2 p(w_j) \approx - \frac{1}{N} \log_2 p(w_j) \quad (5)$$

avec  $j$  représentant le  $j^e$  mot dans les données de test  
et avec  $N$  le nombre de mots dans les données de test

La perplexité est alors définie par la formule (6).

$$PP(W) = 2^{H(W)} \quad (6)$$

Une faible valeur d'entropie croisée pour des données inconnues indique que la distribution observée sur les données inconnues est *proche* de celle observée sur le modèle de langue. De la même manière, une valeur élevée de perplexité indique une mauvaise correspondance entre les données ayant servi à construire le modèle de langue et celles testées.

Pour des facilités d'interprétation nous choisissons d'utiliser la perplexité.

---

4. les mots qui le précèdent

### 3.2 Construction d'un corpus comparable pour la tâche d'étiquetage morpho-syntaxique

Nous instancions la procédure décrite à la section 2.3 comme suit.

Dans un premier temps, nous entraînons un étiqueteur à partir d'un corpus source, dont l'annotation a été validée manuellement.

Ne possédant pas de corpus tiers de référence pour permettre l'évaluation des modélisations que nous construisons, nous utilisons le corpus source, comme corpus de référence en évaluant l'étiqueteur par validation croisée sur 10 strates. Par la suite, nous utilisons aussi le corpus source pour évaluer les étiqueteurs dont les modélisations sont construites à partir d'extraits des différents corpus relais.

Nous utilisons le système construit sur le corpus source pour annoter automatiquement le corpus relais.

A partir de là, nous initions le processus de sélection de énoncés du corpus relais que nous poursuivons jusqu'à l'obtention du corpus comparable désiré ou l'appauvrissement total du corpus relais.

Pour ce faire, nous estimons un modèle de langue sur le corpus source et nous calculons la perplexité du modèle sur chaque phrase du corpus relais. Les scores obtenus nous permettent d'ordonner les phrases entre elles selon une perplexité croissante.

La phase de construction effective des corpus comparables candidats revient à sélectionner et à ajouter incrémentalement les  $n$  premières phrases du corpus relais non encore sélectionnées.

A partir de chaque paquet de phrases construit, nous entraînons un étiqueteur que nous évaluons sur le corpus de référence (ici le corpus source) et dont nous comparons les résultats avec le système entraîné directement sur le corpus source.

Dans les expériences que nous rapportons ci-après nous ne stoppons pas la sélection de nouveaux énoncés quand nous estimons les systèmes comparés non différents. Nous rapportons des observations à différents étapes de la procédure itérative afin d'observer plus finement les courbes de comparabilité entre les corpus.

Pour évaluer les étiqueteurs, nous utilisons les mesures classiques d'évaluation pour la tâche d'étiquetage morpho-syntaxique à savoir : la précision sur les mots (nombre de mots correctement étiquetés sur le nombre de mots total), la précision sur les phrases (nombre de phrases dans lesquelles tous les mots ont été correctement étiquetés par rapport au nombre de phrases total) et la précision sur les mots inconnus (nombre de mots inconnus correctement étiquetés sur le nombre de mots inconnus total (calculée à partir des mots n'apparaissant pas dans l'ensemble d'entraînement)). En pratique, nous ne rapportons dans cet article que la précision sur les mots.

Pour comparer les systèmes entre eux (ceux construits à partir d'extraits d'un corpus relais et celui construit sur le corpus source), nous utilisons un test de Student comme test de significativité. Ce test permet de mesurer si il y a une différence statistiquement significative entre les scores de précision obtenus par les différents systèmes. En pratique, chaque système donne un score de précision sur les partitions du corpus source ayant servi à évaluer les performances d'un système construit sur ce corpus par validation croisée. Chaque système est alors décrit par les scores obtenus pour chaque partition. Le test de Student compare alors les ensembles de scores décrivant deux systèmes pour déterminer si ils sont différents l'un de l'autre de manière significative.

## 4 Cadre expérimental

Dans cette section, nous présentons brièvement le corpus source et les corpus relais utilisé. Comme énoncé précédemment, le corpus source nous sert à la fois pour annoter les corpus relais et pour évaluer les systèmes construits à partir des corpus relais. Nous précisons aussi les implémentations utilisées pour segmenter les corpus en mots, les étiqueter morpho-syntaxiquement et manipuler les modèles de langue. Enfin, nous donnons quelques caractéristiques quantitatives en termes de taille du vocabulaire, nombre de mots et de phrases pour décrire les corpus.

