Extraction de lexique bilingue à partir d'un corpus de traduction : une stratégie par écrémage

Etienne Leblois

CRTT, Université Lyon 2

Abstract

To extract a bilingual lexicon from a translation corpus or aligned bilingual corpus, the standard approach is to select equivalences based on their specificity, an indication of the small probability that the two terms of the equivalence are paired only by chance. An additional approach can be taken from the lexicographer's intuition: when confronted with an unknown pair of languages, he/she will start retrieving first the transcodables, then less likely pairs, thus following a skimming strategy that favors reliability of the extractions over recall. We define ignition loss, a numerical criterion that quantifies this transcodable nature of the candidate equivalence. Equivalence candidates being 2D-characterized by both ignition loss and specificity, the skimming strategy consists in selecting a pair candidate on the convex hull of the plot. Once that pair has been withdrawn from the corpus, we update the elementary statistics and select a next pair, until residual equivalence scores become insufficiently significant. A first implementation of such a skimming strategy is demonstrated and applied to both a simulated corpus and a real corpus of Finnish literature aligned with its French translation.

Keywords: bilingual lexicon, translation corpus, equivalence criteria, Finnish, French.

Résumé

En extraction de lexique bilingue à partir d'un corpus de traduction ou corpus bilingue aligné, l'usage est d'extraire les équivalences candidates sur la base de leur spécificité. La spécificité quantifie la faible probabilité que les deux termes de l'équivalence candidate soient ainsi présents en segment appariés du seul fait du hasard. Un autre critère nous paraît fourni par l'intuition du lexicographe qui face à une paire de langues inconnues commencera par récolter les transcodables, avant de s'attaquer à des paires plus incertaines, dans une stratégie d'écrémage privilégiant la fiabilité des extractions sur le rappel. Nous introduisons la perte au feu, critère numérique permettant de quantifier le caractère transcodable d'une équivalence candidate. Les équivalences candidates étant désormais caractérisées en deux dimensions par leur perte au feu et leur spécificité, on appellera stratégie d'écrémage toute stratégie consistant à sélectionner une paire sur l'enveloppe convexe de ce diagramme, à la retirer, à procéder à la mise à jour des statistiques élémentaires, et à recommencer jusqu'à atteindre des scores insuffisants. On étudie ici le comportement d'une implémentation de cette stratégie, d'une part sur corpus simulés, d'autre part sur un corpus réel constitué de littérature finnoise alignée avec sa traduction française.

Mots-clés : lexique bilingue, corpus de traduction, critères d'extraction, finnois, français.

1. Introduction

Dans le foisonnement des corpus de données textuelles (Habert et al, 1997) suscité par les enjeux spécifiques aux industries de la langue (Pierrel et al., 2003), les corpus de traduction forment une classe particulière. Leur exploitation emblématique est la recherche et la valorisation des trésors d'équivalents de traductions qu'ils recèlent, grâce au travail des traducteurs au fil des décennies (Kraif, 1999, 2001) ; (King, 2003) ; (Martinez et Zimina, 2002) ; Zimina (2004 et 2005), (Tiedemann, 2003). On attend des travaux en cours une aide

réelle aux travaux de lexicographie bilingue (Béjoint et Thoiron, 1996), voire une aide directe au traducteur.

En pratique, la recherche d'équivalences présente plusieurs visages. Si le lemme cherché est spécifié par un lexicographe en train de rédiger une notice, on voudra produire très rapidement une concordance pour ce lemme précis, quitte à ce qu'un certain nombre de propositions erronées s'y glissent. Si on procède à une exploitation générale pour laquelle toute paire est intéressante, la stratégie d'extraction, de sélection et de présentation seront de première importance afin de ne pas générer un immense listing d'approximations. On s'intéresse dans ce qui suit à cette recherche générale.

La section 2 rappellera comment peut être décrite la répartition du vocabulaire au sein d'une langue, et comment ceci interagit avec la richesse lexicale que l'on peut attendre d'un texte d'une longueur donnée. La section 3 introduit quelques notations relatives à un bitexte et rappelle comment se calcule la spécificité d'une équivalence candidate ; sur la base de cette seule spécificité, on en déduit quelques encadrements en matière de potentiel lexical d'un bitexte. La section 4 introduit un critère de perte au feu complémentaire à la spécificité, et présente une stratégie d'extraction de lexique par écrémage s'appuyant sur l'un et l'autre critère. La section 5 montre comment se comporte cette stratégie d'écrémage sur des corpus de traduction idéalisés et réels. La section 6 présente quelques commentaires et perspectives.

2. Rappels de statistique lexicale

2.1. La loi de Zipf-Mandelbrot et son interprétation probabiliste

Zipf a observé que la fréquence d'utilisation d'un mot dans un texte est approximativement inversement proportionnelle à son rang dans la liste de mots du texte triés par fréquence décroissante (Zipf, 1944). Ceci se note $f(r) = K / r$ où K est une constante, r le rang du mot dans l'échantillon trié par fréquence décroissante, et *f(r)* le nombre d'occurrences constatées ou fréquence empirique (Muller, 1977). Le modèle de Zipf a été ensuite élargi par B. Mandelbrot en : $f(r) = K /(a + b.r)^c$, modèle à 4 constantes K, a, b, c dont la loi de Zipf apparaît comme un cas particulier.

Notons que si c>1, la somme des fréquences des n premiers mots est bornée quand n tend vers l'infini. Si c<=1, il faut se fixer pour n une valeur limite N qui sera la taille du lexique. Dans ces deux cas seulement, les fréquences de Zipf-Mandelbrot deviennent assimilables à des probabilités. On se situera par la suite dans ce cadre.

