EXTRACTION AUTOMATIQUE D'ASSOCIATIONS TEXTUELLES À PARTIR DE CORPORA NON TRAITÉS

Gaël Dias[1,2,3], Sylvie Guilloré[1], José Gabriel Pereira Lopes[2]

[1] LIFO – UFR Sciences (Orléans) – F-45061 Orléans – France

[2] UNL – FCT/DI (Lisbonne) – P-2825-114 Caparica – Portugal

[3] UBI – DMI (Covilhã) – P-6200-001 Covilhã – Portugal

Abstract

In this article, we present a new statistically based system that automatically extracts from raw texts multiword lexical units (i.e. idioms, compounds, etc...) and morphological units (i.e. affixes, stems, etc...). On one hand, we define a new mathematical model called the Mutual Expectation that measures the degree of cohesiveness that links together all the components of an n-gram composed of basic textual units. On the other hand, we propose a new acquisition process called the LocalMaxs that elects every n-gram which corresponding association value is a local maximum. The results obtained over a trilingual *corpus* (French-Portuguese-English) transformed in a set of contiguous and non-contiguous word n-grams show that it is possible to extract precisely compound nouns, names and determinants as well as verbal, adverbial, adjectival, conjunctive and prepositional locutions. Analogously, the system retrieves from the set of all the cohesiveness-valued character n-grams, contiguous and non-contiguous bound and free morphemes.

Résumé

Nous présentons dans cet article un système novateur qui permet l'extraction automatique d'unités polylexicales (i.e. idiomes, noms composés, etc...) et morphologiques (i.e. affixes, thèmes, etc...) à partir de *corpora* non traités. Dans un premier temps, nous définissons un nouveau modèle probabiliste, l'Expectative Mutuelle, qui permet de mesurer le degré d'association qui lie entre eux tous les éléments d'un vecteur d'unités textuelles élémentaires. Dans un deuxième temps, nous proposons un nouvel algorithme d'acquisition, le LocalMaxs, qui élit tout vecteur dont le degré d'association correspond à un maximum local. Les résultats obtenus à partir d'un *corpus* trilingue (Français-Portugais-Anglais) montrent qu'il est possible d'extraire avec précision des noms et des déterminants composés ainsi que des locutions verbales, adverbiales, adjectivales, conjonctives et prépositionnelles. De la même façon, il est possible d'identifier des morphèmes libres et liés à partir de la représentation des textes en séquences de caractères.

Mots-clés : Analyse Exploratoire de Données Textuelles Multilingues, Acquisition d'Unités d'Association.

1. Introduction

L'acquisition automatique d'associations textuelles revêt une importance cruciale pour le succès de nombreuses applications dans le domaine du traitement de la langue. D'une part, les unités polylexicales (i.e. suites de mots qui se trouvent plus fréquemment associés qu'ils ne le seraient par le seul fruit du hasard) posent de nombreux problèmes lors des phases de compréhension et de production du langage. Ainsi, l'identification préalable de ces unités bénéficierait la plupart des applications en traduction automatique, extraction d'information et fouille de textes. D'autre part, l'extraction d'unités morphologiques (i.e. suites de caractères qui se trouvent plus fréquemment associés qu'ils ne le seraient par le seul fruit du hasard) telles

