Expanding Lexical Coverage of Parallel Corpora for the EBMT Approach

Jeffrey Allen and Christopher Hogan*
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213 USA
{jeffa, chogan}@cs.cmu.edu

Abstract

In this paper, we discuss a method called TEXT-R that improves lexical coverage in creating parallel corpora that are to be subsequently implemented in an Example-Based Machine Translation (EBMT) system. First, we explain the purpose and importance of the EBMT approach. Second, we indicate how low-density languages can benefit from rapid corpora development using our method as compared to other corpora expansion techniques. Third, an evaluation of these various methods, based on tests for current lexical coverage, projected lexical coverage, and word frequency distribution, clearly indicates that TEXT-R is a valid mechanism for EBMT corpus development. We conclude that this method will improve the quality and speed with which parallel corpora are created, both for our specific purposes and possibly for other corpus-based applications.

Background

EBMT and Rapid-Deployment

The DIPLOMAT project (Frederking and Brown, 1996; Frederking et al., 1997), using Example-Based Machine Translation (EBMT) technology (Brown, 1996; Sumita and Iida, 1991; Sato, 1992; Nirenburg et al., 1994), is designed to test the feasibility of rapid-deployment translation systems. By rapid-deployment we mean developing a machine translation (MT) system that performs initial translations at a useful level of quality between a new language and English within a matter of days or weeks, with continual, graceful improvement to a good level of quality over a period of months. EBMT (also known as translation by analogy (Nagao, 1984)) is an MT technique which builds on the notion of translation memory (TM), a technique used to improve the performance of human translators by capitalizing on the reuseability of already translated sentences, phrases and terms (EAGLES, 1995; Heyn, 1995). EBMT translates by matching the text to be translated against a parallel corpus (bi-text) consisting of examples (usually sentences) in both the source and target languages. After suitable examples are found, additional processing transforms non-exact matches into correct translations. In the simplest case, where the input exactly matches some example in the corpus, no processing is necessary. In an inexact match, EBMT processing may either translate only the parts of the example which match, or it may modify the translation from the corpus in order to make it a correct translation of the input sentence. In contrast to other MT designs (*e.g.* Statistical MT) which also use parallel corpora to perform translation, EBMT does not rely on the relative frequencies of the words in the corpus.

Corpus Development

The initial stage of DIPLOMAT subprojects includes the creation of a large parallel corpus. The procedure normally followed for creating a large, general, balanced monolingual corpus (*e.g.* London-Oslo-Bergen corpus, Brown Corpus, Collins COBUILD, British National Corpus) (Collins COBUILD, 1998; Burnage and Baguley, 1997; Francis and Kucera, 1979) of written and/or spoken data is the following: 'according to standard bibliographic practice ... Dewey Decimal codes were used to balance the domain targets agreed in the corpus design' (Burnage and Baguley, 1997). Some have stated that 'the important point is that real language ... can be extensively analysed using a comprehensive and balanced computerised corpus' (Burnage and Baguley, 1997), yet computational methods are applied only after the collection of the data. It appears that in each case of large corpora collections mentioned above, samples were selected randomly from available sources according to pre-established domain and subject matters. While this method is possibly adequate for so-called "high-density" languages (i.e., those languages for which a large amount and variety of on-line resources are available), these data requirements limit investigation into the opposite kind of language, so-called "low-density" languages, such as Turkish, Kazakh and Swahili, for which the necessary written resources (*e.g.* monolingual and parallel corpora) are almost or completely non-existent. In order to facilitate MT for such languages, given the time constraints for the development of rapid-deployment systems, it is desirable to 1) reduce the data requirements of empirical techniques, and at the same time 2) develop efficient ways of producing parallel corpora. From initial testing on several languages (i.e. Croatian, Haitian Creole, Spanish)

*We wish to extend a special thanks to our translators, Brana Smailagic and Svetlana Vujanovic for their patience and perseverence in translating the myriads of articles we have given them.

we are already convinced that the EBMT approach is efficient in its use of data. We adhere to the suggestions that 'it is desirable to produce a "universal" lexical resource' and that 'the lexical knowledge base for an application needs to be tuned to the application' (Evans and Kilgarriff, 1995, p. 131). In other words, it is preferable to create a universal corpus and subsequently define categories, and their application, according to the content of the corpus that is acquired. We have thus chosen to design an automated process for building such a universal corpus. From our findings, we propose to discuss how to design better corpora for EBMT and why this is even more important for dealing with lower density languages.

In this paper we address the problem of designing a parallel corpus for general-purpose EBMT of English/Croatian. Croatian may be considered to be a medium-density language. Monolingual Croatian corpora and some English/Croatian parallel texts do exist but we have found no large-scale source of parallel corpora along the lines of the Canadian Hansards for English and French. For this reason, it is necessary to augment the few existing resources with additional translated examples. This paper discusses an efficient implementation of this process.

