Workshop Programme

31/05/2008	**Workshop on Comparable Corpora**
9:00	***Welcome and Introduction***
9:15	***Oral Session 1: Some Challenges*** *Translation universals: do they exist? A corpus-based and NLP approach to convergence* Gloria Corpas Pastor, Ruslan Mitkov, Naveed Afzal, Lisette Garcia Moya *Enhancing a Statistical Machine Translation System by using an Automatically Extracted Parallel Corpus from Comparable Sources* Sanjika Hewavitharana, Stephan Vogel
10h15	Coffee break + Poster session 1 *(see list of poster presentations on page iii)*
11:00	***Oral Session 2: Extracting Bilingual Lexicons from Comparable Corpora*** *Translating Named Entities using Comparable Corpora* Iñaki Alegria, Nerea Ezeiza, Izaskun Fernandez *Evaluating Two Different Methods for the Task of Extracting Bilingual Lexicons from Comparable Corpora* Pablo Gamallo Otero *Automatic extraction of bilingual terms from comparable corpora in a popular science domain* Xabier Saralegi, I. San Vicente, A. Gurrutxaga
12:30	***Invited talk (speaker to be announced)***
13h30	Lunch break
14:30	***Oral session 3: Linguistic studies*** *Functional-Typological Approaches To Parallel And Comparable Corpora: The Bremen Mixed Corpus* Christel Stolz, Thomas Stolz *On the use of comparable corpora of African varieties of Portuguese for linguistic description and teaching/learning applications* Maria Fernanda Bacelar do Nascimento, Antónia Estrela, Amália Mendes, Luísa Pereira *Empirical studies on language contrast using the English-German comparable and parallel CroCo corpus* Oliver Culo, Silvia Hansen-Schirra, Stella Neumann, Mihaela Vela
16h00	Coffee break + Poster session 2 *(see list of poster presentations on page iii)*
16h45	***Panel session***
16h45	**Panel:** *Comparable corpora: varying definitions, varying uses*
18h00	End of workshop

Workshop Organisers

Pierre Zweigenbaum	LIMSI, CNRS, Orsay, France
Éric Gaussier	LIG, Université Joseph Fourier, Grenoble, France
Pascale Fung	Department of Electronic & Computer Engineering, University of Science & Technology, Hong Kong

Workshop Programme Committee

Lynne Bowker	University of Ottawa, Canada
Hervé Déjean	Xerox Research Centre Europe, Grenoble, France
Éric Gaussier	Université Joseph Fourier, Grenoble, France
Gregory Grefenstette	CEA/LIST, Fontenay-aux-Roses, France
Pascale Fung	University of Science & Technology, Hong Kong
Natalie Kübler	Université Paris Diderot, France
Tony McEnery	Lancaster University, UK
Emmanuel Morin	Université de Nantes, France
Dragos Stefan Munteanu	Information Sciences Institute, Marina Del Rey, USA
Carol Peters	ISTI-CNR, Pisa, Italy
Reinhard Rapp	Johannes Gutenberg-Universität Mainz, Germany
Serge Sharoff	University of Leeds, UK
Monique Slodzian	INALCO, Paris, France
Richard Sproat	University of Illinois at Urbana-Champaign, USA
Pierre Zweigenbaum	LIMSI-CNRS, Orsay, France

Table of Contents

Poster presentations

Author Index

Foreword

Research in comparable corpora is motivated by the scarcity of parallel corpora. Parallel corpora are a key resource to mine translations for statistical machine translation or for building or extending bilingual lexicons and terminologies. However, beyond a few language pairs such as English-French or English-Chinese and a few contexts such as parliamentary debates or legal texts, they remain a scarce resource, despite the creation of automated methods to collect parallel corpora from the Web. A more fundamental limitation is that translated texts, whatever the skills of translators, are generally influenced by the very translation process and by the language of source texts, so that they may not be fully adequate for the task at hand.

This has motivated research into the use of comparable corpora: pairs of monolingual corpora selected according to the same set of criteria, but in different languages or language varieties. Comparable corpora overcome the two limitations of parallel corpora, since sources for original, monolingual texts are much more abundant than translated texts. However, because of their nature, mining translations in comparable corpora is much more challenging than in parallel corpora. What constitutes a good comparable corpus, for a given task or per se, also requires specific attention: while the definition of a parallel corpus is fairly straightforward, building a comparable corpus requires control over the selection of source texts in both languages.

This workshop aimed to bring together researchers interested in the constitution and use of comparable corpora. Contributions were solicited on the constitution and application of comparable corpora, including the following topics:

Applications of comparable corpora:

- tools for translators;
- tools for language learning;
- cross-language information retrieval;
- cross-language document categorization;
- machine translation;
- monolingual comparable corpora for writing assistance;
- extraction of parallel segments in comparable corpora.

Units aligned in comparable corpora:

- single words and multi-word expressions; proper names; alignment across different scripts.

Constitution of comparable corpora:

- criteria of comparability;
- degree of comparability;
- methods for mining comparable corpora.

We are very glad that this topic attracted papers from both Linguists (Session 3) and Computer Scientists (Session 2). The whole day will therefore be an opportunity to discover the views of both streams of research on a common object, and should give rise to lively and stimulating discussions. The challenging topics of Session 1 will open the workshop, an invited talk will focus the attention at mid-workshop, two poster sessions will keep the discussions going during extended breaks in the morning and afternoon, and a panel discussion will close the day.

Pierre Zweigenbaum, Éric Gaussier, Pascale Fung

Translation universals: do they exist?
A corpus-based and NLP approach to convergence

Gloria Corpas Pastor*, Ruslan Mitkov, Naveed Afzal** and Lisette García Moya*****

* University of Malaga, Email gcorpas@uma.es

** University of Wolverhampton, Email {r.mitkov, n.afzal}@wlv.ac.uk

*** Centre for Pattern Recognition and Data Mining, Santiago de Cuba, Email lisette.garciamoya@gmail.com

Abstract

Convergence is one of the so-called universals in translation studies which postulates that translated texts tend to be more similar than non-translated texts. This paper discusses the results of a project which applies NLP techniques over comparable corpora of translated and non-translated texts in Spanish seeking to establish whether this universal holds. The results of this project do not provide sufficient support to the validity of this universal.

1. Introduction

Studying the characteristics of translated text or more specifically, what distinctive features typically translated texts exhibit and how they differ from original, non-translated texts written by native speakers has been a topic of long-standing interest in translation studies. Initial research goes back to Toury (1995) who put forward the laws of growing *standardization* (sic) and the law of interference, but it was Baker (1993, 1996) who formulated many of the so-called universals and proposed the use of corpora to study these. The universals attracted considerable attention from translation experts but their formulation and initial explanation has been based of intuition and introspection with follow-up corpus research limited to comparatively small-size corpora, literary or newswire texts and semi-manual analysis. In addition, previous research has not provided sufficient guidance as to which are the features which account for these universals to be regarded as valid (Corpas Pastor, 2008).

In this paper we are taking a completely different and innovative approach by employing robust NLP techniques on corpora of translated texts into Spanish and on comparable corpora of non-translated Spanish in order to investigate the validity of translation universal of *convergence*. According to this universal, translated texts tend to be more similar than non-translated texts. The objective of this study is to establish whether this universal is valid with Spanish as target text. To this end, we analyse corpora of translated texts into Spanish and comparable corpora of Spanish non-translated texts. Then we compute similarity between every pair of corpora of translated texts and every pair of corpora of original texts for both languages. The similarity is measured in terms of both *style* and *syntax*.

2. Corpora used

According to the convergence universal, translated texts tend to be more similar than non-translated texts. The objective of this study is to verify whether this universal is valid with Spanish as the target language. To this end, we compare pairs of corpora of translated texts as well as pairs of comparable corpora of original, non-translated Spanish texts in terms of style and syntax with a view to establishing whether translated texts are found to be more similar than non-translated texts. We specifically compiled the following corpora for this experiment:

- Corpus of Medical Spanish Translations by Professionals (MSTP)
- Corpus of Medical Spanish Translations by Students (MSTS)
- Corpus of Technical Spanish Translations (TST)
- Corpus of Original Medical Spanish Comparable to Translations by Professionals (MSTPC)
- Corpus of Original Medical Spanish Comparable to Translations by Students (MSTSC)
- Corpus of Original Technical Spanish Comparable to Technical Translations (TSTC)

As stated above, MSTP is comparable to MSTPC, MSTS is comparable to MSTSC and TST is comparable to TSTC. Comparability was a crucial consideration for this study as otherwise any style or syntax comparison would have been compromised.

We compiled the corpora in such a way that comparability was ensured. Design criteria comprise diatopic, diachronic, diasystematic and domain constraints. All translated texts have British or American English as the source language and peninsular Spanish as the target language. Both corpora of translated and non-translated texts have roughly the same size. MTSP is composed of biomedical translations performed by professional translators (in-house or freelancers working for certified translation companies in Europe). It is a specialised reference corpus as it does not contain whole documents, but fragments composed of the TL segments of translation memories (TMs). Text types range from research papers in journals to clinical essays, textbooks, product description and PILs, users' guides and instructions for surgical equipment. Its comparable corpus of non-translated biomedical Spanish includes a similar selection of text types and topics. It is a mixed corpus, as it contains fragments and whole documents: SL segments of TMs different from the ones used to compile the MTSP, a small corpus of diabetes and an ad-hoc virtual corpus compiled to match MTSP as regards sub-domains, topics, level of communicative specialisation and text types. The other corpus of biomedical Spanish is a specialised textual corpus that contains whole documents, i.e. translations by last-year undergraduates in Translation and Interpreting during the academic years 2004-2205, 2005-2006 and 2006-2007. It comprises almost the same text types and topics as the MTSP, but with a higher proportion of research papers, product descriptions and PILs. The MSTSC is comparable to the MTSP as they share similar design criteria.

Finally, the TST comprises TL segments of TMs of technical and technological domains (telephony, network services, telecommunications, etc.) and the CRATER Spanish subcorpus. It comprises fragments from user manuals, guides and operating instructions, companies press releases and, to a lesser extent, rules and regulations, standards, projects and monographies. The TSTC has been compiled ad-hoc from evaluated electronic sources. After analysing the TST in terms of text types, domains and topics, we have derived a catalogue of index words and search equations. As a result, we have ended up compiling a corpus which is partially comparable to the TST, as it contains whole documents (not just fragments). It should be pointed out that locating this kind of technical documents in peninsular Spanish has proved to be more complicated than finding original medical Spanish, as many texts of this kind are covert translations. We have ensured that only non-translated original technological texts are included by filtering and refining all electronic searches.

The size of the above corpora (no. of tokens) is as follows[1]:

- MSTP: 1,058,122

- MSTS: 780,006

- TST: 1,736,027

- MSPC: 1,402,172

- MSTSC: 1,164,435

- TSTC: 1,986,651.

Therefore, the corpora of translated Spanish and non-translated Spanish are *comparable* on the following grounds:

(i) The pairs of translated and non-translated corpora include roughly the same range of text types and forms

(ii) They belong to the same domains and sub domains

(iii) They exhibit the same level of specialisation and formality

(iv) They are restricted diatopically to Peninsular Spanish

(v) They were produced during the same span of time (2005-2008)

(vi) They are of a similar size (no. of tokens).

3. Methodology

We compared all 3 pairs of translated texts (MSTP-MSTS; MSTS-TST; MSTP-TST) and all 3 pairs of comparable non-translated texts (MSTPC-MSTSC; MSTSC-TSTC; MSTPC-TSTC). If the convergence universal holds, we would expect to find higher similarity for pairs of translated texts.

Previous studies on universals, unfortunately, have not accounted for what exactly classes as evidence in terms of different features for their validity. Therefore we first have to ask the question when a text or a corpus is more or less similar to another text or corpus. It is important to know what the features or parameters of similarity are so that formal empirical studies can be conducted to compare texts in terms of similarity and more specifically to verify whether translated texts 'converge' in that in general are more similar than non-translated texts. In the absence of any such guidelines, the first step to take in this study is to identify features which could be used for measuring similarity of translated or non-translated texts.

We propose to assess to what extent translated or non-translated texts 'converge' on the basis of (i) style

[1] Whereas the size of these corpora is small by today's standards, we should not that any previous corpus analysis on translation universals (e.g. Laviosa's (2002) work on simplification) has covered even smaller data.

(stylistic features) and (ii) syntax (syntactic features). This experiment covers the following style characteristics[2]: lexical richness (type/token) ratio, lexical density, sentence length, use of simple as opposed to complex sentences, use of aspect, discourse markers as well as conjunctions. Unlike any previous corpus-based work on universals (simplification), we perform stemming of each corpus so that the results related to lexical richness are not compromised in that two morphological variants of a word (e.g. *experiment*, *experiments*) are not regarded as two different words. The analysis of general syntactic patterns is unique in that no such previous experiments have been carried out. We perform part-of-speech tagging/shallow parsing[3] for each corpus and compare the sequences of parts of tags which account for the linear syntactic structures. More specifically, vectors of n-grams are compared using cosine and recurrence metrics modelled as permutation tests (Nerbonne and Wiersma, 2006).

3.1 Style comparison

Lexical density: Lexical density is computed as type/token by dividing the number of types by the total number of tokens present in the corpus. Low lexical density involves a great deal of repetition with the same words occurring again and again. On the other hand, high lexical density means that a more diverse form of language is being employed.

Lexical richness: We argue that lexical density is not indicative of the vocabulary variety of an author as it counts morphological variants of the same word as different word types. However, whereas *student* and *students* may technically be separate words and word types, from lexical point of view they represent the same word. To alleviate this inadequacy, we propose a new measure lexical richness, which is computed as the number of lemmas divided by the number of tokens present in the corpus and accounts for the variety of word use by an author. The lemma of every word is automatically returned by the Connexor parser.

Sentence length: Sentence length is a feature deemed to be typical of an individual style. We compute sentence length as the number of tokens in corpus divided by the number of sentences in this corpus. In this study, unlike Study 1, we have opted for not including the parse tree depth as a stylistic feature because (a) the parse three is more a syntactic concept and (b) we believe the parse three depth and sentence length are not completely independent features.

Simple sentences vs. complex sentences: We argue that whether the use of predominantly simple or complex sentences, or balanced combination of both, is a relevant feature for the style of an author. In order to count the number of simple or complex sentences we developed an algorithm to automatically identify the type of sentence by counting the number of finite verbs (and their corresponding verbal constructions) in a sentence; sentences with more than one finite verb are classified as complex. Constrictions such as (HABER, TENER or SER) + Past Participle and ESTAR + Gerund are counted as well. Verbs are detected by the Connexor parser, so are past participles and gerunds. We have computed the proportion of cases where simple or complex sentences are used.

Discourse marker: According to Biber (1988, 1995, 2003), the use of discourse markers is another characteristic of someone's style. To this end, using a list of discourse markers in Spanish, we have extracted and calculated the proportion of both discourse markers from the number of all words in a corpus.

In order to compute similarity between each pair of translated and non-translated texts, two statistical tests (Chi-Square test and T-test) are employed. Chi-square takes all features used and produces a global score of similarity between the corpora analysed. T-test does not provide a global score but instead compares separate features and establishes any statistically significant difference or not.

3.2 Syntax comparison

In this experiment we compare sequences of POS tags between for every pair of corpora. Sequences of POS tags account for the linear syntactic structure of sentences and the idea behind our general methodology consists of comparing any two corpora taking into account n-grams. Previously, n-grams of POS tags have been used to measure syntactic distance and best results have been reported for n=3 (Nerbonne and Wiersma, 2006). The corpora to be compared are represented as frequency vectors of 3-grams and the measures employed for comparison are the cosine as well as the measures R and Rsq which were inspired by the recurrence (R) metric (Kessler, 2001).

4. Results

This section reports the results of the experiments/comparisons described above and seeks to offer insights whether convergence holds as a universal.

4.1 Style comparison

In order to compare the style of translated texts as well as

[2] Some of these features have been adopted from Biber (1993, 1995); other such as the type of sentences, are our own proposals. It is worth noting that the set of stylistic features is language dependent. For example, the use of active or passive voice would have been more interesting for English or German.

[3] Part-of-speech tagging /shallow parsing is performed using Connexor's Machinese.

the style of non-translated texts, we first compute the style features lexical density, lexical richness, the average sentence length, proportion of simple/complex sentences and discourse markers (Table 1).

Features	**MSTP**	**MSTS**	**TST**	**MSTPC**	**MSTSC**	**TSC**
Lexical Density	0.027954	0.052715	0.020679	0.042505	0.041159	0.025529
Lexical Richness	0.016929	0.037709	0.013281	0.029992	0.028905	0.015591
Average Sentence Length	25.256248	28.499456	27.292782	20.702349	26.442412	18.124363
Simple Sentences (%)	0.441768121	0.507205751	0.476949103	0.638889238	0.52120611	0.592110096
Discourse Markers (Ratio)	0.001268941	0.001852604	0.000763805	0.002022331	0.002099085	0.001649655

Table 1: Stylistic features

Next, we compute similarity between each pair of translated and non-translated texts using the results obtained for the above features (lexical density, lexical richness, sentence length, simple sentences proportion, discourse markers) in two statistical tests: Chi-Square test and T-test. The Chi-Square values obtained for each pair of corpus of translated and non-translated texts are as displayed in Table 2.

Translated Corpora

Corpora	**Chi-Square Values**
1MSTP → 2MSTS	0.010622566
1MSTP → 3TST	0.00266151
2MSTS → 3TST	0.023731912
Total	**0.037015988**
Average	**0.012338663**

Non-translated Corpora

Corpora	**Chi-Square Values**
1MSTPC → 2MSTSC	0.059779549
1MSTPC → 3TSC	0.006140764
2MSTSC → 3TSC	0.07122404
Total	**0.137144352**
Average	**0.045714784**

Table 2: Ch-Square values

Finally, we conducted T-tests for statistical significance. In order to conduct T-test, each corpus was divided into small subsets of equal size. For each subset the figures for the above stylistic features are computed and compared with the figures of the corresponding subsets of the corpus being compared.

Features	**Translated Corpora (T-test Values)**			**Non-translated Corpora (T-test Values)**		
	MSTP → MSTS	MSTS → TST	MSTP → TST	MSTPC → MSTSC	MSTSC → TSC	MSTPC → TSC
Lexical Density	0.002545387	0.000123172	0.079875166	0.140348431	0.201151185	0.000748439
Lexical Richness	0.0006604	0.000006.9792	0.140236542	0.140711253	0.015893183	0.00009.71905
Sentence Length	0.011826639	0.522122939	0.202480843	0.145216739	0.002807505	0.368840258
Simple Sentences	0.057465277	0.673936375	0.202830407	0.096465071	0.462960518	0.21217697
Discourse Markers	0.001048007	0.005746253	0.351552034	0.063428055	0.00084074	0.072337471

Table 3: T-Test values

4.2 Syntax comparison

We assess syntax similarity (in our case dissimilarity) between each pair of translated and non-translated texts by comparing sequences of 3-grams of part-of-speech (POS) tags for every pair of corpora. We first run the Connexor parser to identify all POS tags, then collect frequency vectors of 3-grams whose dissimilarity is compared on the basis of the 1-C (C=cosine), R and Rsq measures.

.

Corpora	1-C	R	Rsq
Translated texts			
MSTP - MSTS	0.206015066283	252526.914323	638848591.082
MSTP - TST	0.337626383799	388466.504863	3146471863.13
MSTS - TST	0.176310545152	432725.578482	2643068563.82
Non-Translated texts			
MSTPC - MSTSC	0.0176469276126	98448.0858054	82218137.9687
MSTPC - TSC	0.150912596476	364322.217714	851312764.364
MSTSC - TSC	0.167167511143	372940.61477	1008322991.78

Table 4: Results measuring vector differences

More specifically, for every corpus we build a frequency vector featuring all trigrams of POS tags. For example, the comparison of the frequency vectors of the corpus of all translated texts (MSTP+MSTS+TST) and the corpus of non-translated texts (MSTPC+MSTSC+TSTC) involves a total of 18, 468 different POS.[4] Table 4 below represents the results obtained from comparing the pairs of corpora applying the aforementioned dissimilarity measures. The higher values of the measures employed indicate greater dissimilarity (and less similarity) between two corpora under comparison.

