Xerox TREC-6 Site Report: Cross Language Text Retrieval

Eric Gaussier Gregory Grefenstette David A. Hull
B. Maximilian Schulze*
Xerox Research Centre Europe†
{gaussier,grefen,hull}@xrce.xerox.com

January 9, 1998

Abstract

Xerox participated in the Cross Language Information Retrieval (CLIR) track of TREC-6. This track examines the problem of retrieving documents written in one language using queries written in another language. Our approach is to use a bilingual dictionary at query time to construct a target language version of the original query. We concentrate our experiments this year on manual query construction based on a weighted boolean model and on an automatic method for the translation of multi-word units. We also introduce a new derivational stemming algorithm whose word classes are generated automatically from a monolingual lexicon. We present our results on the 22 TREC-6 CLIR topics which have been assessed and briefly discuss the problems inherent in the cross-language IR task.

1 Introduction

Cross Language Information Retrieval (CLIR) addresses the problem of retrieving documents written in one language using queries written in another language. As document repositories grow in size and distribution, and one can consider the World Wide Web as such a repository, it is becoming more important to find solutions to this long-standing research problem [11].

Xerox participation in TREC-6 is limited to the Cross Language Information Retrieval track. We are interested in learning how well queries translated via a general purpose bilingual dictionary can perform in relation to queries written in the document language. Our experimental technique is to perform baseline monolingual retrieval using the topics provided by NIST (originally in English and manually translated into French and German) for English and French. Then, we produce automatic translations of the queries, attempting to see how closely the monolingual performance level can be approached. We explore three methods for query translation and document retrieval: manually constructed queries which use a weighted boolean model for retrieval, automatic translation of multi-word units, and

*Maximilian Schulze now works for Xerox Imaging Systems in Peabody, MA, USA and can be reached at: bschulze@xis.xerox.com

†Our research center has changed its name. We were formerly known as the Rank Xerox Research Centre.

the application of a new automatic method for derivational stemming. The purpose of this approach is to simulate the results that can be expected by a cross-language retrieval system that translates queries at run-time into the language of the pre-indexed documents stored in the system.

In the following sections, we will present the transformations that a query undergoes during our translation process, a discussion of the weighted boolean alternative to vector space retrieval, a tabular version of partial results from our CLIR runs, and a discussion of the problems that our system encountered in performing these tasks. Our experiments this year are limited to English and French. We do not work with the German topics or the German document collection.

2 Creating a Baseline for CLIR

In order to produce a baseline for comparing translated queries to original language queries, we first create monolingual versions of each target language query using the NIST-supplied translations.

The monolingual transformation of a query follows the following steps. The query text sections (Title, Description, and Narrative fields) are first isolated and then part-of-speech tagged [3, 13]. Our Xerox part-of-speech taggers[1] provide the user with lemmatized forms[10] of the words in the text. Using this tagged text we also extract entire noun phrases [12], as well as the decomposition of complex noun phrases into two word subparts. To create index terms, the individual words extracted from the query as well as the individual words[2] in noun phrases are derivationally stemmed. The stemmed versions of noun phrases are sorted in alphabetical order, as has been done in SMART since [5] in order to eliminate positional variation, and joined with an underscore to form a new index term for the query. Individual words and joined noun phrases are stored separately in newly created fields, which can be weighted in different ways.

This same treatment has been applied to documents for indexing. The addition of phrases derived in this way has improved our average precision for documents having many[3] relevant documents by 7% over baseline retrieval using simple stemmed words in past TREC experiments[9].

3 Translating queries

In this section we show the transformations performed on a sample query during translation. First, as in the monolingual case explained in the last section, the query is part-of-speech tagged, noun phrases are extracted, stopwords are removed, and words and phrases are stemmed. Then, in order to produce a translated version of the query, each of the query terms are expanded (a reversal of the stemming process) to produce all the derivational variants. Each of these variants is looked up in a general language bilingual dictionary. The

[1] http://www.xrce.xerox.com/research/mltt/Tools/pos.html.

