How can lexicographers use a translation corpus?

Raphael Salkie

University of Brighton, England

Introduction

Translation (parallel) corpora are standard tools in several fields, such as translator training, machine translation, contrastive linguistics, and various language engineering applications. One area where one might expect such corpora to be widely used is bilingual lexicography, but in fact such corpora have not been exploited significantly in dictionary compilation – unlike monolingual lexicography, where it would be unthinkable today not to use single-language corpora. Here we discuss why this is so, and we propose a way of assessing which words in such corpora are likely to repay investigation by lexicographers.

1. Bilingual lexicographers should use translation corpora

Compilers of bilingual dictionaries – like most lexicographers – nowadays routinely use corpora as a convenient source of authentic data. It is easy to show that translation corpora could be a useful extra resource in this work. Here is the entry for the French word *partout* in the *Collins-Robert French-English Dictionary* (2002):

partout /paʀtu/ **ADV** everywhere, everyplace *(US)* ◆ ~ **où** everywhere (that), wherever ◆ **avoir mal** ~ to ache all over ◆ **tu as mis des papiers** ~ you've put papers all over the place ◆ **2/15** ~ *(Sport)* 2/15 all ◆ **40** ~ *(Tennis)* deuce

Now consider these examples from the INTERSECT translation corpus (see Salkie 2002 for a description of the corpus):

1. Par exemple, le format des cartes de crédit, des cartes à prépaiement téléphonique et des cartes dites «intelligentes» que l'on retrouve [[partout]] est dérivé d'une Norme internationale ISO.
2. For example, the format of the credit cards, phone cards, and "smart" cards that have become commonplace is derived from an ISO International Standard.

3. Ils avaient même un cimetière fin prêt à proximité et planté de fleurs [[partout]].
4. They even had a cemetery all ready to hand and full of flowers.

Neither of these examples is suggested by the Collins-Robert dictionary. Of course, it is possible that the translations in (1 – 4) are not suitable for a dictionary. They may be rare or even unique, the product of a particularly creative translator dealing with a one-off problem. They may be specific to these particular contexts, as the highlighting suggests: in both examples it is an entire phrase that is the translation unit here, not the individual word *partout*. Dictionaries mostly deal with individual words, and only a limited amount of information about phrases can be included. I am not a lexicographer, so I do not presume to tell compilers of bilingual dictionaries that they must include examples like this. I would, however, like to encourage them to at least consider such examples. In principle, it is possible to use a corpus to find out which translations are frequent and

which are unusual. There is at least the potential here to enrich bilingual dictionaries in interesting ways.

2. Bilingual lexicographers don't use translation corpora

Experts on bilingual lexicography agree that translation corpora are useful resources. Corréard (2005) and Krishnamurthy (2005) both say so, but neither gives an example of a bilingual dictionary that uses a translation corpus. The only instance that I have found is the *Dictionnaire Canadien Bilingue*, which made limited use of the Canadian Hansard corpus (Roberts 1996; Roberts & Cormier 1999). In late 2005, Sue Atkins posted a query on the Euralex list asking for other examples: she had no positive responses.

In other fields, translation corpora are widely used. This table lists some of them:

Field of study	**Example**	**Purpose**
Translation teaching	Zanettin (2001)	Develop teaching materials.
Translation research	Laviosa (2002)	Investigate how translation studies can benefit from corpus analysis.
Language engineering	Véronis (2000)	Design computer programs that can align texts automatically.
Machine translation research	Carl & Way (2003), Hutchins	Example-based Machine

	(2005)	Translation (EBMT).
Contrastive linguistics	Celle (2006)	Compare tense and modality in three languages.
Mainstream linguistics	Mortelmans (2004)	Examine future time reference in German; brief comparison with English.
Bilingual lexicography	??????	Produce bilingual dictionaries.

This situation strikes me as unfortunate.

3. Why don't bilingual lexicographers use translation corpora?

Most of the reasons are severely practical ones. I will discuss these briefly, and then devote more space to the theoretical ones, which are more interesting.

3.1 Availability

Even for the obvious Western European languages, no translation corpora exist which are (a) freely available, (b) textually diverse, (c) easy to use, and (d) of high quality. Several large multilingual corpora meet some of these criteria, notably the JRC-Acquis corpus of EU documents (Steinberger 2008), and the OPUS corpus, which mostly contains computer documentation (Tiedemann 2007). Although large, and clearly useful for certain purposes, the specialised language of these corpora (like Canadian Hansard, which is a bilingual record of debates in the Canadian parliament) means that they are not sufficient on their own to be a complete resource for

lexicographers. Textual diversity is vital, and this is hard to achieve for reasons that we turn to now.