que les morphèmes affixaux, les thèmes et les associations entre préfixes et suffixes constitue une étape obligatoire en amont de l'analyse lexicale. Elle permet entre autres[i] de confronter les formalismes des règles de formation des mots avec la réalité des dictionnaires électroniques et des *corpora*. Nous présentons dans cet article un système qui permet l'extraction automatique d'associations textuelles à partir de *corpora* non traités. Notre approche se base dans un premier temps sur la définition d'un nouveau modèle probabiliste appelé Expectative Mutuelle (Dias et al., 1999a) qui permet de mesurer le degré d'association qui lie entre eux tous les éléments d'un vecteur d'unités textuelles élémentaires (i.e. caractères ou mots graphiques). Dans un deuxième temps, la sélection des vecteurs d'intérêt est réalisée par l'application d'un nouvel algorithme, le LocalMaxs (Silva et al., 1999), qui élit tout vecteur dont le degré d'association correspond à un maximum local. Les résultats obtenus à partir d'un *corpus* trilingue (Français-Portugais-Anglais) transformé en un ensemble de n-grams contigus et non contigus de mots graphiques montrent qu'il est possible d'extraire avec précision des noms et des déterminants composés ainsi que des locutions verbales, adverbiales, adjectivales, conjonctives et prépositionnelles. De la même façon, il est possible d'identifier, à partir de la représentation des textes en séquences contiguës et non contiguës de caractères, des morphèmes libres et liés (Hausser, 1999). Notre méthode se différencie de la majorité des travaux proposés par l'absence de pré-traitement du *corpus*. Ainsi le texte n'est ni lemmatisé ni étiqueté ni épuré au moyen de "stop-listes". Parallèlement, l'Expectative Mutuelle, basée sur une normalisation de la probabilité conditionnelle, permet l'acquisition d'associations n-aires sans recourir aux techniques d'amorçage. Finalement, la sélection des vecteurs d'intérêt repose sur les variations locales des mesures d'associations et non sur la définition de valeurs seuil globales.

2. Préparation des données textuelles

Les méthodes proposées[ii] pour l'extraction automatique d'associations textuelles se basent en grande partie sur l'étude de textes étiquetés (Daille, 1995; Justeson, 1993; Bourigault, 1996) ou bien sur des énoncés épurés à partir de "stop-listes" (Enguehard, 1993). Cependant, le traitement linguistique des textes implique l'introduction de contraintes qui ne sont pas présentes (du moins explicitement) dans leurs versions originales. Par exemple, quel ensemble d'étiquettes doit-on utiliser? quel type de lemmatisation doit-on choisir? peut-on réduire la séquence "*Nations Unies*" à sa forme non fléchie "*Nation Uni*"? est-il possible d'extraire des séquences telles que "*en matière de*", "*toujours plus*", "*tomber des nues*" à partir de patrons syntaxiques?[iii]. Devant l'absence de réponses concrètes à nombre de ces questions, nous avons choisi de n'apporter aucun traitement linguistique au *corpus*. Ainsi, le texte n'est ni lemmatisé ni étiqueté ni épuré au moyen de "stop-listes". L'objectif principal de notre travail est donc d'identifier et d'extraire un ensemble, le plus vaste possible, d'associations textuelles à partir des seules contraintes présentes dans l'énoncé original. A partir du texte non traité, deux unités textuelles élémentaires peuvent être mises en évidence: le caractère et le mot graphique[iv]. Ainsi, pour l'extraction d'unités polylexicales, nous construisons à partir du texte divisé en mots graphiques tous ses n-grams contigus et non contigus calculés pour chaque mot pivot dans une fenêtre de 5 mots à sa gauche et 5 mots à sa droite[v]. Dans ce cas, un n-gram est un vecteur de n mots indéxés positionnellement par rapport au mot pivot du vecteur. En ce qui concerne l'identification des unités morphologiques, nous construisons à partir du texte divisé

[i] Un traitement adéquat de ces séquences est aussi important pour l'étiquetage phonologique des *corpora*.
[ii] Toutes les méthodes proposées n'interviennent que dans le cadre de l'extraction d'unités polylexicales.
[iii] Les deux dernières questions ont un objectif provocateur assumé.
[iv] Un mot graphique est identifié comme étant une séquence de caractères délimitée à droite et à gauche par le caractère "espace" et excluant ce dernier.
[v] Sinclair (Sinclair, 1974) montre que les relations lexicales associent des mots séparés par 5 autres mots.

en caractères l'ensemble de tous ses n-grams contigus et non-contigus à partir d'une fenêtre de 3 caractères à gauche et 3 caractères à droite du caractère pivot[vi]. Nous présentons dans le Tableau (1) un n-gram de mots graphiques ayant comme mot pivot *Traité* et un n-gram de caractères ayant comme caractère pivot *T*, tous deux calculés à partir du texte (1).