Driving Force for our Investigations

The reason for investigating the automation of lexical expansion comes from comments made by translators[1] at our institute who have been translating articles on a daily basis from English into Croatian since May 1996. During the first year, these articles were taken from a single domain, specifically that of the Balkans region. We chose this domain for two reasons: first, we thought that information about current events in the region would be most pertinent for the translation system, and second, that the translators would be most familiar with the Croatian terminology in this area. Near the end of the first year, the translators indicated that they were finding similar and exact sentences to translate over again. In order to avoid duplicating the translation tasks, we decided late in summer 1997 to seek out new texts by manually selecting articles from domains other than that of the Balkans (*e.g.* education, medicine, *etc*). After at least two months of expanding the corpus in this way, we found the hand selection of texts to be a very time consuming task. As a result, we developed in fall of 1997 a lexical search program called TEXT-R (Text Retrieval) which finds maximally different texts by searching for new words that are not found in the source language side of existing parallel corpora. From November 1997 until April 1998, we used TEXT-R on a regular basis to prepare a large English/Croatian parallel corpus for the DIPLOMAT machine translation (MT) project.

While there is much to be said for improving the efficiency of human translators, much of it has been mentioned elsewhere (Language Partners International, 1997; Allen, 1997). We specifically focus on the modifications to the content of a corpus that improve the quality and time factors involved in corpus building. This is possible since EBMT, as mentioned above, does not rely on frequency information. For other systems, in a context where word frequencies are important, the frequencies found in the corpus must correspond closely to those found in naturally occurring text and it is obviously undesirable to artificially modify them in the course of developing the corpus. In an EBMT corpus, however, it is indeed possible to modify these frequencies with few side-effects to the manual translation process or to the functionality of the EBMT system. We demonstrate a certain type of frequency modification in what follows and how this modification significantly improves the development of a parallel corpus.

Description of TEXT-R

EBMT requires at least one example of a word in order to translate that word. Additional examples may offer alternative translations, but are in general unnecessary. Because of this, our primary goal in developing the corpus is expanding lexical coverage. A second goal is to reduce the time required to develop a corpus to a given level of performance. This goal serves the end of rapid-deployment by making our translators more efficient: able to produce the same degree of coverage in a shorter period of time. In order to do this most effectively and efficiently, we have automated the process of finding new texts using the program we call TEXT-R.

TEXT-R simulates the human process of repeatedly selecting from among the English texts available for translation the one which contains the greatest number of words which are not already in the corpus. After one text is selected, the words it contains are added to the corpus, and the process is repeated to select the next text. New texts to translate are output in a particular order which guarantees that words occurring earlier in the order will not be considered as new words when selecting later texts. Texts are therefore translated in an order which maximizes vocabulary acquisition. Except for the specific order, the new texts were translated in the same way as all other texts in the existing corpus.

Before running TEXT-R, a certain amount of preprocessing is performed to remove headers and such nontext as pictures, and to undo MIME-encoding where applicable. Certain classes of words are eliminated entirely: we decided early on not to include words with any capitalized characters, in order to eliminate the possibility of choosing texts based on the proper names they contained. We also removed numbers for the same reason. In the end, only strings of lowercase characters, apostrophe (') or hypen (-) were counted as words for TEXT-R's measurements. Post-processing was performed to divide the articles into sentences and format them properly.

TEXT-R is similar to a host of other greedy algorithms that are used for balancing corpora, primarily in speech recognition (Fisher et al., 1986; Gibbon et al., 1997; Van Santen, 1992). The current application differs from these in that the optimization function is lexical coverage, rather than phonetic balance, and that the computation must be performed over the several

[1] Translator in this paper always refers to human translators. We will use different terminology to refer to Machine Translators.

[2] The larger number of Croatian word types is probably due to the fact that Croatian is more highly inflecting than English.

	old	continue	random	TEXT-R
articles	1771	27	27	27
paired lines	35078	601	735	768
English words (tokens)	845221	13404	13087	13353
English words (types)	27970	2968	3137	4120
Croatian words (tokens)	770786	12490	12624	12425
Croatian words (types)[2]	64959	4461	4746	5554

Table 1: Statistics of old and test corpora

thousand words present in a typical article rather than the several hundred phonemes in a typical sentence.

We did make one mid-course improvement during this experiment. One of the results of using the TEXT-R algorithm was that the articles selected for translation were usually quite a bit longer than those to which the translators were accustomed. In order to reduce the workload on the translators, we asked them to translate only the sentences which contained novel words as determined by TEXT-R[3]. The sentences were automatically highlighted, but were left *in situ*, with the surrounding sentences serving as context. In this way, we were able to further increase the speed with which the corpus is translated.