3.2 Quality

In a monolingual, general-purpose corpus, authenticity is the key design criterion. The point is to find out how language is used by real people, with all the oddities, errors, hesitations, interruptions and other types of noise that distinguish genuine language from invented, artificial examples. In a translation corpus, the key issue is quality. There are innumerable badly translated texts on the web, but for research in linguistics, and for lexicographers, poor quality translations are of no interest. (For translation researchers, inferior translations can shed light on the translation process; but for most other users of translation corpora, only good translations are worth considering). What counts as 'translation quality' is, of course, controversial, but as a bare minimum, it is important to (a) know which is the source language; (b) be sure that the translator was a skilled professional; (c) if possible, have a guarantee that the translation was checked before being published; and (d) know that the translation was published by an organisation which takes quality seriously. Even these criteria do not guarantee quality, and unfortunately using texts which meet these minimal criteria means only using published texts, and this means that they are likely to be copyrighted.

3.3 Copyright

Copyright increases enormously the labour and expense of compiling a translation corpus. Consider the work that would be involved in compiling a quadrilingual corpus of English, French, German and Spanish for lexicographical use. One can imagine starting the fiction sub-corpus by

picking 10 authors in each language whose books are translated into the other three – the kind of books that are found in any large bookshop in the four language areas. The compilers would have to obtain copyright clearance from four publishers for each text by each author, a minimum of 40 publishers and in practice probably more. The journalism sub-corpus would involve similar labours with publishers of newspapers in different countries. The non-fiction section of the corpus might be easier: there are, for example, several international organisations and companies whose web sites contain material in the four languages, and they may not guard their intellectual property as fiercely as publishers of fiction or newspapers. It is clear that a great deal of work would be involved, in what is still a legal grey area, to obtain copyright clearance, for research use, for a diverse range of texts and their high-quality translations. Only a large consortium of committed people could do this, and none appears to be on the horizon.

3.4 Other practical issues

Translation corpora for lexicographical use would need to be edited, aligned and lemmatised, to contain only modern texts (another difficult concept – post-1945 or post-1990?), and to be searchable by concordancing software without needing a team of computer specialists constantly on hand. With each of these criteria, the labour and expense increase still further.

3.5 Creative vs systematic translations

As we noted in the discussion of *partout* above, lexicographers are interested in translation equivalents which are part of the system of a language, rather than in one-off, highly context-bound solutions to translation problems – however interesting and clever these might be. As noted in Salkie (2002), the difference between these two types of translation equivalents is not

always straightforward: we showed that *contain* had a few common French equivalents in the corpus but that in about 50% of cases some other translation strategy was used. The German word *kaum*, on the other hand, had a large number of regular translation equivalents in English, with only a few 'creative' translations. For a corpus to be useful for lexicographers, it needs to be large and diverse enough to distinguish between the two types of translation equivalent without lexicographers needing to examine large numbers of examples manually.

I once asked an expert bilingual lexicographer why they had not used a translation corpus in compiling their latest dictionary. She replied that they had started to use a small journalistic corpus in the two languages, but had come up with such a huge amount of fascinating data that they had reluctantly decided to abandon it: they were spending too much time trying to work out how to handle this rich range of material. Dictionary compilers operate under intense time constraints, and any potential improvement in quality has to be weighed against the work involved.

One solution to this problem of too much interesting data would be to flag its existence in some way for users of the dictionary. Assuming that the practical problems of compiling a translation corpus could be solved, lexicographers certainly would not want to include all or even most of the corpus data in their dictionary. It would, however, be very valuable for some users of a bilingual dictionary if for certain words the dictionary had a special label which meant 'Take a look in the corpus'. For some lexical items, this label would in effect tell the user that the equivalents offered in the entry were only the most common ones, and that it could be worth the user's while to look at other corpus equivalents. These could either be made available on a companion web site, or distributed along with the dictionary in some electronic medium such as a CD.