5. Discussion and Conclusion

The average Chi-Square values[5] of translated texts are smaller than average Chi-Square values of non-translated texts (Table 2) which implies that the translated texts included in our experiment are more similar than non-translated texts with regard to the stylistic features studies. On the basis of the corpora used and the features employed, it appears that the convergence universal holds on this occasion.

- The T-Test values (Table 3) of non-translated texts show that there is no significant difference between any of the above mentioned list of features in MSTPC→ MSTSC pair and the MSTPC→ TSC pair results show that there is a significant difference between lexical density and lexical richness, while in the MSTSC→TSC pair there is a significant difference among 3 features (lexical richness, sentence length and discourse markers). In case of translated texts the T-Test values of the pair MSTP→ MSTS significantly differ in terms of lexical density and discourse markers and of the pair MSTS→TST significantly differ lexical density and lexical richness. There is no significant difference between MSTP→ TST.

- From the T-test results it is clear that whereas the Chi-square test suggests general greater similarity between translated texts, we can make several interesting observations.

- There are non-translated texts which are not statistically different in terms of the chosen stylistic features whereas the corresponding comparable corpora of translated texts different statistically with regard to two stylistic features (see the pairs MSTPC -> MSTSC and MSTP-> MSTS respectively)

- There are non-translated texts which are statistically different in terms of only one stylistic feature whereas the corresponding comparable corpora of translated texts different statistically with regard to two stylistic features (see the pairs MSTPC -> TSC and MSTP-> TST respectively)

- Translated texts could often differ significantly with regard to certain style features (MSTP -> MSTS; MSTS -> TST) of which especially surprising is the lexical density. Whereas difference in the lexical

[4] We compare a total of 8,484 trigrams between MSTP and MSTS, 9,954 trigrams between MSTP and TST and 10,019 between MSTS and TST. We also compare 8,278 trigrams between MSTPC and MSTSC, 13,297 trigrams between MSTPC and TSC and 13,007 between MSTSC and TSC.

[5] The smaller Chi-Square value indicates the bigger similarity between the two corpora.

density between student and professional translators could be somehow acceptable, statistical difference in lexical density between professional translators is unexpected.

Therefore, on the basis of our data and with regard to the style features adopted, whereas convergence appears to be broadly holding, we argue that no definite conclusion can be made that convergence is a clear-cut universal due to the above T-test results. In the case of an absolute, clear-cut universal, one would not have expected results such as the ones stated in (i) and (ii) above.

From Table 4 it is clear that translated texts differ more in terms of syntax for all compared pairs and from the point of view of all measures (1-C, R and Rsq). It is also clear that the difference of syntax is greater between texts of different domains. On the basis of the above results we can conclude that there is no evidence that convergence holds in terms of syntax. In fact, the results from Table 4 even show that translated texts differ more syntactically than non-translated texts on our experimental data.

In general, the results do not provide sufficient support to the convergence universal.

6. References

Baker, M. 1993. “Corpus Linguistics and Translation Studies – Implications and Applications”. In: M. Baker, M.G. Francis & E. Tognini-Bonelli (eds.). 1993. *Text and Technology: In Honour of John Sinclair.* Amsterdam & Philadelphia: John Benjamins. 233-250.

Baker, M. 1996. “Corpus-based Translation Studies: The Challenges that Lie Ahead”. In: H. Somers (ed.). 1996. *Terminology, LSP and Translation: Studies in Language Engineering, in Honour of Juan C. Sager.* Amsterdam & Philadelphia: John Benjamins. 175-186.

Biber, D. 1988. *Variation across Speech and Writing.* Cambridge: Cambridge University Press.

Biber, D. 1995. *Dimensions of Register Variation: a Cross-Linguistic Comparison.* Cambridge: Cambridge University Press.

Biber, D. 2003. “Variation among University Spoken and Written Registers: A New Multi-dimensional Analysis”. In: P. Leistyna & C. F. Meyer (eds.). 2003. *Corpus Analysis. Language Structure and Language Use.* Amsterdam & New York: Rodopi. 47-70.

Corpas Pastor, G. 2008 (In press). *Investigar con corpus en traducción: los retos de un nuevo paradigma.* Frankfurt am Main, Berlin & New York: Peter Lang.

Kessler, B. 2001. *The Significance of Word Lists.* Stanford: CSLI Press.

Laviosa, S. 2002. *Corpus-based Translation Studies. Theory, Findings, Applications.* Amsterdam & New York: Rodopi.

Nerbonne J. and Wiersma, X. 2006. “A Measure of Aggregate Syntactic Distance”. In: J. Nerbonne & E. Hinrichs (eds.) 2006. *Linguistic Distances. Workshop at the joint conference of International Committee on Computational Linguistics and the Association for Computational Linguistics*. Sidney, Australia. 82-90.

Toury, G. 1995. Descriptive Translation Studies and Beyond. Amsterdam: John Benjamins.

Enhancing Statistical Machine Translation with Parallel Data extracted from Comparable Corpora

Sanjika Hewavitharana and Stephan Vogel
Language Technologies Institute, School of Computer Science
Carnegie Mellon University, Pittsburgh, PA 15213, USA.
{sanjika, vogel+}@cs.cmu.edu

Abstract

In this paper, we present the results of experiments to enhance the performance of a baseline statistical machine translation system with an automatically extracted parallel corpus from comparable corpora. We train new translation systems by combining the extracted corpus with different sizes of parallel corpora. We also experiment with combining the phrase tables trained separately from the two resources. These phrase tables are then used in a phrase-based SMT decoder to translate test sets. Results indicate that the extracted corpus helps improve performance when the initial parallel corpus is small. The additional data also helps to reduce the number of unrelated words.

1. Introduction

Parallel corpora are an important resource for many natural language processing tasks, especially statistical machine translation (SMT) where a large amount of training data is required to produce reliable models. However, due to the special effort that is required to create them, which is time consuming and costly, these corpora are limited in quantity, genre and language coverage. Large parallel corpora are only available for a handful of languages including English, French, Chinese and Arabic. Majority of this data comes from parliamentary proceedings, and a limited amount of newswire text is also available. For the majority of other languages, parallel corpora are virtually non-existent.

One potential solution to the bottleneck of data sparseness is to exploit comparable corpora. These are documents that are rough translations of each other, containing overlapping information. Multilingual newswire texts produced by news organizations such as AFP and Reuters are a good example for comparable corpora. These texts often describe the same event in multiple languages in varying degree of detail. They often contain sentences that are fairly good translations of each other; sometimes parallel sentence pairs. The Web is by far the largest source of comparable texts. It contains an ever-expanding body of text in multiple languages that can be mined for comparable documents.

These resources can be used in a number of ways to enhance translation systems. The most straight forward way is to mine for parallel document pairs from the web and other multilingual sources. These parallel documents can then be used directly in training translation models. Most of the time, however, it is hard to find entire documents of parallel text. Comparable documents are much easier to find. Parallel sentence pairs can be identified from comparable documents that convey the same information.

There have been several attempts in the past to exploit bilingual comparable corpora. Many of the early efforts focused on learning bilingual lexical translations from comparable sources (Fung and Yee, 1998). Searching for parallel sentence pairs within comparable news corpora have been attempted by extending parallel sentence alignment algorithms (Zhao and Vogel, 2002) and also using cross-lingual information retrieval techniques (Munteanu and Marcu, 2005). Most of these research attempts were concentrated on large document collections containing newswire and political dialogues. Another related effort had been mining the web for parallel documents by exploiting the similarities in the URL structure, document structure, and other clues (Resnik and Smith, 2003).

In this paper, our primary interest is to investigate if comparable corpora can be effectively used to improve the performance of our SMT system. We will also investigate the effect of comparable corpora when used with different sizes of parallel data to build translation systems. For our experiments, we use a corpus containing parallel sentence pairs automatically extracted from comparable sources. We conduct translation experiments with this corpus as well as a large Arabic-English parallel corpus.

2. Parallel Corpus from Comparable Data

For our experiment, used the *ISI Arabic-English Automatically Extracted Parallel Corpus* (ISI) which was released under the LDC catalog number LDC2007E07. This corpus contains Arabic-English parallel sentences, which were automatically extracted from two large monolingual corpora: Arabic Gigaword and English Gigaword. Documents in these two corpora are mainly from newswire sources such as AFP and Xinhua.

The process used to identify parallel sentence pairs is explained in Munteanu and Marcu (2005). Potential parallel document pairs are first identified using cross-lingual information retrieval methods. A word-overlap filter is then used to select candidate sentence pairs. Finally, a maximum entropy based classifier is used to decide if the sentences in each pair are translations of each other. Table 1 lists the corpus statistics.

	Arabic	English
Sentences	1,124,609	
Words	28,880,558	30,856,669
Vocabulary	532,443	388,761

Table 1: Characteristics of the ISI corpus

3. Experiments

We evaluate the effect of the ISI parallel corpus by directly using it in our SMT system. We train multiple translation systems using a large parallel corpus, the ISI corpus, and combination of both corpora. Following sections give details of the data sources and the experimental setup that was used.

3.1 Data Sources

As our primary parallel corpus, we use a collection of Arabic-English parallel corpora released by LDC which includes data from news genre as well as UN proceedings. The collection has over 100 million words for each language. We use this corpus to train the baseline translation system. ISI corpus is roughly one third of the size of the baseline parallel corpus.

As a pre-processing step, we separate punctuations from words on both sides, and convert English side into lower case. Table 2 lists corpus statistics for both the parallel corpus and the ISI corpus after pre-processing.

	Arabic	English
Parallel Sentences	6,880,398	
Words	101,994,860	117,227,473
Vocabulary	532,330	247,265
ISI Sentences	1,124,609	
Words	30,639,122	35,292,131
Vocabulary	322,403	164,504
Language Model		
Words	-	231,706,912
Vocabulary	-	1,070,392
Dev Sentences	1,056	-
Words	28,293	-
Vocabulary	7,712	-
Test Sentences	1,797	-
Words	41,059	-
Vocabulary	12,067	-

Table 2: Characteristics of the data

To train the language model we use the English side of the primary parallel corpus as well as part of the English Gigaword corpus.

We use two test sets from previous NIST evaluations to evaluate the translations: MT05 as the development test set, and MT06 (NIST-part) as the unseen test set.

3.2 N-gram Coverage

One of the ways the additional training data can help is by providing translations for words that are not already covered by the primary corpus. It can also provide additional long n-grams that match with the test data. This helps to improve translation quality by avoiding erroneous re-orderings produced by the decoder when using a collection of shorter phrases.

We compared the n-gram coverage of our training corpora for the development test set MT05. Coverage was calculated for the primary parallel corpus (Baseline), ISI corpus and both corpora combined (Basline+ISI). Table 3 gives the n-gram matching statistics. N-gram matching percentage is given within parenthesis.

n	# n-grams	Baseline	ISI	Baseline+ISI
1	28,293	28,128 (99.4)	28,047 (99.1)	28,188 (99.6)
2	27,237	23,114 (84.9)	22,284 (81.8)	23,983 (88.1)
3	26,181	13,375 (51.1)	12,693 (48.5)	15,189 (58.0)
4	25,125	5,873 (23.4)	5,917 (23. 6)	7,617 (30.3)
5	24,069	2,459 (10.2)	2,593 (10.8)	3,592 (14.9)
6	23,015	1,117 (4.9)	1,189 (5.2)	1,761 (7.7)
7	21,962	561 (2.6)	548 (2.5)	916 (4.2)

Table 3: N-gram coverage for MT05

To avoid potential overlap between the ISI corpus and the test sets, we removed documents in the ISI corpus which falls within the black-out periods of NIST test sets we used. We also replaced numbers with a tag, in both training data and test sets, to increase the coverage of the n-grams across numbers.

As expected, the large parallel corpus has better n-gram coverage than the smaller ISI corpus. However, the combined corpus has a considerably larger coverage than individual corpora, especially for higher order n-grams. This shows that the additional data has the potential to improve translation quality, if we can reliably identify translation equivalencies.

3.3 Translation System

We generated IBM word alignment models by running GIZA++ (Och and Ney, 2003) with the parallel text. These alignments were then used to extract the phrase table which is identical to Moses (Koehn et al., 2007). An n-gram suffix array language model was used for all the experiments. The phrase table and the language model were then used in a phrase-based SMT decoder to translate the test sets. The decoder performs minimum error-rate training (Och, 2003) on the development set, to find the best scaling factors for the models used.

3.4 Evaluation Results

3.4.1 Translation Experiments

The most straightforward way to evaluate the effect of additional data is to train a translation system with and without it. We trained a baseline translation system using the large parallel corpus. Then we trained a second translation system by combining the ISI corpus and the baseline corpus. Additionally, for comparative purpose, we also generated a translation system using only ISI corpus. All these systems were tested on a development set (MT05) and an unseen test set (MT06). Table 4 shows the translation results in case-insensitive Bleu (Papineni et al., 2002) scores.

	MT05	MT06
Baseline	53.37	40.73
ISI	49.46	33.37
Baseline+ISI	53.35	40.12

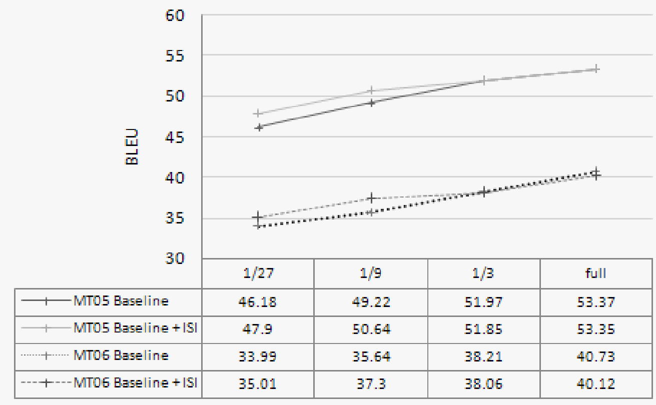
Table 4: Translation results in Bleu for the full corpus

The combined corpus (Baseline+ISI) shows no improvement over the baseline for the development set. We see a drop in the performance for the unseen test set. The system trained only on the ISI corpus (ISI) gives the lowest scores. It is no surprise when considering its relatively small size and the fact that it was automatically extracted from comparable sources. However, for the development set, the score is very close to the baseline.

3.4.1.1. Experiments with different sizes of corpora

We also wanted to observe the effect of automatically extracted sentences with different sizes of parallel corpora. For many languages, the amount of parallel data available is very limited. We therefore use these Ar-En experiments to simulate such poor-resource scenarios. We generated three parallel corpora with 1/3, 1/9 and 1/27 of the original size. The 1/3 corpus with 39 million English words is similar in size to the ISI corpus. The 1/9 and 1/27 corpora (with 13 million and 4 million English words) can be considered as medium and small sized corpora, respectively. These two corpora better match the resource levels for many languages.

For each parallel corpus, we generated a baseline translation system and a second system by adding ISI corpus as well. Results of these experiments are given in Figure 1.



	1/27	1/9	1/3	full
MT05 Baseline	46.18	49.22	51.97	53.37
MT05 Baseline + ISI	47.9	50.64	51.85	53.35
MT06 Baseline	33.99	35.64	38.21	40.73
MT06 Baseline + ISI	35.01	37.3	38.06	40.12

Figure 1: Translation results for different corpus sizes

When adding the ISI corpus, we see significant improvement in performance for both 1/27 and 1/9 corpora. When the parallel corpus and the automatically extracted corpus are of the same size (i.e. 1/3), we see performance starting to degrade, although the difference is not significant.

We looked closely at the word alignments and the phrase tables of each system. Automatically extracted parallel corpus contains sentence pairs with varying degrees of parallelism. Although many of the sentences are aligned fairly accurately, presence of some spurious alignments in the combined corpus has affected the distribution of different scores in the phrase table. One way to remedy the problem is to identify these problematic sentence pairs and remove them from the corpus before training the translation model.

3.4.1.2. Filtering

A simple filtering method based on alignment probability and lexical overlap was used to select sentences from the ISI corpus. Using this method we selected 10% and 20% of ISI corpus, and added it to the parallel data. For fast turnaround time, this was only performed on the 1/3 corpus. Filtering improved the performance, but is still lower than the baseline scores. A more sophisticated filtering approach would be required to identify most promising sentences from the automatically extracted corpus.

	MT05	MT06
1/3	51.97	38.21
1/3+ISI	51.85	38.06
1/3+20% of ISI	51.82	37.85
1/3+10% of ISI	51.98	38.18

Table 5: Translation results after filtering

3.4.2 Phrase Table Combination Experiments

We investigated on combining phrase tables from the parallel and ISI corpora. This was motivated by the strong results we see when using ISI corpus alone. Here, we used individually trained phrase tables from the full baseline and ISI corpora, instead of generating the phrase table from the combined corpus. By doing so, we try to reduce the negative effect on the clean phrase pairs from the baseline system.

We sampled both phrase tables for the test set individually and concatenated the phrases together into one phrase table. Each phrase pair is tagged so that we can identify which phrase table it originated from.

Translation results so far did not show improvement over the baseline. However, as Table 6 shows, the phrase table combination has managed to reduce the number of un-translated words due to the additional phrases found in ISI corpus.

	MT05
Baseline PT	262
ISI PT	488
Baseline PT + ISI PT	215

Table 6: Number of un-translated words

We looked closely at the translation lattices that were generated by the decoder to produce the final translations. This shows that there is still a problem in the combined phrase table. Although the new phrases from the ISI corpus increase the coverage, they do not match very well with the rest of the phrases from the baseline phrase table. This is mainly due to the fact that the two phrase tables were independently trained, and hence have different probability distributions for the same feature. Inside the decoder, phrase pairs from one phrase table is favored, as it has higher value for the feature.

4. Discussion and Conclusions

In this paper, we explained translation experiments to enhance the performance of a baseline statistical machine translation system with a parallel corpus extracted from comparable sources. We conducted several experiments by training new translation systems by combining the extracted corpus with different sizes of parallel corpora. Results show that we get the maximum benefit from automatically extracted data when the initial parallel corpus is small. When the amount of parallel data increases, the benefit diminishes.

The extracted corpus contains sentences with varying degrees of parallelism. Alignment errors in some of the non-parallel sentences contribute to the decrease in performance. Our effort to filter out some of these sentences showed improvements, but did not surpass the baseline system. A more sophisticated filtering would be required to achieve similar benefits in large size systems. We plan to investigate this further in the future.

We also experimented with combining the phrase tables trained separately from the two resources. The new data helped to reduce the number of un-translated words, but did not show improvements in translation performance. Currently we are working on a system where the phrase pairs are extracted from the two corpora separately, but they are scored in a homogeneous way. That would result in phrase pairs that have similar distribution of scores.

It is encouraging to see that we can benefit from automatically extracted data when the initial parallel corpus is small. For vast majority of languages, only a limited amount of parallel data is available, and hence has the potential to benefit from automatically extracted data.

Acknowledgement

This work was supported in part by DARPA under contract HR0011-06-2-0001 (GALE).

5. References

Mona Diab and Philip Resnik. 2002. An unsupervised method for word sense tagging using parallel corpora. In *Proceedings of the 40th Annual meeting of the ACL*, Philadelphia, PA, USA, pages 255-262.

Pascale Fung and Lo Yen Yee. 1998. An IR approach for translating new words from nonparallel, comparable texts. In *Proceedings of the 36th Annual Meeting of the ACL*, Montreal, Canada, pages 414-420.

