Meta
Journal des traducteurs
Translators' Journal

Constructing a Large-Scale English-Persian Parallel Corpus

Tayebeh Mosavi Miangah

Volume 54, Number 1, janvier 2009

URI: https://id.erudit.org/iderudit/029804ar
DOI: https://doi.org/10.7202/029804ar

See table of contents

Publisher(s)

Les Presses de l'Université de Montréal

ISSN

0026-0452 (print)
1492-1421 (digital)

Explore this journal

Cite this note

Mosavi Miangah, T. (2009). Constructing a Large-Scale English-Persian Parallel Corpus. *Meta*, *54*(1), 181–188. https://doi.org/10.7202/029804ar

Article abstract

In recent years the exploitation of large text corpora in solving various kinds of linguistic problems, including those of translation, is commonplace. Yet a large-scale English-Persian corpus is still unavailable, because of certain difficulties and the amount of work required to overcome them.

The project reported here is an attempt to constitute an English-Persian parallel corpus composed of digital texts and Web documents containing little or no noise. The Internet is useful because translations of existing texts are often published on the Web. The task is to find parallel pages in English and Persian, to judge their translation quality, and to download and align them. The corpus so created is of course open; that is, more material can be added as the need arises.

One of the main activities associated with building such a corpus is to develop software for parallel concordancing, in which a user can enter a search string in one language and see all the citations for that string in it and corresponding sentences in the target language. Our intention is to construct general translation memory software using the present English-Persian parallel corpus.

ÉTUDES ET PROSPECTIVES

Constructing a Large-Scale English-Persian Parallel Corpus

RÉSUMÉ

Au cours des dernières années, l'exploitation de grands corpus de textes pour résoudre des problèmes linguistiques, notamment des problèmes de traduction, est devenue une pratique courante. Jusqu'à récemment, aucun corpus bilingue anglais-persan à grande échelle n'avait été constitué, en raison des difficultés qu'implique une telle entreprise.

Cet article présente un projet réalisé en vue de colliger des corpus de textes numériques variés, tels que des documents du réseau Internet, avec le moins de bruit possible. L'utilisation d'Internet peut être considérée comme une aide précieuse car, souvent, il existe des traductions antérieures qui sont déjà publiées sur le Web. La tâche consiste à trouver les pages parallèles en anglais et en persan, à évaluer la qualité de leur traduction, à les télécharger et à les aligner. Le corpus ainsi obtenu est un corpus ouvert, soit un corpus auquel de nouvelles données peuvent être ajoutées, selon les besoins.

Une des principales conséquences de l'élaboration d'un tel corpus est la mise au point d'un logiciel de concordance parallèle, dans lequel l'utilisateur pourrait introduire une chaîne de caractères dans une langue et afficher toutes les citations concernant cette chaîne dans la langue recherchée ainsi que des phrases correspondantes dans la langue cible. L'étape suivante serait d'utiliser ce corpus parallèle pour construire un logiciel de traduction générale.

Le corpus bilingue aligné se trouve être utile dans beaucoup d'autres cas, entre autres pour la traduction par ordinateur, pour lever les ambiguïtés de sens, pour le rétablissement des données interlangues, en lexicographie ainsi que pour l'apprentissage des langues.

ABSTRACT

In recent years the exploitation of large text corpora in solving various kinds of linguistic problems, including those of translation, is commonplace. Yet a large-scale English-Persian corpus is still unavailable, because of certain difficulties and the amount of work required to overcome them.

The project reported here is an attempt to constitute an English-Persian parallel corpus composed of digital texts and Web documents containing little or no noise. The Internet is useful because translations of existing texts are often published on the Web. The task is to find parallel pages in English and Persian, to judge their translation quality, and to download and align them. The corpus so created is of course open; that is, more material can be added as the need arises.

One of the main activities associated with building such a corpus is to develop software for parallel concordancing, in which a user can enter a search string in one language and see all the citations for that string in it and corresponding sentences in the target language. Our intention is to construct general translation memory software using the present English-Persian parallel corpus.