[2] Stopwords are removed using standard IR stopword lists. See ftp://ftp.cs.cornell.edu/pub/smart

[3] More than four relevant documents. With documents with four or less relevant documents, adding phrases improves average precision by 14%.

translations are restemmed using a derivational stemmer for the target language, and the target language version of the initial query is recomposed in a TREC format.

3.1 Sample Treatment

Topic number 7 deals with *sex education.* The French title is *L'education sexuelle.* Let's imagine that this title is the entire French query and that we want to access English documents with it. The steps followed in the source language treatment of this query are the following:

- Part-of-speech tag sequence — *L'/DET éducation/NOUN sexuelle/ADJ*
- Stopword removal — *éducation sexuelle*
- Lemmatisation(inflectional lexicon) — *éducation sexuelle*
- Stemming (derivational lexicon) — *éduquer sexué*
- Noun Phrase extraction, stemming and alphabetical ordering —
  NP extracted: {*éducation/NOUN sexuel/ADJ*}
  Stemmed and ASCII ordered: *sexué_éduquer*

If this title were the entire query, then the monolingual query would consist of the following stemmed index terms: *eduquer sexue sexue_eduquer.* In order to translate this version of the query into English, the following additional steps are performed:

- For each single word, reverse stemming producing the related lemmas
  eduquer — éducation, éducatif, éducateur, éduquer
- Translation of the lemmas
  (English) *education, training, manners, educational, educative, educate, train, bring up*
- Stemming of the translated lemmas:
  educe, train, mannered, mannerism, bring_up
- Filtering of the stems obtained on the basis of their presence in the collection:
  educe, train, mannered, mannerism
- For multiword expressions, generate all possible combinations of the stems produce above
  eduquer_sexue mannerism_sex mannered_sex sex_train educe_sex mannerism_sexual mannered_sexual sexual_train educe_sexual mannerism_sexualiser mannered_sexualiser sexualiser_train educe_sexualiser

The derivational stemming algorithm used in these experiments is based on a new technique to automatically derive an approximation to derivational families using only a lexicon. Since the technique is currently under development, there are a number of problems which still need to be resolved. The algorithm is entirely automatic, meaning that like most traditional stemming algorithms, it will make a number of stemming errors. A manual correction step is planned for the future. On the other hand, this means that the algorithm is nearly language independent (for certain language families, given a lexicon), so we can develop derivational stemmers for new languages relatively easily. This is a key advantage for cross-language text retrieval.

This automatic procedure was followed for all the 25 English and 25 French cross language topics, CL1 – CL25. The resulting queries were fed into a traditional information retrieval system implementing a vector space model for retrieval (a heavily modified version of SMART [2]). As monolingual baselines, the English versions of the NIST English topics were run over the English documents: this is our run XRCE-E2EA; and French versions of the NIST French topics were run over the French documents: XRCE-F2FA. From the NIST English topics, we generated French versions as described above and ran them over the French documents: this is run XRCE-E2FA; and likewise automatically generated English version from the NIST French topics: run XRCE-F2EA. The final A of the above runs stands for Automatic, as opposed to the manually constructed runs described in the next section.

4 Manual Runs - Weighted Boolean Model

The Xerox approach to manual query construction is based on a simplified weighted boolean model. The model assumes that each query can be divided into a number of concepts. Each concept consists of one or more terms, and the terms within a concept are combined using a weighted OR operator. The concepts are then combined using a weighted AND operator. In addition, the user is expected to assign a value of 1, 2, or 3 to each concept to indicate its importance in the query (1 = not important, 2 = important, 3 = mandatory). These values are used internally to adjust the concept weights before applying the weighted AND operator. A concept value of 3 leads to a strict boolean constraint, while lower values relax the constraint, allowing documents which do not contain any terms in the concept to be retrieved with a non-zero weight.[4] A longer description of the probabilistic weighted boolean model is provided in [7]. The detailed mathematical formulation of the operators can be found in [6], which should soon be available.