Testing

Description of Test Conditions

In order to evaluate our approach to increasing lexical coverage in a corpus, we have undertaken several experiments to compare the TEXT-R technique to two other approaches:

1. Continuing as before, translating "Balkans" articles
2. Randomly choosing articles to translate

These approaches are described below.

Old corpus

Recalling that the current work is intended to be an extension to an already existing corpus, we include statistics about that corpus as a baseline. We will call this pre-existing corpus the 'old' condition. Some statistics of the old corpus are shown in the first column of Table 1. We developed this corpus over a period of about a year and a half, and have edited it extensively for quality of translation.

Three test corpora

In fall of 1997 we undertook a study of three possible extensions to the old corpus. The three extensions were designed so as to be roughly comparable: including the same number of articles (topics) and approximately the same number of words. The first of these extensions was to continue doing what we were doing before: translating documents in the 'Balkans' domain. We will call this extension 'continue' since it continues the previous methodology. The second extension we considered concurrently was to choose articles by randomly selecting them from among all new articles available on a particular day. We will call this extension 'random'. For the third extension we used the TEXT-R software to choose articles. This extension we will call 'TEXT-R'. Each of these experiments produced approximately 13,000 English words in 27 articles. The actual figures are presented in Table 1.

Since we have set out to measure the interactions between the old corpus and the three experimental corpora, we are most interested in the combination of each of them with the old corpus. In what follows, we will refer to combinations of these corpora by combining their names with a '+': old+continue, old+random and old+TEXT-R.

old:	tokens	types	avg. rate %
English	845221	27970	3.31 %
Croatian	770786	64959	8.43 %
old+continue:	tokens	types	avg. rate %
English	858625	28128	3.28 %
Croatian	783276	65433	8.35 %
old+random:	tokens	types	avg. rate %
English	858308	28513	3.32 %
Croatian	783410	66185	8.45 %
old+TEXT-R:	tokens	types	avg. rate %
English	858574	28983	3.38 %
Croatian	783211	66811	8.53 %

Table 2: Word count results for old and test corpora

Word Count Results

In order to get an impression of the progress of the various experiments, we examined the number of new words, both in English and Croatian, acquired by each of the experiments, and the lexical growth rates achieved by each. Table 2 lists these results together with those for the old corpus repeated from Table 1. The 'avg. rate %' is the number of new words acquired per 100 words translated. The baseline is, of course, the old corpus, with a rate of 3.31 % on the English

[3] We did try another solution: changing the objective function to $\frac{\text{number of new words in article}}{\text{total number of words in article}}$. This, however, resulted in extremely bad choices of articles to translate, so we did not follow up.

side. If we were to continue translating in-domain articles, we would end up reducing our acquisition rate to 3.28 %. The random method is an improvement over the old method since its rate of acquisition is higher than the baseline although only marginally (3.32 % against 3.31 %). The TEXT-R method is clearly the best, with a rate far higher than the baseline (3.38 % against 3.31 %). Most importantly, the trends that are observed in the English side of the corpus also hold for the Croatian side, demonstrating that we can increase the lexical coverage of a foreign language by translating English documents with increased lexical coverage.

The graphs in Figure 1 (next page) show the cumulative acquisition of new words over time[4]. The left-hand graph shows the results of the old corpus, with the three experiments in the far upper right. Overall, the graph of the old corpus shows signs of flattening, something we would like to avoid as our intent is to continue to increase lexical coverage.

Because the results of the experiments are difficult to see in the first graph, the second one is an enlarged view of the last 60,000 tokens of the earlier graph. In the latter graph, the trends of the three experiments are clearly seen. The continue corpus, as expected, continues more or less the trend of the old corpus. On the other hand, the random corpus has a substantially greater slope than either the baseline or the continue corpus. The TEXT-R trend, however, is steeper than even that of the random corpus.

The analysis of acquisition rates in this section has shown that while all three acquisition methods continue to acquire new words, the continue method actually causes the average acquisition rate to decrease. This bodes ill for continued high levels of lexical acquisition. The random method does improve on the rate in the old corpus, although hardly noticeably. The TEXT-R method significantly increases the rate of change of acquisition.

EBMT Results

While the improvement in lexical acquisition rates presented in the previous section is desirable in theory, improved knowledge resources may not always result in improved applications. In this section, we test the resulting corpora as the knowledge base for EBMT.

In evaluating EBMT, we consider primarily coverage: how much of a test corpus is matched against the source side of the EBMT corpus, and secondly how many words are eventually provided with translations. This second number is necessarily smaller than the first and reflects the failure of the system to find a correspondence between the two languages in the corpus.