Pascale Fung and Percy Cheung. 2004. Mining very non-parallel corpora: Parallel sentence and lexicon extraction via bootstrapping and EM. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Barcelona, Spain, pages 57–63.

Dragos Stefan Munteanu and Daniel Marcu. 2005. Improving Machine Translation Performance by Exploiting Non-parallel Corpora. *Computational Linguistics*, 31(4):477-504.

Dragos Stefan Munteanu and Daniel Marcu. 2006. Extracting parallel sub-sentential fragments from non-parallel corpora. In *Joint Proceedings of the ACL and the International Conference on Computational Linguistics*, Sydney, Australia.

Franz Josef Och and Hermann Ney. (2003). A systematic comparison of various statistical alignment models, *Computational Linguistics*, 29(1):19-51.

Franz Josef Och. (2003). Minimum error rate training in statistical machine translation, In *Proceedings of the 41st Annual Meeting of the ACL*, Sapporo, Japan, pages 160-167.

Kishore Papineni, Saleem Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL*, pages 311–318.

Philip Resnik and Noah Smith. 2003. The web as a parallel corpus. *Computational Linguistics*, 29(3):349–380.

Bing Zhao and Stephan Vogel. 2002. Adaptive parallel sentence mining from web bilingual news collection. In *2002 IEEE International Conference on Data Mining*, Maebashi City, Japan, pages 745-748.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, et al. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. *Annual Meeting of the ACL*, demonstration session, Prague, Czech Republic, June.

Translating Named Entities using Comparable Corpora

Iñaki Alegria, Nerea Ezeiza, Izaskun Fernandez

IXA NLP Group
University of the Basque Country
Donostia, Basque Country
i.alegria@ehu.es, n.ezeiza@ehu.es, izas.fernandez@gmail.com

Abstract

In this paper we present a system for translating named entities between different language pairs, using comparable corpora. We present the different experiments we have tried, where we have translated entities from Basque into Spanish, and from Spanish into English. The aim of this experiments is twofold: on the one hand, we want to validate the strategy we propose to translate Basque named entities into Spanish taking advantage of comparable corpora; on the other hand, we want to prove that this approach is applicable to different language pairs and that the performance is reasonable.

1. Introduction

Person, location and organization names, are the main types of named entities (NEs), and they are expressions commonly used in all kinds of written texts. Recently, these expressions have become indispensable units for many applications in the area of information extraction, as well as for many searching engines. We can find many tools dealing with the identification and classification of named entities (CoNLL[1]) for specific languages. But, there is less published research on NEs translation. Luckily the interest is increasing considerably in the last years as we will see in the following section.

Our main goal is to build a multilingual NE database, which can be very useful for translation systems, multilingual information extraction tools (i.e. Question Answering) or multilingual systems in general. Since getting the information for that multilingual NE database was a complex task, we decided to work in the field of NEs' translation; furthermore, we wanted too design a system for translating those expressions between different language pairs.

If we look at the works published about NE translation, we can distinguish 3 types of systems: systems based on parallel corpora, which are the most widely used; the ones based on comparable corpora; and finally, the ones that only use the web as an open corpus.

As we have mentioned before, most of the related works use parallel corpora. However, and as it is widely known, obtaining parallel corpora is not an easy task, and it becomes harder when one of the languages in the pair is a minority language, as it is the case of Basque. Nevertheless, we can use comparable corpora to solve the problem of lacking parallel corpora. Comparable corpora are those datasets which are written in different languages but are not translations of one another, thus, they cannot be aligned. But they are supposed to deal with similar subjects and to be written in similar styles. Compiling that kind of corpora is much easier than obtaining parallel ones, although sometimes it is not possible to get neither of them. In this case, we can use the web as a multilingual corpus, in order to search for possible entity translations.

For this work, we obtained the comparable corpora with the NEs tagged from the Hermes project[2](news databases: cross-lingual information retrieval and semantic extraction). All the entities have been automatically identified and classified. Those datasets are newspaper articles borrowed from different newspapers of the same year but they are not translations of one another. Anyway, the articles from different newspapers deal

[1] http://www.cnts.ua.ac.be/conll2003/ner/

[2] http://nlp.uned.es/hermes/

with similar topics and news: international news, sports, politics, economy, culture, local issues and opinion articles, but with different scopes.

The Basque corpus has 40,648 articles with 9,655,559 words and 142,464 NEs from *Euskaldunon Egunkaria*, a newspaper entirely written in Basque; the Spanish corpus has 16,914 articles with 5,192,567 words and 106,473 NEs from the news agency *EFE*[3]; and finally, the English dataset has also been borrowed from *EFE*, and has 16,942 articles 3,631,335 words and 49,768 NEs.

As we can see, there are much more articles in the Basque corpus than in the others. And, even the Spanish and English corpora have similar amount of articles, the Spanish set has twice the number of NEs in the English set. However, we assume that they share common NEs and they could be an interesting resource for the NE translation task.

For our experiments, we have used two comparable datasets, one for the Basque-Spanish language pair, and another for the Spanish-English pair.

Besides these two datasets, we have also used some other information sources in order to develop the language independent NEs translation system:

- A finite-state transducer based on edit distance (Kukich, 1992), simulating simple cognates and transliteration transformations (Al-Onaizan *et al.*, 2002b) in a language independent way;
- A bilingual dictionary for the corresponding language pair;
- An element rearrangement module for language pairs that follows different syntactic patterns.

The paper is structured as follows. Section 2 presents the related works. Section 3 presents the experimental settings. In section 4 we describe the development of the NE translation system using a limited amount of linguistic knowledge. In section 5, we present the results of the experiments, and finally, section 6 presents some conclusions and future work.

2. Related Works

Recently, considerable research effort has been focused on machine translation systems (MT) and their improvement. But most of the MT systems translate named entities without any specific treatment. That is the reason why most systems will translate the Spanish form *escuela de derecho de Harvard* into *school of the right of Harvard* instead of *Harvard Law School* which is the correct English form, as Reeder argues (Reeder, 2001). So besides being a good way to obtain multilingual NE information, NE translation can be considered a helpful task for MT improvement.

Concerning the resources, despite the difficulty to get bilingual parallel corpora for many languages, most NE translation systems work with parallel datasets. Furthermore, those bilingual corpora are aligned at paragraph or even at phrase level. For example, Moore's work (Moore, 2003) uses bilingual parallel English-French aligned corpora, and he obtains a French form for each English entity applying different statistical techniques.

Although comparable corpora has been less studied, there are some known systems designed to work with them as well; Such as the system that translates entity names from Arabic to English (Al-Onaizan *et al.*, 2002a), and the Chinese-English translation tool presented in ACL 2003 (Chen *et al.*, 2003).

The main goal of both systems is to obtain the equivalent English form, taking Chinese and Arabic respectively as source language. Two kinds of translations can be distinguished in both systems: direct/simple translations and transliterations (Al-Onaizan *et al.*, 2002b). However, the techniques used by each tool are different. Frequency based methods are used in Chinese-English translations, while in the Arabic-English language pair, a more complex combination of techniques is applied.

Similar techniques are applied at (Sproat *et al.*, 2006) and (Tao *et al.*, 2006), in which transliterate English-Chinese named entities using comparable corpora. The former combines a supervised phonetic transliteration technique and a phonetic frequency correlation approach, while the latter combines those techniques, but applying the pho-

[3] EFE is a news agency with delegations in Madrid and Miami

netic approach in an unsupervised way, where the distance is determined by a combination of substitution, insertion and deletion of characters.

Finally, we also want to mention the work (Poliquen *et al.*, 2005) which is integrated at the news analysis system NewsExplorer[4]. This research tries to extract person names from multilingual news collections to match name variants referring to the same person, and to infer relationships between people based on the co-occurrence information in related news.

In this paper, we present the research carried out for translating entity names using comparable corpora. We consider this method language independent, even though a bilingual dictionary is required, because we don't use any language dependent linguistic rule for the translation process. We have applied our method to Basque-Spanish and Spanish-English language pairs. We have also compare our results to the ones obtained with a language dependent NE translation system (Alegria *et al.*, 2006).

3. Experimental settings

When we started working at the NE translation task, we designed a language dependent tool for translating NEs from Basque to Spanish using comparable corpora. That system used linguistic information for both transliteration and entity element rearrangement. We tested this system using a set of the most common entities, and we obtained interesting results, with about a 78.7% F-score.

Since our goal is to obtain not only bilingual, but also multilingual NE information, and bearing in mind that designing a system for each language pair in a language dependent way is very expensive, we decided to experiment designing a relatively language independent tool following a similar strategy, and using comparable corpora and bilingual dictionaries. Firstly, we tested this tool in the Basque-Spanish language pair, in order to validate the methodology, and we compared it to the language dependent tool. We saw that the performance was even better than we expected and, it obtained an F-score of 77.5%, which is quite close to the performance of the language dependent tool.

For this reason, we wanted to see if the tool could be really applied to other language pairs, and hence be useful for extracting multilingual NE information without an exhaustive linguistic modelling of other languages. So we tried the same experiment in the Spanish-English language pair.

As we have mentioned before, we have used two main resources for our experiments: comparable corpora and bilingual dictionaries. We have already described the corpora in the introduction. Concerning the bilingual dictionaries, we have used a set of 74,331 Basque words with their corresponding Spanish translations for Basque-Spanish experiments, while for Spanish-English experiments this resource contains 73,784 entries.

For evaluation purposes, we have used similar corpora, but extracted from different years. For each language pair, Basque-Spanish and Spanish-English, we have extracted 200 most frequent NEs in the source language and we have translated them manually.

In order to carry out an evaluation based on correct NEs, since the NEs were automatically treated, we verified that all the entities were correctly identified, because if the original entity was not a correct expression, the translation system could not probably propose a correct translation.

4. System Description

As we have mentioned before, we have applied a similar strategy to that used in the language dependent system for the design of the language independent NE translation tool.

The system uses 4 main modules: a grammar for transliteration combined with a bilingual dictionary for those words that cannot be translated only applying transliteration but also need some translation; an element rearranging module for the construction of the whole entity from components, which will treat the possible different syntactic structures between both languages, as it happens in the Basque-Spanish pair; and finally a searching module to decide which candidate is the most suitable. This architecture is described in Figure 1. In the following subsections we will present each module in detail.

[4] http://press.jrc.it/NewsExplorer/entities/en/1.html

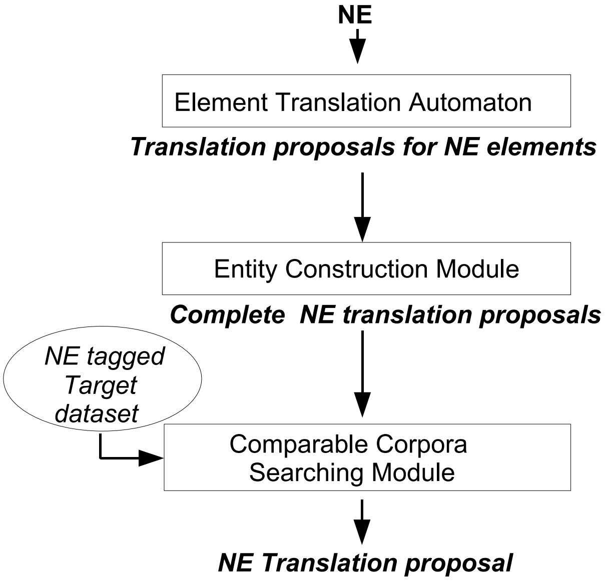

NE
Element Translation Automaton
Translation proposals for NE elements
Entity Construction Module
Complete NE translation proposals
NE tagged Target dataset
Comparable Corpora Searching Module
NE Translation proposal

Figure 1: System Architecture

4.1. Entity element translation module

The entity translation module has two main components: a transliteration finite-state automaton; and a bilingual lexicon.

We have used two main resources to automatically generate the transliteration rules: an edit distance (Kukich, 1992) based on a finite state grammar and a lexicon of the target language. Since this process is automatic it can be applied to any other language pair that uses similar alphabets.

The edit distance grammar uses the typical character based edit operations: insertion, deletion and replacement of a character in a word. Each operation is implemented as a rule in *XFST* (Beesley and Karttunen, 2001).

There is no specific rule in the grammar for switching adjacent characters, because that transformation can be simulated just combining the deletion and insertion operations mentioned above.

So this module will be able to obtain the translations of some of the NEs applying transliteration. For example, for the Basque-Spanish language pair, the system will transliterate *Kuba* into *Cuba*, replacing *K* with the *C* character; for the Spanish-English language pair, the system will transliterate *Constitución* into *Constitution*, replacing the second *c* with *t* and *ó* with *o*.

Since each rule can be applied *n* times for each word, the set of all translated words that we obtain after applying rules independently and combining them, is too extent. In order to reduce the output proposal-set, the system combines the grammar with a lexicon of the target language, and it restricts the transformation rules to at most two applications per word, avoiding the generation of words with more than two transformations (see Figure 2).

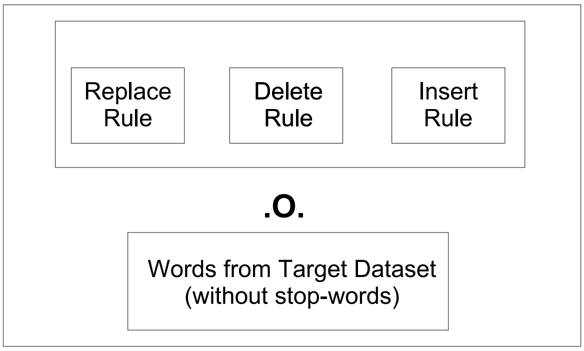


Figure 2: Transliteration automaton generation

We have generated three transliteration automata (TA) combining the mentioned resources:

- An automaton that copies the input word into the output (TA Max-transformations=0)
- An automaton generating words with at most one transformation (TA Max-transformations=1)
- An automaton generating words with at most two transformations (TA Max-transformations=2)

For the experiments, the target lexicons have been constructed using all the words from each target training set, excluding grammatical words such as prepositions, articles, etc., and using stop-lists[5].

However, there are some translations that cannot be obtained applying only transliteration rules. The system uses a source-target bilingual dictionary, converted into an automaton for those words. This automaton is combined with the three transliteration automata mentioned before. The application strategy is shown in Figure 3.

[5] http://www.lc.leidenuniv.nl/awcourse/oracle/text.920/a96518/astopsup.htm

The system firstly tries to obtain a translation proposal applying the zero-transformations TA to the input entity element. When the element is not found in the target lexicon, it applies the bilingual dictionary, and so forth.

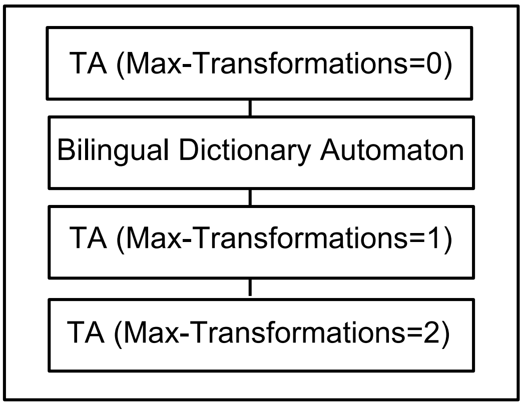

TA (Max-Transformations=0)
Bilingual Dictionary Automaton
TA (Max-Transformations=1)
TA (Max-Transformations=2)

Figure 3: Element Translation Strategy

So this module is able to translate not only the transliterated words in the comparable corpora, but also, the words that cannot be translated using transformation knowledge and that need information from a bilingual dictionary, such as 'Erakunde' vs. 'Organización'[6].

Since we have considered these datasets comparable, we assume that most of the source words would have their corresponding translation in the target dataset, in order to verify the correctness of the final translation automaton's output.

4.2. Entire Entity Construction

Since we want to build a language independent system that works just having two different language datasets, we don't want to use further linguistic information to combine syntactically the entity components. But we cannot ignore the possibility of having different syntactic patterns between languages, and this makes necessary to include some treatment for element rearrangement. This happens, for example in the Basque-Spanish language pair; Entity constituents may occur in different positions in both languages, so this module is applied before searching for translation candidates in the comparable corpora.

We might use many approaches to order elements, but we have chosen the simplest one: combining each proposed element with the rest, considering that each proposal can appear in any position within the entity. Thus, the system will return a large list of candidates, but it will include the correct one, if the independent translation of all the elements has been done properly.

Although in some cases prepositions and articles are needed to obtain the correct target form, the translation candidates for the whole entity will not contain any element apart from the translated words of the original entity. So, we will take into account the lack of these elements in the following step.

4.3. Comparable Corpus Search

Once the system has worked out all possible translation candidates for the whole entity, the following step consists on selecting the most suitable proposal. For that purpose, the system searches for them in the target language dataset, where entities are tagged.

Every translation proposal obtained from the previous step will be searched in the target dataset and each proposal will be positioned at a ranked list according to its frequency in the training corpus. Thus, the most repeated entities in the corpus will appear at the top of the list, being the most suitable translation proposals.

So briefly, the system takes a NE in source language as input, applies the translation module to each element, then it constructs the entire entity translation candidates, and finally it searches for them in a comparable corpora in order to obtain the most suitable ones, as described in Figure 4.

```
Input: NE composed of n element.
Output: An ordered list of NE  translations
Initialize: $proposals_i$ (i=1..n)
foreach $element_i$ ; i=1..n do
  $proposals_i$ = Apply element translation
               automaton to $element_i$
end for
entity_translations = entity construction($proposals_{1..n}$)
ranked_entity_translations = Searching
```

Figure 4: NE Translation Tool

[6] Organization

5. Experiments

As we have mentioned before, we have used a set of 200 most frequent NEs for each language pair, both Basque-Spanish and Spanish-English for evaluation.

We have used three evaluation measures to present the results of the experiments:

- $Precision = \frac{correctly\_translated\_NEs}{Translated\_NEs}$
- $Recall = \frac{correctly\_translated\_NEs}{All\_NEs}$
- $F - score = \frac{2*Precision*Recall}{Precision+Recall}$

When we compared the results for the Basque-Spanish pair of the language independent system with the ones obtained with the language dependent system (Alegria *et al.*, 2006), we saw that although the latter gets almost a 1.3% better performance, the performance of the language independent system could be considered a good approach with no need of exhaustive linguistic structure study.

However, we wanted to measure the performance of the Spanish-English language pair as well to verify if the results could be considered similar. The results of both experiments are shown in Table 1.

Lang. Pair	Pr.	R.	Fs
eu-es	82.02%	73%	77.5%
es-en	75.15%	62%	67.94%

Table 1: Language Independent System results

Observing these results, it seems that the system works considerably worse on the second language pair. In order to know the reason of that significant loss, we have reviewed all the supposed incorrect translations. We have observed that 26 of those translations were considered bad translations, because the frequency of the source NE form was higher than the one of the target form. This could be due to writing errors done by non-native speakers in the English EFE dataset. For example, when the system translates the Spanish form *Italia* into English, it creates a list of candidates where both *Italy* and *Italia* are generated. Then, as we have seen, it searches the candidate list at the comparable corpora and it ranks that list using frequency information on the corpus. Since in the English corpus *Italia* occurs more often than the correct form *Italy*, the former will be proposed as the most suitable translation, although the latter is the correct one. So when we evaluate this translation we see that an incorrect translation is proposed. Nevertheless, the error happens due to errors at the target corpus and not because of the bad performance of the language independent translation tool.

So, we can conclude that the system is very sensible to the target dataset correctness. And so, we guess that, if those 26 NE forms have their corresponding correct English form, the system would translate them correctly, and the results would be 5% better than the results for the Basque-Spanish pair.

6. Conclusions and Further Work

We have presented an approach for the design and development of a language independent NE translation system in order to obtain NE multilingual information, using comparable corpora, which seems to work well for different language pairs that have similar alphabets and writing habits.