What this would be doing would be expanding the range of phrase and sentence expressions which some dictionary entries include along with single-word equivalents. The *partout* entry above has two of these: *to ache all over*, and *you've put papers all over the place*. By supplying access to a wider range of such equivalents from the corpus, bilingual dictionaries would enable keen users to explore some of the richness of the two languages.

It is, I grant, an open question how many dictionary users would have the time and the enthusiasm to follow up invitations of this kind. Many users of bilingual dictionaries just want to find the solution to their problem quickly and move on. I believe, however, that lots of people using a second language rely heavily on their bilingual dictionary for help in finding the best translation, both from and into the second language. For them, the dictionary is a treasure chest of interesting contrasts, which they regard as a valuable ally. Using a translation corpus could considerably increase the amount of treasure in the chest.

3.6 Which words are likely to repay investigation?

Anyone who has used a translation corpus knows that only certain words throw up interesting equivalents which go beyond those that a bilingual dictionary would suggest. Words denoting concrete objects and actions are not likely contenders: they will either have single equivalents which apply in all cases, or a small range of options which are well covered in existing dictionaries. The same is true of proper names. For other words, it would be useful if there was a way to predict in advance whether it was likely to be worth the lexicographer's while to look them up in a translation corpus.

Word frequency could be a step towards guiding lexicographers' attention. Words which occur very rarely in running text are unlikely to occur frequently in a translation corpus, and can be safely ignored. At the other end of the scale, very common words are also likely to be of little lexicographical interest: the most common words in a language are usually function words, which have a lot of grammar but only a little graspable meaning. They are properly dealt with at length in grammars of a language, rather than in dictionaries.

That leaves words of medium frequency as the most likely candidates for fruitful investigation in a translation corpus. If we could specify the range of this medium frequency more precisely, then we would have narrowed the search-space for lexicographers considerably.

To see if this is possible, we conducted an experiment with the French-English part of the INTERSECT translation corpus. The full corpus, which has about 1.6 million words in each language, is unbalanced, in that some texts are much longer than others, and that some types of text are better represented than others. We therefore constructed a balanced subset of the corpus, containing about 300,000 words in French and English. Texts were divided into 7 fiction and 8 non-fiction samples, with each sample containing 20,000 words. This enabled us to construct meaningful frequency counts. In order to facilitate comparison with other corpora, I have given below both the absolute frequency of the items discussed and the frequency per million words.

The INTERSECT corpus is far from ideal: it fails to meet many of the minimal criteria discussed above for a lexicographically workable corpus. It does, however have three advantages: it exists, it

contains a wide variety of text-types, and it is easy to construct frequency counts and concordances using Michael Barlow's admirable ParaConc software (cf. Barlow 2007). Although some of the corpus data is old-fashioned or otherwise of poor quality, the corpus is usable to show how such an experiment can work.

We used ParaConc to make a frequency list of words in French occurring 20 or more times in the balanced corpus. The corpus is not lemmatised, so the results were lemmatised manually in order to make the frequency counts meaningful: for languages like French which have a rich morphology (particularly for verbs), a frequency count based on word forms would be misleading.

This produced about 1250 lexemes. We selected the 25th word from the top, then the 50th and so on down to the 1250th word. We then examined the English equivalents of these 50 lexemes, noting any 'surprising' translations that were not found in the *Collins-Robert* (in the main entry for the lexeme or in the examples). In the table below we use the technical term 'nice surprises' to refer to these translations. If a word yields a high number of 'nice surprises', then at least in principle it is worth looking to see if these are of lexicographical interest. Our hypothesis is that most of the nice surprises will cluster around the middle of the frequency range.

4. Results

The table below shows the results of this experiment.

Rank	Word forms	Freq of word	Freq of lexeme	Freq per million	Nice surprises

		forms		words	
25	**vous** (n/a)		1136	3787	0
50	**cela** **ça** (n/a)	218 190	408	1360	0
75	**document** **documents**	174 113	287	957	10
100	**texte**		220	733	8
125	**trois**		196	653	8
150	**ministre** **ministres**	133 27	160	533	2
175	**maison** **maisons**	115 24	139	463	7
200	**plusieurs**		123	410	8
225	**ville** **villes**	88 21	109	363	5
250	**suite**		101	337	17
275	**particulier** **particuliers** **particulière**	51 18 25	94	313	5
300	**résultats** **résultat**	59 28	87	290	5
325	**cent**		79	263	2
350	**société**		75	250	2