As a test corpus, English material was randomly drawn from the same newswire service as the original translation materials, with a uniform distribution across articles. The corpus was edited slightly to remove non-language material, primarily lists of stocks and sports scores, which would artifically inflate the coverage of the translation system.[5] The following table lists characteristics of the corpus:

old:	Count	Rate (%)
Tokens matched	52116	75.80 %
Tokens translated	12040	17.51 %
old+continue:	Count	Rate (%)
Tokens matched	52261	76.01 %
Tokens translated	12109	17.61 %
old+random:	Count	Rate (%)
Tokens matched	52989	77.07 %
Tokens translated	12082	17.57 %
old+TEXT-R:	Count	Rate (%)
Tokens matched	53285	77.50 %
Tokens translated	12173	17.71 %

Table 3: EBMT results for old and test corpora

182	articles
2988	lines (sentences)
58558	English words (tokens)
8907	English words (types)
68752	tokens[6]

Again the old corpus provides the baseline, as shown in Table 3. We will be primarily interested in the 'Tokens matched' figure, since it reflects how well our corpus covers the test data. Nevertheless, we should not ignore the 'Tokens translated' figure, since it reflects the effectiveness of our techniques in a real system.

We now compare the three experimental corpora, shown in Table 3. In this case again, TEXT-R outperforms both the continue and the random methods, as measured both by tokens matched and by tokens translated. Although the increase is not large in either case, the number of tokens matched has increased by a significant amount (1.5 %), especially considering the small amount of text in the experimental corpora.

One odd element of the data is that the number of tokens translated in the old+random corpus (17.57 %) is actually less than that of the old+continue corpus, even though the number of matching tokens is 1.06 % higher. The quality of EBMT is not monotonic in the quality of its corpus, as is illustrated by these numbers. This is probably due to several speed enhancements to the EBMT algorithm which ignore parts of the corpus once a sufficient number of matches have been found.

Long-term application of TEXT-R Method

After the initial experiments suggested that TEXT-R was producing desirable results, we incorporated it as our primary method for acquiring new translation material. This section describes the results of 5 months

[4] We measure time in terms of number of words translated in order to avoid discrepancies in time due to editing and translator tasks.

[5] Numbers are recognized by the system and translated without relying on a bilingual corpus.

[6] The EBMT system includes punctuation as separate tokens when translating, with the result that this number does not match the number of words. All results in this section will be presented in terms of EBMT tokens, rather than whitespace-separated words.

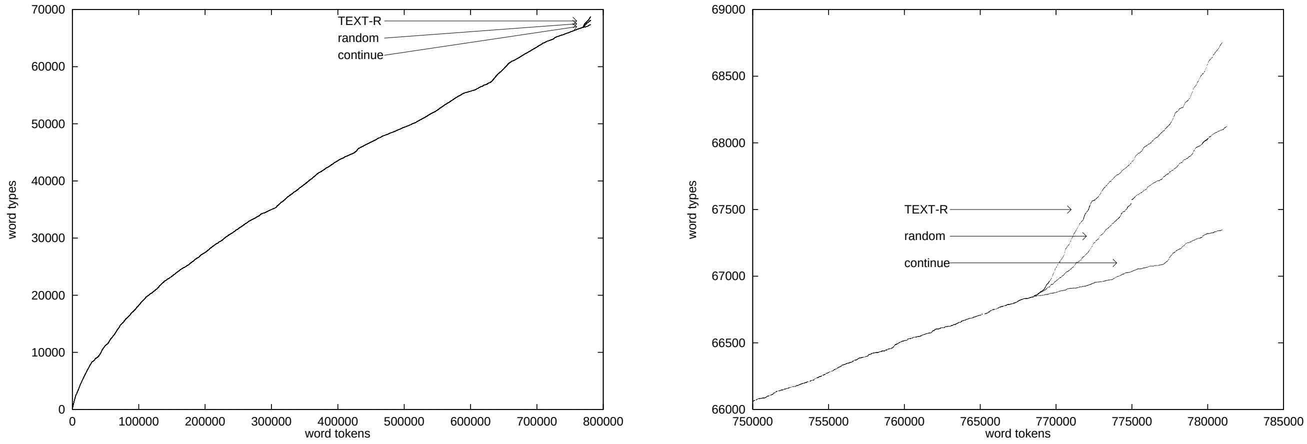

Figure 1: Acquisition of New Words over time (Normal & Zoomed view)

TEXT-R+:	tokens	types	avg. rate %
English:	211992	24627	11.62 %
Croatian:	199734	44073	22.07 %

old+TEXT-R+:	tokens	types	avg. rate %
English:	1057213	40515	3.83 %
Croatian:	970520	88210	9.09 %

Table 4: Statistics of TEXT-R+ and old+TEXT-R+

of translation work completed using the new methodology.