MOTS-CLÉS/KEYWORDS

alignment, concordancing, parallel corpus, translation memory

1. Introduction

A corpus is simply defined as a large collection of linguistic evidence, mainly naturally occurring data either written texts or a transcription of recorded speech. Corpora can be exploited for a range of research purposes in a number of disciplines. In recent years large monolingual, comparable and parallel corpora have played a crucial role in solving problems of computational linguistics, such as part-of-speech tagging (Brill 1995), word sense disambiguation (Mosavi Miangah and Delavar Khalafi 2005), language teaching (Aston 1997; Leech 1997; Nesselhauf 2004), phrase recognition (Cutting *et al.* 1992), information retrieval (Braschler and Schauble 2000) and statistical machine translation (Brown *et al.* 1990). Corpus-based linguistics has provided accurate descriptions of languages, and these descriptions of structure and use have many applications in theoretical linguistics, translation tasks and language teaching. There are different kinds of corpora for different kinds of applications. Parallel corpora have texts in one language with the corresponding translations in some other language or languages.

In this paper we present our work on constructing and using English-Persian parallel corpora to support research in fields such as English-Persian bilingual lexicography, developing translation memory software, English-Persian cross-language information retrieval, and statistically-based machine translation from English into Persian. We have tried to design a program to automatically align the corpus at sentence level, but here, our main concern is to introduce the procedures and techniques for developing an online parallel corpus of English and Persian texts in various domains. This corpus is extendable: more and more parallel sentences in the languages may be added, and it will be provided free to those interested in language and translation matters,

especially translation trainees. One of the main activities associated with building such a corpus is developing software for parallel concordancing, in which a user can enter a search string in one language and see all citations for that string in the search language as well as corresponding sentences in the target language. Aligned bilingual corpora have in fact proved useful in many tasks, including machine translation (Brown *et al.* 1990; Sadler 1989), sense disambiguation (Brown *et al.* 1991a; Dagan *et al.* 1991; Gale *et al.*1991), cross-language information retrieval (Davis and Dunning 1995; Landauer and Littman 1990; Oard 1997) and bilingual lexicography (Klavans and Tzoukermann 1990); Warwick and Russell 1990).

Our underlying hypothesis is that exploiting the present English-Persian parallel corpus in building a translation memory system – a database of previously translated units (sentences, phrases and words), along with a set of other translation tools – will result in improvements to the translation process, in speed, consistency and quality.

Many researchers have tried to create corpora of different sizes for different purposes. Resnik (1998), for instance, presents a method of automatically finding parallel translated documents on the Web, which he calls STRAND (Structural Translation Recognition for Acquiring Natural Data). The method is conceptually simple, fully language independent, and scalable. Given a pair of languages, a *candidate generation* module first generates pairs (Uniform Resource Locator URL1, URL2) identifying World Wide Web pages that may be parallel translations. Next, a language independent *candidate evaluation* module behaves as a filter, keeping only those candidate pairs that are likely to actually be translations. Optionally, a third module for *language independent filtering* applies additional filtering criteria that depend upon language-specific resources. The end result is a set of candidate pairs that can reliably be added to the Web-based parallel corpus for the two languages in question (Resnik 1998).

Le Sun, Song Xue, Weimin Qu, Xiaofeng Wang, and Yufang Sun report on their attempts to construct a large-scale Chinese-English parallel corpus, totaling more than 500,000 sentence pairs. Then they attempt to align their corpus at the sentence level by an "improved" length-based algorithm (described in Sun *et al.* 1999). They also design a bilingual concordance tool for discovering facts during the translation between Chinese and English. Besides listing the keywords with the contexts in which they appear, corresponding translated sentences are also presented (Sun, Lee *et al.* 2002).

Chris Callison-Burch and Miles Osborne introduce two methods for the automatic creation of parallel corpora, as opposed to attempting to gather new translations from the Web. In fact they try to examine the use of existing translations as a resource to bootstrap more training data, and to create data for new language pairs (Callison-Burch and Osborne 2003). Their method seems to be applicable and useful for language pairs for which extensive parallel corpora do not exist.

A typical, and of course the earliest, parallel corpus is the Canadian Hansard corpus, consisting of transcripts of debates from the Canadian Parliament in the country's official languages, English and French. However, in recent years there has been an increase in the number of parallel corpora for various language pairs and of various sizes. For example, an English-Norwegian Parallel Corpus (ENPC) has been created at the University of Oslo (Johansson 1997), and the Translation Corpus of English and German has been compiled at the Technical University of Chemnitz-Zwickau in Germany. The Japanese NTT Communication Science Laboratories have carried out research into tagging and aligning several collections of Japanese and English texts, while the Thai Internet Education Project has collected and studied on Thai/English parallel texts and developed a toolkit to work on them. (Fan and Xunfeng 2002).