## 4.1 Données

**Le corpus arboré de Paris 7** (P7T) (Abeillé et al., 2003; Abeillé & Barrier, 2004), alias le *French Treebank*<sup>5</sup>, se compose d'articles journalistiques issus du journal Le Monde couvrant la période 1989 à 1993. Ce corpus offre une analyse multi-niveaux (lexicale, morphologique et syntaxique) Sa licence propriétaire autorise une exploitation sous condition à des fins de recherche.

Ce corpus nous sert de corpus source.

**Wikinews** figure parmi les projets de la Wikimedia Foundation<sup>6</sup>. Il s'agit d'un recueil de dépêches et de reportages d'actualité écrit par ses utilisateurs. La version francophone de Janvier 2013 compte plus de 28 000 articles d'actualité et couvre une période s'étalant de Janvier 2005 à nos jours. Les textes sont exploitables sous licence<sup>7</sup> Creative Commons Attribution 2.5 (CC-BY 2.5) (les textes antérieures à Septembre 2005 sont dans le domaine public) qui permet à l'utilisateur d'utiliser, de modifier et de diffuser la ressource et ses modifications comme il le souhaite moyennant l'obligation d'en citer l'auteur.

**L'Est républicain** est un corpus de type journalistique mis à disposition par le CNRTL<sup>8</sup>, et composé d'articles du quotidien régional du même nom. La version 0.3 met à disposition les éditions intégrales des années 1999, 2002 et 2003 exploitable sous license Creative Commons (CC BY-NC-SA 2.0 FR).

**Europarl** Ce corpus est constitué de textes multilingues alignés issus des actes du Parlement Européen<sup>9</sup> préparés par (Koehn, 2005) pour l'entraînement de systèmes statistiques de traduction. La section en français de la version 7 (mai 2012) couvre une période s'étalant de 1996 à 2011. Ces textes sont libres de reproduction<sup>10</sup>.

Les corpus Wikinews, Est Républicain et Europarl nous servent de corpus relais. Les corpus Wikinews et Est Républicain présentent la particularité d'être du genre journalistique à l'instar du corpus source (extraits du Monde).

## 4.2 Segmentation des mots

Afin de comparer des étiqueteurs, il est important que ceux-ci aient été entraînés sur des textes segmentés en mots de la même manière. La question de segmentation se pose surtout sur le traitement des mots composés.

Le corpus P7T, qui constitue notre corpus source, fait reposer ses annotations sur une segmentation en mots composés. Afin de permettre à un système automatique de reproduire au plus près la segmentation du P7T nous avons réalisés un certain nombre d'adaptations (Hernandez & Boudin, 2013) pour ne considérer comme «mots» que les unités graphiques ne contenant pas d'espace. La segmentation de Wikinews et d'Europarl reposent sur l'utilisation d'un même segmenteur<sup>11</sup> qui met en oeuvre ce principe. Nous l'avons étendu pour traiter l'Est Républicain qui contient davantage d'entités spécifiques.

La table 1 rapporte la taille du vocabulaire, le nombre de mots et le nombre de phrases pour chacun des corpus. La table 2, quant à elle, rapporte le taux de recoupement entre les différents vocabulaires des différents corpus ( $\frac{|V_1 \cap V_2|}{|V_1 \cup V_2|}$ ).

## 4.3 Jeu d'étiquettes morpho-syntaxiques et étiqueteur

Le jeu de catégories morpho-syntaxiques que nous utilisons est celui mis au point par (Crabbé & Candito, 2008), contenant 28 catégories qui combinent différentes valeurs de traits morpho- syntaxiques du P7T. Outre le fait que ce jeu soit plus

5. <http://www.llf.cnrs.fr/Gens/Abeille/French-Treebank-fr.php>

6. <http://wikimediafoundation.org>

7. <http://dumps.wikimedia.org/legal.html>

8. <http://www.cnrtl.fr/corpus/estrepubicain>

9. <http://www.statmt.org/europarl>

10. «Except where otherwise indicated, reproduction is authorised, provided that the source is acknowledged.» [http://www.europarl.europa.eu/guide/publisher/default\\_en.htm](http://www.europarl.europa.eu/guide/publisher/default_en.htm)

11. <https://github.com/boudinfl/kea>

	<b>P7T</b>	<b>Wikinews</b>	<b>Est Républicain</b>	<b>Europarl</b>
Taille du vocabulaire	34 677	75 175	382 342	129 093
Nombre de mots	629 788	2 535 396	36 478 209	61 396 216
Nombre de phrases	23 539	87 461	1 947 360	1 967 951

TABLE 1 – Taille du vocabulaire, nombre de mots et de phrases selon les corpus P7T, Wikinews, Est Républicain et Europarl.