Previous experiments [7] have found that the weighted boolean model is particularly effective for cross-language text retrieval, as it addresses two important problems in query translation. A primary source of error in CLIR is translation ambiguity (a source language term can can have multiple unrelated target language translations). The boolean AND operator provides a natural form of disambiguation, since it is likely that correct translations will cooccur much more often in documents than incorrect translations. This approach has significant advantages when compared to other corpus-based and user-based disambiguation strategies. The search corpus itself is used for disambiguation, so domain relevance is guaranteed and no additional reference corpora are required. User knowledge is incorporated implicitly in the query construction process, so no additional user effort specifically for disambiguation is required. The boolean model also performs automatic normalization of

[4] In cases where the retrieved set was less than 1000 documents, strict boolean constraints were systematically relaxed until at least 1000 documents were returned. This step was taken solely to optimize performance for TREC evaluation, since there is no advantage to returning fewer than 1000 documents.

term importance. For example, a common source language term may have one or more rare target language translations which receive large term weights. The boolean operators make sure that rare terms and terms with many synonymous translations do not dominate during retrieval (since they may well be incorrect translations). These problems can also be addressed with other disambiguation strategies, but other methods tend to add substantial overhead to the query translation process.

For the TREC-6 CLIR experiments, two manual query sets were built, one in English, one in French. Each query set was constructed independently by a native speaker of the language who took about 2 hours to build the entire query set (slightly less than 5 minutes per query). The queries were used for both monolingual and cross-language retrieval (English to French or French to English). The English searcher is an expert with the weighted boolean system while the French searcher was using the system for the first time, although he has several years of experience as a researcher in information retrieval. Both searchers are authors of this paper and the queries were not generated in a controlled experimental setting. The manual queries average 10-12 words per topic, as opposed to 24-26 words per topic for the automatic versions. We should note that no documents were examined in the course of writing these queries, so there is no use of relevance feedback or similar techniques for term selection.

English: Is wine consumption/production rising or decreasing world-wide?
English Boolean:

3	wine wines
2	consume consumption produce production
2	increase decrease rate curve forecast future
1	world international

French Boolean:

3	vin
3	consommation production consommation_vin production_vin
2	augmentation diminution diminuer croissance croître decroître
2	monde pays

Table 1: Natural language and structured versions of query CL19.

The manually generated queries for topic CL19 are presented in Table 1. For this topic, both searchers used more or less the same concept structure for their queries, although this is often not the case. The concept weights are different, however, with the French searcher requiring that either production and consumption occur in the document for it to be retrieved.

5 TREC-6 Results for 22 Topics

Xerox submitted 8 runs to the CLIR track, consisting of all combinations of the factors: monolingual and cross-language, English and French, and Manual and Automatic. We did not submit any German runs due to time pressure but hope to work with the German

	Run	Man(WTB)	Auto	Run	Man(WTB)	Auto
original	E2E(ML)	0.437/0.595	0.411/0.572	F2F(ML)	0.407/0.562	0.394/0.540
revised	F2E(CL)	0.239/0.350	0.222/0.358	E2F(CL)	0.242/0.337	0.167/0.290
original	F2E(CL)	0.218/0.332	0.185/0.285	E2F(CL)	0.195/0.278	0.163/0.265

A2B - A = query language, B = document language (E = English, F = French)
ML = monolingual, CL = cross-language
Man(WTB) = manual (weighted boolean), Auto = automatic
#/# - Avg uninterpolated precision / Avg precision at 5, 10, 15, 20 docs (AP20)

Table 2: Average/high precision score table averaged over 22 evaluated topics.

data next year. The results presented here are averages of the 22 topics[5] which had been evaluated at the time the final version of this paper was being written. At the time of the TREC conference, only 13 topics had been evaluated. To see evaluation scores for the 13-topic set, please consult our draft paper in the notebook distributed at the conference. The average performance figures for these topics for both average uninterpolated precision and precision averaged at 5, 10, 15, and 20 documents retrieved (AP20) are presented in Table 2. In general, we feel that the high-precision measure is more appropriate for the cross-language retrieval task, due to the added burden of translating/glossing/reading documents in a foreign language.