2.2. Implication de la loi de Zipf-Mandelbrot quant au contenu en lexique d'un texte unilingue

Nous n'avons pas su trouver dans la littérature ou développer de modèle explicite reliant les paramètres de la loi de Zipf-Mandelbrot et la longueur du texte qu'elle décrit. Nous utiliserons donc la simulation pour faire ressortir que la distribution des mots du lexique et leur apparition au fil d'un texte sont liées.

On étudie le vocabulaire de textes unilingues simulés. La répartition des fréquences est estimée par la formule de Mandelbrot, calée dans le cas de la langue finnoise sur des données du Center for Computer Science de l'Université d'Helsinki (CSC dans la suite). Sur les 10 000 mots les plus fréquents d'un corpus de référence, on trouve a=2.8096, b=33.479, c=0.88202 ; K est calculé pour une hypothèse de troncature du lexique à 100 000 mots.

On simule des textes de 100, 1000, 10 000 etc. jusqu'à 10 millions de mots. Pour chacun de ces textes on évalue pour combien de lemmes on dispose de plus de k occurrences. Ceci donne la figure 1a.

On montre également que l'accroissement du vocabulaire rencontré au long d'un texte, de rapide au début, devient de plus en plus lent (1b), et que corrélativement le rapport des formes distinctes sur le nombre de forme totales diminue (1c). La loi de Zipf-Mandelbrot présente là un comportement tout à fait attesté sur textes réels, et que l'on illustre en figure 2 sur un ensemble de textes finnois réels, l'évaluation étant ici basée sur les formes non lemmatisées.

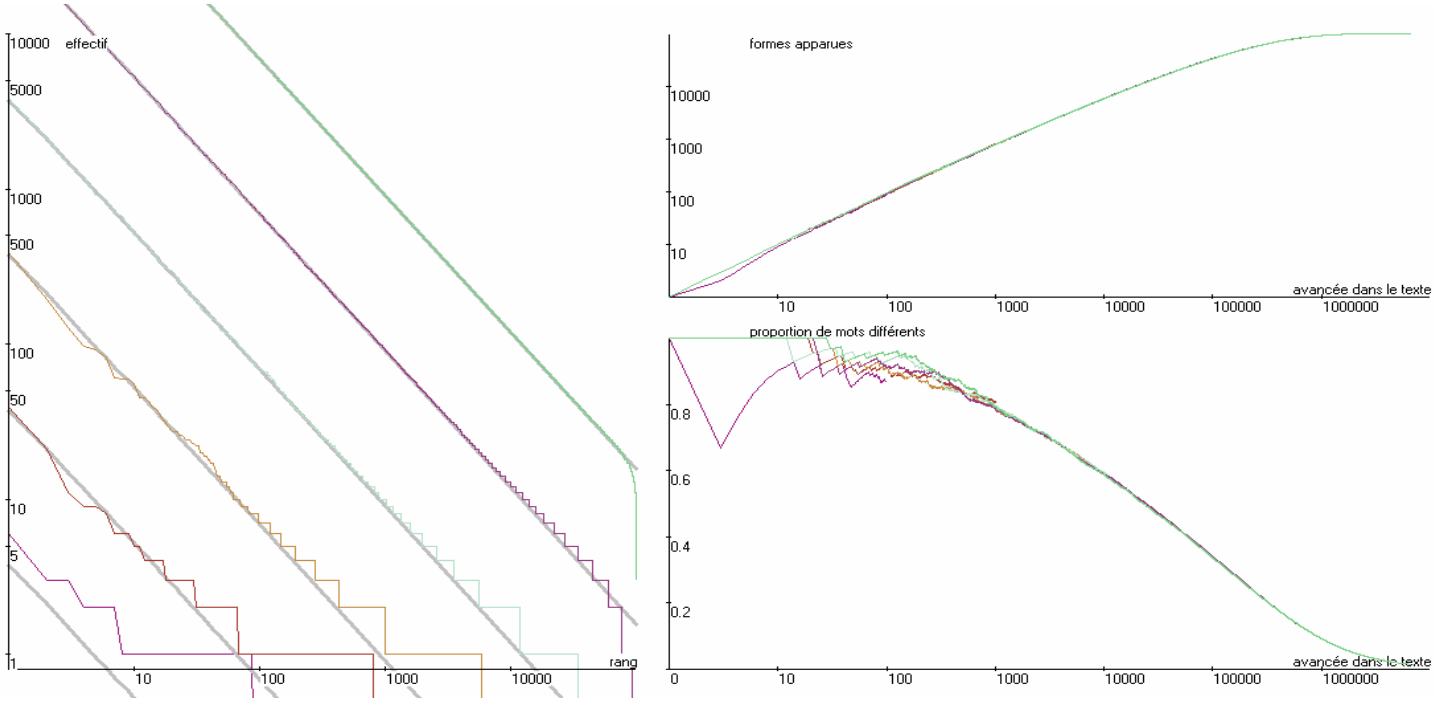

Figure 1 : (a) à gauche, fréquence calculée et simulée des lemmes apparus triés par fréquence décroissante (diagramme de Zipf). (b) en haut à droite : le nombre de lemmes apparus depuis le début du texte, évidemment plafonné à la taille du lexique. (c) en bas à droite : proportion de lemmes différents en fonction du nombre de mots.