	<b>Wikinews (75 175)</b>	<b>Est Républicain (382 342)</b>	<b>Europarl (129 093)</b>
<b>P7T (34 677)</b>	.27 (23 093 / 86 749)	.08 (29 831 / 387 188)	.20 (27 329 / 136 433)
<b>Wikinews</b>		.13 (52 068 / 405 439)	.30 (46 628 / 157 622)
<b>Est Républicain</b>			.17 (75 213 / 436 213)

TABLE 2 – Taux de vocabulaires en commun entre les corpus P7T, Wikinews, Est Républicain et Europarl.

complet que les catégories du P7T, qui elles sont au nombre de 13, les auteurs montrent que les performances d'un étiqueteur entraîné sur de telles annotations sont meilleures.

En ce qui concerne l'étiqueteur morpho-syntaxique que nous avons utilisé pour nos expériences, il s'agit de la version 3.1.3 du *Stanford POS Tagger* (Toutanova et al., 2003). Ce système utilise un modèle par maximum d'entropie, et peut atteindre des performances au niveau de l'état-de-l'art en français (Boudin & Hernandez, 2012; Hernandez & Boudin, 2013). Nous utilisons un ensemble standard 16 de traits bidirectionnels sur les mots et les étiquettes.

#### 4.4 Estimation de modèles de langue et calcul de la perplexité

Les modèles de langue que nous estimons sont d'ordre 5 (historique de 4 mots) et sont construits en utilisant la technique de lissage communément utilisée de (Kneser & Ney, 1995). En pratique, nous utilisons la bibliothèque *berkeleylm*<sup>12</sup> (Pauls & Klein, 2011).

Pour calculer la perplexité nous utilisons la bibliothèque *kylm*<sup>13</sup>.

## 5 Expériences

Dans les sections qui suivent nous rapportons les résultats de systèmes d'étiquetage entraînés à partir de différents corpus relais. Nous les évaluons sur les données du P7T+.

### 5.1 Performance d'un étiqueteur état-de-l'art

A titre de comparaison, la table 3 rapporte les résultats d'un système état de l'art (à savoir le *Stanford POS Tagger*) évalué par validation croisée en 10 strates sur le P7T+. Ces résultats peuvent être interprétés comme la performance maximale que peut obtenir un système lorsqu'il est entraîné sur des données qui ont été manuellement validées. La précision moyenne est de 96,93% sur les tokens. Nous renvoyons à FIXME pour évaluation plus exhaustives en terme de précision sur les phrases et en tenant compte des mots inconnus.

### 5.2 Score de précision et test de significativité selon les corpus relais

Dans cette section nous rapportons pour différents corpus les scores de précision et de significativité (via le test de Student aussi appelé t-test) obtenus à partir de modélisations fondées sur un nombre croissant de phrases ordonnées

12. <https://code.google.com/p/berkeleylm>

13. <http://www.phontron.com/kylm/>



	Précision	Min. - Max.	Écart type
Tokens	96,93	96,55 - 97,28	0,219

TABLE 3 – Scores de précision sur les tokens du *Stanford POS tagger* calculés à partir du P7T+ en validation croisée en 10 strates. Le minimum, le maximum et l'écart type des scores calculés sur les 10 strates sont également reportés.

soit aléatoirement<sup>14</sup> soit sur le score de perplexité vis-à-vis du corpus source.

Les figures 1, 2 et 3 rapportent ces scores respectivement pour les corpus relais Wikinews, Est Républicain et Europarl.