Après multes négociations, le Traité de Maastricht a été ratifié par tous les Etats Membres ***(1)***

n-gram	**Séquence associée**
[*Traité* –1 *le* +1 *de* +2 *Maastricht* + 4 *été*]	*le Traité de Maastricht _____ été*[vii]
[*T* -1 # +1 *r* +2 *a* +3 *i*] [viii]	*#Trai*

Tableau (1): Exemples de n-grams

3. Expectative Mutuelle

Les modèles statistiques proposés dans la littérature (Salem, 1987; Church et Hanks, 1990; Gale, 1991; Smadja, 1993; Dunning, 1993; Smadja, 1996; Shimohata, 1997) ne sont définis que pour les 2-grams et ne permettent ainsi que l'acquisition d'associations binaires. Pour les associations de plus de deux unités textuelles élémentaires, l'acquisition requiert un travail complémentaire où les paires d'association acquises initialement jouent le rôle d'amorce. Parallèlement, la plupart des modèles mathématiques sont sensibles à l'occurrence d'unités textuelles élémentaires fréquentes[ix] et leur normalisation aboutit à la définition de valeurs de cohésion incohérentes. Afin de résoudre ces problèmes, nous définissons un nouveau modèle probabiliste appelé Expectative Mutuelle (Dias et al., 1999a) qui mesure le degré d'association qui lie entre eux tous les éléments d'un n-gram (i.e. $\forall n,\ n \geq 2$) et permet ainsi d'acquérir des associations n-aires sans recourir aux techniques d'amorçage. L'Expectative Mutuelle (EM) est basée sur la notion d'Expectative Normalisée (EN)[x].

3.1. L'expectative normalisée

Nous définissons l'expectative normalisée existant entre n unités textuelles (UTs) comme étant l'expectative moyenne de voir apparaître une UT dans une position donnée sachant que les autres (n-1) UTs apparaissent dans le texte contraintes par leurs positions. L'idée de base est d'évaluer le coût de la perte d'une UT dans un n-gram. Ainsi, plus une suite d'UTs est figée et témoigne d'une forte cohésion, moins cette séquence accepte la perte d'un de ses constituants et plus la valeur de l'expectative normalisée doit être élevée. Le concept sous-jascent à l'expectative normalisée est celui de la probabilité conditionnelle définie par l'équation (1).

$$p(\mathrm{X} = \mathrm{x} \mid \mathrm{Y} = \mathrm{y}) = \frac{p(\mathrm{X} = \mathrm{x}, \mathrm{Y} = \mathrm{y})}{p(\mathrm{Y} = \mathrm{y})} \qquad \textbf{(1)}$$

La probabilité conditionnelle mesure l'expectative d'apparition de l'événement X=x sachant que l'événement Y=y conditionne son apparition.

Considérons le n-gram $[p_{11}\ u_1\ p_{12}\ u_2\ p_{13}\ u_3\ \ldots\ p_{1i}\ u_i\ \ldots\ p_{1n}\ u_n]$ dans lequel p_{11} vaut zéro, u_1 identifie l'unité pivot et p_{1i} représente la distance signée entre l'unité u_i et l'unité pivot.

[vi] Le choix d'utiliser une fenêtre de 3 mots a été déterminé expérimentalement. En effet, nous ne connaissons aucune étude qui mette en évidence les distances entre caractères dans les relations morphologiques.
[vii] L'espace (i.e. "_____") correspond au saut d'un mot graphique dans le texte. Dans ce cas, l'occurrence "*a*".
[viii] Pour des raisons de représentation, le caractère "espace" est identifié par le caractère '#'.
[ix] En conséquence, Daille (Daille, 1995) et Enguehard (Enguehard, 1993) ne considèrent que les occurrences des mots pleins pour l'évaluation des forces de cohésion.
[x] Nous donnons le nom d'Expectative Normalisée (EN) à la normalisation de la probabilité conditionnelle.

L'extraction une à une des UTs de ce n-gram correspond à la définition de n événemements que nous illustrons dans le Tableau (2).

(n-1)-gram[xi]	**UT extraite**
[_____ p_{12} u_2 p_{23} u_3 ... p_{2i} u_i ... p_{2n} u_n] [xii]	p_{11} u_1
[$p_{11}u_1$ _____ p_{13} u_3 ... p_{1i} u_i ... p_{1n} u_n]	p_{12} u_2
...	...
[p_{11} u_1 p_{12} u_2 p_{13} u_3...$p_{1(i-1)}$ $u_{(i-1)}$ _____ $p_{1(i+1)}$ $u_{(i+1)}$...p_{1n} u_n]	p_{1i} u_i
...	...
[p_{11} u_1 p_{12} u_2 p_{13} u_3 ... p_{1i} u_i ... $p_{1(n-1)}$ $u_{(n-1)}$ _____]	p_{1n} u_n

Tableau 2. Extraction une à une des UTs d'un n-gram.