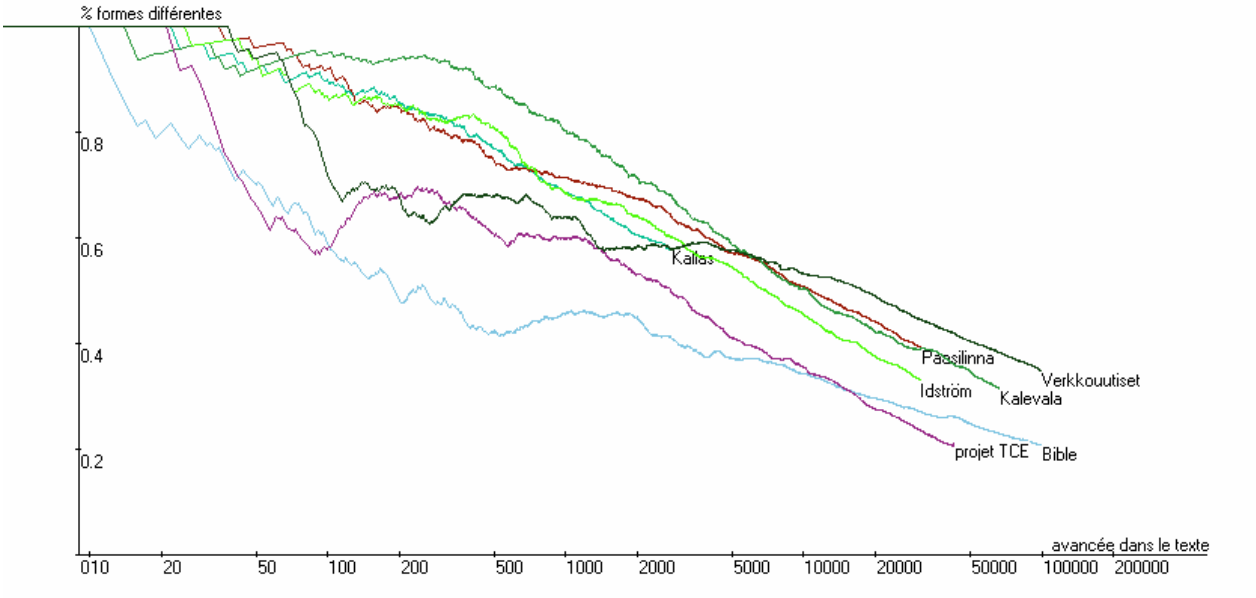

Figure 2 : proportion de formes nouvelles, en fonction de l'avancée dans les textes. Les textes littéraires (Idström, Kallas, Kalevala, Paasilinna) ont une bonne richesse de vocabulaire. Le texte de l'ex-projet de traité constitutionnel pour l'Europe (TCE) est plus répétitif (de par son objet, parce qu'il utiliserait une terminologie contrôlée, ou une grammaire plate nuisant en finnois à la variété morphologique ? cela n'a pas été étudié) Un texte composite tel la Bible résiste mieux sur le long terme. Verkkouutiset est un site de « brèves » en ligne, dont la rubrique « kotimaa », soit les pages nationales, présentent un taux de formes nouvelles durablement élevé, en raison probablement plus de son caractère ouvert sur le chaos de la vie humaine que des moyens littéraires mobilisés.

3. Bitexte et spécificité

3.1. Structure statistique élémentaire d'un bitexte

Au sein d'un bitexte structuré en nseg segments, considérons une paire de mots, l'un en langue 1, l'autre en langue 2, en équivalence traductionnelle partielle.

On adopte les notations suivantes :

pp : nombre de segments où m1 apparaît en langue 1 et m2 apparaît en langue 2.

pn : nombre de segments où m1 apparaît en langue 1 et m2 n'apparaît pas en langue 2.

np : nombre de segments où m1 n'apparaît pas en langue 1 et m2 apparaît en langue 2.

nn : nombre de segments où m1 n'apparaît pas en langue 1 et m2 n'apparaît pas en langue 2.

La fréquence d'apparition du mot m1 au sein de la langue 1 est (pp+pn)/nseg, la fréquence d'apparition du mot m2 au sein de la langue 2 est (pp+np)/nseg.

3.2. La spécificité

La spécificité attribuée à une paire veut caractériser le fait que les deux mots se présentent trop souvent en face l'un de l'autre pour que le hasard soit seul en cause (Oakes, 1998).

Avec les notations introduites ci-dessus, la spécificité s'écrit $S = \sqrt{pp} - \frac{(pp+pn)(pp+np)}{nseg\sqrt{pp}}$

La spécificité n'a qu'une probabilité 0,1 de dépasser 1.28 du seul fait du hasard, une probabilité 0,001 de dépasser 3.08 du seul fait du hasard, etc. Dès lors une valeur élevée de spécificité peut caractériser une équivalence candidate.

Les logiciels d'extraction auront pour tâche de répertorier les équivalences pertinentes, dont la spécificité est supérieure à un seuil fixé, et de les documenter par une concordance.

Supposons maintenant une situation symétrique, où pp/(pp+pn)=pp/(pp+np)=ρ, au sein d'un bitexte de *nseg* segments tous de longueur L (présentant donc un volume de N=*nseg*.L formes par langue). On note p le rapport (pp+pn)/N, c'est la probabilité du mot m1 en langue 1 (et aussi de m2 en langue 2).

Exprimée en fonction du nombre de segments et de la longueur de ces segments, la spécificité devient $S = \sqrt{Np}.(1 - pL/\rho)$

3.3. Relation entre le volume du corpus et nombre d'équivalences extractibles sur la base de leur spécificité

De l'expression qui précède on déduit que le volume de bitexte nécessaire pour détecter une paire m1 ⇔ m2 biunivoque est au moins égal à S^2/p, p étant la probabilité d'occurrence des mots m1 et m2 en langue, et S le seuil de spécificité requis.