En abscisse de chaque figure, on note le nombre de phrases considéré pour la construction d'une modélisation. En ordonnée sur la gauche se trouve une échelle de scores du test de Student variants de 0 à 1 et en ordonnée sur la droite se trouve une échelle de scores de précision centrée sur les scores les plus élevés variants de 92% à 98% environ. Cette double échelle en ordonnée permet de confronter ces deux types de scores. Les courbes en pointillées (et bleues) représentent les scores de précision tandis que les courbes en ligne pleine (et rouges) correspondent aux scores du test de Student.

Il y a pour chaque figure deux courbes en pointillée et pleines. Celles marquées d'un point vide concernent les scores du test de Student et de précision obtenus à partir de modélisations fondées sur un nombre croissant de phrases ordonnées aléatoirement (ordre original). Celles marquées d'un point plein concernent les scores du test de Student et de précision obtenus à partir de modélisations fondées sur un nombre croissant de phrases ordonnées sur le score de perplexité vis-à-vis du corpus source.

Dans la table 4, nous rapportons différents facteurs de nombres de phrases calculés entre différents corpus. Plus exactement nous comparons les corpus obtenus à la valeur maximale du test de Student (pour notre échantillonnage) avec les corpus relais dont ils sont issus et le corpus source. Nous comparons aussi la différence du nombre de phrases entre les corpus triés et non triés.

Sur les trois figures 1, 2 et 3, on peut voir que la précision de tous les systèmes augmentent avec la quantité de phrase. Les élévations observées sur les courbes du test de Student correspondent à des scores où le test considère qu'il n'y a pas de différences statistiquement significatives entre les systèmes entraînés sur les corpus relais et celui sur un corpus source.

Il est à noter que les sommets de ces courbes sont sur la base de notre échantillonnage. Ils ne correspondent pas aux maximums absolus (probablement voisins).

Concernant le corpus relais Wikinews (figure 1), nous observons que le score de précision est meilleur pour le corpus relais sans ordre a priori que sur le corpus relais ordonné sur le score de la perplexité. La différence entre ces scores s'amenuisent à mesure que l'on considère plus de phrases dans les corpus et finit par se confondre à partir d'une certaine quantité de phrases. Concernant la significativité des résultats, sur la base de notre échantillonnage, nous ne pouvons différencier les deux corpus candidats au maximum du test de Student. La courbe traduit une légère préférence pour le corpus ordonné selon la perplexité mais la différence est minime.

Concernant le corpus relais Est Républicain (figure 2), les scores sont tout autre. Concernant la précision des systèmes, malgré une convergence qui se précise avec le nombre de phrases considéré, on observe qu'un système, construit en sélectionnant ses instances d'entraînement selon un score de perplexité avec le corpus source, obtient une précision plus haute qu'avec un système entraînés avec des phrases sans ordre a priori. Concernant la significativité, on constate qu'un système qui se fonde sur un ordre des phrases découlant d'un score de perplexité croissant requiert moins de phrases qu'un système qui ne se fonde pas sur un ordre a priori des phrases (1,85 fois la taille du corpus source pour le corpus trié contre 3,48 pour celui non trié). L'une des différences fondamentales avec le corpus relais Wikinews est que le corpus relais Est Républicain est beaucoup plus grand.

Cette expérience montre que le filtrage des phrases selon un score de perplexité qu'un corpus source peut avoir sur elles constitue un critère effectif de sélection.

Les scores observés à la figure 3 (corpus relais Europarl) corroborent globalement ceux observés à la figure 2. On note qu'il faut des quantités de données beaucoup plus importantes pour observer des résultats. On attribue cette caractéristique au fait que ce corpus relais n'est pas du même genre que les corpus source et relais Wikinews et Est Républicain. Dans

14. L'ordre original n'étant pas trié est considéré comme étant sans a priori.



cette expérience, on constate encore qu’avec des phrases sélectionnées selon un score de perplexité observé par un corpus source, une quantité moins importante est requise avec celles-ci pour obtenir plus rapidement des scores de précision avec une différence statistiquement non-significatives. De même la précision sur mots est globalement toujours plus élevée.