Or, chaque événement correspond à une probabilité conditionnelle. Par conséquent, chaque n-gram est associé à n probabilités conditionnnelles. Par définition, l'expectative normalisée mesure l'expectative moyenne incarnée par les n probabilités conditionnelles qui résultent de la décomposition d'un n-gram en n (n-1)-grams. Dans le cadre de la normalisation de la probabilité conditionnelle, nous introduisons la notion d'Unité Moyenne d'Expectative (UME) définie comme étant la moyenne arithmétique de toutes les probabilités conjointes des n (n-1)-grams contenus dans le n-gram (équation (2))[xiii] i.e. la moyenne arithmétique des dénominateurs des n probabilités conditionnelles[xiv].

$$UME([p_{11}u_1\ p_{12}\ u_2 ...\ p_{1i}\ \ u_i ...\ p_{1n}\ \ u_n]) = \frac{1}{n}\left(p([p_{12}u_2 ...\ p_{2i}\ u_i ...\ p_{2n}\ u_n]) + \sum_{i=2}^{n} p\left(\left[p_{11}u_1 ...\ \hat{p}_{1i}\ \ \hat{u}_i ...\ p_{1n}\ u_n \right]\right)\right) \quad \textbf{(2)}$$

Ainsi, l'expectative normalisée d'un n-gram est introduite comme étant une probabilité conditionnelle "juste" qui utilise le concept d'UME et est définie par l'équation (3).

$$EN([p_{11}u_1 ...p_{1i}\ u_i ...\ p_{1n}\ u_n]) = \frac{p([p_{11}u_1 ...p_{1i}\ u_i ...\ p_{1n}\ u_n])}{UME([p_{11}u_1 ...p_{1i}\ u_i ...\ p_{1n}\ u_n])} \quad \textbf{(3)}$$

3.2. L'expectative mutuelle

L'un des critères les plus importants pour l'identification d'associations textuelles est la fréquence. Or, l'expectative normalisée mesure le degré de cohésion qui lie les constituants d'un n-gram mais ne rend pas compte de l'hypothèse formulée précédemment. Certains de cette supposition, nous déduisons qu'entre deux n-grams ayant la même expectative normalisée, il est plus probable que le plus fréquent des deux corresponde à une association textuelle pertinente. Ainsi, nous définissons l'expectative mutuelle par l'équation (4).

$$EM([p_{11}u_1 ...p_{1i}\ u_i ...\ p_{1n}\ u_n]) = p([p_{11}u_1 ...p_{1i}\ u_i ...\ p_{1n}\ u_n]) * EN([p_{11}u_1 ...p_{1i}\ u_i ...\ p_{1n}\ u_n]) \quad \textbf{(4)}$$

L'expectative mutuelle permet donc de mesurer le degré de cohésion de tout n-gram sans être limitée aux associations binaires. Ainsi, il est possible de classer tout n-gram (i.e. $\forall n,\ n \geq 2$) suivant son degré de pertinence.

[xi] L'espace (i.e. "_____") correspond au mot extrait du n-gram.

[xii] Ce n-gram est obtenu par la transformation suivante: $p_{2i} = p_{1i} - p_{12}$ (pour i=3 à n).

[xiii] L'accent circomflexe "^" correspond à une convention fréquemment utilisée en Algèbre qui consiste à écrire un "^" au-dessus du terme omis d'une suite indexée de 2 à n.

[xiv] En effet, les numérateurs restent inchangés d'une probabilité à l'autre. Ainsi, la normalisation peut être réalisée par le calcul du centre de gravité des dénominateurs.