Le degré de certitude souhaité (et donc la possibilité de faire confiance à la détection automatique) et la rareté des mots (donc l'étendue du vocabulaire cherché) jouent très fortement sur le volume de corpus nécessaire. Par exemple, en se fixant une spécificité de 4, et en s'intéressant aux 10 000 mots répertoriés par le CSC, la probabilité du dernier étant de 0.00000741, on constate qu'il convient de disposer au moins d'un corpus parallélisé de 2,2 millions de mots par langue pour bâtir un dictionnaire de 10 000 mots par voie automatique.

Accessoirement, on notera la nécessité pour L d'être très inférieure à rho/p : la segmentation doit être assez fine pour que les mots les plus fréquents ne soient pas présents partout, faute de quoi on ne parviendra pas à en dire quoi que ce soit. Sans que cette restriction soit dramatique, il est intéressant ne noter qu'elle ne dépend pas du volume du corpus.

Cependant, une spécificité élevée ne garantit pas que la raison qui fonde la co-occurrence soit celle de l'équivalence traductionnelle. Deux mots m1 et m2a peuvent être mis en relation par la spécificité simplement parce que m2a co-occure en langue 2 avec m2b. Egalement, de part la forme des distributions de Zipf-Mandelbrot, il arrive qu'un mot très fréquent en langue 2 fasse concurrence à l'équivalent véritable d'un mot de langue 1. En fait, les listes d'équivalences candidates basées sur la seule spécificité ne sont pas directement exploitables.

4. Perte au feu et extraction par écrémage

4.1. Perte au feu

Introduisons un élément de diagnostic complémentaire, la perte au feu.

Dans le cas d'un transcodable (situation idéale du strict point de vue terminographique), le mot m1 de la langue 1 se traduit toujours et exclusivement par m2 en langue 2. On a donc pn=0 et n =0, et la spécificité sera maximale pour la valeur de pp fixée.

Elle vaut dans ce cas $S = Sx = \sqrt{pp}\,(1 - pp/nseg)$

Nous proposons d'appeler perte au feu la différence 100%(Sx-S)/Sx. Une perte au feu faible sera un indice de transcodabilité. En considérant les valeurs de perte au feu sur un bitexte concret, on constate que les pertes au feu inférieures à 1% correspondent effectivement à des transcodables ; que des pertes au feu d'environ 5% correspondent à des manières différentes d'exprimer les choses ; mais que des équivalences présentant des pertes au feu supérieures à 20% sont souvent erronées. Enfin, si les distributions de deux lemmes sont indépendantes, cas où la significativité devient nulle, la perte au feu est totale.

La perte au feu a donc une interprétation relativement simple, distincte de celle apportée par la significativité statistique. Nous n'avons pas non plus trouvé que la perte au feu puisse être reliée simplement à l'information mutuelle. Il semble qu'il s'agisse d'un indicateur nouveau.

4.2. Stratégie d'extraction par écrémage

Pour chaque version linguistique, on détermine et trie les différents lemmes représentés ; pour chaque lemme, on recueille les numéros des segments où il apparaît. Ceci génère un profil de 0 et de 1 pour chacun des lemmes de V1 et chaque lemme de V2.

Pour chaque paire de lemmes, on procède au comptage des pp co-apparitions, des pn et np discordances, et bien sûr nn vaut nseg-np-pn-pp. On en déduit la spécificité et la perte au feu pour chaque paire candidate.

Les équivalences candidates étant caractérisées par deux critères sont représentées par un point dans le plan, chaque critère étant figuré sur un axe.

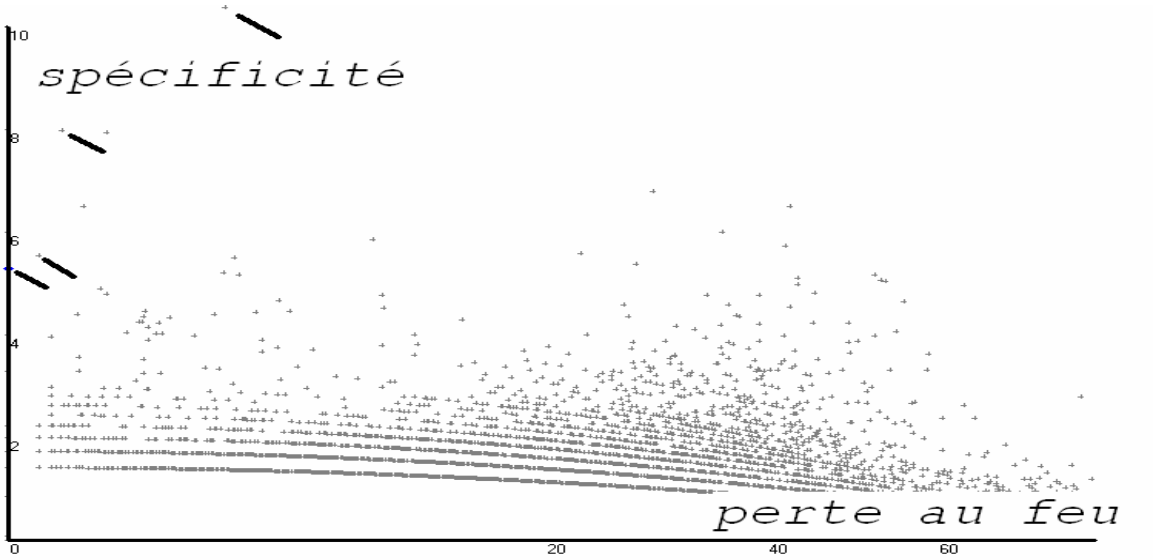

Figure 2 : représentation bidimensionnelle des équivalences candidates. Les quatre paires désignées d'un trait fort sont non-dominées et peuvent être sélectionnées. Pièce de théatre : la reine C. (Kuningatar K), de Ruohonen.