This corpus, produced by TEXT-R over an extended period of time will be called TEXT-R+. The size of this corpus is as follows:

207	articles
10557	lines
211992	English words (tokens)
199734	Croatian words (tokens)

Word Count Results

We analyze the TEXT-R+ corpus in the same manner as described earlier. In order to demonstrate the significant difference that TEXT-R makes, we will illustrate the numbers with both the TEXT-R+ corpus by itself and together with the old corpus (old+TEXT-R +). These data are presented in Table 4. Notably, the acquisition rate of the TEXT-R+ corpus (11.62 %) is remarkably higher than anything else.

The progress of the TEXT-R+ corpus, together with the old corpus and the two other test corpora is shown in Figure 2. This graph shows that we have achieved a new height in the acquisition of new vocabulary.

EBMT Results

Perhaps the most surprising results of this paper may be found when we replicate the EBMT experiments for the TEXT-R+ corpus. In order to show this more clearly, we first present the results of using only the TEXT-R+ corpus as the basis for EBMT. For convenience, we repeat the figures for the old corpus.

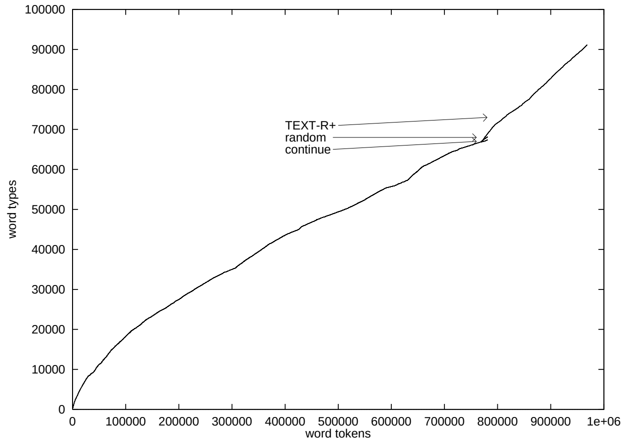

Figure 2: Acquisition of New Words over time (old+TEXT-R+)

old:	Count	Rate (%)
Tokens matched	52116	75.80 %
Tokens translated	12040	17.51 %

TEXT-R+:	Count	Rate (%)
Tokens matched	47140	68.57 %
Tokens translated	13790	20.06 %

In five months of using the TEXT-R algorithm to select texts to translate, we have managed to acquire a vocabulary that performs almost as well (68.57 % vs. 75.80 %) on a test text as did the old corpus, which itself took $1\frac{1}{2}$ years to develop. This represents a significant improvement in translator efficiency. Furthermore, the vocabulary that we acquired with TEXT-R actually produced more translations using EBMT than the old corpus did (20.06 % vs. 17.51 %). This speedup and increase in translation coverage clearly justifies the TEXT-R method as a way of gathering texts for the EBMT translation system.

We now consider the results of combining the TEXT-R+ corpus with the old corpus.

	Count	Rate (%)
Tokens matched	55611	80.89 %
Tokens translated	13100	19.05 %

While the data show that the tokens matched rate has increased, as expected, to 80.89 %, the number of tokens translated actually decreases to 19.05 %. This surprising result is probably due to the non-monotonicity of EBMT mentioned earlier.

Impact of TEXT-R on translators

We would like to describe in this section the impact that the TEXT-R program has on the translators who work with the texts that it automatically collects.

During the initial phase using the old collection method, the translators became very familiar with the types of texts to be translated and were able to work autonomously with bilingual reference materials that are readily available for English/Croatian. Assistance from native English speakers at the translation laboratory was seldomly needed. As we transitioned to the phase of expanding the corpus in fall of 1997 by manually selecting texts, we began to encounter a few questions per week from the translators, yet the additional topics (*e.g.* health, education, cinema, family, leisure) did not necessarily increase the difficulty of the translation task, except for the domain of American sports. We then noticed a significant increase in questions asked by the translators starting with the testing of TEXT-R in November 1997; even now in April 1998 we continue to receive several questions per day from each of our two Croatian translators. This is to be expected because TEXT-R began to include domains that have large numbers of new words (*e.g.* medical, pharmaceutical, sports) that have a low frequency of occurrence in the texts. This, combined with the fact that we mark only the sentences that contain new words in order to limit the translation task to these sentences and thus reduce the appearance of highly repetitive material in long texts (*e.g.* function words, common nouns, common verbs), resulted in a higher level of translation difficulty.