To construct a new NE translation system, it is necessary to collect NE tagged comparable corpora for source and target languages, and also a bilingual source-target dictionary. The next step would be to extract the list of words (excluding stop-words) in the target dataset to generate the word translation automata using the general transliteration grammar already developed (as shown in Figure 2). Then the bilingual dictionary must be combined with the TAs obtained in the previous step (as shown in Figure 3). And finally, NEs in the target corpus must be extracted and stored along with their frequency, in order to select the most suitable translation among all the candidates.

Another way to select the most suitable NE translation is to use the web instead of the target dataset, as in (Moore, 2003). Nevertheless if we used the web, the system would be considerably slower due to the size of the resource, and consequently the answer time would be higher.

Another important issue is how to represent and link all this multilingual information to answer to a single language question in different language.

And finally, we want to improve the NE systems, including the translation system presented in this paper, and using the multilingual information we are collecting from all the comparable corpora.

7. Acknowledgement

This work has been partially funded by the Spanish Ministry of Education and Science (OpenMT: Open Source Machine Translation using hybrid methods,TIN2006-15307-C03-01), and the Local Government of the Basque Country (AnHITZ 2006: Language Technologies for Multilingual Interaction in Intelligent Environments, IE06-185), and partially supported by the Spanish Ministry of Education and Science (FIT-340000-2007-157 carried out on Tekniker). We would like to thank Itsaso Esparza for her help during the writing process of the paper.

8. References

Alegria I., Ezeiza N., Fernandez I. 2006. *Named Entities Translation Based on Comparable Corpora*. Proceedings of Multi-Word-Expressions in a Multilingual Context Workshop in EACL 2006.

Al-Onaizan Y., Knight K. 2002. *Translating Named Entities Using Monolingual and Bilingual Resources*. Proceedings of ACL 2002.

Al-Onaizan Y., Knight K. 2002. *Machine Transliteration of Names in Arabic Text*. Proceedings of ACL 2002.

Beesley K.R., Karttunen L. 2003. *Finite State Morphology:Xerox Tools and Techniques*. CSLI

Chen H., Yang C., Lin Y. 2003. *Learning Formulation and Transformation Rules for Multilingual Named Entities*. Proceedings of the ACL 2003 Workshop on Multilingual and Mixed-language Named Entity Recognition.

Kukich K., 1992. *Techniques for automatically correcting word in text. ACM Computing Surveys* Vol. 24 No. 4 377-439

Moore R. C., 2003. *Learning Translations of Named-Entity Phrases from Parallel Corpora*. Proceedings of EACL 2003.

Poliquen B., Steinberger R., Ignat C., Temnikova I., Widiger A., Zaghouani W., Žižka J. 2005. *Multilingual person name recognition and transliteration*. CORELA - COgnition, REpresentation, LAnguage, Poitiers, France, CERLICO. ISSN 1638-5748, 2005, vol. 3/3, no. 2, pp. 115-123.

Reeder F., 2001. *The Naming of Things and the Confusion of Tongues*. MT Evaluation: Who Did What To Whom Workshop on MT Summit VIII.

Sproat R., Tao T., Zhai C. 2006. *Named Entity Translation with Comparable Corpora*. Proceedings of the 21st International Conference on Computational Linguistic and 44th Annual Meeting of the ACL 2006.

Tao T., Yoon S., Fister A., Sproat R., Zhai C. 2006. *Unsupervised Named Entity Translation Using Temporal and Phonetic Correlation*. Proceedings of the 2006 EMNLP.

Evaluating Two Different Methods for the Task of Extracting Bilingual Lexicons from Comparable Corpora

Pablo Gamallo Otero

Departamento de Língua Espanhola
Universidade de Santiago de Compostela
Galiza, Spain
pablogam@usc.es

Abstract

In this paper, two different approaches to extract bilingual lexicons from comparable corpora are evaluated and compared. One uses syntactic contexts, and the other windows of tagged words. On a Spanish-Galician comparable corpus of 2 × 10 million words, syntactic contexts produce significantly better results for both frequent and less frequent words.

1. Introduction

In the last ten years, some methods have been proposed to acquire bilingual lexicons from non-parallel and comparable corpora. A non-parallel, comparable corpus (hereafter ”comparable corpus”) consists of sets of documents in several languages dealing with a given topic or domain, but in which the documents have been composed independently of each other in the different languages. As comparable texts are much easier to collect than parallel corpora, especially for minority languages and for a given domain, there is a growing interest in acquiring bilingual lexicons from comparable corpora. Indeed, they are more abundant, less expensive, and easily available via web than parallel texts. The main assumption underlying the approaches using comparable corpora is that a word in the target language is a candidate translation of a word in the source language, if the former tends to co-occur with expressions that are also translations of expressions co-occurring with that word in the source language. That is, the associations between a word and its context seed words are preserved in comparable texts of different languages.

The main contribution of this paper is to describe and compare two different approaches for extracting bilingual lexicons from comparable corpora. One of the tested approaches uses as contexts syntactic dependencies that can be extracted for each word in a corpus by robust parsers. The other approach uses the classic windowing technique around each word. Both techniques are applied to the same non-parallel, comparable corpus. A somehow related evaluation was performed by (Grefenstette, 1993), but on a monolingual corpus. According to the experiments we will describe later, the dependency-based method provides much better results than the windowing approach, very especially if only the top translation candidate is considered. In addition, further experiments will be performed to compare the efficiency of different similarity measures.

The paper is organized as follows: Section 2. introduces some comparable-based strategies to learn translation equivalents. Then, sections 3. and 4. describe a window and a syntax based method, respectively. The former is inspired by the Rapp approach (Rapp, 1999), and the later relies on a very simple dependency parser. Finally, in Section 5., some experiments will be performed against the same comparable corpus in order to evaluate several features of the 2 methods described in the previous sections.

2. Some Related Work

There is a growing interest in approaches focused on extracting word translations from comparable corpora (Fung and McKeown, 1997; Fung and Yee, 1998; Rapp, 1999; Chiao and Zweigenbaum, 2002; Dejean et al., 2002; Kaji, 2005; Gamallo, 2007; Saralegui et al., 2008). Most of them share a standard strategy based on context similarity. This strategy can be described as follows: a word w_2 in the target language is a candidate translation of w_1 in the source language if the context expressions with which w_2 co-occurs tend to be translations of the context expressions with which w_1 co-occurs. The basis of the method is to find the target words that have the most similar distributions with a given source word. The starting point of this strategy is a list of bilingual expressions that are used to build the context vectors of all words in both languages. This list is usually provided by an external bilingual dictionary. In Gamallo (2007), however, the starting list is provided by bilingual correlations previously extracted from a parallel corpus. In Dejean (2002), the method relies on a multilingual thesaurus instead of an external bilingual dictionary. In all cases, the starting list contains the “seed expressions” required to build context vectors.

There exist other approaches to bilingual lexicon extraction which do not use a starting list of seed expressions (Fung, 1995; Rapp, 1995; Diab and Finch, 2001). Yet, Fung (1995) failed to reach an acceptable accuracy rate for actual use, Rapp (1995) had strong computational limitations, and Diab et al. (2001) was applied only to non-parallel texts in the same language.

As far as the standard approach is concerned, works mainly differ in the coefficients (Dice, Jaccard, Cosine, City-Block, Lin ...) used to measure the similarity between context vectors. One of the contributions of this paper is to evaluate the efficiency of these coefficients to extract

This work has been supported by the Galician Government, within the project EXTRALEX, ref: PGIDIT07PXIB204015PR.

translation equivalents from comparable corpora. Moreover, works based on the standard approach also differ in the way they define word contexts. Most of them model contexts as a window of words of size N (*window-based paradigm*). Another technique (*syntax-based paradigm*) defines contexts by means of dependency relationships (Gamallo, 2007). The two techniques are very similar except that in one case a partial syntactic analysis is performed. As have been said, the main contribution of this paper is to evaluate and compare the results of each technique against the same comparable corpus.

3. Window-Based Method

The first technique for extracting bilingual lexicons does not perform any kind of syntactic analysis, but simply consider some window of words as forming the context of the compared words. We follow the method described in Rapp (1999), which is one of the most cited works on this topic.

3.1. Building Context Vectors

It is assumed that there is a small bilingual dictionary available at the beginning. The entries of the dictionary are considered as the starting list of seed words. Texts in both languages are lemmatized and POS tagged, and function words are removed. Then, for each lemma we build a context vector whose dimensions are seed words in different window positions with regard to the lemma. For instance, if we have chosen the window size 2, we compute a first context vector of lemma A whose dimensions are the seed words co-occurring 2 positions to the left of A. We also compute a second vector counting co-occurrences between A and the seed words appearing 1 position to the left of A. The same for the 2 positions following lemma A. Finally, we combine the 4 vectors of length *n* (where *n* is the size of the seed lexicon) into a single vector of length *4n*. This method takes into consideration word order to define contexts.

Each vector dimension of a lemma takes as value the number of co-occurrences between the lemma and a seed word in a given window position. Besides simple context frequency, additional weights can be considered, namely, a statistical degree of association between the lemma and each seed word. In the experiments described later, we will make use of log-likelihood ratio. This procedure is performed on the two monolingual texts.

3.2. Vector Similarity

Given a context vector defining a lemma of the source language, we compute a similarity score for each target vector. Then, a ranking list is built according to this score. The lemmas represented by the best-ranked target vectors are considered candidate translations of the given source lemma. We used several similarity coefficients for comparing pairs of vectors: *city-block* (Rapp, 1999), *cosine* (Fung and McKeown, 1997; Fung and Yee, 1998; Chiao and Zweigenbaum, 2002; Saralegui et al., 2008), *lin* (Lin, 1998a), and two different versions of both *jaccard* and *dice*. This way, the similarity of two lemmas, w_1 and w_2, is computed as follows:

$$\text{city-block}(w_1, w_2) = \sum_j |A(w_1, c_j) - A(w_2, c_j)|$$

$$\text{cosine}(w_1, w_2) = \frac{\sum_j A(w_1, c_j)A(w_2, c_j)}{\sqrt{\sum_j (A(w_1, c_j))^2} \sqrt{\sum_k (A(w_2, c_k))^2}}$$

$$\text{diceMin}(w_1, w_2) = \frac{2\sum_j \min(A(w_1, c_j), A(w_2, c_j))}{\sum_j A(w_1, c_j) + \sum_k A(w_2, c_k)}$$

$$\text{diceProd}(w_1, w_2) = \frac{2\sum_j A(w_1, c_j)A(w_2, c_j)}{\sum_j (A(w_1, c_j))^2 + \sum_k (A(w_2, c_k))^2}$$

$$\text{jaccardMin}(w_1, w_2) = \frac{\sum_j \min(A(w_1, c_j), A(w_2, c_j))}{\sum_j \max(A(w_1, c_j), A(w_2, c_j))}$$

$$\text{jaccardProd}(w_1, w_2) =$$

$$\frac{\sum_j A(w_1, c_j)A(w_2, c_j)}{\sum_j (A(w_1, c_j))^2 + \sum_k (A(w_2, c_k))^2 - \sum_i A(w_1, c_i)A(w_2, c_i)}$$

$$\text{lin}(w_1, w_2) = \frac{\sum_{c_i \in C_{1,2}} (A(w_1, c_j) + A(w_2, c_j))}{\sum_j A(w_1, c_j) + \sum_k A(w_2, c_k)}$$

Where $A(w_1, c_j)$ is an association value of a vector of length n, with j, i, and k ranging from 1 to n. In our experiments, the association value stands for either the simple co-occurrences of lemma w_1 with a contextual seed word c_j, or the log-likelihood ratio between the lemma and its context. For both $jaccardProd$ and $diceProd$ metrics, the association values of two lemmas with the same context are joined using their product (Chiao and Zweigenbaum, 2002; Saralegui et al., 2008), while for $jaccardMin$ (Grefenstette, 1994; Kaji and Aizono, 1996) and $diceMin$ (Curran and Moens, 2002; van der Plas and Bouma, 2004; Gamallo, 2007) only the smallest association weight is considered. As regards lin coefficient, the association values of common contexts are summed (Lin, 1998a), where $c_j \in C_{1,2}$ if only if $A(w_1, c_j) > 0$ and $A(w_2, c_j) > 0$.

4. Syntax-Based Method

The second technique to extract translation equivalents relies on the identification of syntactic dependencies. So, context vectors will be provided with syntactic information.

4.1. Partial Parsing with Regular Expressions

As in the previous method, monolingual texts are lemmatized and POS tagged. Then, instead of searching for windows positions around lemmas, we make use of regular expressions to identify syntactic dependencies. Regular expressions represent basic patterns of POS tags which are supposed to stand for binary dependencies between two

Dependencies	Patterns of POS tags
(green$_5$, $mod_<$, jacket$_6$) (big$_{10}$, $mod_<$, $d$dog$_{11}$)	*R$_1$: $s/(\mathbf{A_i})(\mathbf{N_j})/\mathbf{N_j}/$
()	*R$_2$: $s/(\mathbf{N_i})(\mathbf{N})_\mathbf{j}/\mathbf{N_i}/$
(man$_2$, $with_3$, jacket$_5$)	*R$_3$: $s/(\mathbf{N_i})(\mathbf{P_k})(\mathbf{N})_\mathbf{j}/\mathbf{N_i}/$
(see$_6$, $obj_>$, dog$_{11}$)	R$_4$: $s/(\mathbf{V_i})(?:D_k\vert R_n)*(\mathbf{N})_\mathbf{j}/\mathbf{V_i}/$
(see$_6$, $obj_<$, man$_2$)	R$_5$: $s/(?:D_k)*(\mathbf{N_i})(?:R_n)*(\mathbf{V})_\mathbf{j}/\mathbf{V_j}/$
()	R$_6$: $s/(\mathbf{V_i})(?:R_n)*(\mathbf{P_k})(?:\vert D_m\vert R_r)*(\mathbf{N})_\mathbf{j}/\mathbf{V_i}/$

Table 1: Dependency triplets and patterns of POS tags

lemmas. Our experiments are focused on dependencies with verbs, nouns, and adjectives. Our parsing strategy consists of a sequence of syntactic rules, each rule being defined by a specific pattern of tags that stands for a binary dependency. This strategy is implemented as a finite-state cascade (Abney, 1996). Let's take an example. Suppose our corpus contains the following tagged sentence:

a_D$_1$ man_N$_2$ with_P$_3$ a_D$_4$ green_A$_5$ jacket_N$_6$ see_V$_7$ yesterday_R$_8$ a_D$_9$ big_A$_{10}$ dog_N$_{11}$

The aim is to identify dependencies between lemmas using basic patterns of POS tags. Dependencies are noted as triplets: $(head, rel, dependent)$. The first column of Table 1 shows the 5 triplets generated from the sentence above using the patterns appearing in the second column. Patterns are organized in a sequence of substitution rules in such a way that the input of a rule R_n is the output of a rule R_m, where $m \leq n$. A rule substitutes the POS tag of the head word (right side) for the whole pattern of tags representing the head-dependent relation (left side). The first rule, R_1, takes as input a string containing the ordered list of all tags in the sentence:

D$_1$N$_2$P$_3$D$_4$A$_5$N$_6$V$_7$R$_8$D$_9$A$_{10}$N$_{11}$

The left pattern in this rule identifies two specific adjective-noun dependencies, namely "A$_5$N$_6$" and "A$_{10}$N$_{11}$". As a result, it removes the two adjective tags from the input list. Then, rule R_3 is applied to the output of R_1. The left pattern of this rule matches "N$_2$P$_3$D$_4$A$_5$" and rewrites the following ordered list of tags:

D$_1$N$_2$V$_7$R$_8$D$_9$N$_{11}$

This list is the output of the following applicable rule, R_4, which produces "D$_1$N$_2$V$_7$". Finally, rule R_5 is applied and gives as result only one tag, V$_7$, which is associated to the root head of the sentence: the verb "see". As this verb does not modify any word, no rule can be applied and the process stops. This is in accordance with the main assumption of dependency-based analysis, namely, a word in the sentence may have several modifiers, but each word may modify at most one word (Lin, 1998b). In sum, each application of a rule, not only rewrites a new version of the list of tags, but also generates the corresponding dependency triplet. So, even if we do not get the correct root head at the end of the analysis, the parser generates as many triplets as possible. This strategy can be seen as partial and robust parsing, as faster as identifying contextual words with a window-based technique.

The 5 triplets in Table 1 where generated from 4 substitution rules, each matching a type of dependency: adjective-noun, noun-prep-noun, verb-noun, and noun-verb. The sentence analysed above does not contain triplets instantiating noun-noun and verb-prep-noun dependencies. Wildcards $(? : D|R)*$ stand for optional determiners and adverbs, that is, they represent optional sequences of determiners or/and adverbs that are not considered for triplets. Rules with an asterisk can be applied several times before applying the next rule (e.g., when a noun is modified by several adjectives). Subscript numbers allow us to link tags in the patterns with their corresponding lemmas in the sentence.

To represent triplets, we use 4 types of binary relations: prepositions, left modifiers (noted as $mod_<$), right objects ($obj_>$), and left objects ($obj_<$). The latter two are generic dependencies between verb and nouns. They are likely to be specified with further linguistic information. For instance, a left object can be seen as a *direct object* if there is a passive form of a transitive verb; otherwise the left object is a *subject*. As we are not provided with information on transitivity, our list of dependencies does not contain subjects nor direct objects. Furthermore, long-distance dependencies are not taken into account. This is because rules are organised in such a way that they resolve attachment ambiguities by "Minimal Attachment" and "Right Association". Finally, relative clauses are also considered. However, for the sake of simplicity, Table 1 does not show the rules dealing with this phenomenon.

Note that the patterns of tags in Table 1 work well with English texts, but they are so generic that they can be used for many languages. To extract triplets from texts in Romance languages such as Spanish, French, Portuguese, or Galician, at least, 2 tiny changes are required: to provide a new pattern with dependent adjectives at the right position of nouns ($mod_>$), and to take as the head of a noun-noun dependency the noun appearing at the left position. Our main grammar only contains 10 generic rules suitable for Romance languages while the English grammar was provided with 9 rules. The linguistic knowledge required is then very low. The experiments that will be described later were performed over Spanish and Galician text corpora.

4.2. Lexico-Syntactic Contexts

The second step of our syntax-based method consists in extracting lexico-syntactic contexts from the dependencies and counting the occurrences of lemmas in those contexts. This information is stored in a collocation database. The extracted triplets of our example allow us to easily build the collocation database depicted in Table 2. The first line of

Lemmas	Lexico-Syntactic Patterns and freqs.
man	$< (\text{see}, obj_<, N), 1 >$ $< (N, with, \text{jacket}), 1 >$
see	$< (V, obj_<, \text{man}), 1 >$ $< (V, obj_>, \text{dog}), 1 >$
big	$< (\text{dog}, mod_<, A), 1 >$
dog	$< (N, mod_<, \text{big}), 1 >$ $< (\text{see}, obj_>, N), 1 >$
green	$< (\text{jacket}, mod_<, A), 1 >$
jacket	$< (N, mod_<, \text{green}), 1 >$ $< (\text{man}, with, N), 1 >$

Table 2: Collocation database of lemmas and lexico-syntactic contexts

the table describes the entry "man". This noun occurs once in two lexico-syntactic contexts, namely that representing the left position ($obj_<$) of the verb "see", $(\text{see}, obj_<, N)$, and that denoting the noun position being modified by the prepositional complement "with a jacket". The second line describes the entry "see", which also occurs once in two different lexico-syntactic contexts: $(V, obj_<, man)$ and $(V, obj_>, dog)$, i.e., it co-occurs with both a left object, "man", and a right object: "dog". The remaining lines describe the collocation information of the remaining nouns and adjectives appearing in the sentence above.