Les paires candidates indiscutables à la sélection seront celles situées sur l'enveloppe du nuage du côté des fortes spécificités et des faibles pertes au feu (sur la figure, ce sont les quatre paires repérées en haut à gauche). En effet, pour toute paire située plus à l'intérieur du nuage, un meilleur choix existe en spécificité (à perte au feu fixée), en perte au feu (à spécificité donnée), et souvent sur les deux critères. Par contre aucune des quatre possibilités soulignées ne domine les autres.

On choisit donc une des équivalences non-dominées, que l'on supprime du corpus. Ceci amène à une mise à jour de tout calcul impliquant l'un ou l'autre des membres de l'équivalence qui vient d'être retirée, et donc à un ajustement du nuage, avant itération.

Notons que l'écrémage se fait dans un ordre donné, et que chaque décision prise influence la suite. Dès lors il convient de tronquer les résultats dès les premières erreurs manifestes, car la suite sera pire.

Le choix de l'individu à extraire parmi les individus de l'enveloppe reste un réglage à faire. Dans la suite nous avons asservi les indicateurs de perte au feu tolérée et de spécificité requise l'un à l'autre, en commençant à une perte au feu tolérée de 1% pour une spécificité requise de 4, et en évoluant par petite touches jusqu'à 75% de perte au feu tolérée pour une spécificité requise de 1.3. Ces seuils finaux semblent fort peu sélectifs, mais ils s'appliquent sur un bitexte déjà fortement épuré, et d'expérience les résultats obtenus restent pertinents.

L'idée de cet algorithme d'écrémage nous est née d'une description faite par (Al-Onaizan et al, 2002) de l'attitude spontanée de personnes confrontées à des bitextes en langues inconnues et qui comme premier réflexe cherchent les transcodables évidents afin d'alléger le travail ultérieur, aussi naturellement qu'un amateur de puzzle commence par rechercher les bords, le ciel, etc, parmi les pièces qui se présentent à lui.

5. Application

5.1. Application à un corpus de bitextes simulés

La stratégie proposée a été évalué sur un corpus de bitextes simulés. Pour construire un bitexte on se donne :

1) une langue V1 avec un vocabulaire de n1 mots distribués selon la loi de Zipf-Mandelbrot ;

2) une langue V2 avec un vocabulaire de n2 mots ;

3) un « dictionnaire bilingue », pour lequel on suppose que chaque mot de 1 pourra être rendu par trois mots différents pris au hasard dans 2 (le fait que plusieurs mots de 1 puissent se traduire par des mots de 2 identiques n'a été ni exclu, ni cherché). La probabilité des trois équivalents en 2 d'un mot de 1 est tirée au hasard, la première p1 sur]0,1[, la deuxième p2 sur]0,1-p1[, la troisième p3 valant 1-p1-p2.

4) On se donne un corpus de nseg segments de lg mots chacun en V1, que l'on « traduit » mot à mot à l'aide du dictionnaire (la traduction est probabilisée, aucune notion de contexte en langue 1 n'est prise en compte) ; ceci donne des textes en V2 réputés former avec leurs homologues corpus de traduction.

On a procédé ainsi à la simulation de 40 textes bilingues basés sur des lexiques de taille variée.

L'algorithme d'écrémage est ensuite appliqué. Pour chaque bitexte on relève diverses statistiques, dont celles illustrées ci-dessous et relatives soit au bitexte lui-même, soit au lexique extrait. Et puisque le dictionnaire bilingue est connu par construction, on peut vérifier pour chaque équivalence proposée si elle est correcte.

Individu	Signification	Règle de calcul	n°10	n°31	n°35	n°36
volume	volume du corpus en nombre de formes par langue		49600	95961	151056	3504
n1	nombre de lemmes en L1		828	74	720	601
n2	nombre de lemmes en L2		911	74	683	575
lg	longueur des segments, en nombre de formes		20	29	36	12
ns	nombre de segments		2480	3309	4196	292
nt	nombre total d'équivalences proposées		1698	259	1950	231
ne	nombre d'équivalences erronées		6	53	47	1
précision		1-ne/nt	0.996	0.795	0.976	0.996
e1	rang de la première erreur		1269	109	1662	155
confiance	position relative de la première erreur	e1/nt	0.747	0.421	0.852	0.671
rappel		(nt-ne)/(3*n1)	0.681	0.928	0.881	0.128
rend.mat.	rendement du matériau	(nt-ne)/volume	0.0341	0.0021	0.0126	0.0656

Une analyse en composantes principales de ces résultats permet de montrer simplement les corrélations existant entre différentes variables et de repérer des bitextes typiques des différents cas de figure.

On constate que :

- Les deux premiers axes factoriels représentent ensemble près de 80% de la variance des données analysées. Notons qu'en conséquence de deux valeurs propres très proches, les figures pourraient éventuellement être tournées ensemble d'un angle quelconque mais égal autour de leur point central.

- La confiance et la précision que l'on peut attendre sont liées avant tout à la richesse en vocabulaire de chaque version linguistique au sein du bitexte.