Ordre des phrases dans corpus relais	Wikinews		Est Républicain		Europarl		Fusion
	aléa	trié	aléa	trié	aléa	trié	
# de phrases <i>au maximum du t-test</i>	76 524	76 524	82 000	43 730	349 844	262 383	43 730
Facteur <i>diviseur</i> / corpus relais	1.14	1.14	23.75	44.53	7.50	5.62	91.53
Facteur <i>multiplicateur</i> / corpus source	3.25	3.25	3.48	1.85	14.86	11.14	1.85
$\neq$ entre # de phrases aléa et trié	1		1.87		1.33		

TABLE 4 – Facteurs multiplicateur et diviseur calculés sur le nombre de phrases respectivement par rapport au corpus source et aux corpus relais, pour la valeur maximale du test de Student. Sont comparés les facteurs lorsque les phrases sont ordonnées aléatoirement ou selon la perplexité. Est donné le rapport entre le nombre de phrases non triées et triées pour le maximal du test de Student. Est aussi indiqué le nombre de phrases de l’échantillon pour lequel le score maximal du test de Student a été atteint.

### 5.3 Significativité obtenu à partir d’un corpus relais fusionné et ordonné

La figure 4 rapporte les scores de significativité obtenus à partir de systèmes construits sur un corpus relais résultant de la fusion des corpus relais Wikinews, Est Républicain et Europarl. Les courbes décrivant les scores du test de Student obtenues sur les corpus relais ordonnés Wikinews et Est Républicain sont rappelés. Celle de Europarl ne sont pas visibles à cette échelle.

Nous notons que le système construit sur le corpus fusionné requiert globalement le même nombre de phrases que celui construit à partir du meilleur corpus composant le corpus fusionné. Nous notons néanmoins que même si les deux courbes sont proches (courbe noire avec point rempli et courbe rouge avec point carré rempli), le système construit seulement à partir de l’Est Républicain obtient des scores plus importants rapidement.

Nous attribuons cela à une limitation de notre approche pour la construction des corpus comparables. En effet en l’état, la sélection des phrases à considérer se fait seulement sur la base du corpus source et de la phrase analysée. Elle ne prend pas en compte les phrases déjà sélectionnées et celles potentiellement sélectionnables.

Ainsi, pour une même quantité de phrases, nous pensons que dans le corpus résultant de la fusion il y a plus de phrases avec une perplexité faible vis-à-vis du corpus source mais globalement probablement moins de phrases qui couvrent l’étendu des phénomènes du corpus source par rapport au corpus construit à partir de l’Est Républicain.

Pour vérifier cette hypothèse nous calculons la perplexité d’un modèle de langue construit à partir du corpus source testé sur le corpus résultant de la fusion et sur celui de l’Est Républicain au plus haut du score du test de Student. Nous calculons aussi la perplexité sur le corpus source à partir de modèles de langues construits sur chacun de ces corpus comparables candidats. La table 5 rapporte ces résultats. On peut lire que le modèle de langue construit sur le corpus source P7T «reconnaît» plus facilement le corpus extrait de la fusion que celui issu de l’Est Républicain (perplexité de 267 contre une perplexité de 1192). On peut lire aussi que le modèle de langue construit à partir du corpus extrait de l’Est Républicain «reconnaît» mieux le corpus source P7T que le corpus issu de la fusion (perplexité de 308 contre une perplexité de 417). Ces observations confirment notre hypothèse de différences de représentativité entre ces corpus.

Perplexité ↗	Fusion	Est Républicain	P7T
<b>P7T</b>	267	1192	
<b>Fusion</b>			417
<b>Est Républicain</b>			308

TABLE 5 – Scores de perplexité réciproque entre le corpus source P7T et les corpus comparables candidats extraits des corpus relais ordonné fusion et Est Républicain au score du test de Student le plus haut.

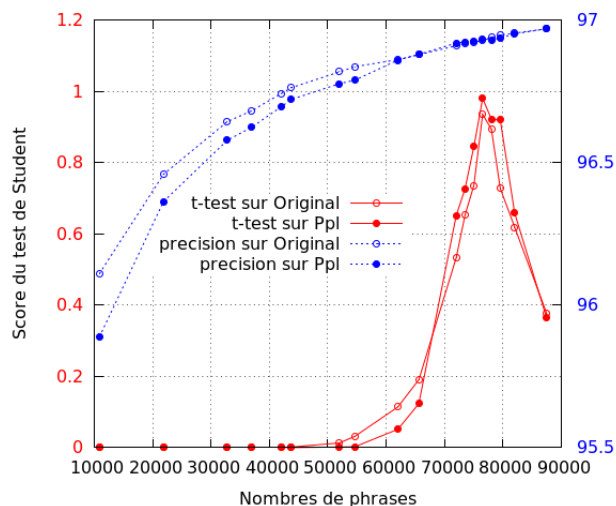


FIGURE 1 – Corpus Wikinews : Scores de précision et de t-test obtenus à partir de modélisations fondées sur un nombre croissant de phrases ordonnées soit aléatoirement (ordre original) soit sur le score de perplexité vis-à-vis du corpus source.