4. Processus d'Acquisition

La plupart des méthodes proposent des valeurs limites globales (ou seuils) qui permettent de définir si un vecteur est d'intérêt ou non. Ces seuils font l'objet d'un ajustement qui est crucial pour la réussite des expérimentations statistiques (Church et Hanks, 1990; Smadja, 1993; Dunning, 1993; Daille, 1995; Smadja, 1996; Shimohata, 1997). Il s'agit d'un compromis entre des valeurs (de fréquence ou de mesure d'association) assez permissives pour que la collecte soit importante (taux de rappel) et des valeurs pas trop généreuses pour que le résultat soit précis (taux de précision). Malheureusement, cette approche se révèle peu fiable et peu flexible. En effet, les résultats dépendent de l'expérimentation, et, suivant la longueur, le type, le domaine et la langue du *corpus*, il est nécessaire de réajuster les valeurs des seuils. Ainsi, nous introduisons un nouvel algorithme, le LocalMaxs (Silva et al., 1999), qui ne dépend d'aucun seuil (pré-établi ou mesuré par expérimentation) et qui élit tout vecteur dont le degré d'association correspondant est un maximum local. Soient, une mesure d'association, *assoc*, un n-gram, *W*, l'ensemble de tous les (n-1)-grams contenus dans *W*, Ω_{n-1}, l'ensemble de tous les (n+1)-grams contenant *W*, Ω_{n+1} et une fonction *taille* qui rend la longueur d'un n-gram *W* donné en argument, alors:

$\forall x \in \Omega_{n-1}$
$\forall y \in \Omega_{n+1}$
W est unité d'association si
$(taille(W)=2 \wedge assoc(W) > assoc(y)) \vee (taille(W)\neq 2 \wedge assoc(W) \geq assoc(x) \wedge assoc(W) > assoc(y))$

5. Résultats

Les résultats obtenus à partir d'un *corpus* trilingue (Français-Portugais-Anglais) d'environ 300000 mots graphiques montrent qu'il est possible d'extraire avec précision des noms composés, des déterminants composés ainsi que des locutions verbales, adverbiales, adjectivales, conjonctives et prépositionnelles contigus ou non. Les résultats mettent également en évidence l'extraction de syntagmes récurrents (Annexe 1). De la même façon, à partir de la représentation du même *corpus* en une suite de caractères, il est possible d'extraire des morphèmes affixaux contigus et non contigus ainsi que des thèmes (Annexe 2).

Dans le cadre de l'extraction d'unités polylexicales, l'Expectative Mutuelle a été comparée avec plusieurs modèles mathématiques appliqués au LocalMaxs: l'Information Mutuelle spécifique (Church et Hanks, 1990), le coefficient Dice (Smadja, 1996), le ϕ^2 (Gale, 1990) et le coefficient de vraisemblance (Dunning, 1993)[xv]. Dans tous les cas, l'Expectative Mutuelle a mis en évidence un taux de précision inégalé (Dias, et al. 1999b). Nous présentons dans le Tableau (3) le taux de précision obtenu pour chaque langue avec une fenêtre de 10 mots et nous le comparons avec celui obtenu avec une fenêtre de 4 mots. Les résultats mettent en évidence la nécessité d'un filtrage d'autant plus important que la fenêtre est plus large.

	Français	**Portugais**	**Anglais**
Fenêtre de 10 mots	62.13%	61.33%	58.35%
Fenêtre de 4 mots	86.59%	90.16%	90.35%

Tableau 3: Taux de Précision des unités extraites

Particulièrement, contrairement à ce que propose Smadja (Smadja, 1993), notre système montre qu'il n'est pas nécessaire de recourir à un post-traitement des résultats pour l'extraction d'unités polylexicales non contiguës. Ainsi, un grand nombre d'unités polylexicales contenant plusieurs interruptions sont identifiées. Dans ce cas, une interruption correspond à un groupe

[xv] D'autres modèles ont été testés: coefficients de Cramer et de Pearson (Bhattacharyya, & Johnson 1974).

d'au moins deux mots différents, souvent synonymes et interchangeables. Par exemple, l'unité "*transport de _____ dangereuses*" met en évidence une interruption correspondant à l'occurrence des deux synonymes "*matières*" et "*substances*".