- Cette confiance et précision s'opposent à des grandeurs descriptives de la structure du bitexte qui sont le nombre de répétitions (d'un même mot dans le corpus) et la confusion (rendue possible par des segments trop long) (le bitexte n°10 est un bon exemple).

- Le volume du bitexte (volume, travail, nombre et longueur des segments) s'oppose au rendement tant du matériau que du travail d'extraction (bitexte n°36).

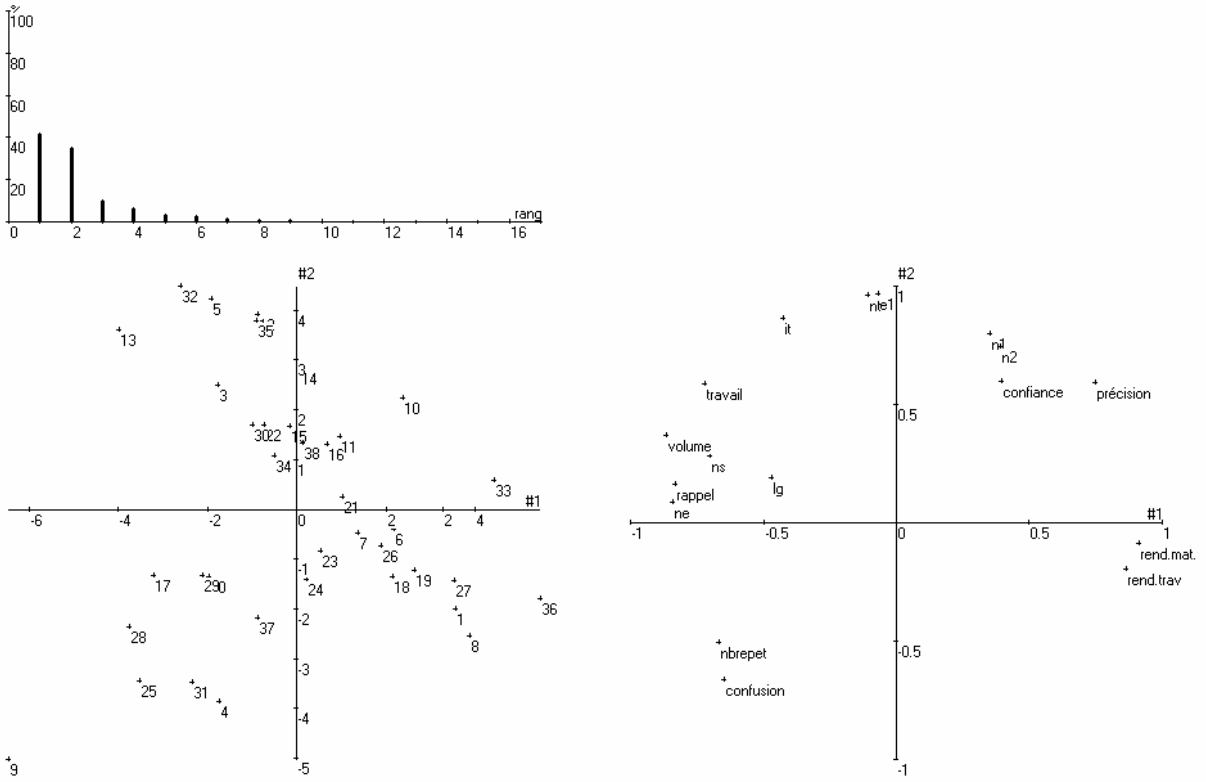

Figure 4 : analyse par ACP des caractéristiques de bitextes simulés et des résultats d'extractions correspondants. En haut la distribution de la variance sur les axes factoriels. En bas à gauche le plan des individus, en bas à droite le plan des variables. Sur cette technique et d'autres, beaucoup d'excellents ouvrages existent, par exemple (Saporta, 1990)

Ainsi pour avoir beaucoup d'équivalences et pouvoir leur faire confiance, il faut disposer d'un grand corpus, finement divisé, basé sur des textes au vocabulaire riche et diversifié, ceci permettant de diminuer fortement la confusion et si possible les répétitions ; il faut ensuite procéder à un écrémage très progressif, donc travailler beaucoup (exemple numéro 35, sauf la longueur des segments, un peu grande). Un corpus volumineux mais pauvre en vocabulaire, avec des segments trop longs facilitant confusion et répétitions, produira une information pauvre et peu précise (bitexte numéro 31).

Toutes les constatations partielles ici faites ne sont pas forcément cohérentes entre elles quant on les examine dans le détail. Ceci est lié au fait que l'outil d'analyse des données utilisé est linéaire alors que les dépendances réelle entre les variables étudiées ne le sont pas. En interprétation d'expériences et de simulation, on pourrait essayer le recours à des techniques non linéaires (Jodouin, 1994). Avant cela, la spécification d'un modèle probabiliste devrait pouvoir, dans certains cas, permettre des calculs analytiques.

On retiendra que la création de corpus simulés permet de comparer différentes stratégies d'extraction, de tester des variantes d'implémentation et de pister les erreurs et inefficacités de codage. Plus fondamentalement, mais sous réserve de la pertinence du modèle adopté pour les bitextes, elle permet d'évaluer a priori des performances à attendre en situation réelle.

A ce titre : la stratégie d'écrémage proposée détecte correctement de très nombreuses équivalences, et notamment les diverses traductions alternantes possibles, ceci malgré la distribution de type Zipf-Mandelbrot simulée pour les mots en langue 1.