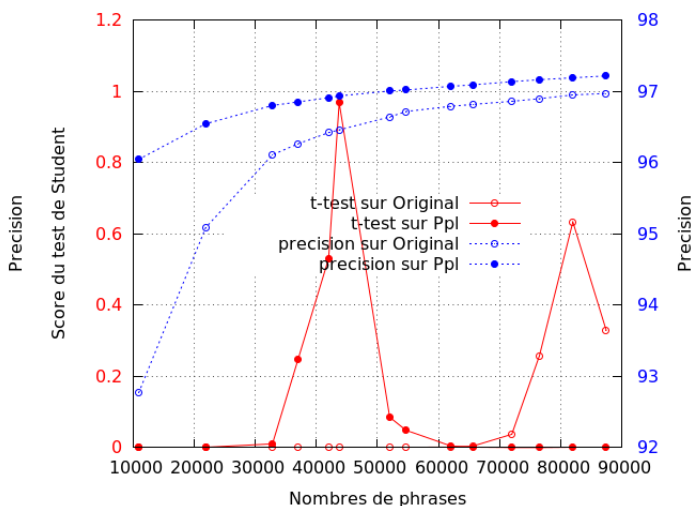


FIGURE 2 – Corpus Est Républicain : Scores de précision et de t-test obtenus à partir de modélisations fondées sur un nombre croissant de phrases ordonnées soit aléatoirement (ordre original) soit sur le score de perplexité vis-à-vis du corpus source.

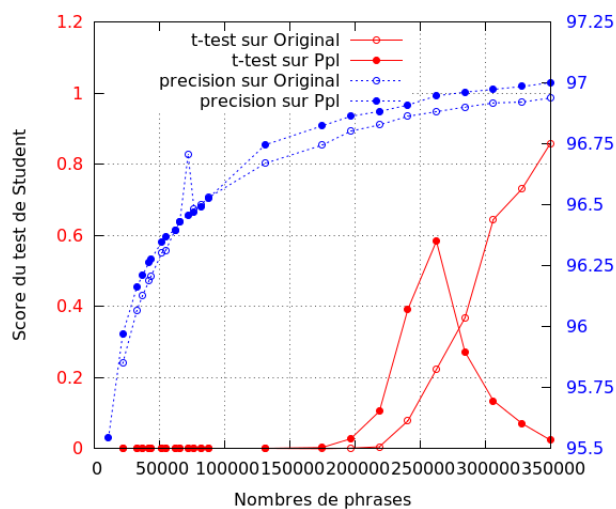


FIGURE 3 – Corpus Europarl : Scores de précision et de t-test obtenus à partir de modélisations fondées sur un nombre croissant de phrases ordonnées soit aléatoirement (ordre original) soit sur le score de perplexité vis-à-vis du corpus source.

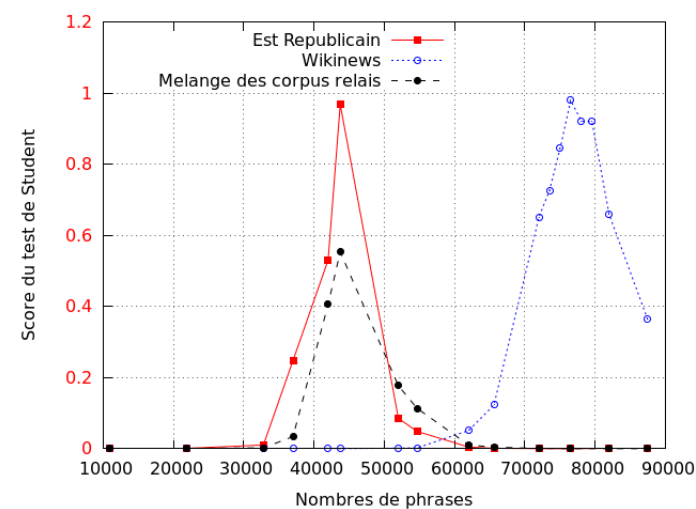


FIGURE 4 – Corpus mélangés : Scores de t-test obtenus à partir de modélisations fondées sur un nombre croissant de phrases ordonnées sur le score de perplexité vis-à-vis du corpus source. Sont rappelés les scores des corpus Wikinews et Est Républicain.