Given the increasing difficulty of texts chosen by TEXT-R, evidenced by a constant flow of questions from translators, we believe that an ongoing implementation of this new algorithm requires that translators have a mature knowledge of the source language. Although access to English monolingual and English/Croatian bilingual reference materials is important, definitions for much of the new vocabulary (e.g. terminology in advertising and technical domains, colloquial expressions) collected by TEXT-R can be found through Internet search engines and domain-specific Internet discussion groups which have already proven to be noteworthy sources for vocabulary that is not yet available in the above-mentioned traditional references. Training on these Internet resources is crucial for translating texts that are furnished by TEXT-R.

TEXT-R Corpora Characteristics

While the corpora produced by TEXT-R appear to be effective in improving EBMT, the question arises about how useful such corpora may be for other uses.

There is one primary difference between the way in which EBMT uses corpora and the way in which many other technologies use corpora. The goal of EBMT is to find examples of certain phenomena in a corpus. However, whether a phenomenon occurs once in a corpus or many times is largely irrelevant to EBMT, particularly if it is translated the same way each time. This is fundamentally different from the way in which other systems use corpora: most uses of corpora involve computing statistics from the corpus, and using those statistics in some way. Examples of this are Language Modelling (Rosenfeld, 1994; Brown and Frederking, 1995) and Statistical Machine Translation (Brown et al., 1990). Because EBMT computes no counts over a corpus, any repeated phenomena are merely redundant. However, to a system that does compute statistics, relative frequency is one of the most important pieces of information available. Since TEXT-R may alter these frequencies, it may produce corpora unsuitable for these methods.

There is a second way in which modifications to the frequencies in a corpus may affect the way that it is used. Some users of corpora treat low-frequency phenomena differently from the way they treat high-frequency phenomena. For example, in the IBM work on Statistical Machine Translation (Brown et al., 1993), all words occurring only one time are replaced by a special unknown English word or unknown French word token. This was done in order to eliminate "some of the typographical errors that abound in the text" (Brown et al., 1993, p. 283). Nevertheless, if most items of the new vocabulary we have introduced to our corpus occur only one time, doing this will eliminate the gains we have adduced.

In this section we will therefore address the degree to which the corpus we have produced corresponds to what other consumers of corpora have come to expect from a corpus.

Zipf's law

One well-known measure of the naturalness of a language is Zipf's law. Actually, Zipf proposed two laws governing "the frequency distribution of words in the stream of speech" (Zipf, 1935, p. 39). These are described below:[7]

- **"rank-frequency" law**
  Order all words from most frequent to least frequent. The most frequent word has rank 1, the second most frequent word rank 2, and so on. The graph of $log(frequency)$ (y-axis) versus $log(rank)$ (x-axis) approximates a straight line with slope -1.

- **"number-frequency" law**
  Let f_n be the number of tokens with frequency n. Then the plot of $log(n)$ (y-axis) versus $log(f_n)$ (x-axis) approximates a straight line with slope -0.5.

Zipf himself notes (Zipf, 1935, p. 43) that the "number-frequency" law should hold better for low-frequency words than for high-frequency words, motivating him to investigate the "rank-frequency" law for

[7] Of these two, the "rank-frequency" law is by far the more well-known, and is usually the one referred to as "Zipf's law". Nevertheless, the "number-frequency" law, which is more accurate for low-frequency words, is possibly of greater interest to us.