Bilingual corpora for high density languages such as English or French are very extensive, and the results are encouraging because of the easy accessibility of the texts in these languages in digital form, including Websites. However, when a low or medium density language such as Persian is one of the languages involved in a bilingual corpus, the problem is much more difficult because of the shortage of digitally stored materials and detectable parallel pages on the World Wide Web.

2. The collection of parallel texts

There are guidelines for constructing a parallel corpus, e.g., a corpus should be balanced and representative; i.e., it should contain texts from different domains and different genres in reasonable proportions. It should be a reasonable reflection of language use. In constructing an English-Persian corpus we found it difficult to build a perfectly balanced one, since there are few electronic English-Persian bilingual texts available. We, therefore, decided to collect as many text pairs as we could, provided they were of good quality. We were not able to use sampling techniques for the same reason, so the full texts were included in the corpus.

Unlike Resnik (1998), who used an automatic method for extracting parallel material from the Web, we tried to do the task manually. We tried to gather pages from the Web that were potential translations of each other by searching documents

in one language which have links containing the name of another language. For instance, if an English Web page contains a link such as "Persian page" or "Persian version," the page associated with this link is taken to be a potential translation of that English page.

The Internet is a cumulative source of language data potentially available to everyone, everywhere, for every purpose and in great quantity. However, there are many problems in extracting parallel corpora in English and Persian from the Web, such as the following. As we mentioned earlier, the number of parallel English-Persian pages on the Internet is relatively small. This problem is compounded by the fact that some of these pages are not downloadable because of their special formats, particularly in the Persian side. That is, not all Persian translations of English pages can be copied or downloaded. Some of them are images and others are written in a special type of Portable Document Format (PDF) which cannot be converted into Text format. For example, in the present experiment, about ten percent of the pages obtained from the Web in Persian for which an English translation was available could not be entered into the bilingual corpus for this reason. The other problem encountered when extracting parallel pages from the Web was that some texts in one language were not exact translations of the other language. Quite frequently we encountered translations with some omissions or some insertions, that is, they were unduly freely translated. In such cases it took a long time to distinguish passages constituting exact translations and to separate them from the noisy ones. Lastly, there were some parallel pages on the Web which we could not access because of subscription restrictions. That is, certain parallel pages, such as scientific articles and legal documents, accessible only to members.

Despite these problems, we collected as many well-matched texts in English and Persian as possible. Most Web pages extracted for this purpose had Hyper Text Markup Language (HTML) format. However, some of them were PDF, in which case they had to be converted to text format, which in turn has its own difficulties.

In sum, although the availability of bilingual texts involving Persian is subject to some limitations due to the low density of this language around the world and the unavailability of texts in some specific genres and domains, we succeeded in collecting a relatively large number of texts, totaling over 4,860,000 words in English and Persian. The procedure for finding parallel Web pages is as follows. First we write the query "Persian version" or "to Persian" or "click here for Persian version" in a search engine such as Google, Yahoo or AltaVista, through which many different links are released. Then all of the links in which the query words appear are checked to find the pages in which "Persian version" or "to Persian" acts as a link to the Persian translation of that page, most probably originally in English. These pages can be considered as parallel. In the case of searching on a Persian search engine, such as GooglePersian, or Irandoc the query may be as follows: "English version" or "to English" or "click here for English version." Applying this procedure, we were able to extract English and Persian Web pages which were translations of each other. The vast majority of these texts were collected from the Internet and cover a variety of domains, such as current affairs, literature, interviews, instruction manuals, religion, pedagogy, and offline digital material. Table 1 presents the different types of texts and their absolute and relative sizes in our English-Persian parallel corpus.