375	**concernant**		69	230	4
400	**grâce**		66	220	13
425	**te** (n/a)		61	203	0
450	**nature**		55	183	7
475	**conscience**		50	167	2
500	**malgré**		47	157	4
525	**cadre**		44	147	6
550	**partout**		41	137	13
575	**vingt**		39	130	0
600	**contrôle**		37	123	2
625	**importance**		36	120	4
650	**rivière**		35	117	0
675	**production**		34	113	2
700	**rythme**		33	110	2
725	**moitié**		32	107	0
750	**vaste**		31	103	1
775	**sud**		30	100	0
800	**réunion**		29	97	1
825	**garçon**		28	93	0
850	**beauté**		27	90	2
875	**pain**		27	90	0
900	**gabarit**		26	87	0
925	**répertoire**		26	87	1

950	**généralement**		25	83	1
975	**accès**		24	80	1
1000	**libre**		24	80	1
1025	**bande**		23	77	0
1050	**époque**		23	77	1
1075	**proposition**		23	77	2
1100	**emploi**		22	73	3
1125	**télégraphique**		22	73	0
1150	**faveur**		21	70	2
1175	**salon**		21	70	0
1200	**banque**		20	67	0
1225	**égard**		20	67	2
1250	**sérieux**		20	67	0

Most of the 'nice surprises' are in the top half of the table, between *document* (3rd in the list, 957 occurrences per million words) and *partout* (22nd place, 137 occurrences p.m.w.). Our hypothesis is not confirmed. The recommendation to French-English bilingual lexicographers: concentrate your attention on lexemes occurring between 100 times and 1000 times p.m.w.

5. Different kinds of surprise

As well as listing the frequency results, it is interesting to look at some of the equivalents in detail, still with the aim of guiding lexicographers towards those that may be of most interest to them. The number in parentheses after some examples indicates how often that kind of translation was

found.

5.1 Synonyms

Some of the nice surprises are just synonyms of the expected equivalents.

Plusieurs

5. Enfin, le dernier, mais non le moindre, le Comité des bénévoles, dirigé par Edna Wilson, continuera, avec l'aide de la coordonnatrice des bénévoles, Helen Elliott, à superviser [[plusieurs]] activités auxquelles participent nos bénévoles.
6. And, last but not least, the Volunteer Committee under Edna Wilson, with the help of Coordinator of Volunteers Helen Elliott, will continue to oversee the wide variety of activities in which our volunteers are involved.

7. L'idole à [[plusieurs]] bras, la danse de mort, ne sont point des allégories du monde en perpétuelle transformation.
8. The idol of the many arms, the dance of death, these are not at all allegories of the perpetual flux of the universe.

9. Après l'isolement des principes actifs de l'écorce de quinquina, [[plusieurs]] tentatives de synthèse de la quinine étaient restées vaines.
10. Various attempts at synthesizing quinine were made soon after the isolation of the active principles of cinchona, but all of them failed.

grâce (à)

11. Les racines réduites en poudre du Ch'ang shan (Dichroa febrifuga) dont on se sert en Chine depuis au moins 2000 ans ont d'indubitables effets médicinaux [[grâce]] à la présence d'un alcaloïde, la fébrifugine, qui n'a été isolé et analysé que récemment
12. The powdered roots of Ch'ang shan (Dichroa febrifuga), used in China for at least 2000 years, have an undoubted medicinal effect, owing to the presence of an alkaloid, febrifugine, isolated and analysed only recently.

13. Or, si le nombre de demandeurs d'emploi a progressé de 5,1 % en un an, le pouvoir d'achat a augmenté de manière non négligeable : [[grâce]] à la modération des prix, il atteindra 1,7 % cette année dans le secteur privé.
14. Now while the-number of those seeking work has gone up 5.1 per cent in a year, the increase in purchasing power has not been negligible: as a result of prices being held down, it will be I7 per cent in the private sector-for those who have kept their jobs.

15. L'emploi du proguanil se répandit vers la fin de la Deuxième Guerre mondiale, au moment où, [[grâce]] à la mépacrine, la plupart des problèmes militaires posés par la forte incidence palustre dans les régions tropicales étaient devenus moins aigus.
16. Proguanil came into wider use at the end of the Second World War, by when most of the military problems related to a high malaria incidence in tropical areas had become less urgent because of the availability of mepacrine.