5.2. Application à un corpus de traduction réel

Nous disposons d'un corpus constitué de textes en versions finnoise et française. Ce corpus se compose pour un quart de textes littéraires finnois traduits en français, alignées non pas automatiquement selon (Gale & Church, 1993), mais de façon supervisée, les erreurs d'alignement étant détectées à l'aide d'un algorithme détaillé en pages 223 et suivantes de (Crochemore et al, 1997) ; pour un deuxième quart de textes littéraires autres préalignés (la Bible), pour moitié enfin de prose administrativo-politique également préalignée. Le total est d'environ 2 millions de mots par langue, pour 200 000 segments (Leblois 2006).

L'application pratique de l'algorithme d'écrémage a fourni les résultats suivants :

		Version 1				Version 2								
		Formes	Inconnues	Comprises	Lemmes	Formes	Inconnues	Comprises	Lemmes	Nseg	Equiv.	V1	V2	Durée (s)
Krohn	Tainaron, chap. 19	1401	173	1228	747	2075	188	1888	644	80	109	100	93	5
Ruohonen	Kuningatar K	7285	774	6491	2003	10583	1632	8951	1811	1571	546	472	422	43
Marx & Engels	Manifeste du parti communiste	7376	616	6760	2146	11939	1018	10921	1846	465	868	657	603	84
Haavikko	Kullervon tarina	8140	1170	6970	1563	12968	1537	11451	1217	1342	563	482	414	32
Krohn	Doña Quijotte	11053	1482	9571	3055	16823	2491	14132	2589	1214	930	732	725	104
Kallas	Sudenmorsian	12406	2596	9810	3531	19486	2108	17378	2654	1088	884	722	684	112
Ala-Harja	Tom Tom Tom	30004	3019	26985	5769	53343	6314	47029	4584	4341	1922	1603	1430	397
Idström	Kirjeitä Trinidadin	30856	3719	27137	5614	50135	5848	44287	4594	2847	1774	1491	1300	357
Paasilinna	Paratiisisaarenvankit	31457	4075	27382	6683	50262	4964	45278	4492	2452	1741	1446	1295	440
Katz	Kun iso-isä...	39583	6483	33110	7474	61851	8471	53380	5892	3236	2525	2058	1865	736
Projet de traité constitutionnel		43277	6460	36817	4408	70164	9759	60405	3836	2296	2503	1844	1683	420
Lehtolainen	Ensimmäinen murhani	49730	7480	42250	8183	77859	10588	67271	5691	3098	2402	1944	1745	688
Waltari	Nuori Johannes	115244	15912	99332	13963	184348	18917	165431	8722	9868	4169	3271	2960	2064
La Bible		544205	104709	439496	53245	773343	119536	653807	41644	31170	7287	6109	4291	45736
Débats du parlement européen		796516	186930	611586	33990	1237111	244562	992529	23896	129778	10045	8618	7657	23027
« La totale »		1791628	354680	1436948	101670	2731063	451271	2279792	71662	198492	19967	17218	14216	28034

Tableau 2 - performance de l'algorithme d'écrémage, appliqué à l'extraction automatique de vocabulaire sur bitextes finnois-français

Par exemple, l'extraction d'un lexique bilingue par écrémage du bitexte constitué du roman historique « Nuori Johannes » de M. Waltari et de sa traduction réalisée par J.-L. Moreau fournit les résultats suivants : en langue 1, le finnois, on observe 115 244 formes en tout, dont 99 332 sont rattachées à 13 963 lemmes, 15 912 formes restant non-interprétées. La langue 2 est le français. Le bitexte comprend 9 868 segments. L'extraction de vocabulaire fournit 4 169 équivalences potentielles concernant 3 271 entrées en langue 1 et 2 960 entrées en langue 2.

5.2.1. Quel volume extrait, quel rappel ?

Le nombre d'équivalences proposées dépend avant tout du volume du corpus soumis à l'analyse. Il augmente approximativement comme la racine carrée du volume du texte. On retiendra comme ordre de grandeur de dix mille propositions pour un matériau présentant un million de formes dans chaque version linguistique.

Le rappel est le taux de restitution de la caractéristique cherchée par l'algorithme étudié. Nous cherchons des équivalences, linguistiquement au niveau de la lexie et n'en connaissons pas le nombre dans les bitextes réels soumis. On peut au moins fournir le ratio entre le nombre de lemmes figurant dans chaque version linguistique et le nombre d'entrées dans le lexique bilingue créé, donc la proportion de lemmes documentés : environ 25%.

5.2.2. Dépendance de l'extraction à la qualité de la lemmatisation

L'extraction rapproche des distributions non de formes, mais de lemmes. La capacité d'extraction sera bien évidemment fonction de la qualité de la lemmatisation dans chacune des deux langues concernées.

5.2.3. Difficultés liées à un partage morphologie/syntaxe différent selon les langues

Prenons un exemple : il est d'usage en exploitation lexicographique d'ignorer les mots les plus fréquents et réputés les moins intéressants, fournis au code sous forme d'une liste, ou « stoplist », fixée a priori. Peut-on étendre ce concept à une extraction bilingue ?

Si l'on met les mots les plus fréquents du français en stoplist (et, à, le, un, être, ...), on constate immédiatement que les équivalents finnois désormais orphelins (ja, olla) tendent à s'associer avec n'importe quoi. Si donc un mot est mis en stoplist dans une langue, son équivalent doit impérativement être mis en stoplist de l'autre langue. Mais comment mettre en place des stoplists cohérentes, alors qu'en finnois et en français le partage entre morphologie et syntaxe diffère ?