TABLE 1

Distribution of text types in the English-Persian parallel corpus

Text types	No. of files	words (thousand)	Percent
news	165	499	10.25
reports	102	710	14.58
articles	53	266	5.46
literature	372	1104	22.68
interviews	54	97	1.99
courses	42	65	1.34
offline digital material	265	973	19.98
manuals	142	132	2.71
religious texts	298	1023	21.01
Total	**1493**	**4869**	**100**

As Table 1 shows, the majority of texts are literary ones. Our literature sources were selected from various kinds of short stories, especially English translations of Saadi's *Gulistan*, a large collection from a well-known Iranian poet, Saadi (originally in Persian), as well as translations of works of other poets and writers, such as Shakespeare and Tolstoy, which have been translated into Persian. These bilingual materials can be found through the Web with sufficient time and effort. As for religious texts, the Bible and the Quran are the two main sources for which there are various kinds of translations in different languages. Many of these translations are freely accessible on the Internet.

There are certain Websites specifically designed to teach English to Persian speakers or Persian to English speakers. On such sites, different sentences, along with their translations, are available and intended to teach translation methods or certain grammatical points. The offline digital materials

are mainly extracted, the translation assignments prepared by translation trainees and checked for accuracy by their instructors.

3. Corpus preparation

Raw texts extracted from the above sources must be preprocessed to enter the corpus. The preprocessing database is composed of about 1500 files, each containing at least 57 words and at most 30,699 words. Downloading, format conversion, and text normalization are among the labor-intensive steps in preparing the corpus. Some HTML pages which were irrelevant, as well as all figures, tables and pictures, were removed from the texts before they entered into the corpus. Moreover, some parallel pages in PDF format were not convertible to Text format and hence were discarded from the corpus. In some cases where a sentence or part of a sentence was not translated, untranslated parts were deleted. Moreover we tried to manipulate the corresponding paragraphs such that they consisted of corresponding sentences. The resulting movements of sentences or of parts thereof to other sentences in the translations was another problem here. It was resolved by personal checking and manual correction before the texts were entered into the corpus. After verification, they were uniformly encoded into Extendible Mark-up Language (XML) format (using XML encoding tools), in order for the corpus to be application-independent and easier to exchange via the Internet. The programming language was ASP, VBSCRIPT, using the databank ACCESS. Persian is of course a language among those with rich morphology. We have a program for stemming morphological variants of Persian words (Mosavi Miangah 2006), but we have not incorporated this program into our corpus because it requires revisions. This will certainly be done eventually.

4. Aligning the parallel corpus

The very first requirement for any bilingual corpus is alignment. Alignment can be done at the paragraph, sentence, or word level. Paragraph alignment is a rather easy task, since the boundaries are usually clearly marked, except for the merging or splitting of some paragraphs in translations, but it seems not to be very useful for further exploitation of parallel corpora in research. Word alignment is the process of locating corresponding word pairs in two languages. It is in fact a very difficult task, and requires complicated algorithms. So most parallel corpora are aligned in terms of sentences. Reviewing the literature on aligning parallel corpora, we found four main approaches to the problem of alignment at the sentence level: word length-based (Gale and Church 1991), character length-based (Brown *et al.* 1991), dictionary- or translation-based (Chen 1993, Melamed 1996, Moore 2002), and partial similarity-based (Simard and Plamondon 1998). In this experiment, the alignment of sentences was done entirely manually. Although we could have used automatic methods, we preferred to align sentences manually in order for the accuracy to reach 100%. The present corpus is intended to be used in tasks for which high accuracy in aligning is crucial, as will be explained in the next section.

The corpus software has several components. One enables us to add more sentences with their translations. No limitation was placed on the length of sentences. They vary in length from two to over fifty words. Two other components are an option for editing previous records and one for deleting unwanted records (previously entered) from the corpus. The system has two types of search, simple and advanced. The simple search enables the user to enter a word, a phrase or a sentence in English or Persian as a query. The user then has three choices depending on whether the desired results must correspond exactly, or whether they must correspond approximately, or both. In the first case, the result corresponds exactly to the query. In the second case, it contains similar expressions (in terms of graphic forms) to these in the query. In the third case, every word of the given expression occurring in corpus sentences is shown. It should be noted that the searches are carried out simultaneously in the English and Persian texts. In the advanced search, the user can search any expression separately in the English or Persian corpus, and the results will be shown only in the selected language. In this sense, the corpus acts as two monolingual corpora.

Searching can be either monolingual or bilingual on request. Moreover, the results of each query can be selected, copied, pasted, and saved for further applications in other situations. So far the corpus has been aligned only at the sentence level, but our intention is to align the entire corpus at the word level, using automatic tools.