Compare these examples, which are also more frequent and go beyond synonymy:

17. [[Grâce]] à l’outil loupe, vous pouvez agrandir l’image lue en résolution complète.
18. By using the magnifying glass tool, you can zoom in on the scanned image at full resolution. (3)

19. [[Grâce]] à ces outils, vous pouvez visualiser des pages et identifier les zones de texte et d’images spécifiques à traiter.
20. With these tools, you can view pages and identify specific text and image areas to process. (4)

21. Le traitement avancé de l’information qui concerne l’optimisation du comportement fonctionnel [[grâce]] à la combinaison architecturale du matériel et du logiciel.
22. Advanced information processing that addresses the optimization of functional behaviour through the architectural combination of hardware and software. (5)

Compare also this translation of *textes*, which is also clearly not synonymy:

23. Dans le cas de [[textes]] contenant une grande proportion de termes techniques, tels que des rapports scientifiques et des spécifications de produits, le système vous permet également de préparer et de charger un dictionnaire utilisateur.
24. For jobs that contain a great deal of technical terminology, such as scientific reports and product specifications, the system also lets you prepare and load a user dictionary.

Lexemes where the ‘surprises’ include many synonyms are *résultat* (findings, outcomes), *nature*

(character, essence) and *partout* (*partout dans le monde* :: all parts of the world; *partout dans le pays* :: across the country). Arguably it is not the best use of space in a bilingual dictionary to indicate, for instance, that *plusieurs* can be translated by *several*, *various*, *a wide variety of*, and so forth. These cases are unlikely to be of interest to lexicographers.

5.2 Omissions

Often the source lexeme is simply omitted:

25. Dans chaque cas, quelle est la [[nature]] précise des travaux exécutés ou du service rendu?
26. In all cases what was the exact work or service performed?

27. Plusieurs CD successifs peuvent être examinés jusqu'à ce qu'un consensus soit atteint sur le contenu technique du [[document]].
28. Successive committee drafts may be considered until consensus is reached on the technical content.

29. [[Plusieurs]] avant- projets successifs peuvent être examinés jusqu'à ce que le groupe de travail ait acquis la certitude d'avoir élaboré la meilleure solution technique au problème considéré.
30. Successive working drafts may be considered until the working group is satisfied that it has developed the best technical solution to the problem being addressed.

31. *3) L'article 3 est remplacé par le [[texte]] suivant:

32. *3) Article 3 shall be replaced by the following: (14)

33. A... s’écarte de la voiture bleue et, après un dernier regard en arrière, se dirige de son pas décidé vers la porte de la [[maison]].
34. A... walks away from the blue car and, after a last look back, heads towards the door with her decisive gait.

35. Il s’y maria et eut [[trois]] fils.
36. He married and had sons there.

37. Michel nous a reçus sans témoigner de joie; très simple, il semblait craindre toute manifestation de tendresse; mais sur le seuil, d’abord, il embrassa chacun de nous [[trois]] gravement.
38. Michel showed no signs of pleasure as he welcomed us; he was very simple and seemed afraid of any demonstrations of tenderness; but on the threshold, he stopped and kissed each one of us gravely.

39. En fonction de l’équipement de votre machine, vous pouvez pétrir les pâtes avec l’un des [[trois]] accessoires suivants:
40. Depending on which version of the Multipractic you have, you can make dough with:

This poses tricky problems for lexicographers. Here again my suggestion would be to decide which of these cases are lexicographically interesting, and then to have a special label, directing users to the corpus and indicating that often the word is not translated and has no equivalent in the

other language. Teachers of second languages would probably be very grateful if this kind of guidance were easily available to learners.

5.3 Reformulations

In many cases, we find reformulations which do not contain an equivalent of the source lexeme:

41. La désignation ISO de la sensibilité des pellicules, parmi bien d'autres normes [[concernant]] le matériel photographique, a été adoptée mondialement, facilitant singulièrement les choses pour l'utilisateur.
42. The ISO film speed code, among many other photographic equipment standards, has been adopted worldwide making things simpler for the general user.

43. *21 Parce que les sages-femmes avaient eu la crainte de Dieu, Dieu fit prospérer leurs [[maisons]].
44. *21 And because the midwives feared God he gave them families.