Nous préférons l'observation selon laquelle les associations erronées du type que l'on voulait écarter par usage de stoplists présentent une forte perte au feu. La stratégie d'écrémage proposée y sera donc peu sujette.

Retenons qu'en matière d'extraction de lexique bilingue il faut se défier de certaines solutions développées en contexte unilingue ; et plus encore que la différence de structure entre les langues, si elle est ignorée, se retournera contre nous.

6. Conclusions

En langue, il existe un continuum entre ce que la terminologie linguistique désigne comme termes, quasi-termes, groupement lexicalisés, collocations manifestes et simples affinités lexicales.

Le bi-terme ou élément transcodable sera toujours un des pôles de l'équivalence traductionnelle que nous étudions. Pour caractériser la distance à ce pôle nous avons introduit une notion de « perte au feu », indicateur du caractère bi-univoque d'une équivalence candidate. En complément de la spécificité, la perte au feu permet de sélectionner les équivalences pour lesquelles le risque d'erreur est le plus faible, et de les retirer du bitexte en une stratégie d'écrémage qui dégage l'horizon, permettant de s'intéresser ensuite à des choses plus floues en ayant déjà mis à l'abri ce qui pouvait l'être.

Une stratégie d'écrémage se doit d'être prudente : si une décision erronée est prise, l'écrémage de la paire postulée laisse orphelin des éléments dont la présence embrouillera ensuite le processus, provoquant des décisions erronées en chaîne.

Le nombre d'équivalences proposées est de l'ordre de 10 000 pour un million de mots. Le rappel semble modeste, tout en étant supérieur à ce que donne la simple sélection sur la base des spécificités : 25% des lemmes identifiés en V1 sont documentés par une équivalence au moins. La précision est excellente.

Pour aller plus loin, la première chose est d'améliorer dans chaque langue la qualité de la lemmatisation. Au-delà, il faudra des améliorations conceptuelles, qui iront de l'amélioration de la métrique d'association (position relative des formes dans les segments) à une meilleure

définition des objets cherchés (incluant notamment la possibilité qu'une lexie soit rendue non par un lemme isolé, mais par une collocation).

Références

Al-Onaizan Y., Germann U. et al. (2002). Translation with scarce bilingual resources. *Machine translation* 17: 1-17.

Béjoint H. et Thoiron P. (éd.). (1996). *Les dictionnaires bilingues*. Aupelf-Uref, Duculot, Louvain-la-Neuve, 256 pages.

CSC : http://www.csc.fi/kielipankki/aineistot/naytaAineistot.phtml?lang=fi_FI Catalogue de 97 corpus gérés par le *Centre de calcul scientifique de l'Université d'Helsinki* (en finnois)

Gale W.A. and Church K.W. (1993). A program for aligning sentences in bilingual corpora. *Computational Linguistics* 19(1): 75-102.

Granger S., Lerot J. et al. (eds.). (2003). *Corpus-based approaches to contrastive linguistics and translation studies. Approaches to translation studies*. Amsterdam.

Habert B., Nazarenko A., Salem A. (1997). *Les linguistiques de corpus.* Armand Colin, 240 pages.

Jodouin J.-F. (1994). *Les réseaux de neurones, principes et définitions.* Hermès, 124 pages, ISBN 2-86601-435-9.

King P. (2003). *Parallel concordancing and its applications*. S. Granger: 157-167.

Kraif O. (2002). Méthodes de filtrage pour l'extraction d'un lexique bilingue à partir d'un corpus aligné. *Lexicometrica,* 22 pages.

Kraif O. (1999). Identification des cognats et alignement bi-textuel : une étude empirique. *TALN,* 10 pages.

Kraif O. (2001). *Constitution et exploitation de bitextes pour l'aide à la traduction. Sciences du langage*. Nice, Université de Nice Sophia Antipolis, 547 pages + annexes.

Leblois E. (2006). Réalisation d'un corpus bilingue finnois-français et de sons système d'exploitation. Mémoire de master LEA, spécialité Lexicologie et Terminologie Multilingue et Traductologie, Université Lyon 2.

Muller Ch. (1977). *Principes et méthodes de statistique lexicale.* Hachette université, 206 pages.

Oakes M. P. (1998). *Statistics for corpus linguistics.* Edingburgh textbooks in empirical linguistics. Edingburgh university press.

Pierrel J.-M. (éd). (2000). *L'ingéniérie des langues.* Hermès Sciences éditions, 360 pages, ISBN 2-7462-0113-5.

Renault : http://www.unicaen.fr/ufr/homme/linguistique/ressources/finnois/index.html ressources linguistiques unitex pour le finnois, CRISCO, Université de Caen & CNRS

Saporta G. (1990 ; 2e éd. 2006). *Probabilités, analyse des données et statistiques.* Editions Technip, 622 pages.

Tiedemann J. (2003). Recycling translations, thèse de doctorat, Acta Universitatis Upsaliensi, 132 pages.

Vergne J. (2004). Découverte locale des mots vides dans des corpus bruts de langues inconnues, sans aucune ressource. *Journées internationales d'Analyse statistique des données textuelles* 7: 8 pages.

Zimina M. (2002) Repérages lexicométriques des équivalences à basse fréquence dans les corpus bilingues. *Lexicometrica*, 27 pages.

Zimina M. (2004). Alignement textométrique des unités lexicales à correspondances multiples dans les corpus parallèles. *Journées internationales d'Analyse statistique des données textuelles* 7: 8 pages.