5. Finding translation equivalents

One of the main applications of parallel corpora is to find different possible equivalents of certain words or collocations. That is, aligned translation units are simply displayed on the screen, offering the translator a range of similar contexts from a corpus of past translations. In some cases we may refer to a parallel corpus to verify the equivalent(s) provided by bilingual dictionaries, since it is believed that parallel corpora provide information that bilingual dictionaries do not usually contain.

Using a bilingual dictionary for selecting a translation equivalent, the translator will decide on the appropriateness of different possible equivalents, based on their definitions or the few examples given by the dictionary, while a parallel corpus offers the best possible translation equivalent, based on real-world evidence gained from past translations. It is said that dictionaries take a synthetic approach to lexical meaning (by means of definition), while parallel corpora take an analytic approach (by means of multiple contexts).

Finding appropriate and natural equivalents for different types of collocations is a difficult task, especially in a non-native language, and parallel corpora can be of great help in this respect. In some cases the aim is to confirm the translation equivalent of a collocation which is the same as the majority of occurrences. As Figure 1 shows, the most frequently occurring equivalent for the collocation "Web hosting" is میزبانی وب , with a frequency of 12 in the corpus.

However, there are other cases in which the translator needs to see all possible equivalents of a certain expression in each language, and then make the best decision, based on similar or identical contexts in which the expression is found. Four different English translations of the word مرجع have been displayed in Figure 2.

This shows the actual translations of the same expression chosen by translators according to the

FIGURE 1

Display of some lines generated by our corpus for the search word "Web hosting"

Quick Search (*) w eb hostir GO! Show All Advance Search

Add

entext (*)	frtext (*)			
After registering your domain name, you need to select a **Web hosting** service	پس ازثبت نام مورد نظر خود، نیاز به اجاره یک سرویس میزبانی وب دارید	View	Edit	Delete
Now that you have registered your domain name, and chosen your **Web hosting** package, you need to design your web site's pages	حال که هم نام دامنه خود و هم سرویس میزبانی وب مورد نیاز خود را انتخاب نمودهاید، نوبت به طراحی صفحات سایت شما میرسد	View	Edit	Delete
First, you should register a domain name for your web site. Then, you need to choose a **Web hosting** service, and at last, you need to design the pages for your site	ابتدا باید برای سایت خود، یک نام دامنه ثبت نموده، سپس یک سرویس میزبانی وب را برای آن در نظر گرفته، و در آخر باید صفحات سایت شما طراحی گردد	View	Edit	Delete
Resellers **Web Hosting** Packages	خدمات میزبانی وب برای عمدهفروشان	View	Edit	Delete
Web Hosting Account Types	انواع سرویسهای میزبانی وب	View	Edit	Delete
All of our **Web hosting** packages include the following features	تمام بسته های میزبانی وب ما شامل امکانات و مشخصات زیر میباشند	View	Edit	Delete
All of the above services can be added to any of our **Web hosting** packages	خدمات فوق قابل اضافه شدن به تمام سرویسهای میزبانی وب میباشند	View	Edit	Delete
Choosing the **Web hosting** package	انتخاب نوع سرویس میزبانی وب	View	Edit	Delete
Different **Web hosting** account types offered by Farda Technology are listed below.	انواع سرویسهای میزبانی وب فردا به شرح زیر میباشد	View	Edit	Delete
It enables our customers to use our **Web hosting** services in a more secure manner	امکان استفاده از خدمات میزبانی وب به صورت ایمن تری را در اختیار مشتریان ما قرار میدهد	View	Edit	Delete
Farda Technology is in the process of implementing a free advertisement plan for its **Web hosting** customers	شرکت فردا، دستاندرکار بررسی و پیاده سازی یک طرح رایگان برای تبلیغات برای مشتریان خدمات میزبانی وب خود میباشد	View	Edit	Delete
This provides the best **Web hosting** services with the lowest prices	این بهترین خدمات میزبانی وب را با ازرانترین قیمت فراهم میآورد	View	Edit	Delete

Figure 2

Display of some lines generated by our corpus for the search word "مرجع"