In general, our monolingual performance is very good. While we don't have the results for the 22-topic set for all participants, our system scored at or near the top in monolingual performance for the 13-topic set in English and French. We are using the basic Okapi TREC-3 term-weighting formula and no additional query expansion other than what is provided by the derivational morphology. This seems to indicate that the derivational stemming is working well for monolingual retrieval. We have heard from other groups that traditional methods for query expansion are not helpful (or are even harmful) on this query set. We feel that derivational expansion is likely to be more robust than traditional query expansion techniques (it is actually just the inverse of stemming), so this may be part of the reason for our success. We have not yet run internal comparisons of our TREC runs with our standard stemming algorithm based on inflectional morphology.

Readers will note that Table 2 has two different lines for cross-language retrieval, original and revised. The original line represents our TREC-6 submission. The revised line gives the performance after correcting a bug in our query translation algorithm. This bug caused some terms in the translated queries to be stemmed in a different way than the documents. Any term which suffered from this problem was useless for retrieval, since it didn't match any terms in the documents. Fortunately, the number of terms affected was small, but correcting the bug does result in a noticeable increase in our cross-language evaluation scores across the board.

Our revised cross-language scores are roughly 60% of monolingual performance. This ratio is nearly identical to what we have found in our previous studies [8]. Therefore, our

[5] Actually the results are based on only 21 topics, since topic 8 has no relevant topics in English and topic 22 has no relevant documents in French. Each of these topics is excluded when averaging scores over the respective language.

new translation strategy for phrases doesn't seem to be giving us any dramatic improvement in cross-language performance. When we have the chance to test the performance of our system without this feature, we'll have a more direct measure of its impact. This ratio is low compared to the numbers reported by many other TREC-6 CLIR participants at the conference. However, many of these systems started with a much lower monolingual baseline, so in absolute terms the performance of our cross-language system is reasonable. We recognize that there is still a lot of room for improvement. Note that we perform no disambiguation of our dictionary entries. For the weighted boolean model (manual runs), we believe that there is no advantage to disambiguation, but we are currently exploring several corpus-based disambiguation strategies to be used with our vector models (automatic runs).

The manually constructed topics perform slightly better than the automatic topics in all cases, but there is no evidence that the weighted boolean model (used with the manual runs) is more effective for cross-language retrieval. Therefore, we find no experimental confirmation of the advantages we hypothesized for the weighted boolean model. Only the difference between monolingual retrieval and cross-language retrieval is statistically significant. In our previous study on English-Spanish cross-language retrieval using the Elnorte collection [7], we found a similar difference in monolingual performance but a much larger difference in cross-language performance. However, it is important to highlight the difference in experimental set-up between the two experiments. The English-Spanish dictionary was cleaned up specifically for the experiments, which was not the case this year. This is reflected in a much smaller difference between monolingual and cross-language performance in the Spanish experiments.

Note that the structure of our topics makes comparisons between the manual monolingual and cross-language runs difficult. This is because the monolingual results and the cross-language results for a single document collection are based on queries written by different searchers, which means that variation due to query formulation is confounded with variation due to query translation. Similarly, monolingual and cross-language runs using the same manual queries are applied to completely different document collections, which means that variation due to the type and density of relevant documents and the document language is confounded with variation due to query translation. In addition, comparisons between the weighted boolean model and the vector space model are more tenuous because of the differences in query formulation strategies. In our previous experiments, we used exactly the same query terms (without boolean structure) for measuring performance of the vector space model.