Notice we always extract 2 complementary lexico-syntactic contexts from a triplet. For instance, from $(\text{man}, with, \text{jacket})$, we extract:

$(N, with, \text{jacket})$ $\quad(\text{man}, with, N)$

This is in accordance with the notion of co-requirement defined in (Gamallo et al., 2005). In this work, two syntactically dependent words are no longer interpreted as a standard "predicate-argument" structure, where the predicate is the active function imposing syntactic and semantic conditions on a passive argument, which matches such conditions. On the contrary, each word in a binary dependency is perceived simultaneously as a predicate and an argument. In the example above, $(\text{man}, with, N)$ is seen as an unary predicate that requires nouns denoting parts of men (e.g. jackets), and simultaneously, $(N, with, \text{jacket})$ is another unary predicate requiring entities having jackets (e.g. men).

4.3. Building Syntax-Based Context Vectors

In this approach, the seed expressions used as cross-language contexts are not bilingual pairs of words as in the window-based approach, but bilingual pairs of lexico-syntactic contexts. The process of building a list of seed syntactic contexts consists of two steps: first, we generate a large list from an external bilingual dictionary Second, this starting list is used to build the context vectors of the lemmas appearing in the comparable corpus.

To show how we generate bilingual correlations between lexico-syntactic contexts using bilingual dictionaries, let's take an example. Suppose that an English-Spanish dictionary translates the noun "import" into the Spanish counterpart "importación". To generate bilingual pairs of lexico-syntactic contexts from these two nouns, we follow basic linking rules such as: (1) if "import" is the left object of a verb (i.e, if it is the subject of the verb), then its Spanish equivalent, "importación", is also the left object; (2) if "import" is modified by an adjective at the left position, then its Spanish equivalent is modified by an adjective at the right position; (3) if "import" is restricted by a prepositional complement headed by the preposition *in*, then its Spanish counterpart is restricted by a prepositional complement headed by the preposition *en*. The third rule needs a closed list of English prepositions and their more usual Spanish translations. For each entry (noun, verb, or adjective), we only generate a subset of all possible lexico-syntactic contexts. Table 3 depicts the contexts generated from the bilingual pair "import-importación" by making use of 6 basic linking rules for English-Spanish. As regards the other language pairs, we use a very similar set of rules. The human effort required to develop such rules is very low.

The second step consists in building a context vector for each lemma appearing in the comparable corpus. Vector dimensions are constituted by those contexts of the collocation database created above that also appear in the list of bilingual contexts generated from the external dictionary. For instance, if (import, of, N) both occurs in the corpus (i.e, it is in the collocation database), and belongs to the list of bilingual pairs, then it must be taken as a dimension in a context vector.

Finally, vector similarity between lemmas is computed as in the window-based approach.

5. Experiments and Evaluation

Three experiments were performed in order to evaluate three different parameters of the extraction techniques described in this paper: First, the quality of dependency relationships was compared to the linguistic relevance of relations between words co-occurring the same window. Second, we compared the efficiency of different similarity coefficients. And third, we evaluated the accuracy of both the syntax and the window based approaches described above.

5.1. Experiment 1

We first evaluated the triplets generated by our dependency-based parser. For this purpose, we manually analysed a Spanish text containing 200 dependency triplets. We considered only those types of dependencies likely to be identified by our parser, namely, prepositional complements, left and right verbal objects, and nominal modifiers. Among the verbal complements and objects, we also include the relationships between a noun and the main verb in a relative clause modifying the noun. As in Lin (1998b), the gold standard dependencies are called *key*. On the other hand, the triplets generated by our parser from the same text are called *answer*. Once the key and the answer are both represented as dependency triplets, we can compare and calculate *precision* and *recall*. Precision is the percentage of dependency relationships in the answer that are also found in the key. Recall is the percentage of dependency relationships in the key that are also found in the answer.

Table 4 summarizes the evaluation results considering the different types of dependency relationships. The total precision is 74% while recall reaches 64%. These results are not far from baseline dependency parsers for English. For instance, in Lin (1998b), if we only consider the precision

English	**Spanish**
(import, $of\|to\|in\|for\|by\|with, N$)	(importación, $de\|a\|en\|para\|por\|con, N$)
($N, of\|to\|in\|for\|by\|with$, import)	($N, de\|a\|en\|para\|por\|con$, importación)
($V, obj_{>}$, import)	($V, obj_{>}$, importación)
($V, obj_{<}$, import)	($V, obj_{<}$, importación)
($V, of\|to\|in\|for\|by\|with$, import)	($V, de\|a\|en\|para\|por\|con$, importación)
(import, $mod_{<}, A$)	(importación, $mod_{>}, A$)

Table 3: Bilingual correlations between contexts generated from the translation pair: import-importación.

Dependency type	Precision	Recall
modification	78%	94.5%
left object	67%	45%
right object	90%	79%
pp attachment	68%	55%
Total	**74%**	**64%**

Table 4: Evaluation of different types of dependency relations.

Type of strategy	Precision	Recall	F-Meas.
Dependency-Based	74%	64%	69%
Window-Based	32%	91%	47%

Table 5: Evaluation of dependency and window based relationships.

of dependencies such as subject, complement, pp attachment, and relative clause, the average score is 76%, with 70% of recall.

The linguistic relevance of dependency triplets was compared to that of window-based contexts. For this purpose, we computed precision and recall of the relationship between window-based contexts and their co-occurrence lemmas. More precisely, we used the same Spanish text to generate an answer consisting of binary relations between lemmas and their context lemmas within a window of size N (where $N = 2$, see Section 3.). Here, types of dependencies cannot be taken into account. So, if a relationship between a lemma and a context lemma is instantiated by one of the specific dependencies in the key, then such a relation is considered to be correct. Results are depicted in Table 5 . We used the same key as in the previous evaluation.

These results show that a rudimentary dependency parser allows us to extract much more precise contexts than a window-based strategy. However, the latter reaches a greater recall. Regarding computational efficiency, the two strategies turned out to be similar. Identifying dependency triplets takes the same time as extracting window-based contexts: about $9,000$ words per second, using a 2.33GHz CPU. We will see in the third experiment which contexts are more significant for translation equivalents extraction.

5.2. Experiment 2

The aim of the second experiment was to compare the efficiency of several similarity metrics in the task of bilingual lexicon extraction. Each metric was combined with two weighting schemes: simple occurrences and log likelihood. The strategy used here was the window-based method described in Section 3.. For each source lemma, we obtain a ranked list of 10 target lemmas considered as their translation equivalents.

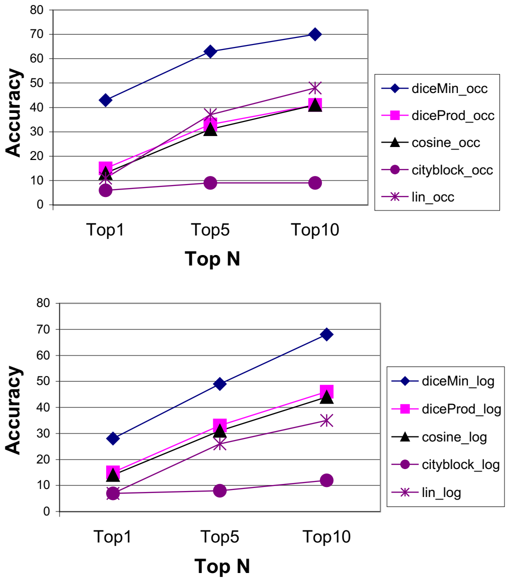


Figure 1: Percentile rank of the measures weighted with occurrences and log-like

5.2.1. Training Corpus and Bilingual Dictionary

The experiment was performed on a Spanish and Galician comparable corpus being constituted by news from on-line journals published between 2005 and 2006. As the Spanish corpus, we used $10,5$ million words of two newspapers: *La Voz de Galicia* and *El Correo Gallego*, and as Galician corpus 10 million words from *Galicia-Hoxe*, *Vieiros* and *A Nosa Terra*. The Spanish and Galician texts were lemmatized and POS tagged using a multilingual free software: Freeling (Carreras et al., 2004). Since the orientation of the newspapers is quite similar, the two monolingual texts can be considered as more or less comparable. The bilingual dictionary used to select seed words is the lexical resource integrated in OpenTrad, an open source machine translation system for Spanish-Galician (Armentano-Oller et al., 2006). The dictionary contains about $25,000$ entries.

Table 6: Syntax-Based Approach

Cov(%)	**Nouns** (74, 205 cntxs)			**Adjs** (13, 047 cntxs)			**Verbs** (39, 985 cntxs)		
	acc-1	**acc-10**	**freq**	**acc-1**	**acc-10**	**freq**	**acc-1**	**acc-10**	**freq**
50	.87	.89	$> 1,221$	.95	.97	$> 1,239$	.99	.99	$> 3,290$
80	.60	.72	> 123	.71	.76	> 187	.89	.94	> 770
90	.38	.45	> 28	.58	.63	> 49	.84	.94	> 266

Table 7: Window-Based Approach

Cov(%)	**Nouns** (128, 504 cntxs)			**Adjs** (94, 669 cntxs)			**Verbs** (111, 007 cntxs)		
	acc-1	**acc-10**	**freq**	**acc-1**	**acc-10**	**freq**	**acc-1**	**acc-10**	**freq**
50	.49	.80	$> 1,221$	.72	.86	$> 1,239$	.62	.84	$> 3,290$
80	.26	.51	> 123	.43	.70	> 187	.56	.78	> 770
90	.14	.36	> 28	.27	.51	> 49	.47	.65	> 266

5.2.2. Evaluation

To evaluate the efficiency of the different coefficients in the process of extracting bilingual lexicons, we elaborated an evaluation protocol with the following characteristics. A random sample of 200 test adjectives was selected from a list of adjectives occurring in the Spanish corpus. This list consists of those adjectives whose frequency achieves 80% of the total occurrences of adjectives in the corpus (80% of coverage). At this level of coverage, we computed 3 types of accuracy: *accuracy-1* is the number of correct translation candidates ranked first divided by the number of test lemmas. Then, *accuracy-5* and *accuracy-10* represent the number of correct candidates appearing in the top 5 and top 10, respectively, divided by the number of test lemmas. Indirect associations are judged to be incorrect.

5.2.3. Results

Figure 1 shows results using 7 different metrics combined with two types of weighted context vectors: simple occurrences and log-likelihood. In sum, we performed 14 different experiments. As the scores obtained using jaccard and dice coefficients were very similar, for the sake of simplicity, only dice scores ($diceMin$ and $diceProd$) are depicted in the figure.

These results show that the use of log-likelihood improves slightly *cityblock*, *cosine*, and *diceProd*, compared to the use of simple occurrences. However, *diceMin* (and so *jaccardMin*) as well as *lin* get better scores when simple occurrences are considered. On the other hand, there is a significant difference between *diceMin* compared to the other coefficients, regardless of the weight employed. With *diceMin*, 70% of the adjectives find their correct translation within the top 10 words, which is much better than the score achieved by lin_{occ} (49%), the second better coefficient. The reason of such a difference is that the product (or the sum as in *lin*) of association values maximizes odd similarities whereas the choice of the smallest value minimizes them. This is in accordance with the results obtained by (Curran and Moens, 2002) and (van der Plas and Bouma, 2004) Finally, the distance coefficient *city-block* seems to be unsuitable for this type of data.

5.3. Experiment 3

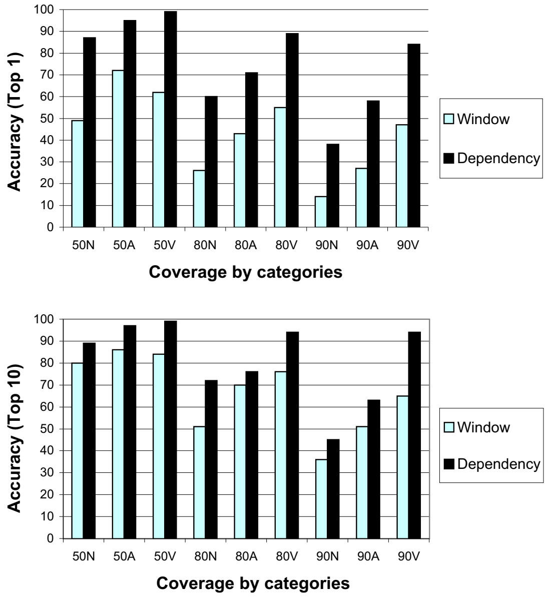


Figure 2: Comparison of accuracy between the two approaches considering both top 1 (above) and top 10 (below) translation equivalents

The aim of the third experiment was to compare the accuracy of both window and syntax based methods to extract bilingual lexicons. For this purpose, we used the same comparable corpus and bilingual dictionary as in the previous experiment. Similarity measure was computed with the most effective metric/weight combination: $diceMin_{occ}$. The evaluation protocol was more elaborated. We evaluated both *accuracy-1* and *accuracy-10* at three levels of coverage: 50%, 80%, and 90%, taking into account three POS categories: nouns, adjectives, and verbs. As nouns, we included proper nouns constituted by both mono and multi-word lemmas. Results are depicted in two tables: 6 and 7. They convey information on accuracy of three POS categories at different levels of coverage. They also show the number of contexts (i.e., vector size) used to define the lemmas of each category. Notice the number of syntactic contexts is much smaller than the number of contexts based on windows. As the size of context vectors in the syntactic approach is not very large, the process of computing similarities turns out to be more efficient. In addition, in order to analyze the impact the frequency has on the results, we include lemma frequencies of each category at each level of coverage. For instance, the nouns evaluated at 80% of coverage have more than 123 occurrences in the source corpus. This is not far from the usual threshold used in related work, where only words with frequency > 100 are evaluated.

It can be seen in tables 6 and 7 that the approach based on syntactic contexts (i.e., dependencies) works much better than that based on the windowing technique, at whatever level of coverage and for the three POS categories. The reason is that syntactic dependencies allow us to define finer-grained contexts which are semantically motivated. It can also be seen that the differences between both approaches are more significant when we only consider *accuracy-1* (see Figure 2): for instance, .87 against .49 percent considering nouns at 50% of coverage. If we look among the top 10 ranked lemmas (*accuracy-10*), differences are not so important: .89 against .80.

6. Conclusion

In this paper, we described and compared two techniques focused on bilingual extraction from comparable corpora. The syntax-based method produced better results than the window-based technique for very frequent ($> 1,221$), less frequent (> 123), and low frequent (> 28) nouns, adjectives, and verbs. In addition, the former method is more computationally efficient since it defines and uses smaller context vectors. On the other hand, the syntactic method can be seen as a knowledge-poor strategy (as the window-based approach), because our partial parsing relies on few generic regular expressions. Moreover, as the generic knowledge underlying the parsing technique is used to identify basic dependencies for the same family of natural languages, our syntax-based strategy turns out to be almost as language-independent as any windowing technique. Finally, we compared many similarity coefficients and discovered that two specific versions of Dice and Jaccard, *diceMin* and *jaccardMin*, are the best suited metrics for this specific task.

7. References

Steven Abney. 1996. Part-of-speech tagging and partial parsing. In Ken Church, Steve Young, and Gerrit Bloothooft, editors, *Corpus-Based Methods in Language and Speech.* Kluwer Academic Publishers, Dordrecht.

Carme Armentano-Oller, Rafael C. Carrasco, Antonio M. Corb-Bellot, Mikel L. Forcada, Mireia Ginest-Rosell, Sergio Ortiz-Rojas, Juan Antonio Prez-Ortiz, Gema Ramrez-Snchez, Felipe Snchez-Martnez, and Miriam A. Scalco. 2006. Open-source portuguese-spanish machine translation. In *Lecture Notes in Computer Science, 3960*, pages 50–59.

X. Carreras, I. Chao, L. Padró, and M. Padró. 2004. An open-source suite of language analyzers. In *4th International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal.

Y-C. Chiao and P. Zweigenbaum. 2002. Looking for candidate translational equivalents in specialized, comparable corpora. In *19th COLING'02*.

James R. Curran and Marc Moens. 2002. Improvements in automatic thesaurus extraction. In *ACL Workshop on Unsupervised Lexical Acquisition*, pages 59–66, Philadelphia.

H. Dejean, E. Gaussier, and F. Sadat. 2002. Bilingual terminology extraction: an approach based on a multilingual thesaurus applicable to comparable corpora. In *COLING 2002*, Tapei, Taiwan.

Mona Diab and Steve Finch. 2001. A statistical word-level translation model for comparable corpora. In *Proceedings of the Conference on Content-Based Multimedia Information Access (RIAO).*

Pascale Fung and Kathleen McKeown. 1997. Finding terminology translation from non-parallel corpora. In *5th Annual Workshop on Very Large Corpora*, pages 192–202, Hong Kong.

Pascale Fung and Lo Yuen Yee. 1998. An ir approach for translating new words from nonparallel, comparable texts. In *Coling'98*, pages 414–420, Montreal, Canada.

Pascale Fung. 1995. Compiling bilingual lexicon entries from a non-parallel english-chinese corpus. In *14th Annual Meeting of Very Large Corpora*, pages 173–183, Boston, Massachusettes.

Pablo Gamallo, Alexandre Agustini, and Gabriel Lopes. 2005. Clustering syntactic positions with similar semantic requirements. *Computational Linguistics*, 31(1):107–146.

Pablo Gamallo. 2007. Learning bilingual lexicons from comparable english and spanish corpora. In *Machine Translation SUMMIT XI*, Copenhagen, Denmark.

Gregory Grefenstette. 1993. Evaluation techniques for automatic semantic extraction: Comparing syntactic and window-based approaches. In *Workshop on Acquisition of Lexical Knowledge from Text SIGLEX/ACL*, Columbus, OH.

Gregory Grefenstette. 1994. *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publishers, USA.

Hiroyuki Kaji and Toshiko Aizono. 1996. Extracting word correspondences from bilingual corpora based on word co-occurrence information. In *16th Conference on Computational Linguistics (Coling'96)*, pages 23–28, Copenhagen, Danmark.

Hiroyuki Kaji. 2005. Extracting translation equivalents from bilingual comparable corpora. In *IEICE Transactions 88-D(2)*, pages 313–323.

Dekang Lin. 1998a. Automatic retrieval and clustering of similar words. In *COLING-ACL'98*, Montreal.

Dekang Lin. 1998b. Dependency-based evaluation of minipar. In *Workshop on Evaluation of Parsing Systems*, Granada, Spain.

Reinhard Rapp. 1995. Identifying word translations in non-parallel texts. In *33rd Conference of the ACL'95*, pages 320–322.

Reinhard Rapp. 1999. Automatic identification of word translations from unrelated english and german corpora. In *ACL'99*, pages 519–526.

X. Saralegui, I. San Vicente, and A. Gurrutxaga. 2008. Automatic generation of bilingual lexicons from comparable corpora in a popular science domain. In *LREC 2008 Workshop on Building and Using Comparable Corpora*.

Lonneke van der Plas and Gosse Bouma. 2004. Syntactic contexts for finding semantically related words. In *Meeting of Computational Linguistics in the Netherlands (CLIN2004)*.

Automatic Extraction of Bilingual Terms from Comparable Corpora in a Popular Science Domain

X. Saralegi, I. San Vicente, A. Gurrutxaga

Elhuyar R&D
Zelai Haundi kalea, 3. Osinalde Industrialdea, 20170 Usurbil. Basque Country
{xabiers, inaki, agurrutxaga}@elhuyar.com

Abstract

In the literature several approaches have been proposed for extracting word translations from comparable corpora, almost all of them based on the idea of context similarity. This work addresses the aforementioned issue for the English-Basque pair in a popular science domain. The main tasks our experiments focus on include: designing a method to combine some of the existing approaches, adapting this method to a popular science domain for the English-Basque pair, and analyzing the effect the comparability of the corpora has on the results. Finally, we evaluate the different prototypes by calculating the precision for different cutoffs.

1. Introduction

In the literature several strategies have been proposed for extracting lexical equivalences from corpora. Most of them are designed to be used with parallel corpora. Although these kinds of corpora give the best results, they are a scarce resource, especially when we want to deal with certain language pairs and certain domains and genres. As a solution to this limitation the first algorithms (Rapp 1995, Fung 1995) were developed for automatic extraction of translation pairs from comparable corpora. These kinds of corpora can be easily built from the Internet.