Quick Search (*) مرجع GO! Show All Advance Search

Add

entext (*)	frtext (*)			
As such, this work will appeal to the specialist as well as the general reader, and it will undoubtedly prove to be an invaluable reference source for all teachers and students concerned with Ismaili history and thought for many years to come.	از این رو، این اثر هم برای متخصصین و هم برای خوانندگان عام اثری جالب خواهد بود و بدون تردید تا سالیانی دراز **مرجعی** گرانبها برای تمامی معلمین و دانشجویانی که در زمینه‌ی تاریخ و تفکر اسماعیلی کار می‌کنند، به شمار خواهد آمد .	View	Edit	Delete
Eagle's Nest contains a wealth of information and resources; it is essential reading for scholars, students and others with an interest in medieval or Ismaili history.	آشیانه‌ی عقاب سرشار از اطلاعات و منابع است؛ این کتاب **مرجعی** ضروری برای محققان، دانشجویان و دیگر کسانی است که به تاریخ قرون وسطا یا تاریخ اسماعیلیه علاقه‌مند هستند .	View	Edit	Delete
Peter Willey is an authority on the Ismaili castles of Iran and Syria, spending nearly a lifetime discovering and investigating them.	پیتر ویلی **مرجعی** در زمینه‌ی دژ های اسماعیلی ایران و سوریه است و تقریبا عمر خود را به کشف و تحقیق درباره‌ی آن‌ها گذرانده است .	View	Edit	Delete
The students engaged with the contributions of great Muslim thinkers such as al-Khwarizmi, Avicenna, and Nasir al-Din Tusi, whose works are still a point of reference for scientists and scholars around the world.	دانشجویان با سهم اندیشمندان بزرگ مسلمان از قبیل خوارزمی، ابن سینا و نصیر الدین طوسی پرداختند که آثارشان هنوز **مرجعی** برای دانشوران و پژوهش‌گران در اطراف جهان به شمار می‌رود .	View	Edit	Delete

context, concrete data which is not found directly in bilingual dictionary.

6. Other applications

Parallel corpora are invaluable resources for natural language processing as well as computational linguistics. What follows are examples of the applications of such corpora.

6.1. Translation memory systems

Translation memory can be considered a specialized kind of parallel corpus, in which parallel texts in a database are stored and then an expression (a word, phrase or sentence) can be retrieved, with its previous translation equivalents, when it needs to be translated again. Translation memory systems are new essential tools for translators and compiling large comparable corpora is the first step in developing them. Sentences and their translations can be easily extracted from the corpus and put into the translation memory. Of course, parallel corpora compiled for specific domains are the most useful, but it is not easy to create domain-specific bilingual corpora of very large sizes. Our intention is to construct general translation memory software, using the present English-Persian parallel corpus. The goal is a kind of parallel concordance from which translators can automatically retrieve fragments of past translations that totally or partially match a current segment to be translated.

6.2. Statistical machine translation

A further application of parallel corpora is as the main database for statistical machine translation systems. In fact, the training corpus for all statistical machine translation systems is a bilingual, sentence-aligned corpus (Brown *et al.* 1993). Brown *et al.* have produced a probabilistic machine translation system trained on an aligned French-English corpus. Their system tries to find the most probable translation sentence in the target language, given a sentence in the source language, using a trigram language model based on three-word sequences and a translation model derived from the word-level alignment of their English and French parallel corpora. That is, their system relies on parallel corpora sufficient to train the translation model (Brown *et al.* 1991).

6.3. Lexicography

The corpus can be used to find terminology, phraseology and collocations. It is parallel corpora from which empirical data are extracted for study. Lexicography of course involves improving, updating and classifying words or expressions, as well as gathering collocations, sub-categorization requirements, and definitions. The importance of corpus studies in dictionary compiling can be deduced from the number of dictionary publishers investing in corpus technology, and it is easy to find corpus-based monolingual as well as bilingual dictionaries. Bilingual lexicographers can also find possible equivalent(s) of expressions (e.g., slang, neologisms, etc.) that are not in existing bilingual dictionaries.

An aligned parallel corpus can be used to retrieve multiword units, making it easier to build a multilingual dictionary or specialized terminology databases. The availability of different types of corpora has eased the task of lexicographers, in that they can access all the examples of the usage of a word or phrase in a few seconds. Consequently, compiling and revising dictionaries is theoretically at least more precise and faster.