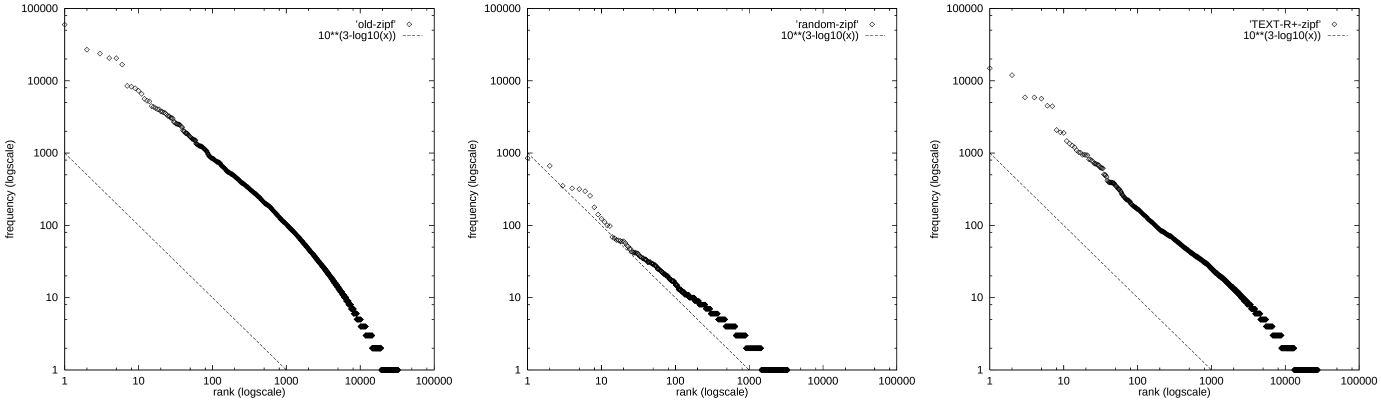

Figure 3: The "rank-frequency" law for three corpora

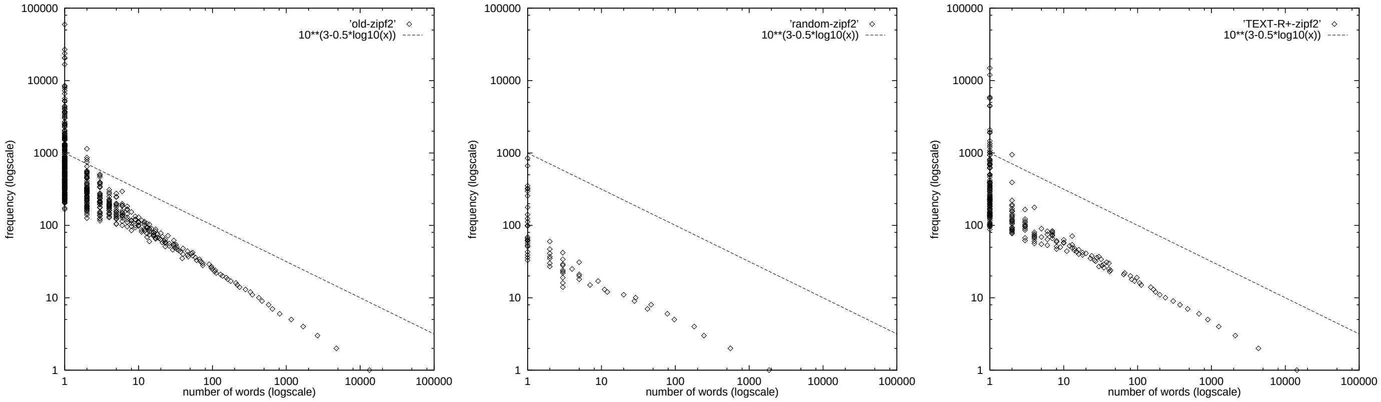

Figure 4: The "number-frequency" law for three corpora

high-frequency words. Zipf's laws have been confirmed in a number of languages, including Chinese, Japanese and several European languages (Landini, 1997).

In order to determine the degree to which our corpus adheres to Zipf's laws, it is necessary to compare our corpus to one that represents unadulterated text. In this case, we want to compare the corpus produced with the TEXT-R algorithm (TEXT-R+) to ones that were generated without it (old, random). In this way we can determine whether any deviations from the law by TEXT-R+ indicate abnormally distributed text.

The relevant graphs for the "rank-frequency" law are plotted in Figure 3. All graphs are to the same scale, and a line of slope -1 is provided for comparison. We are interested only in the slope of the plotted curve. If the slope is nearly identical to -1 (the slope of the line in the chart), the data follow Zipf's law.

From these graphs, it is clear that the TEXT-R+ corpus obeys the "rank-frequency" law at least as well as the other two corpora do, if not better. The deviation in the high-frequency items is well-understood *cf.* (Mandelbrot, 1965), and does not undermine basic trend.

Figure 4 shows the graphs of the "number-frequency" law. In this case, the comparison is with a line of slope -0.5, also shown in the graphs.

This figure confirms that the TEXT-R+ graph follows the -0.5 slope better than the other two graphs. Most of the points on these graphs represent lower frequency words, and the points in Figure 3 represent higher frequency words. Therefore we can be confident that our TEXT-R+ corpus does not deviate significantly from what is expected of a corpus of this size with respect to either high or low frequency ranges.

In this section, we have used Zipf's laws to investigate the makeup of the TEXT-R+ corpus with respect to the frequency distributions of its words. We have applied tests that measure both the high-frequency and low-frequency words, and found the statistics to be at least as close to the ideal law as those from unadultered corpora. We therefore have reason to believe that our corpus is adequate for other corpus-based applications. However, nothing can substitute for careful testing using the target application.

It is our belief that we were able to produce a corpus with a significantly greater rate of new words without affecting the frequency characteristics because translations were always performed on entire sentences or entire articles, never on individual words. For this reason, our statistics probably approximate those of an accumulation of random sentences, rather than an accumulation of random words or an accumulation of frequency-manipulated words.