The techniques proposed for the extraction task are mainly based on the idea that translation equivalents tend to co-occur within similar contexts. An alternative is to detect translation equivalents by means of string similarity (cognates). Nevertheless, none of these techniques achieve the precision and recall obtained with the parallel corpora techniques.

This work focuses on the Basque-English pair and popular-science domain. Taking this scenario as the starting point, we channeled our efforts towards designing a hybrid approach to the methods proposed in the literature, adapting it to the scenario, and designing a measure to compute the comparability of a corpus. The results of the techniques applied to comparable corpora depend on the degree of comparability of a corpus. Hence, a proper measure is a determining factor to evaluate the adequacy of a corpora for terminology extraction.

2. Comparable Corpora

Comparable corpora are defined as collections of documents sharing certain similar characteristics and written in more than one language. In bilingual lexicon extraction some of these characteristics depend on the lexicon type we aim to extract. Thus, achieving a high degree of comparability with regard to these characteristics is very important, since context similarity techniques will be more effective. The more similar the corpora are, the higher the comparability between the collocated words of the equivalent translations (Morin et al. 2007).

In order to guarantee this comparability fully, we believe a global measure that takes different aspects relating to global comparability into account needs to be designed.

This work focuses on bilingual comparable corpora in popular science, that is, the domain is ‘science’ and the type of discourse is ‘news for non-specialized readers’. Besides these two main aspects, there are other characteristics that are related to the degree of comparability, such as distribution of topics and publication dates. All of them can be measured in order to estimate the global comparability of the corpora. Our hypothesis is that the comparability correlates with both the presence of word translations and the comparability of their contexts or collocates.

We introduce a method to compute the similarity between corpora, based on the Earth Movers Distance (EMD) (Rubner et al. 1997). This measure has been used to compute document similarity (Wang and Peng 2005). Section 4.1 further explains our strategy behind using this measure.

3. Identification of Equivalents

3.1. Context Similarity

The main method is based on the idea that the same concept tends to appear with the same context words in both languages, that is, it maintains many collocates. It is the same hypothesis that is used for the identification of synonyms. There are various approaches for implementing this technique. Problems arise with low frequency words, polysemous words and very general words, because they are difficult to represent. The representativity of the context vectors depends on the representativity of the corpus. However, we are only interested in the comparability of the context vectors, so while the representativity of the corpus is a significant problem, it is nevertheless a secondary one. The methods based on context similarity consist of two steps: modeling of the contexts, and calculation of the degree of similarity using a seed bilingual lexicon (Rapp 1999, Fung 1998).

The majority of the methods for modeling are based on the “bag-of-words” paradigm. Thus, the contexts are represented by weighted collections of words. There are several techniques for determining which words make up the context of a word: distance-based window, syntactic based-window (Gamallo 2007). Different measures have been proposed for establishing the weight of the context words with regard to a word: Log-likelihood ratio (LLR), Mutual Information, Dice coefficient, Jaccard measure, frequency, tf-idf, etc. Another way of representing the contexts is by using language models (Shao et al. 2004).

After representing word contexts in both languages, the proposed algorithms compute for each word the similarity between its context vector and all the context vectors in the other language by means of measures such as Cosine, Jaccard or Dice. According to the hypothesis, the correct translation should be ranked in the first positions. To be able to compute the similarity, the context vectors are put in the same space by translating one of them. This translation can be done by using dictionaries or statistical translation models.

3.2. Cognates

Another technique proposed in the literature is the identification of translations by means of cognates (Al-Onaizan and K.Night 2002). This method could be appropriate in a science domain where the presence of cognates is high. In fact, using a Basque-English technical dictionary we were able to calculate automatically that around 30% of the translation pairs were cognates. Dice coefficient or LCSR (Longest Common Subsequence Ratio) measures are proposed for computing string similarity.

4. Experiments

4.1. Measuring the Comparability Degree of Corpora

The degree of comparability between two corpora depends on several features of their texts (document topics, publication dates, genre, corpus size, etc.), and certain criteria must be adopted to tackle the problem of measuring comparability. Besides, the criteria depend on the target of the task and the methodology used to achieve that target. Our objective is to extract bilingual terminology from popular science texts by using a method based on comparing contexts of words. Therefore, we need a method to guarantee a minimum amount of comparable contexts of translation equivalents.

There are few works in the literature on this topic, and they do not deal with the impact of comparability on terminology extraction. Among them, (Kilgarriff 1998) evaluates certain measures and concludes that techniques based on word frequency information perform better. These techniques extract lists of the most frequent *n* words appearing in both corpora, and then these are compared by means of Hypothesis Tests. While (Kilgarriff 1998) uses raw word lists, (Rayson & Garside 2000) also tests POS tag lists and semantic tag lists.

We aim to find a measure which can tell how similar two corpora are; what is meant by *similar* is that the corpora are semantically alike on a document level. The more similar the documents are, the more similar the contexts of the words should be and hence, the performance of the term extraction process is expected to improve.

The method we propose in order to obtain a degree of comparability between two corpora takes the document as a unit for comparison. Let us say that the corpus C_1 (Basque) has *m* documents eu_i (where $i \in 0..m$) and the corpus C_2 (English) has *n* documents en_j (where $j \in 0..n$). Document similarity is computed for all of the inter-corpora document pairs, using *Dokusare, a* tool for cross-lingual similarity measuring described in (Saralegi and Alegria 2007). As a result, we obtain a *n*x*m* matrix (*DM*), where each $d_{i,j}$ entry corresponds to the content similarity between eu_i and en_j. This matrix is passed as a parameter to the EMD, which calculates the global similarity score.

$$\begin{array}{cc} & \begin{array}{ccccc} en_1 & .. & en_j & .. & en_m \end{array} \\ DM = & \left\{ \begin{array}{ccccc} d_{11} & .. & d_{1j} & .. & d_{1m} \\ .. & .. & .. & .. & .. \\ d_{i1} & .. & d_{ij} & .. & d_{\mathfrak{I}} \\ .. & .. & .. & .. & .. \\ d_{n1} & .. & d_{nj} & .. & d_{nm} \end{array} \right\} \begin{array}{c} eu_1 \\ .. \\ eu_i \\ .. \\ eu_n \end{array} \end{array}$$

Where DM is the matrix storing distance between documents computed using Dokusare.

$$p_j = en_j$$
$$q_i = eu_i$$
$$P = \{(p_1; w_{p_1}), \ldots, (p_m; w_{p_m})\} = \{(en_1; 1/m), \ldots, (en_m; 1/m)\}$$
$$Q = \{(q_1; w_{q_1}), \ldots, (q_n; w_{q_n})\} = \{(eu_1; 1/n), \ldots, (eu_n; 1/n)\}$$

We want to find a flow $F = [f_{ij}]$ with f_{ij} being the flow between p_i and q_j, which minimizes the overall cost

$$WORK(P; Q; F) = \sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} d_{ij}$$

constraints:

$$f_{ij} \geqslant 0, 1 \leqslant i \leqslant m; 1 \leqslant j \leqslant n$$
$$\sum_{j=1}^{n} f_{ij} \leqslant w_{pi}; 1 \leqslant i \leqslant m$$
$$\sum_{i=1}^{m} f_{ij} \leqslant w_{qi}; 1 \leqslant j \leqslant n$$
$$\sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} = min(\sum_{i=1}^{m} w_{pi}; \sum_{j=1}^{n} w_{qj})$$

The EMD is defined as the work normalized by the total flow:

$$EMD(P; Q) = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} d_{ij}}{\sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij}}$$

4.2. Term Extraction from Comparable Corpora

4.2.1. Preprocess

We needed to identify the words we considered to be meaningful for our process, that is, content-words. POS tags were used for this task. Treetagger is the tagger we chose to tag the English corpus and Eustagger in the case of the Basque corpus. Only nouns, adjectives and verbs are regarded as content words. In our experiments, adverbs were found to produce noise. Proper nouns also produced noise due to a cultural bias effect. Both were removed.

4.2.2. Vector-contexts Construction

We established a window depending on the POS of the word being focused on. The window size was determined empirically: 10 words for Basque (plus and minus 5 around a given word) and 14 for English (plus and minus 7). Furthermore, our experiments showed that using punctuation marks to delimit the window improved the results. Therefore, we also included this technique in our system.

We calculated the weight of the words within the context by means of the absolute frequency, LLR, Dice coefficient or Jaccard measure, and then contexts were modeled in a vector space. The best results were achieved by using the LLR. In addition, experiments were conducted combining the LLR with a distance factor between the center word *x* and the word *y disfactor(x,y),* for which the weight was being calculated:

$$LLRmod\,(x\,,y) = LLR\,(x\,,y) * disfactor(x\,,y)$$

The distance factor increases hyperbolically when the average distance between *x* and *y* decreases. We adopted this strategy to penalize the words farther from the center word, because the farther two words are from each other the weaker their relation is.

4.2.3. Context Vector Translation

To compute the translation of a Basque word, we translated its context vector in order to make it comparable with English context vectors. A bilingual Machine Readable Dictionary (MRD) was used for this purpose. If a word had more than one translation, we included all of them in the translated context vector, since the English equivalents were not sort by frequency of use. Our hypothesis is that the probability of concurrence of wrong translations in an English context-vector is low, and consequently, the first positions of the similarity-ranking are not distorted. In the case of the cosine distance, vectors were normalized before translation in order to prevent the noise produced by hypothetically wrong translations. Otherwise, the recall of the MRD determines the representativity of the context vector. In our experiments with a general dictionary, the average translation recall by vector was 55%. The higher the recall the greater the possibilities of finding the right translation for a word, because context vectors held more detailed information about the word in question.

To increase the recall of our translated vectors, we try to find equivalents not included in the dictionary by means of cognates. For all the Out Of Vocabulary (OOV) words, we looked for cognates among all the context words in the target language. The identification of these cognates is made by calculating the LCSR between the Basque and English context words. Before applying the LCSR, we processed some typographic rules to normalize equal phonology n-grams (e.g., *ph→f phase=fase*) or regular transformation ones (e.g., *-tion→-zio*, *action=akzio*) in both equivalent candidates. The candidates that exceeded a certain threshold (0.8, determined after several tests) were taken as translations.

4.2.4. Context Similarity Calculation

To obtain a ranked list of the translation candidates for a Basque word, we calculated the similarity between its translated context vector and the context vectors of the English words by using different similarity measures (Dice coefficient, Jaccard measure and Cosine). The best results were obtained with cosine. Furthermore, to prevent noise candidates, we pruned those that had a different grammatical category from that of the word to be translated.

4.2.5. Equivalent Similarity Calculation

In addition to context similarity, string similarity between source words and equivalent candidates is also used to rank candidates. LCSR is calculated between each source word and its first 100 translation candidates in the rank obtained after context similarity calculation. LCSR is applied in the same way as in context vector translation.

When used in combination with context similarity, LCSR data is used as the last ranking criteria. The candidates that exceeded the 0.8 threshold are ranked first, the remaining candidates not changing their positions in the rank. A drawback to this method is that cognate translations are promoted over the translations based on context vector similarity.

5. Evaluation

5.1. Building Test Corpora

We built two corpora with different characteristics in order to analyze the effect that comparability has on the results. The sources of the documents were science information web-sites. Zientzia.net (Basque), Sciam.com. AlphaGalileo, BBC News, ESA, EurekAlert!, NASA, New Scientist, news@nature, and ScienceNOW (English).

Zientzia.net and Sciam.com are quite similar with respect to the distribution of topics and register, so we chose them to build the first corpus (test corpus A). A correlation between topic and date was expected and for that reason we downloaded only all news items between 2000 and 2008. Moreover, other types of documents like articles, dossiers, etc. were rejected in order to maintain the same register throughout the corpus. Finally, the HTML documents were cleaned and converted into text using Kimatu (Saralegi & Leturia 2007). The size of this corpus was 1,092 million tokens for Basque and 1,107 for English. The distribution of the documents among the domains was comparable (table 1).

We built a second corpus (test corpus B), aiming for a lower comparability degree. We tried to unbalance important characteristics for the comparability degree like distribution among dates, topics and sources. We took the test corpus A as a starting point and randomly removed 1,000 documents from each language. In order to produce the bias we introduced 1,000 Basque news items from Zientzia.net belonging to the 1985-2000 period, and 1,000 English news items from the sources other than Sciam belonging to the 2007-2008 period. All new HTML documents were also cleaned and converted into text by Kimatu. The size of this corpus was 1,106 million tokens for Basque and 1,319 for English.

Domain	Sciam	Zientzia.net
Health, Mind & Brain	15.99%	14.85%
Space	9.83%	9.17%
Technology & Innovation	8.53%	15.40%
Biology	16.29%	28.35%
Earth & Environment, Archaeology & Paleontology	22.25%	17.88%
Physics, Chemistry, Math	11.15%	5.95%
History of Science, Society & Policy	15.96%	8.41%

Table 1: Domain distribution of documents for test corpus A.

The degree of comparability was computed using the EMD for both corpora. The value obtained for test corpus B was higher than the one obtained for the test corpus A. However, it was not as high as we expected. We are aware that these are only relative values, since there is no reference or threshold to compare them with. Anyway, the EMD value obtained in both cases is far from 0, which would indicate the maximum comparability degree. These high values are partly due to the rigorousness of Dokusare for calculating content similarity.

corpus	#word		#doc		EMD
	eu	en	eu	en	
Test corpus A	1,092K	1,107K	2,521	2,900	0.84
Test corpus B	1,106K	1,319K	2,521	2,900	0.86

Table 2: Characteristics of test-corpora

5.2. Tests

For the automatic evaluation of our system, we need a list of Basque-English equivalent terms occurring in each part of the corpora and which are not included in the dictionary used for the translation of content words in the construction of context vectors. To build that list, firstly we take all the Basque content words obtained in the preprocess step for the two built corpora. Secondly, those words are searched in the Basque-English Morris dictionary[1], and, for all the Basque words not included in that dictionary, we randomly select 200 pairs of words that reached a minimum frequency (10) and which appeared in two terminology Basque-English dictionaries (*Elhuyar Science and Technology Dictionary*[2] and Euskalterm terminology bank[3]).

This enabled us to estimate the precision automatically. In order to analyze the impact the frequency has on the results, we divide this set in two subsets. The first one includes words of high frequency (>50), and the other one, medium-low frequency words (within the 10-30 frequency range).

We also analyze the effect that the dispersion of the source test-words across the domains has on the precision of the system. Some scholars have pointed out the existence of a general academic vocabulary (Coxhead 2000) or a *lexique scientifique transdisciplinaire* (Drouin 2007). Those kinds of words are widely used in science-domain texts but do not belong to a specific domain. Therefore, the contexts of those words are, in principle, more heterogeneous than the contexts of specialized terms, and it is reasonable to suppose that they will be more difficult to extract. To analyze this effect, we calculated the correlation between the position of the target word in the ranking and the dispersion of the source word across the domains. We measured this dispersion by computing the coefficient of variation (CV) of the frequency of the source word across the domains. The reference domain list is the one used in Zientzia.net to classify news:

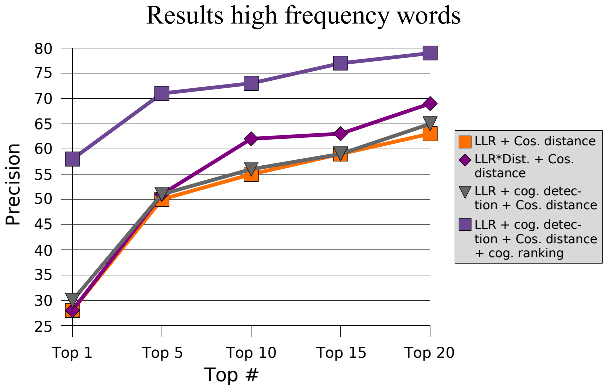
- Biology
- Space
- Physics, Chemistry, Math
- Computer science
- Earth sciences
- Environment
- Health
- Technology
- General

We analyzed different variables: the comparability of the corpus, the modeling of the contexts, and the way to combine the different approaches.

- Comparability: we processed the two test corpora in order to analyze the effect of the degree of comparability has on the results
- Modeling of contexts: Association Measures (AM), techniques to reduce OOVs
- Combining methods: context similarity, cognates

5.3. Results

Figures 1 and 2 show the results for both test corpora.

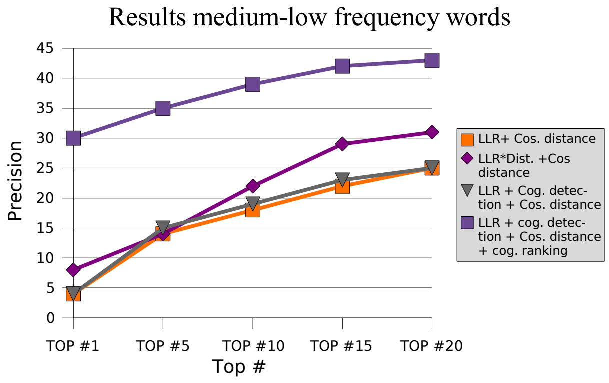


a)

[1] English/Basque dictionary including 67,000 entries and 120,000 senses.

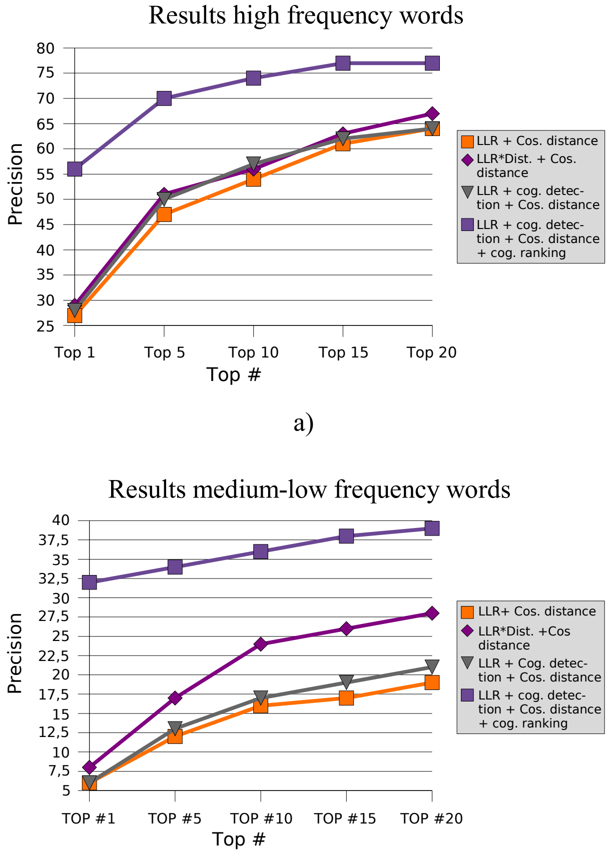
[2] Encyclopaedic dictionary of science and technology including 15,000 entries in Basque with equivalences in Spanish, French and English.

[3] Terminological dictionary including 100,000 terms in Basque with equivalences in Spanish, French, English and Latin.



b)

Figure 1: Precision results for test corpus A. Context similarity (cosine) combined with and without cognates detection during the vector translation phase (LCSR>0.8) and/or the ranking phase. Weighting the words in context vectors according to their distance from the centre word is also presented here.



a)

b)

Figure 2: Precision results for test corpus B.

In general, the precision obtained for the test corpus A is slightly better than the one obtained with the test corpus B. Although the difference is small, we can observe the influence of the degree of comparability on the precision. Another aspect that should be evaluated is the relation between the degree of comparability and the recall. As we mentioned in section 1, our hypothesis is that the comparability degree correlates with both the presence of word translations and the comparability of their contexts. In any case, more experiments must be carried out to deeper analyze these relations.