6.4. Cross-language information retrieval

Cross-language information retrieval involves a query in one language and searching document collections in one or more other languages. Traditional approaches to cross-language information retrieval used bilingual dictionaries ("dictionary-based" methods), but experiments show that using bilingual corpora for automated extraction of equivalents results in superior performance (Savoy 2003). There are several cases in which users need a reliable retrieval system. Some users formulate a query in one language and want their retrieved documents to be in another language. Some may need information in various languages. They usually can read documents in other languages in which they cannot formulate a query (Oard and Resnik 1999). The need therefore arises for a parallel corpus for cross-language information retrieval.

6.5. Language learning

Examples from a corpus provide learners with the kinds of sentences that they will encounter when using the language in real life situations. By searching for key words in context, the learners can induce rules of grammar or usage and lexical features. On the basis of their observation of patterns in authentic language they can question putative rules. In the case of vocabulary, they can be critical of dictionary entries. They can also compare texts produced by native and non-native speakers of a language, as to the position of lexical items in context, by means of a concordancer. Furthermore, learners can use a corpus to work with multi-meaning and multi-usage words, in that they are given concordances for a single word, and they can try to group them according to usage.

Teachers also can benefit from parallel corpora. Using the examples originally taken from a variety of sources, they can construct appropriate exercises in the form of cloze tests, for example, targeting collocations, points of grammar, or vocabulary usage. After all, they can find the most appropriate equivalent for certain words or collocations.

7. Conclusion

In this paper we have described a general method for collecting, building, and aligning a parallel corpus for English and Persian. The corpus so created is open; that is, more material can be added as the need arises. Naturally, the richer the corpus is in terms of the volume of data and its variety, the more useful it will be for solving linguistic problems. This is a work in progress and there is great room for enhancing the potential of the corpus. New ways may be found to obtain more parallel texts. As we have mentioned, the present corpus will be aligned at word level in the near future, making the corpus a database for a translation-memory system. At that stage the corpus will be of even greater help to translators between English and Persian. A parallel concordance tool is a future goal in this direction.

Tayebeh Mosavi Miangah
Payame Noor University of Yazd, Yazd, Iran
mosavit@pnu.ac.ir

ACKNOWLEDGEMENTS

The research project reported here is partly supported by the Research Affairs Office of Payame Noor University of Iran.

REFERENCES

Aston, G. (1997): "Enriching the learning environment: Corpora in ELT," in Wichmann, A., Fligelstone, S.

Braschler, M. and P. Schauble (2000): "Using corpus-based approaches in a system for multilingual information retrieval," *Information Retrieval* 3, pp. 273-284.

Brill, E. (1995): "Unsupervised learning of disambiguation rules for part of speech tagging," *2nd Workshop on Large Corpora*, Boston, USA.

Brown, P., Cocke, S., Della Pietra, V., Della Pietra, S., Jelinek, F., Lafferty, J., Mercer, R. and P. Roosin (1990): "A Statistical Approach to Machine Translation," *Computational Linguistics* 16-2, pp. 79-85.

Brown, P., Lai, J. C., and R. L. Mercer (1991): "Aligning Sentences in Parallel Corpora," proceedings of the *29th Annual Meeting of the Association for Computational Linguistics* (ACL-91), pp. 169-176.

Brown, P., Della Pietra, S., Della Pietra, V., and R. Mercer (1993): "The mathematics of machine translation: parameter estimation," *Computational Linguistics* 19, pp. 263-312.

Callison-Burch, C. and M. Osborne (2003): "Bootstrapping Parallel Corpora," proceedings of the HLT-NAACL 2003 Workshop *Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, pp. 44-49.

Chen, S. (1993): "Aligning Sentences in Bilingual Corpora Using Lexical information," proceedings of the *31st Annual Meeting of the Association for Computational Linguistics* (ACL-93), pp. 9-16.

Cutting, D., Kupiec, J., Peterson, J. and P. Sibun (1992): "A practical part of speech tagger," proceedings of the *3rd Conference on Applied Computational Linguistics*, pp. 133-140.

Dagan, I. and A. Itai (1994): "Word sense disambiguation using a second language monolingual corpus," *Association for Computational Linguistics* 20-4, pp. 563–596.

Davis, M. and T. Dunning (1995): "A TREC evaluation of query translation methods for multilingual text retrieval," proceedings of the *4th Text Retrieval Conference (TREC-4)*, NIST, pp. 483-497.

Fan, M. and X. Xunfeng, (2002): "An evaluation of an online bilingual corpus for the self-learning of legal English," *System* 30-1, pp. 47-63.