45. En attendant, [[grâce]] aux techniques les plus perfectionnées " qui font toute la différence sur le plan de la rentabilité " on récupère 95 % du minerai, et " trois ou quatre ans devraient permettre de rentrer dans nos frais ".
46. The latest techniques, which are vital to profitability, make it possible to extract 95 per cent of the ore. Pliley expects the mine to break even within 3 or 4 years. (14)

47. De la cruauté en somme, mais juste ce qu'il faut, une cruauté qu'on peut embrasser, insidieuse amertume comme celle des vins du Rhin, agréable [[malgré]] soi.
48. Cruelty in fact, but just the right amount of it, a cruelty to be kissed, a sharp insidious quality like that taste of Rhine wine which one somehow can't help rather liking.

49. Le groupe des décrocheurs compte un peu plus d'enfants de familles monoparentales pauvres ou de parents chômeurs que celui des diplômés, mais la majorité vit dans un [[cadre]] familial normal et financièrement viable.
50. Although they were somewhat more likely than the youths who actually graduated to come from poor single-parent families or have parents who were unemployed, the majority came from financially viable two-parent families. (11)

51. Des recherches très sérieuses ont démontré qu'au moins 24 pour cent des Canadiens de 18 ans et plus sont incapables de lire un [[texte]] simple ou d'effectuer les opérations arithmétiques élémentaires.
52. Well-grounded research has shown that a minimum of 24 per cent of Canadians aged 18 and over are functionally illiterate and/or unable to do simple arithmetic.

53. La Sixième Commission devrait s'efforcer de développer le droit international en la matière en énonçant de nouvelles règles de [[nature]] à favoriser les relations de bon voisinage entre Etats.
54. The Sixth Committee should seek to develop the relevant international law by establishing new norms that would be conducive to good-neighbourly relations between States

55. Les quatre partis d'opposition ... demandent " la mise en place d'un climat de détente politique de [[nature]] à redonner confiance au peuple ".
56. The four opposition parties are asking for " political détente to restore the people's confidence", Youssoufi said.

57. La République démocratique allemande attache une importance particulière à ... toutes les mesures concrètes de [[nature]] à développer les relations de bon voisinage entre Etats.
58. The German Democratic Republic attached particular importance to ... to concrete action to develop good-neighbourly relations between States.

59. Depuis notre centenaire, nous avons vu quelques améliorations de [[nature]] à dissiper les soupçons de bien des Canadiens à l'égard des hommes politiques.
60. Since our centennial year we have seen some improvements which should help reduce the suspicions that many Canadians harbour about politicians.

Here again a special flag in the dictionary might be a good solution.

5.4 Other equivalents

Some of these are listed here. For these examples, it is hard to determine in advance what a lexicographer might decide about them. This is where dictionary compilers can use their skill and experience to make the best use of the data.

61. Les bibliothèques nationales, et d'autres établissements, ont assumé une responsabilité internationale, … pour rassembler, organiser, conserver et mettre à la disposition des gens toutes sortes de [[documents]] favorisant la réflexion, la prise de décision, la prise de mesures, la création et le plaisir.
62. National libraries, and others, have assumed an international responsibility, … for gathering, organizing, preserving and making available all sorts of records as the basis for reflection, decision, action, creation and enjoyment. (3)

63. Kusturica, envolé lui-même de sa Yougoslavie natale vers les États-Unis, rescapé de la tempête qui massacre sa [[ville]] Sarajevo, s'apprête à tenter un improbable tour de force.
64. With this prologue in the form of a dream sequence, the director-who himself took off from his native Sarajevo to the United States well before tragedy engulfed Yugoslavia-tells us he is about to embark on a most improbable tour de force.

65. Nous nous proposons, en les publiant, de préciser les mouvements de deux sensibilités, et de suggérer à ceux qui les liront des réflexions [[particulières]] sur la vie de leurs sens et de leur esprit, qui peut sembler singulière.
66. By publishing them, we propose to delineate the developments of two sensibilities, and to suggest to those who read them some arresting thoughts on the seemingly unusual sensuous and spiritual lives of these two men.

67. En tant que pays, la fortune nous a relativement souri dans l'ensemble, sur le plan économique, [[malgré]] les politiques adoptées par le gouvernement depuis 1967.

68. Relative good fortune, economically, has been ours as a country, largely independently of the policies of the government from 1967 to this time.