	Fr $\rightarrow$ En	En $\rightarrow$ Fr
ML-CL $\leq$ 0.10	5	5
0.10 $<$ ML-CL $<$ 0.20	10	4
ML-CL $\geq$ 0.20	6	10
CL = 0.0	1	1

Table 3: Topic comparison of automatic monolingual (ML) and cross-language (CL) retrieval as a function of absolute difference in AP20.

With the small topic sample, it may make more sense to look at performance differences on individual topics. Table 3 splits up topics according to the absolute difference in AP20 between automatic monolingual and cross-language runs. Note that the manual runs are not included for the reasons mentioned in the previous paragraph. There are only 19 topics in the English-French column because the two topics with an average precision of 0.0 for monolingual French retrieval are deleted. From this table, we note that the performance of our system seems to be worse from English to French than from French to English. This corresponds with the fact that our English to French dictionary is less complete. We are relieved to note that there is only one query in each direction where performance is disastrously poor (no relevant documents in the top 20 when relevant documents are found in monolingual retrieval).

6 Problems Inherent in CLIR

A rapid evaluation of performance over the 22 topics evaluated by NIST reveals a number of weaknesses in the techniques or resources that we employed in the CLIR track.

6.1 Dictionary Coverage

Using a bilingual dictionary as the only source of translation alternatives puts the burden of discovering at least one correct translation equivalent on that resource. The same can be said, of course, for using parallel texts. When a word is not found in the dictionary, some default strategy should be used. Our strategy is to pass the original source word through as the target translation. In the case of proper names written in the same way in both languages, this strategy should be successful. Another (more robust) strategy implemented by Davis [4] tries to discover potential cognates in the target index using an edit distance up to two characters. This technique would capture both slight alterations of proper name spellings, as well as some true cognates if the languages are etymologically close.

We used a version of the *Oxford University Press/Hachette English-French, French-English* for the submitted TREC runs. Working directly from the SGML-markupped version of the dictionary, we derived an electronic version, limited to single word expressions. The SGML markup, in the version we had, was not always consistent, leading to a large quantity of noise in our electronic dictionary, especially in common words whose entries were the most complex, as well as patches of silence. For example, we discovered that our electronic version was missing an entry for the word *potato*, the main subject of topic CL17. We did not correct this error for our submitted TREC runs, and our experiments should be considered what one can expect when using a noisy dictionary[6].

6.2 Translation of Non-compositional Phrases

As used in these experiments, our dictionary performs only single word translation, modulo derivational variants. Before accessing the dictionary, we derivationally stem the word to be looked up, then generate all other lemmatised form of the same word, and concatenate the dictionary entries for all of these words. An equivalent strategy would be to derivationally

[6] We have begun to replace this dictionary with a cleaner though less complete multilingual dictionary available through the European Language Resources Association. http://www.icp.grenet.fr/ELRA.

normalize all the head words in the dictionary and conflate the translations of all the words stemming to the same form.

When we use this technique to translate phrases, such a technique only works if the phrases themselves are compositional phrases across the languages. For example, *éducation sexuelle* (CL7) can be translated word-by-word, modulo derivational variation, to *sex education*. But *ours en peluche* (CL25) cannot be translated compositionally into *teddy bear*, since *peluche* only translates to *plush* and *fluffy*. One solution to this problem is to have an exhaustive list of the non-compositional phrases of a language, with their translations, and a mechnanism [1] for recognizing instances of these expressions. We plan to incorporate this in future experiments for non-compositional expressions that are contained in our translation dictionaries.