Gale, W. and K. Church (1991) "Identifying Word Correspondences in Parallel Text," *Fourth Darpa Workshop on Speech and Natural Language,* Asilomar.

Johansson, S. (1997): "Using the English Norwegian parallel corpus – a corpus of contrastive analysis and translation studies," Lewandowska-Tomaszczyk, B. and J. Melia (eds.), PALC- 97 *Practical Applications in Language Corpora*, Lodz University Press, pp. 282–296.

Klavans, J. L. and E. Tzoukermann (1989): "Movement Verbs in English-French Translation: A Corpus-based Approach," *Proceedings of the Sixth Israeli Conference of Artificial Intelligence and Computer Vision*, Tel Aviv, Israel.

Landauer, T. K. and M. L. Littman (1990): "Fully automatic cross-language document retrieval using latent semantic indexing," proceedings of the *6th Annual Conference of the UW Centre for the New Oxford English Dictionary and Text Research*, UW Center for the New OED and Text Research, Waterloo, Ontario, pp. 31-38.

Leech, G. (1997): "Teaching and language corpora: a convergence," Wichmann, A., Fligelstone, S., McEnery, T. and G. Knowles (eds.), *Teaching and language corpora*, New York, Addison Wesley Longman, pp. 1-23.

McEnery, T. and G. Knowles (eds.), *Teaching and Language Corpora*, New York, Addison Wesley Longman, pp. 51-64.

Melamed, I. D. (1996): "Automatic Detection of Omissions in Translations," *Proceedings of the 16th International Conference on Computational Linguistics*, Copenhagen, Denmark

Moore, R. C. (2002): "Fast and accurate sentence alignment of bilingual corpora," Proceedings of the *5th AMTA Conf: Machine Translation: From Research to Real Users*, Langhorne, PA, Springer, pp. 135–244.

Mosavi Miangah, T. and A. Delavar Khalafi (2005): "Word sense disambiguation using target language corpus in a machine translation system," *Literary and Linguistic Computing* 20-2, pp. 237-249.

Mosavi Miangah, T. (2006): "Automatic lemmatization of Persian words," *Journal of Quantitative Linguistics* 13-1, pp. 1-15.

Nesselhauf, N. (2004): "Learner corpora and their potential for language teaching," in McHardy Sinclair, J. (ed.), *How to use corpora in language teaching*, Amsterdam, John Benjamins, pp. 125-152.

Oard, D. W. (1997): "Cross-language text retrieval research in the USA," *3rd DELOS Workshop*, European Research Consortium for Informatics and Mathematics.

Oard, D. W. and P. Resnik (1999): "Support for interactive document selection in cross-language information retrieval," *Information Processing and Management* 35-4, pp. 363-379.

Resnik, P. (1998): "Parallel strands: A preliminary investigation into mining the Web for bilingual text," proceedings of the *3rd Conference of the Association for Machine Translation in the Americas* (AMTA-98), in *Lecture Notes in Artificial Intelligence,* Langhorne, pp. 28-31.

Resnik, P. (1999): "Mining the Web for bilingual text," Proceedings of the *37th Meeting of the ACL*, Maryland, pp. 527-534.

Sadler, V. (1989): "Working with analogical semantics: disambiguation techniques," *DLT*, Dordrecht, Foris Publications.

Savoy, J. (2003): "Cross-language information retrieval: experiments based on CLEF 2000 corpora," *Information Processing and Management* 39, pp. 75-115.

Simard, M. and P. Plamondon (1998): "Bilingual Sentence Alignment: Balancing Robustness and Accuracy," *Machine Translation* 13-1, pp. 59-80.

Sun, L., Du, L., Sun, Y. and Y. Jin (1999): "Sentence Alignment of English-Chinese Complex Bilingual Corpora," proceedings of the workshop *MAL'99*, pp. 135-139.

Sun, L., Xue, S., Qu, W., Wang, X. and Y. Sun (2002): "Constructing of a Large-Scale Chinese-English Parallel Corpus," Proceedings of the *3rd workshop on Asian Language Resources and International Standardization*, pp. 1-8.

Warwick, S. and G. Russell (1990): "Bilingual Concordancing and Bilingual Lexicography," EURALEX 4th International Congress, Málaga, Spain.