Conclusion

We conclude that text corpora can be enhanced effectively by increasing lexical coverage using a vocabulary filtering mechanism such as the TEXT-R algorithm. This method allows us to avoid the tapering effect and the minimal increase of lexical acquisition that are associated with unintelligent text selection. Despite the increase in difficulty of the translation task that results from this corpus expansion technique, our findings clearly indicate that there is a significant speed-up in the creation of parallel corpora by using it. In addition, the implementation of the parallel corpora into our EBMT system will allow for a wider range of translation capabilities. TEXT-R is one method that significantly improves the development of rapid-deployment EBMT systems, such as the DIPLOMAT project, and may possibly be used for other applications as well.

References

Allen, J. (1997). Translator training course. Caterpillar Incorporated Translation Services.

Brown, P., Cocke, J., Della Pietra, S., Della Pietra, V., Jelinek, F., Lafferty, J., Mercer, R., and Roossin, P. (1990). A statistical approach to machine translation. *Computational Linguistics*, 16:79–85.

Brown, P., Della Pietra, S., Della Pietra, V., and Mercer, R. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Brown, R. (1996). Example-based machine translation in the Pangloss system. In *Proceedings of the Sixteenth International Conference on Computational Linguistics*, pages 169–174 (vol 1), Copenhagen.

Brown, R. and Frederking, R. (1995). Applying statistical English language modeling to symbolic machine translation. In *Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation*, pages 221–239.

Burnage, G. and Baguley, G. (1997). The British National Corpus. Available at http://info.ox.ac.uk /bnc/using/papers/gblibs.html.

Collins COBUILD (1998). Questions and answers. Available at http://titania.cobuild.collins.co.uk /boe_info.html.

EAGLES (1995). EAGLES Evaluation of Natural Language Processing Systems: Final Report. Available at http://www.issco.unige.ch/ewg95. EAGLES document EAG-EWG-PR.2.

Evans, R. and Kilgarriff, A. (1995). MRDs, standards and how to do lexical engineering. In *Proceedings of the Second Language Engineering Convention*, pages pp. 125–132.

Fisher, W., Doddington, G., and Goudie-Marshall, K. (1986). The DARPA speech recognition research data base: Specifications and status. In *SPEECH RECOGNITION: Proceedings of a Workshop (sponsored by DARPA)*, pages 93–99.

Francis, W. and Kucera, H. (1979). *Brown Corpus Manual.* Linguistics Department, Brown University.

Frederking, R. and Brown, R. (1996). The Pangloss-Lite machine translation system. In *Proceedings of the Conference of the Association for Machine Translation in the Americas (AMTA).*

Frederking, R., Rudnicky, A., and Hogan, C. (1997). Interactive speech translation in the DIPLOMAT project. In *Proceedings of the Spoken Language Translation workshop at Association for Computational Linguistics (ACL 97).*

Gibbon, D., Moore, R., and Winski, R., editors (1997). *EAGLES Handbook of Standards and Resources for Spoken Language Systems.* Mouton de Gruyter, Berlin & New York.

Heyn, M. (1995). Present and future needs in the CAT-world. In *Proceedings of the Dublin University Conference 95.*

Landini, G. (1997). Zipf's laws in the Voynich manuscript. Available at http://sun1.bham.ac .uk/G.Landini/evmt/zipf.htm.

Language Partners International (1997). An introduction to Computer Aided Translation (cat). Available at http://www.languagepartners.com/catintro.htm.

Mandelbrot, B. (1965). Information theory and psycholinguistics. In Wolman, B. and Nagel, E., editors, *Scientific Psychology*, pages pp. 550–562. Basic Books.

Nagao, M. (1984). A framework of a mechanical translation between Japanese and English by analogy principle. In Elithorn, A. and Banerji, R., editors, *Artificial and Human Intelligence.* NATO Publications.

Nirenburg, S., Beale, S., and Domashnev, C. (1994). A full-text experiment in example-based machine translation. In *New Methods in Language Processing*, Manchester, England.

Rosenfeld, R. (1994). *Adaptive Statistical Language Modeling: A Maximum Entropy Approach.* Ph.D. thesis, Carnegie Mellon University.

Sato, S. (1992). CTM: An example-based translation aid system using the character-based best match retrieval method. In *Proceedings of COLING-92.*

Sumita, E. and Iida, H. (1991). Experiments and prospects of example-based machine translation. In *Proceedings of 29th ACL Meeting*, pages 185–192, University of California, Berkeley, California.

Van Santen, J. (1992). Diagnostic perceptual experiments for text-to-speech system evaluation. In *Proceedings of the International Conference on Spoken Language Processing*, pages 555–558. ICSLP.

Zipf, G. (1935). *The Psycho-Biology of Language: An Introduction to Dynamic Philology.* Houghton Mifflin Company.