In addition to derivational variants, it would be useful to include close synonyms to palliate word choice variations. For example, the French version of CL6 dealing with *air pollution* uses the rare term *pollution de l'atmosphère* whose derivational variants appear in the corpus one-half as frequently as the more common *pollution de l'air*. Similarly, *organic farming* (CL12) was described in the French version as *agriculture écologique* rather than the more common, compositionally translatable *agriculture biologique* (which appears 75 times in the French newsire versus just 6 occurrences for *agriculture écologique*).[7]

6.3 Stemming

In a general language dictionary, headwords often correspond to only one member of a derivational family. To recover other related words, we have implemented a derivational stemmer for the languages that we have been treating. Since this derivational stemmer is in its first iteration, we have seen a number of words that are over-stemmed, introducing noise into our results. The query translation program is complex, first generating derivational variants, then translating the variants, and finally stemming the resulting translations. We have discovered a bug in the last stemming step, leading to some differences between how words are stemmed in the collection and how they are stemmed in the translated queries. This results in a substantial over-generation of target language terms, since we make no attempt to disambiguate during the translation process. In most cases, this is not harmful, but we need to explore more carefully whether disambiguation might be helpful. We used the first experimental version of the derivational stemmer, and it can certainly benefit from further work.

7 Conclusion

Our conclusions at this point can only be partial. We still need to perform more comprehensive tests to measure the impact of each different component in our system, and the small topic set will limit our ability to measure significant differences. Though cross language retrieval is feasible in many cases using the existing technology, many steps in the treatment that are specific to cross language information retrieval remain to be mastered: proper level of derivational stemming, production of translation alternatives, translation of

[7] Since one of the translations of *biologique* is organic and one of the translations of *agriculture* is farming.

unknown words, translations of non-compositional phrases, proper handling of the weighting of retained translation alternatives. The work done so far only highlights a few areas where possibilities of improvement is evident, and more experimentation has to be performed to clarify the respective importance of each translation step.

References

[1] Daniel Bauer, Frederique Segond, and Annie Zaenen. Locolex: The translation rolls off your tongue. In *Proceedings of the ACH/ALLC '95*, Santa Barbara, California, July 11–15 1995.

[2] Chris Buckley. Implementation of the smart information retrieval system. Technical Report 85-686, Cornell University, 1985. SMART is available for research use via anonymous FTP to ftp.cs.cornell.edu in the directory /pub/smart.

[3] Doug Cutting, Julian Kupiec, Jan Pedersen, and Penelope Sibun. A practical part-of-speech tagger. *Proceedings of the Third Conference on Applied Natural Language Processing*, April 1992.

[4] Mark Davis. New experiments in cross-language text retrieval at NMSU's computing research lab. In *The 5th Text Retrieval Conference (TREC-5)*, 1997. To appear.

[5] Joel L. Fagan. *Experiments in Automatic Phrase Indexing for Document Retrieval: A Comparison of Syntactic and Non-Syntactic Methods*. PhD thesis, Cornell University, September 1987.

[6] David A. Hull. A probabilistic model for the approximate matching of boolean constraints. Being cleared for publication, 1997.

[7] David A. Hull. Using structured queries for disambiguation in cross-language information retrieval. In *AAAI Spring Symposium on Cross-Language Text and Speech Retrieval*, 1997. To appear.

[8] David A. Hull and Gregory Grefenstette. Querying across languages: A dictionary-based approach to multilingual information retrieval. In *Proc. of the 19th ACM/SIGIR Conference*, pages 49–57, 1996.

[9] David A. Hull, Gregory Grefenstette, Eric Gaussier, B. Maximilian Schulze, Hinrich Schütze, and Jan Pedersen. Xerox TREC-5 site report:routing, filtering, nlp, and spanish tracks. In D.K. Harman, editor, *The Fifth Text REtrieval Conference (TREC-5)*. U.S. Department of Commerce, 1997. NIST Special Publication 500.

[10] Lauri Karttunen, Ronald M. Kaplan, and Annie Zaenen. Two-level morphology with composition. In *Proceedings COLING '92*, pages 141–148, Nantes, France, August 23–28 1992.

[11] Gerard Salton. Automatic processing of foreign language documents. *Journal of the American Society for Information Science*, 21:187–194, 1970.