We have observed that combining the identification of cognates in the list of equivalents with context similarity (as proposed in section 4.2.5) improves the precision of the final rank. The high presence of these kinds of translations explains this improvement.

The detection of cognates in the translation of the context-vectors slightly outperforms translation based exclusively on dictionaries. Besides, the use of the distance factor together with the LLR also improves precision slightly, specially in the case of medium-low frequency words. This fact can be explained on the ground that co-occurrence data could not be enough to estimate correct association degree for the context words.

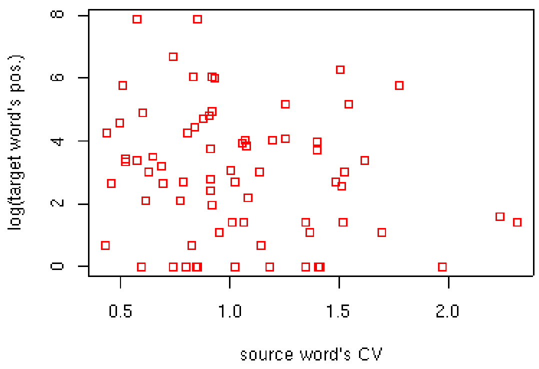


Figure 3: Dispersion diagram for source word’s CV and target word’s rank position

Figure 3 shows some results of the experiments done to measure the influence of the domain specificness of a source word on the rank position of the target word (corresponding to the LLR+Cos. distance experiment of Figure 1. a). There is no statistically significant correlation, contrary to our initial suspicion. There is no clear relation between the heterogeneity of the context of a word and its domain specificness, and therefore we could conclude that this factor does not have a significant effect on extraction based on context similarity calculation. Nevertheless, we think that a deeper analysis needs to be conducted in order to characterize difficult words, e.g. by analyzing the dispersion of frequency across the senses.

6.Conclusions

We've developed the first experiments towards terminology extraction from comparable corpora integrating different existing techniques and adapted them for a new language pair. The combination of the cognates detection in the final ranking as well as in the translation process of the context vectors seems suitable for corpora of science domain where the presence of cognates is high. On the other hand, our corpora are relatively small by current standards, and this leads to a significant decrease in the recall, since very few words reach the minimum frequency threshold necessary to obtain good precision in context similarity based extraction. In fact, in our test corpora only around 18% of the unknown source words

(Basque) reaches a frequency of 10. So, the maximum recall we could obtain is low.

As for the building of corpora, we have analyzed the importance of taking into account certain criteria in order to build comparable corpora for the terminology extraction task. Specifically, we have analyzed the effect that data and domain distribution also have on the degree of comparability and on the precision of the extraction process. The experiments we carried out showed a small effect. This could be due to the fact that the bias we induced in the test corpus B was not strong enough. Besides, we presented a new measure to quantify the degree of comparability, based on the EMD. Nevertheless, only preliminary experiments were conducted with this measure, and so further tests need to be done in order to tune it and ensure its reliability.

7. Future Work

We plan to build bigger corpora for the next experiments. To tackle the problems less-resourced languages like Basque have, we plan to use the Internet as the source of corpora as SIGWAC[4] suggests. So we are currently designing methods for building comparable corpora from the web.

Otherwise, we plan to extend our experiments to other languages, like Spanish, German and French.

In order to improve the extraction process, on the one hand, techniques for correct translation selection based on monolingual co-occurrences models will be integrated into the context vector translation process. On the other hand, we are planning to experiment with probabilistic models to represent contexts.

References

Coxhead, A. (2000). “A new Academic Word List.” In *TESOL Quarterly*, 34.

Déjean, H, Gaussier, E & Sadat, F. (2002) “An Approach Based on Multilingual Thesauri and Model Combination for Bilingual Lexicon Extraction.” In *COLING 2002.*

Druoin, P. (2007). “Identification automatique du lexique scientifique transdisciplinaire.” In Tutin, A. (Ed.) *Lexique des écrits scientifiques.* Revue Françaíse de Linguistique Appliqué. Volume XII-2

Fung, P. (1995) "Compiling Bilingual Lexicon Entries from a Non-Parallel English-Chinese Corpus." In *Proceedings of the Third Workshop on Very Large Corpora*, p.173-183, Boston, Massachusetts.

Fung, P. and Lo Yuen Yee (1998) “An IR Approach for Translating New Words from Nonparallel Comparable Texts.” In *COLING-ACL* 1998: 414-420.

Gamallo, P. (2007) “Learning Bilingual Lexicons from Comparable English and Spanish Corpora.” In *Proceedings of Machine Translation Summit XI*, Copenhagen, Denmark, pp. 191-198.

Kilgarriff, A., Rose, T. (1998) “Measures for corpus similarity and homogeneity.” *In Proc. 3rd Conf. on Empirical Methods in Natural Language Processing* (EMNLP-3). Granada, Spain, June: 46-52.

Morin, E., Daille, B., Takeuchi, K. and Kageura, K. (2007) “Bilingual Terminology Mining - Using Brain, not brawn comparable corpora.” In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, June 2007, Prague, Czech Republic, ACL p. 664-671.

Rapp, R. (1995) “Identifying word translations in non-parallel texts.” In *ACL*, p.320-322.

Rapp, R. (1999) “Automatic identification of word translations from unrelated English and German corpora.” In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics*, p.519-526, June 20-26, 1999, College Park, Maryland.

Saralegi, X. and Alegria, I. (2007) “Similitud entre documentos multilingües de carácter técnico en un entorno Web.” In *SEPLN 2007.* Sevilla. p.71-78.

Saralegi, X. and Leturia, I. (2007) “Kimatu, a tool for cleaning non-content text parts from html docs.” In *Building and exploring web corpora, Proceedings of the 3rd Web as Corpus workshop. Louvain-la-Neuve, Belgium: Presses universitaires de Louvain,* pp. 163—167.

Shao, Li and Ng, Hwee Tou (2004) “Mining New Word Translations from Comparable Corpora.” In Proceedings of the 20th *International Conference on Computational Linguistics* (COLING 2004). (pp. 618-624). University of Geneva, Geneva, Switzerland.

Rubner, Y., Guibas, L.J. and Tomasi, C. (1997) “The earth mover’s distance, multi-dimensional scaling, and color-based image retrieval.” In *Proceedings of the ARPA Image Understanding Workshop,* New Orleans, LA, May 1997, pp. 661-668.

Wan, X. and Peng, Y. (2005) ”The Earth Mover’s Distance as a Semantic Measure for Document Similarity.” In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, pp.301– 302.

Rayson, P. and Garside, R. (2000) “Comparing corpora using frequency profiling.” In *Proceedings of the workshop on Comparing Corpora (38th ACL),* Hong Kong, pp. 1–6.

[4] The Special Interest Group of the Association for Computational Linguistics (ACL) on Web as Corpus

Functional-Typological Approaches to Parallel and Comparable Corpora: The Bremen Mixed Corpus

Christel Stolz & Thomas Stolz
University of Bremen, FB 10: Linguistics, PF 330 440, D-28 334 Bremen/Germany
E-mail: cstolz@uni-bremen.de, stolz@uni-bremen.de

Abstract

This paper addresses a fundamental problem of contemporary crosslinguistic research as e.g. in functional typology. It is argued that the extant descriptive resources have too many disadvantages to qualify as sole reference works in our search for empirically well-founded generalisations across languages. It has become common knowledge in typological circles that working with text corpora is a much more promising way to identify similarities and dissimilarities of languages. However, recent experience suggests that neither parallel corpora nor comparable corpora are sufficient when it comes to determining the absence/presence of a given phenomenon or delimiting its domain. The usual corpora employed in language comparison are such that they exclude many languages from a potential sample because a given text or genre is not attested. Moreover, wherever suitable texts are handy, it is often the case that the are artificial in the sense that they are translated and thus do not necessarily reflect native speakers' first choice. We advertise the Mixed Corpus approach which includes a language-specific corpus which is exempt from genre-specific constraints but serves as the most direct access to natural/original language material in a given language.

1. From problems to solutions

Recently, functional typologists have begun to work more extensively with corpus data because the extant descriptive material of the world's languages has been shown once too often to be insufficient when it functions as the only or major empirical source in large-scale crosslinguistic investigations (Cysouw/Wälchli 2007). Many reference grammars do not cover the whole range of linguistically interesting phenomena. Moreover, they often paint a picture of a given language's structure that is dictated by the ephemeral methodology, theory or model the authors adhere to at the time of writing the grammar. In a manner of speaking, this dependence upon linguistic fashions is responsible for disagreement among the description of one and the same language by specialists representing different linguistic creeds. Furthermore, there are also numerous factual errors and misinterpretations which add to the growing dissatisfaction of functional typologists with their "traditionnal" data bases.

The task of functional typologists requires them to make inductively verified statements and generalisations about the occurrence/non-occurrence of certain properties and combinations thereof in the languages of the world on the basis of an analysis of the said properties in as large a sample of languages as possible. If it is true that one cannot blindly rely on what is said in descriptive grammars and the other usual second-hand sources, functional typologists have to get their hands on first-hand sources i.e. original language products provided by competent native speakers of the languages under scrutiny. Questionnaires and narration-based corpora (of the *Pear-Story* kind) cannot fill all the gaps in our knowledge of language(s) as both ways of collecting data have their specific merits and flaws (Dahl 2007). Suffice it to say that employing questionnaires always implies the potential pitfall of inadvertent manipulation of the native speaker informants by the researcher. Recording free discourse and/or narrations of picture-book stories may lead to multi-lingual corpora which are too diverse both structurally and semantically to allow for direct comparison because one cannot be sure that the data at hand are compatible with one another. In short, there is too much left to the informed but nevertheless subjective interpretation by the professional linguist.

This forms the backdrop of the on-going change in functional typology from grammar-based matrix-typological methodology towards corpus-based cross-linguistic research (cf. the various texts assembled in Cysouw/Wälchli [eds.] 2007). With a view to guaranteeing that what we compare is comparable in the first place, functional typologists have been trying hard to build up so-called parallel corpora based on an original literary text and its translations in other languages. Since for a typologist it is of utmost importance that one's sample is sizeable and at the same time genetically, areally and typologically balanced, insurmountable methodological problems arise because the very few extant texts with a wide distribution over languages are all biased in one way or the other and thus fail to meet the standards of quality imposed by our discipline itself. Ideally, crosslinguistic investigations need data from hundreds of languages. However, there is as yet only one text – the Bible – which boasts of an indisputably great number of translations (de Vries 2007). Unfortunately, this text like other holy texts is not fit for typological research as anachronistic language/hagiolectal style, differrent originnals (Hebrew/Aramaic, Greek, Latin, etc.) from which it is translated, restricted competence of translators who are not native speakers of the language into which the text is translated and many more factors render the Bible largely unsuited for typological purposes.

With this knowledge in the back of our minds, we have tried to build up parallel corpora based on different originnal texts and their translations (Stolz 2007). The Bremen research team thus created the largest parallel corpus so far which comprises the originnal and actually about 150 translations of Antoine de Saint-Exupéry's *Le Petit Prince*. With an average number of 1,650 sentences per text, however, this data base can by no means satisfy anybody's linguistic curiosity as many phenomena simply do not occur in this short text. Functional-typological work is mostly about qualities although strictly quantitative questions are not completely ruled out (for instance, if one tries to establish statistically-based markedness relations, etc.). If certain qualities fail to be attested in a short text, this does not automatically translate into their absence from the language as such. The partial inadequacy of *Le Petit Prince* for crosslinguistic in-depth studies impelled us to create a second parallel corpus of a much larger size.

This time we opted for J.K Rowling's *Harry-Potter* series which with seven volumes of about 3,000 pages altogether has the enormous advantage of exceeding the length of the previous sample text by the score. Thus, the probability that one will encounter a given phenomenon in this text is significantly higher than with *Le Petit Prince*. However, this advantage of *Harry Potter* is counterbalanced by the relative scarcity of translations. Presently, volume I of the series is available only in 60 languages of which 38 are modern languages of Europe (plus two extinct Classical languages, Latin and Old Greek). Since the other twenty translations are practically all into languages of Asia (the only exception being Egyptian Arabic), the language sample based on *Harry Potter* is heavily biased areally and also genetically as members of the Indo-European phylum clearly outnumber all other phyla. *Le Petit Prince* covers a wider range of areas and phyla because there is the occasional translation into languages of the Americas, Africa and Oceania including some Creole languages. Nevertheless, the sample is still Eurocentric as exactly two thirds of the translations are European, far more than half of the translations are Indo-European. These biases make it difficult or even impossible to put forward universals or type-oriented classifications of the world's languages. This however does not mean that these corpora are not valuable for typological research.

In point of fact, there is a branch of language typology that benefits a lot from the work with our parallel corpora, namely areal typology – or more precisely, the areal typology of Europe. Hitherto, areal linguistics did not have a corpus-linguistic component because the phenomena to be investigated were looked at according to the principles of dialectology and linguistic geography i.e. the presence or absence of a feature was stated for a given variety and than marked on a map. Our dialectological tradition is at its best in the realm of phonology. Literary parallel corpora however do not lend themselves readily to research on phonological problems, be they of a segmental or a suprasegmental nature. In contrast to traditional areal linguistics, corpus-based studies have the potential of revealing the areality of many a phenomenon from morphosyntax, semantics, etc. which so far have escaped being noticed by areal linguists. Unfortunately, this is no reason to be too enthusiastic about our pan-European corpora. It is true that with up to 100 languages from Europe, the sample is large enough to allow for generalisations. These generalisations however suffer from the usual problems connected to the imponderable vicissitudes of the translation process. We do not want to repeat too many of the well-known arguments against doing comparative linguistics on the basis of translations. It is common knowledge that translations cannot replace independent originals in terms of naturalness. Translations may be influenced too much by the wording of the original and thus result in artificial or even downright incorrect versions of a given language.

Already early on in our typological enterprise, we noticed that the parallel corpora at hand are far from being optimal even for a relatively restricted task such as the areal typology of Europe. Native speakers of a variety of our sample languages complained about the supposedly bad quality of some of the translations, they blamed the translators for fancy idiosyncrasies and artificially over-long sentences, etc. At least some of these complaints aim at grammatical phenomena and thus affect directly the reliability of the text within the framework of functional typology. How many of these problematic cases are induced by the translator's attempt to copy the French or English original too closely is a question we cannot answer yet (which is not particularly relevant for the issue discussed here, anyway). With a view to avoiding the negative effects of the translation process, the most reasonable solution is to work with original texts for each of the sample languages.

Of course, typological research can be conducted successfully only if there is some common empirical ground for all of the sample languages, meaning: the multi-lingual corpus must consist of texts which are equivalents of each other at least on a number of parameters. Thus, we are in need of so-called comparable corpora – at least this was what we thought initially. However, what exactly is a comparable corpus and how do we build up a comparable corpus which is qualitatively superior to our previous parallel corpora? A reminder: functional typologists do not necessarily go for huge corpora in the sense of voluminous texts. Their aim is a corpus which allows them to compare as many languages as possible – and if this cannot be achieved with longish texts, then we must make do with short ones. Even with short texts we still have the problem to determine what counts as "comparable" in a comparable corpus. One thing is clear: the multi-lingual corpus should be made up of texts of roughly the same size in order to guarantee that our findings are based on segments of similar extension. Furthermore, the same number of texts should be taken into consideration for each and every sample language. These texts should cover the same range of genres, they should stem from the same period, they should not be confined to the oeuvre of one author only. These and still other criteria impose severe restrictions already on the internal make-up of the multilingual corpus. If for instance, only one of our sample languages lacks material for a given genre, this genre is counted out for the entire sample because otherwise we would have an element of incomparability in the supposedly comparable corpus.

Since our primary goal is a large population of sample languages, the arguments in favour of including a given language in the sample despite the fact that it does not fulfill the criterion of providing suitably comparable texts are usually felt to be stronger than most of the methodological reservations – no matter how well-founded the latter happen to be. Thus, in a way, functional typologists are corpus linguists only half-heartedly. Our Bremen project team of course know about the problems a methodologycally non-reflected approach might create. Thus, we decided to take the bull by the horns and build up a corpus which satisfies all interested parties. The solution to our problems is the Mixed Corpus Approach.

2 The Mixed Corpus Approach

The Mixed Corpus Approach (which is relatively well-known also from translation studies and work on terminology but has not been fully integrated into functional typology yet) starts from the idea that both parallel and comparable corpora retain an element of artificiality which might distort the picture of language(s) to an extent that is no longer tolerable at least for functional typologists. Too much depends on the availability of certain texts or text types and thus the strict application of whatever criteria define parallel and/or comparable corpora can have detrimental effects such as the exclusion either of languages which are not equipped with the necessary texts or of phenomena which do not show up in those texts/text types which are readily available. Nevertheless, we want to work on the basis of texts in lieu of grammars or the like in order to get a better understanding of the workings of language structures and last but not least to discover hitherto unknown phenomena. How is it possible to combine the mathematical rigour of corpus linguistics with the ideals of functional typology? Wouldn't any attempt result in a contradiction in terms?

The Mixed Corpus Approach shows that the goals and procedures of corpus linguistics and functional typology can accommodate one another. This is so because the Mixed Corpus Approach integrates the positive aspects of three kinds of corpora. For each language of our sample, we provide three sub-corpora, namely

- texts belonging to one or more parallel literary sub-corpora [in our case, these are the above mentioned originals and translations of *Le Petit Prince* and *Harry Potter* = some 3,000 pages],
- texts belonging to a (presumably literary) comparable sub-corpus [take for instance five exemplars of three different genres, namely adventure stories addressing a readership aged 10-16, life reminiscences, and local history = some 3,000 pages; note that this choice of genres is not meant to be carved in stone for ever, other combinations of genres might turn out to be more promising than this one and the one we have been working with (cf. below)],
- texts belonging to a (presumably literary) language-specific sub-corpus [this sub-corpus contains traditional stories, legends, tales, etc. = (ideally) some 1,500 pages].

The third component differs qualitatively from the first two in so far as the language-specific sub-corpus contains only texts which are full-blown originals in the object language – in terms of both authorship and genre, meaning: the texts assembled in this sub-corpus must be products of the creativity of a native speaker of the language. Moreover, they must belong to a culturally fully established genre. Thus, take-overs from foreign genres – say, thrillers or science fiction – are counted out as members of the third sub-corpus. That this restriction might prove to be too strong is a latent danger because in some speech-communities (for instance, among the Mordvins), prose as such is marginal in the traditional culture where poetry, riddles, song lyrics dominate. Shamanistic or mantic texts may be the only prose-like genres accessible, if at all.

With the above triple basis we avoid losing what research based on parallel and/or comparable corpora has on offer whereas we add a third component which serves inter alia as a check for artificial vs natural data. To build up the sub-corpora, we had to do a lot of handiwork ourselves which included inter alia the age-long scanning and typing of badly printed books and the subsequent manual correction procedure. The texts were then manually aligned – which was another drawn-out process which had to be interrupted frequently to allow for the manual search for certain phenomena. The Mixed Corpus is strictly confined to intra-net use within our project group because negotiations with the copyright holders, especially Gallimard for *Le Petit Prince* and Bloomsbury for *Harry Potter* never came to an end.

The parallel corpus facilitates the formulation of preliminary hypotheses about the distribution of certain phenolmena over the sample languages and also allows for statements as to the probability with which a given phenomenon will be attested in the other components of the Mixed Corpus. We do not stop at this point because we know that the parallel literary corpus has the above mentioned shortcomings which preclude that generalisations be based solely on the evidence drawn from this component of the Mixed Corpus. The next step consists of widening the scope over texts belonging to a comparable corpus. The preliminary hypotheses are checked against the data in a variety of selected original texts. According to the new findings, the original ideas have to be revised and then restated in a new modified version. The analysis of the comparable corpus also helps to identify which of the phenomena observed in the parallel literary corpus can be attributed to the influence exerted on the translator by the original text version. It also indicates which other phenolmena are likely candidates for the status of "natural" categories of the language under inspection.