## 6 État de l’art

L’anonymisation est un moyen de transformation d’un corpus source qui dénature principalement les entités nommées et qui malgré cette perte d’information peut permettre la diffusion du dérivé et son exploitation dans des tâches autres que de la reconnaissance d’entités nommées (Medlock, 2006).

Nous nous situons dans un contexte où nous ne pouvons réutiliser tout ou partie d’un corpus source.

(McEnery & Xiao, 2007) définit la notion de corpus comparable dans le contexte d’études comparatives inter-lingues. Selon les auteurs, un corpus comparable se définit comme un corpus dont les composantes ont été collectées en utilisant la même base d’échantillonnage, le même équilibre et la même représentativité par exemple la même proportion de textes de même genres dans les mêmes domaines collectés à la même période dans différentes langues.

Nous partageons avec les auteurs ces principes de représentativité de phénomènes dans les mêmes proportions entre des corpus comparables. Nous pensons néanmoins que la contrainte sur les langues peut être levée pour offrir une définition plus large.

L’approche classique que l’on retrouve pour construire des corpus comparables monolingues consiste à utiliser des moteurs de recherche et à utiliser les documents du corpus source comme requêtes (Wang & Callison-Burch, 2011; Bo et al., 2011). Le supposé théorique sous-jacent est qu’un corpus comparable avec un autre partage un contenu informationnelle (thématique) en commun. Nous pensons qu’il s’agit d’une restriction limitative de la définition de corpus comparable.

Dans le contexte de la tâche d’extraction terminologique bilingue, (Déjean & Éric Gaussier, 2002; Bo et al., 2011) définissent une mesure de comparabilité fondée sur le lexique partagé par les corpus comparés. A nouveau, cette définition illustre une importance prépondérante à l’information lexicale comme base à la comparabilité. Elle est néanmoins pertinente dans le cadre de cette tâche.

Comme ces auteurs nous pensons qu’une définition de la notion de comparabilité est en lien avec une application. Par notre approche, nous souhaitons néanmoins ne pas nous attacher à l’observation privilégiée de certains traits linguistiques. Notre définition se veut plus générale et non spécifique à une application particulière.

## 7 Discussion et perspectives

Dans cet article nous avons proposé une définition générale et opérationnelle de la relation de la comparabilité entre des corpus monolingues annotés. Nous avons entre autres fourni une mesure de la relation de comparabilité ancrée dans un contexte applicatif mais indépendante d’un domaine en particulier, à savoir la comparaison en termes de test statistique de significativité. Nous avons aussi énoncé une procédure de construction d’un corpus comparable.

Nos expérimentations de ces propositions se sont réalisées autour de la construction d’un corpus comparable annoté morpho-syntaxiquement. Nous avons montré notamment que la mesure de la perplexité définie dans la théorie de l’information constitue un moyen de prioriser les phrases à sélectionner pour construire un corpus comparable. Nous montrons néanmoins que la procédure pour l’exploiter a son importance et qu’elle doit tenir compte du corpus source et du corpus relais dans sa globalité.

Au sujet de la propriété 3, nous formulons la relation de comparabilité comme n’étant pas nécessairement une relation symétrique. En effet, un corpus reconnu comparable à un autre peut intégrer des phénomènes relatifs à une tâche non observés dans un corpus source. Le seul problème qu’il pourrait y avoir à cela est le fait de ne pas retrouver ces phénomènes dans le corpus de référence. Il en découle une difficulté à évaluer et à interpréter leurs rôles dans les résultats du système et dans les annotations produites.

Au sujet du processus de construction du corpus comparable, l’idée est de favoriser la sélection de phrases d’un corpus relais qu’un modèle de langue, construit sur un corpus source, prédit avec le moins de surprise. Notre approche de la sélection des phrases est sans supervision macroscopique. Les résultats de l’expérience décrite à la section 5.3 tendent à montrer cette faiblesse dans notre approche. En effet, rien ne garantit qu’un tel processus de sélection ne conduise pas à des problèmes de redondance et de représentativité faussée par rapport à la réalité. L’utilisation du critère de pertinence marginale maximale (*Maximal Marginal Relevance*) est une piste possible au filtrage de la redondance tout en maintenant une représentativité des phénomènes (Carbonell & Goldstein, 1998).