[12] Anne Schiller. Multilingual finite-state noun phrase extraction. In *Workshop on Extended finite state models of language*, Budapest, Hungary, Aug 11–12 1996. ECAI'96.

[13] Anne Schiller. Multilingual part-of-speech tagging and noun phrase mark-up. In *15th European Conference on Grammar and Lexicon of Romance Languages*, University of Munich, Sept 19–21 1996.

A CLIR TRACK QUESTIONNAIRE: XEROX

```
--------------------------------------------------------------------------------
CLIR TRACK QUESTIONNAIRE:
--------------------------------------------------------------------------------


1. OVERALL APPROACH:
----------------------

  1.1 What basic approach do you take to cross-language retrieval?
      [X] Query Translation
      [ ] Document Translation
      [ ] Other, _______________________


  1.2 Were manual translations of the original NIST topics used as a
      starting point for any of your cross-language runs?
      [ ] No
      [X] Yes, XRCECLF2EA  XRCECLE2FA

  1.3 Were the automatically translated (Logos MT) documents used
      for any of your cross-language runs?
      [X] No
      [ ] Yes, _______________________

  1.4 Were the automatically translated (Logos MT) topics used
      for any of your cross-language runs?
      [X] No
      [ ] Yes, _______________________



2. MANUAL QUERY FORMULATION:
------------------------------

  2.1 If query formulation involved manual effort, how fluent was the
      user in the source (query) language?
      [X] native speakers,
          XRCECLF2EM  XRCECLE2EM  XRCECLF2EM  XRCECLF2FM

  2.2 If query formulation involved manual effort, how fluent was the
      user in the target (document) language?
      [ ] does not apply
```

3. USE OF MANUALLY GENERATED DATA RESOURCES:

3.1 What kind of manually generated data resources were used?
- [X] Dictionaries
- [ ] Thesauri
- [X] Part-of-speech Lists
- [X] Other, part-of-speech taggers, lemmatizers, noun phrases extractors

3.2 Were they generated with information retrieval in mind or were they taken from related fields?
- [ ] Information Retrieval
- [ ] Machine Translation
- [X] Linguistic Research
- [X] General Purpose Dictionaries
- [ ] Other, _______________________

3.3 Were they specifically tuned for the data being searched (ie. with special terminology) or general-purpose?
- [ ] Tuned for data; Please specify _______________________
- [X] General purpose

3.4 What amount of work was involved in adapting them for use in your information retrieval system.
- [X] None
- [ ] _______________________

3.5 Size
- [X] 38,000 entries
- [X] 1 MBytes compressed

3.6 Availability? - Please also provide sources/references!
- [ ] Commercial
- [X] Proprietary, Oxford University Press
- [ ] Free
- [ ] Other, _______________________

4. USE OF AUTOMATICALLY GENERATED DATA RESOURCES: None
----

5. GENERAL

5.1 How dependent is the system on the data resources used? Could they easily be replaced if better sources were available?
    [ ] Very dependent, _______________________
    [ ] Somewhat dependent, _______________________
    [X] Easily replacable, _______________________
    [ ] Don't know

5.2 Would the approach used potentially benefit if there were better data resources (e.g. bigger dictionary or more/better aligned texts for training) available for tests?
    [X] Yes, a lot, _______________________
    [ ] Yes, somewhat, _______________________
    [ ] No, not significantly, _______________________
    [ ] Don't know

5.3 Would the approach used potentially suffer a lot if similar data resources of lesser quality (noisier dictionary, wrong domain of terminology) were used as a replacement?
    [ ] Yes a lot, _______________________
    [X] Yes, somewhat, _______________________
    [ ] No, not significantly, _______________________
    [ ] Don't know

5.4 Are similar resources available for other languages than those used?
    [X] Yes, ELRA dictionaries for English-French-German-Spanish-Italian
    [ ] No