En ce qui concerne l'identification d'unités morphologiques, nous ne présentons pas de résultats de précision dans la mesure où il est difficile de déterminer ce qui est un morphème affixal ou un thème de ce qui ne l'est pas. Une étude exhaustive au cas par cas est donc nécessaire. Le lecteur devra cependant retenir que les résultats sont globalement peu informatifs pour les associations entre deux caractères mais que le nombre d'unités d'intérêt augmente rapidement avec le nombre de caractères en association.

6. Conclusion

Dans cet article, nous avons proposé un analyseur statistique qui extrait à partir d'un texte brut un ensemble d'associations textuelles contiguës et non-contiguës sans recourir aux méthodes d'amorçage ni à la définition de valeurs seuil globales. Ainsi, nous avons conjugué une nouvelle mesure d'association fondée sur le concept d'expectative normalisée, l'Expectative Mutuelle, avec un nouveau processus d'extraction basé sur un algorithme de maxima locaux, le LocalMaxs. En particulier, l'Expectative Mutuelle permet de caractériser la structure des associations complexes sans se limiter aux associations binaires et le LocalMaxs permet d'éviter la définition de valeurs seuil globales dans le processus d'acquisition. Par ailleurs, l'introduction de la mesure d'Expectative Mutuelle a permis une amélioration sensible dans le cadre de l'extraction d'unités polylexicales, comparativement aux modèles mathématiques couramment mentionnés dans la littérature: l'Information Mutuelle spécifique, le coefficient Dice, le ϕ^2 et le coefficient de vraisemblance. Les résultats obtenus à partir d'un *corpus* trilingue (Français-Portugais-Anglais) montrent qu'il est possible d'extraire avec précision des noms et des déterminants composés ainsi que des locutions verbales, adverbiales, adjectivales, conjonctives et prépositionnelles. De la même façon, il est possible d'identifier des morphèmes libres et liés à partir des textes représentés en suites de caractères. Cependant, un certain nombre de problèmes quantitatifs subsistent. Les unités d'association qui n'apparaissent que deux fois dans le *corpus* sont la source principale d'imprécision. De même, il est nécessaire que l'analyse repose sur un nombre élevé d'unités textuelles pour atteindre des taux de précision acceptables. Finalement, notre analyseur tend à élir une proportion trop grande de 3-grams; ceci se révélant une source d'imprécision supplémentaire. Afin de résoudre un certain nombre de ces problèmes, nous travaillons actuellement sur des variantes de normalisation.

Annexe 1 : Unités Polylexicales

Anglais

EM	Fréquence	Unité Polylexicale
0.000104686	3	Peace Accord
0.000104686	3	Court of Justice
0.000131371	4	fall within the competence of
0.000174477	5	one or more of the
0.000174477	5	Turkish and Kurdish political refugees

Français

EM	Fréquence	Unité Polylexicale
0.000118537	4	mettre au point
0.00011007	4	il y a
0.000421063	23	en matière de
0.000543256	19	droits de l'homme
6.42075e-05	2	l'exposition aux rayonnements non ionisants

Portugais

EM	Fréquence	Unité Polylexicale
0.000244309	12	direitos humanos
0.000105115	5	Conselho de Ministros
0.0001339	6	em conformidade com
4.35175e-05	3	por exemplo ,
0.000127244	5	é da competência

Anglais

Unité non Contiguë	Occurrences
to allow ___ ___ to	- to allow foreign observers to - to allow the ICRC to visit
the ___ of equal ___ ___ men and women	- the principle of equal pay for men and women - the question of equal treatment of men and women

Français

Unité non Contiguë	Occurrences
ne ___ pas	- ne sont pas - ne relèvent pas
données relatives ___ ___ ___ indésirables	- données relatives aux demandeurs d'asile indésirables - données relatives à ces étrangers indésirables

Portugais

Unité non Contiguë	Occurrences
transporte de ___ perigosas	- transporte de matérias perigosas - transporte de substâncias perigosas
um sistema ___ de ___ dos	- um sistema geral de reconhecimento dos - um sistema único de selecção dos

Annexe 2 : Unités Morphologiques

Anglais

EM	Fréquence	Unité Lexicale	Occurrences
0.000849731	3841	i v e	c o m p e t i t i v e # c o m p e t i t i v e n e s s
0.00327337	7029	i n g #	p l a n n i n g # d o i n g #
0.00117897	2398	E u r o	# E u r o f e d o p ` E u r o s p o r t e l l o ´
0.000133473	314	i n g s #	s a v i n g s # u n d e r t a k i n g s #
0.000208764	450	a l l y #	e t h n i c a l l y # s p e c i f i c a l l y #