69. Vivax est apparemment associée à l'absence dans les globules rouges des individus de ces groupes, des déterminants de Duffy, qui sont [[généralement]] présents dans d'autres ethnies.
70. Thus the partial insusceptibility of black ethnic groups to infection with P. vivax is apparently associated with the absence in these populations of the Duffy red blood cell determinants that are common in other ethnic groups.

71. Les normes sont des accords documentés contenant des spécifications techniques ou autres critères précis destinés à être utilisés systématiquement en tant que règles, lignes directrices ou définitions de caractéristiques pour assurer que des matériaux, produits, processus et services sont aptes à leur [[emploi]].
72. Standards are documented agreements containing technical specifications or other precise criteria to be used consistently as rules, guidelines, or definitions of characteristics, to ensure that materials, products, processes and services are fit for their purpose.

6. Conclusion

Although our hypothesis was not confirmed, we are able to unearth some advice for bilingual lexicographers:

- The range of lexemes occurring between 100 times and 1000 times p.m.w. in a translation corpus may be most fruitful.

- Think about how wide a range of synonyms to suggest for some lexemes.

- Examine lexemes which are often omitted in translation.

- Consider which reformulations are frequent enough and systematic enough to be worth a mention in the dictionary

- For lexemes where inspection of corpus data might be useful for dictionary users, provide a link in their dictionary entries to carefully selected data in the corpus, which would be available separately.

It is regrettable that translation corpora have been around for about two decades but that practical and theoretical problems have prevented their use in bilingual lexicography, where their potential is vast. This paper has offered a way of solving one of the theoretical problems. I strongly hope that the practical problems can also be overcome, so that large, high-quality, copyright-cleared, easy to use, freely available, textually diverse translation corpora for many languages can be created in the not too distant future.

References

Barlow, M. (2007) ParaConc: a bilingual or multilingual concordancer. Available on the web: http://athel.com/product_info.php?products_id=30 . Accessed July 2008.

Carl, M. & A. Way (eds.). (2003) *Recent Advances in Example-Based Machine Translation.* Dordrecht: Kluwer.

Celle, A. (2006) *Temps et Modalité, l'anglais, le français et l'allemand en Contraste*. Bern: Peter Lang.

Corréard, M.-H. (2005) Bilingual lexicography. In K. Brown (ed.) *Encyclopedia of Language and Linguistics*, 2nd Edn., Vol. 1, 787 796. Oxford: Elsevier.

Hutchins, J. (2005) 'Example-based machine translation: a review and commentary'. *Machine Translation* 19(3-4): 197 211.

Krishnamurthy, R. (2005) Corpus lexicography. In K. Brown (ed.) *Encyclopedia of Language and Linguistics*, 2nd Edn., Vol. 3, 250 254. Oxford: Elsevier.

Laviosa, S. (2002) *Corpus-based Translation Studies: Theory, Findings, Applications.* Amsterdam, Rodopi.

Mortelmans, T. (2004) 'The status of the German auxiliary *werden* as a "grounding predication"'. In O. Letnes and H. Vater (eds.), *Modalität und Übersetzung / Modality and Translation.* (FOKUS: Linguistisch-Philologische Studien 29), 33 56. Trier: Wissenschaftlicher Verlag.

Roberts, R.P. (1996) Parallel text analysis and bilingual lexicography. Available on the web: http://www.dico.uottawa.ca/articles-fr.htm. Accessed July 2008.

Roberts, R.P. & Cormier, M.C. (1999) L'analyse des corpus pour l'élaboration du Dictionnaire canadien bilingue. Available on the web: http://www.dico.uottawa.ca/articles-fr.htm. Accessed July 2008.

Salkie, R. (2002). 'Two types of translation equivalence'. In B. Altenberg & S. Granger (eds.), *Lexis in contrast*, 51 7. Amsterdam: John Benjamins.

Steinberger, R. (2008) The JRC-Acquis Multilingual Parallel Corpus. Available on the web: http://langtech.jrc.it/JRC-Acquis.html. Accessed July 2008.

Tiedeman, J. (2007) OPUS - an open source parallel corpus. Available on the web: http://urd.let.rug.nl/tiedeman/OPUS/. Accessed July 2008.

Véronis, J. (2000) *Parallel Text Processing: Alignment and Use of Translation Corpora.* Dordrecht: Kluwer.

Zanettin, F. (2001) ‘Swimming in Words: Corpora, Translation, and Language Learning’. In Aston, G. (ed). *Learning with corpora*, 177-197. Houston, TX: Athelstan.