A court terme, nous envisageons de reprendre la comparaison des corpus construits et d’observer leur différences par

rapport aux corpus source et relais en volume de mots et non en volume de phrases comme dans les expériences de la section 5.2. Les comparaisons seront ainsi plus précises.

A moyen terme, nous souhaitons comparer différentes mesures d'ordonnement des phrases à sélectionner notamment pour chercher à mieux couvrir les propriétés des corpus comparables que nous avons définies.

Enfin, comme le laisse entrevoir l'expérience de la section 5.3, nous aimerions étudier la possibilité de construire une mesure de comparabilité fondée sur le rapport de scores de perplexités calculés réciproquement à partir de deux corpus distincts.

## Références

- ABEILLÉ A. & BARRIER N. (2004). Enriching a french treebank. In Actes de la conférence LREC, Lisbonne.
- ABEILLÉ A., CLÉMENT L. & TOUSSENEL F. (2003). Building and using Parsed Corpora, chapter Building a treebank for French. Language and Speech series : Kluwer, Dordrecht.
- BO L., GAUSSIER E., AIZAWA A. N. ET AL. (2011). Clustering comparable corpora for bilingual lexicon extraction. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, p. 473–478.
- BOUDIN F. & HERNANDEZ N. (2012). Détection et correction automatique d'erreurs d'annotation morpho-syntaxique du french treebank. In Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 2 : TALN, p. 281–291, Grenoble, France : ATALA/AFCP.
- CARBONELL J. & GOLDSTEIN J. (1998). The use of mmr, diversity-based reranking for reordering documents and producing summaries. In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '98, p. 335–336, New York, NY, USA : ACM.
- CRABBÉ B. & CANDITO M. (2008). Expériences d'analyse syntaxique statistique du français. In Actes de la 15ème conférence sur le Traitement Automatique des Langues Naturelles (TALN), Avignon, France.
- DÉJEAN H. & ÉRIC GAUSSIER (2002). Une nouvelle approche à l'extraction de lexiques bilingues à partir de corpus comparables. Lexicometrica, Numéro thématique "Alignement lexical dans les corpus multilingues", p. 1–22.
- HERNANDEZ N. & BOUDIN F. (2013). Construction d'un large corpus écrit libre annoté morpho-syntaxiquement en français. In Actes de la 20e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2013), p. 160–173, Les Sables d'Olonne, France.
- KNESER R. & NEY H. (1995). Improved backing-off for m-gram language modeling. In International Conference on Acoustics, Speech, and Signal Processing.
- KOEHN P. (2005). Europarl : A parallel corpus for statistical machine translation. In MT Summit.
- MCENERY A. M. & XIAO R. Z. (2007). Parallel and comparable corpora : What are they up to ?, In G. JAMES & G. ANDERMAN, Eds., Incorporating Corpora : Translation and the Linguist. Translating Europe. Multilingual Matters.
- MEDLOCK B. (2006). An introduction to nlp-based textual anonymisation. In Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC-2006), Genoa, Italy : European Language Resources Association (ELRA). ACL Anthology Identifier : L06-1110.
- NIELSEN M. (2011). Reinventing Discovery : The New Era of Networked Science. Princeton, N.J. Princeton University Press.
- PAULS A. & KLEIN D. (2011). Faster and smaller n-gram language models. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies - Volume 1, HLT '11, p. 258–267, Stroudsburg, PA, USA : Association for Computational Linguistics.
- TOUTANOVA K., KLEIN D., MANNING C. & SINGER Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In Proceedings of the 3rd Conference of the North American Chapter of the ACL (NAACL 2003), p. 173–180 : Association for Computational Linguistics.
- WANG R. & CALLISON-BURCH C. (2011). Paraphrase fragment extraction from monolingual comparable corpora. In Proceedings of the 4th Workshop on Building and Using Comparable Corpora : Comparable Corpora and the Web, p. 52–60, Portland, Oregon : Association for Computational Linguistics.