Français

EM	Fréquence	Unité Lexicale	Occurrences
0.000155018	472	é g a l	# é g a l e i l l é g a l e s
0.000110404	376	t i f #	p r é v e n t i f # c o n s u l t a t i f #
2.88237e-05	73	h u m a	i n h u m a i n e s h u m a n i t a i r e
0.000684857	1453	e u r s #	d e m a n d e u r s # d o n a t e u r s #
4.44447e-05	100	# r é d u	# r é d u c t i o n s # r é d u i r e

Portugais

EM	Fréquence	Unité Lexicale	Occurrences
0.0012789	4663	# i n	# i n d e p e n d e n t e # i n d e s e j a d o s
1.67443e-05	64	g o z	# g o z a m r e g o z i j a - s e
0.00026779	732	v e l #	a p l i c á v e l # f a v o r á v e l #
0.000223916	465	p o s i ç	# p o s i ç ã o d i s p o s i ç ã o
5.64065e-05	199	s # a n o s #	t r ê s # a n o s m u i t o s # a n o s[xvi]

Références

Bhattacharyya G. and Johnson R. (1977). *Statistical Concepts and Methods*. John Wiley.

Bourigault D. (1996). Lexter, a Natural Language Processing Tool for Terminology Extraction. In *Proc. of 7th EURALEX International Congress.*

Church K. and Hanks P. (1990). Word Association Norms Mutual Information and Lexicography. *Computational Linguistics*, vol. 16 (1):23-29.

Daille B. (1995). Study and Implementation of Combined Techniques for Automatic Extraction of Terminology. In *The Balancing Act Combining Symbolic and Statistical Approaches to Language*. MIT Press.

Dias G., Guilloré S. and Lopes G. (1999a). Language Independent Automatic Acquisition of Rigid Multiword Units from Unrestricted Text corpora. In *Proc. of Traitement Automatique des Langues Naturelles (6th Conference on natural language processing).*

Dias G., Guilloré S. and Lopes G. (1999b). Multilingual Aspects of Multiword Lexical Units. In Spela Vintar editors, *Proc. of Workshop Language Technologies –Multilingual Aspects*. Ljubljana, Slovenia.

Dunning T. (1993). Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, vol. 19-1.

Enguehard C. (1993). Acquisition de Terminologie à partir de Gros Corpus. In *Proc. of Informatique & Langue Naturelle*.

Gale W. (1991). Concordances for Parallel Texts. In *Proc. of Seventh Annual Conference of the UW Center for the New OED and Text Research, Using Corpora*. Oxford.

Hausser R. (1999). Three Principled Methods for Automatic Word form Recognition. In UniPress editors, *Proc. of VEXTAL*, Venezia, Italy.

Justeson J. (1993). Technical Terminology: Some Linguistic Properties and an Algorithm for Identification in Text. Research Report RC 18906 (82591) 5/18/93. IBM.

Salem A. (1987). *La pratique des segments répétés*. Klincksieck. Paris.

Shimohata S. (1997). Retrieving Collocations by Co-occurrences and Word Order Constraints. In *Proc. of ACL-EACL'97.*

Silva J., Dias G., Guilloré S. and Lopes G. (1999). Using LocalMaxs Algorithm for the Extraction of Contiguous and Non-contiguous Multiword Lexical Units. In Springer-Verlag, *Proc. of 9th Portuguese Conference in Artificial Intelligence.*

Sinclair J. (1974). English Lexical Collocations: A study in computational linguistics. In J. M. Sinclair, *Lexis and Lexicography*. Uni Press.

Smadja F. (1993). Retrieving Collocations From Text: XTRACT. *Computational Linguistics*, vol.19 (1): 143-177.

Smadja F. (1996). Translating Collocations for Bilingual Lexicons: A Statistical Approach. *Computational Linguistics*, vol.22 (1).

[xvi] Cette unité est un morphème qui définit le phonème de liaison entre le pluriel de l'adjectif et son nom associé.