Given that the comparable corpus is restricted to only a small set of genres or typical texts, it remains to be seen whether the distribution of phenomena within a given language (and beyond) is determined by stylistic factors or other. Whether or not a phenomenon is dependent upon genre and the like is a question that can be answered only if the tight bonds of parallel and comparable corpora are overcome. To this end, a third component is called for – a sub-corpus which does justice to language-specific conditions and circumstances i.e. this corpus should comprise those texts which, for instance, native speakers of a given language consider typical representatives of products in their language (cf. above). What is important for the latter as it distinguishes the language-specific sub-corpus from the remaining two sub-corpora is its complete independence of the availability of equivalent texts in the other sample languages. If necessary, one may impose limits upon the minimal and/or maximal size of the language-specific sub-corpus in order to allow for comparative quantitative studies to be carried out including all three types of sub-corpora.

3. Achievements

Admittedly, this is but a sketch of the intricate character of the Mixed Corpus Approach. In recent years, we have employed this approach rather successfully in a number of large-scale typological research projects. Without the Mixed Corpus Approach, quite a few of our discoveries would not have been possible because neither the parallel corpora nor the comparable corpora provide the necessary wealth of data to draw definitive conclusions from. The Mixed Corpus Approach has proved to be feasible especially in those of our studies which are expressly devoted to the areal typology of Europe. This effect is causally related again to the biases described above for parallel and comparable corpora. If a Mixed Corpus contains a parallel literary sub-corpus, it is this sub-corpus which determines how many and also what languages will be part of the sample. In other words: a Mixed Corpus is only as good as its most restricted sub-corpus happens to be. This is a perhaps only minor methodological handicap which results from the general principles of the design of the Mixed Corpus Approach. At the moment, we do not see how this can be remedied. On the other hand, the Mixed Corpus Approach is still vastly superior in comparison to both the parallel corpus and the comparable corpus approaches because the third component serves as their corrective.

Since the employment of the Mixed Corpus has developed in three major steps over time (outlined below), there was at first no prescribed procedure according to which the three components of the Mixed Corpus had to be looked at in a fixed chronological order. However, this initially relative freedom has proved to have serious methodological disadvantages. The biggest problem is caused by the constant need of checking in a criss-crossing manner between the three sub-corpora. We have change this situation to the better by imposing a certain order, viz. we normally start with the parallel sub-corpora and then proceed to the comparable sub-corpus. The final part of the research is based on the language-specific sub-corpus. Note that any other order would be fine too provided it is kept constant throughout the project to be carried out.

To demonstrate how exactly research can be conducted if one applies the Mixed Corpus Approach, we like to refer to three of our typological projects, viz.

- COMITATIVES: In this project (Stolz/Stroh/Urdze 2006), we investigate the distribution profiles of so-called comitatives and instrumentals worldwide. The study includes three major case studies (on Icelandic, Maltese and Latvian) and a separate corpus-study based on *Le Petit Prince* (European languages only). The case studies contain elements of comparable corpora and language-specific corpora whereas the chapter on *Le Petit Prince* reflects a genuine parallel literary corpus. [It is shown that the putative universal according to which comitatives and instrumentals are cognitively the same cannot be upheld in this over-simplifying form because the vast majority of the world's languages keep the two categories formally apart.]
- POSSESSION: In our second project (Stolz/Kettler/Stroh/Urdze 2008), we look at possession splits with special focus on the situation in the languages of Europe. This time we consistently operate on the basis of the Mixed Corpus Approach as throughout the entire study we employ two parallel corpora – *Le Petit Prince* and *Harry Potter* – alongside elements of a comparable sub-corpus and a language-specific sub-corpus. We admit that the demarcation line separating the comparable sub-corpus and the language-specific sub-corpus is blurred more often than not. [We demonstrate that the alienability correlation is grammatically relevant in many European languages although it was commonly believed that these languages were exempt from possession splits.]
- REDUPLICATION: The third project (Stolz/Ammann/Urdze in preparation) which we terminated only a few weeks back inquires into the supposed absence of total reduplication in the languages of Europe. In contrast to the two prior studies, this one applies the Mixed Corpus Approach much more rigorously insofar as we neatly separate our parallel corpora (again *Le Petit Prince* and *Harry Potter*) not only from the comparable sub-corpus but also from the language-specific sub-corpus. We painstakingly define the make-up and size of the comparable sub-corpus (one original text of 150-200 pages for each of the following genres: history, folklore, texts used for primary education, journalistic prose). [The project results are such that (a) Europe can be shown to be far less reduplication-phobic than expected and (b) there are formerly unknown/neglected types of total reduplication that have to be taken into account.]

To demonstrate what can be done with the Mixed Corpus approach, we conclude with a sideways glance at our latest project, the one dedicated to total reduplication. In reduplication research, a construction like Italian *nero nero* 'very black' instantiates total reduplication as the adjective *nero* 'black' is used twice to convey the notion of intensity. This and similar patterns are widely used in languages spoken around the Mediterranean whereas they are practically absent from European languages outside this region. This distribution is largely corroborated by the first of our standard parallel sub-corpora, *Le Petit Prince*, which shows that north of the Alps, total reduplication occurs at best occasionally (text frequency n ≤ 3). Only a closer look at the second parallel text, *Harry Potter* (vol. I) reveals however that the absence of total reduplication is compensated for by the relatively frequent employment of so-called syndetic constructions like English *on and on* where two identical instances of one word form part of a coordinating construction with a fixed meaning. This correlation is shown in table 1 in which the languages are ordered according to decreasing percentages of syndesis vs increasing percentage of asyndesis (= proper total reduplication). The abbreviation *abs* = absolute.

language	syndesis		asyndesis		total
	abs	%	abs	%	abs
English	64	100%	0	0%	64
Faroese	47	100%	0	0%	47
Norwegian	38	100%	0	0%	38
Danish	37	100%	0	0%	37
Swedish	35	100%	0	0%	35
Dutch	29	100%	0	0%	29
Islandic	23	100%	0	0%	23
Portugiese	17	100%	0	0%	17
Finnish	15	100%	0	0%	15
Croatian	15	100%	0	0%	15
Slovenian	15	100%	0	0%	15
Spanish	15	100%	0	0%	15
Polish	7	100%	0	0%	7
German	4	100%	0	0%	4
Low German	71	98.61%	1	1.39%	72
French	41	97.6%	1	2.4%	42
Latvian	43	93.5%	3	6.5%	46
Serbian	11	91.66%	1	8.34%	12
Rumanian	55	90,12%	6	9,88%	61
Czech	27	90%	3	10%	30
Estonian	9	90%	1	10%	10
Galego	34	83%	7	17%	41
Bulgarian	8	80%	2	20%	10
Albanian	50	74.62%	17	25.38%	67
Irish	25	67.56%	12	32,44%	37
Macedonian	4	66.66%	2	33.34%	6
Catalan	53	61.27%	33	38.73%	86
Lithuanian	19	59.37%	13	41.63%	32
Georgian	59	53,63%	51	46,37%	110
Ukrainian	12	44,44%	15	55,56%	27
Russian	19	44,18%	24	55,82%	43
Italian	16	37,20%	27	62,80%	43
Welsh	5	31,25%	11	68,75%	16
Slovak	12	30,76%	27	69,10%	39
Greek	7	29,1%	17	70,90%	24
Hungarian	5	17,85%	23	81,15%	28
Basque	12	3,44%	337	96,56%	349
Turkish	1	1,13%	88	98.87%	89

Table 1: Syndetic vs asyndetic constructions in Harry Potter, vol. I

The discovery of the inverse correlation of reduplicative syndesis and asyndesis would not have been possible on the basis of our first parallel text, Le Petit Prince, whose limited size simply does not allow for a sufficient number of occurrences of the phenomena under review. What cases of syndesis there are in *Le Petit Prince* do not accumulate in any noteworthy amount. On the larger textual basis of *Harry Potter* however the syndesis-asyndesis dichotomy becomes significant (in the non-technical sense of noticeable) and thus gives the linguist food for thought. We emphasise that to reach this conclusion, it is sufficient to check just one volume of the *Harry Potter* series (in the above case the most widely translated vol. I).

We then checked the language-specific sub-corpus includeing a variety of languages for which only one parallel text is available. For Udmurt, Tatar and Kazakh, for instance, the check of *Le Petit Prince* yields relatively low values of total reduplication without any noticeable increase on the side of syndetic constructions. Udmurt has exactly 15 tokens with ten types of total reduplication whereas there is not a single attestation of syndesis in the text under scrutiny. This seems to run counter to our assumption that syndesis and asyndesis are in a kind of complementary distribution across the languages of Europe. In addition, the absolute frequency of total reduplication in the sample text suggests that Udmurt behaves like a language from the Mediterranean basin and its immediate hinterland although Udmurt is located far off the regional hotbed of total reduplication, namely in the northern Eurasian territory of Russia.

Since this was at odds with our earlier hypotheses about the areal distribution of the feature in Europe, we had a closer look at the comparable sub-corpus for Udmurt and other non-Indo-European languages of the former USSR. In this comparable sub-corpus, the type and token frequency of proper total reduplication dropped dramatically. Surprisingly, this drastic decrease of asyndesis did not go hand in hand with the expected increase of syndesis. It soon became clear that this situation was caused mainly by the make-up of the comparable sub-corpus. This sub-corpus contained similar texts for practically all Euroasian languages of our sample, namely technical descriptions of bee-keeping, the history of the local dependency of the Communist Party and short stories about the heroic resistance of the people of the USSR against the German invaders during the 2nd World War. These texts were not directly translated from the Russian, nor were the slavishly copied from one of the Eurasian language to the other(s). However, the comparable sub-corpus contained exclusively texts which are not fully original language products as they depend on foreign master-versions because the genre to which they belong does not form part of the traditional inventory of texts.

However, a closer look at the language-specific sub-samples for Udmurt, Tatar, Kazakh and a variety of other languages from the European East reveals that proper total reduplication abounds in the original literature such that it goes far beyond our expectations. For Udmurt alone, we found 127 tokens (= 81 types) of total reduplication on

154 pages of contemporary narratives. There are eight times as many types and tokens as in the Udmurt translation of *Le Petit Prince*. Without the language-specific sub-corpus Udmurt would have passed as a language with unspectacular frequencies of reduplicative constructions. It is the third sub-corpus which clearly shows that Udmurt counts among the languages with the most pronounced predilection for total reduplication in Europe. What is more, the language-specific sub-corpus contains hardly any evidence of syndesis – and thus, high type and token frequency of total reduplication and near absence of syndesis corroborate the tendency captured by table 1.

The Udmurt findings made us investigate the givens in a variety of Eurasian languages for which none of the parallel texts is available (Chuvash, Bashkir, Mordvin, Mari, Komi, etc.). The situation there is practically identical to the picture painted for Udmurt: the comparable sub-corpora display a very low turn-out for total reduplication without any noticeable increase of syndesis. This is easily explained by the similar internal structure of the sub-corpora as to text types. On this basis, these Eurasian languages would have counted as not particularly reduplication-friendly. At the same time, they would have put our assumptions about the syndesis-asyndesis correlation in jeopardy because of their avoidance of syndesis. In the language-specific sub-corpus, however, the values for type and token frequencies rose in such a way that Chuvash, for instance, occupies the second highest rank position as to type and token frequencies in the entire sample (Basque being number 1). As with Udmurt, syndesis remains a marginal phenomenon in Chuvash and thus the tendency documented in table 1 is again corroborated by the language-specific sub-corpus. These observations have made us revise our original map of the areal linguistics of reduplication in Europe such that the most recent version shows that total reduplication is not only strong in the south but also in the east and thus languages which disfavour total reduplication occupy only a small area of the continent, meaning: the represent the marked option as it is more "normal" for a European language to make use o total reduplication.

4. Conclusions

The above description of the Mixed Corpus Approach still needs to be refined. At the present stage, however, it should be clear already that in the kind of crosslinguistic research typologists are conducting nowadays, all three components of the Mixed Corpus are necessary ingredients to guarantee the maximum of empirical richness which is required in functional typology.

It cannot be denied that the Mixed Corpus Approach itself is in dire need of further elaboration and refinement as it has grown slowly out of methodologically variable previous approaches all of which failed to meet the high expectations of the researchers. One problem which remains to be solved is posed by what we like to all "enforced literacy" i.e. the top-down approach to creating a written register for a traditionally oral culture. Are the texts we assemble for a language-specific sub-corpus of a language of this kind in any way reliable data sources? It might be advisable to add a fourth sub-corpus to the list, namely a sub-corpus which consists entirely of transcribed oral texts, preferably spontaneously produced monologues. With this addition, we are confident that in the not too distant future, the Mixed Corpus Approach will develop into a more generally employed tool in our discipline.

Acknowledgments

We are grateful to the members of our research team at the University of Bremen: Andreas Ammann, Maksym Gorshenin, Sonja Kettler, Nataliya Levkovych, Cornelia Stroh, Aina Urdze, and Elisabeth Verhoeven. The University of Bremen kindly supported our initiative to create the world's largest parallel corpus.

References

Cysouw, M., Wälchli, B. (Eds.) (2007). *Parallel texts*. Focus issue of *Sprachtypologie und Universalienforschung* 60 (2). Berlin: Akademie Verlag.

Cysouw, M. & Wälchli, B. (2007). Parallel texts: using translation equivalents in linguistic typology. *Sprachtypologie und Universalienforschung* 60 (2), pp. 95-9.

Dahl, Ö. (2007). From questionnaires to parallel corpora in typology. *Sprachtypologie und Universalienforschung* 60 (2), pp. 172-81.

De Vries, L. (2007). Some remarks on the use of Bible translations as parallel texts in linguistic research. *Sprachtypologie und Universalienforschung* 60 (2), pp. 148-57.

Stolz, T. (2007). *Harry Potter* meets *Le Petit Prince* – On the usefulness of parallel corpora in crosslinguistic investigations. *Sprachtypologie und Universalienforschung* 60 (2), pp. 100-17.

Stolz, T., Ammann, A., Urdze, A. (In preparation). *Total reduplication – the areal linguistics of a universal*. [to be ready by end of 2008]

Stolz, T., Kettler, S., Stroh, C., Urdze, A. (2008). *Split possession. An areal-linguistic study of the languages of Europe* (= Studies in Language Companion Series 101). Amsterdam, Philadelphia: John Benjamins.

Stolz, T., Stroh, C., Urdze, A. (2006). *Comitatives and related categories. A typological study with special focus on the languages of Europe* (= Empirical Approaches to Language Typology 33). Berlin, New York: Mouton de Gruyter.

On the use of comparable corpora of African varieties of Portuguese for linguistic description and teaching/learning applications

Maria Fernanda Bacelar do Nascimento, Antónia Estrela, Amália Mendes, Luísa Pereira

Centro de Linguística da Universidade de Lisboa, University of Lisbon, Portugal

Av. Prof. Gama Pinto, 2, 1649-003 Lisboa – Portugal

www.clul.ul.pt

E-mail: fbacelar.nascimento@gmail.com, antonia.estrela@clul.ul.pt, amalia.mendes@clul.ul.pt, luisa.alice@clul.ul.pt

Abstract

This presentation focuses on the use of five comparable corpora of African varieties of Portuguese (AVP), namely Angola, Cape Verde, Guinea-Bissau, Mozambique and Sao Tome and Principe, for multiple contrastive linguistic analyses and for the production of teaching and learning applications. Five contrastive lexicons have been corpus-extracted and further annotated with POS and lemma information and have been crucial to establish for each variety a core and peripheral vocabulary. Studies on AVP-specific morphological processes and on variation in verb complementation will also be discussed. These are first steps towards an integrated description of the five varieties and towards the elaboration of teaching and learning materials to be used by teachers of students from those five African countries with Portuguese as official language.

1. Comparable corpora of African varieties of Portuguese

Compared with the quantity of empirical studies on European Portuguese (EP) and Brazilian Portuguese (BP), developed from corpora and lexicons, the shortage of studies on other varieties of Portuguese is mostly due to the lack of Language Resources (LR).

The Center of Linguistics of the University of Lisbon (CLUL) recently compiled five comparable corpora of the Portuguese Varieties spoken in the five countries which have Portuguese as official language - Angola, Cape Verde, Guinea-Bissau, Mozambique and Sao Tome and Principe. The corpora are available at CLUL's webpage for online query.

The five corpora, which constitute the Africa Corpus, are around 640,000 words each and have the same percentage of spoken and written subparts (c. 25,000 spoken words (4%) and c. 615,000 written words), as shown in Table 1. The written corpus is divided in newspapers (50%), literature (20%) and miscellaneous (26%).

For the task of corpus constitution, some samples of written and spoken materials of already existing corpora compiled at CLUL during the last 30 years were reused, while new recordings were specifically made for this project and new texts were collected.

Countries	Spoken	Written	Total
Angola	27.363	613.495	640.858
Cape Verde	25.413	612.120	637.533
Guinea-Bissau	25.016	615.404	640.420
Mozambique	26.166	615.297	641.463
Sao Tome and Principe	25.287	614.563	639.850
Total	**129.245**	**3.070.879**	**3.200.124**

Table 1. Africa Corpus: constitution and dimension per variety

The five corpora are thus comparable in size, in chronology and in broad types and genres. However, it was not possible to attain comparability at a more granular level. Compiling written materials for each African country proved a difficult task during the project and even more difficult when trying to assure comparable data. Compiling comparable corpora was already a challenge for our group in previous experiences involving European initiatives (like the PAROLE corpora due to the large number of languages involved) and was even more obvious in the case of these African countries. The fact that we only considered texts written by native people living in those five countries made it even more difficult to assure the necessary materials.

Besides the limitation in finding and compiling adequate materials, time was also an important factor, due to the short duration of the project. A follow-up of this work is under way and will assure broader coverage and a more fine-grained comparability of the five corpora.

The newspapers selected are publications with wide national coverage and, regarding fiction, poetry was avoided and only native authors or authors who lived all their lives in the countries were selected. We included in the corpus few texts that are strongly marked, like the case of the African author Mia Couto, whose writings present a high level of lexical creativity and are thus not representative of his AVP.

Since some of the texts collected proved to belong to very different subtypes and genres, it made it difficult to devise specific categories that would accommodate this diversity. This lead us to posit the category "miscellaneous", which corresponds, in fact, to a large collection of heterogeneous texts from different kinds, such as literary or social magazines, computer policies, official documents, religious discourse, political interventions, tourism information, university web pages, academic works, law, national constitution, army information and some short poetry texts. This broad category came to represent a large percentage of the written corpus.

The spoken corpus includes recordings (dialogues and conversations) of spontaneous language on much diversified topics and also recordings from TV and radio programs. Some previous recordings were used and new ones were made, some by researchers and teachers resident in the five African countries, using their recorder, and some by our own team. The main objective of the recordings was essentially to provide materials for lexical, morphosyntactic and syntactic studies. Since this goal did not require an extreme acoustic quality, the fact that not all recordings were made with high quality equipment was not crucial.

These recordings were orthographically transcribed, following criteria defined according to the project objectives. In what concerns orthographic transcription, in this project, no specific marks for overlaps were used. The orthographic transcription included punctuation signs that usually received the same value they have in writing, but giving special importance to their prosodic marker function, so to transmit, even in a rudimental way, the spoken language rhythm. As a general rule of orthographic transcription, the team transcribed the entire corpus according to the official orthography.

New words following regular patterns of derivation posed no problems for transcription and other cases were transcribed as closely as possible to their pronunciation respecting the Portuguese orthography, and in some cases confirmed with native speakers. Foreign words were transcribed in the original orthography when they were pronounced closely to the original pronunciation. When the foreign words were adapted to the Portuguese pronunciation, they were transcribed according to the entries of the reference dictionaries or according to the orthography adopted in those dictionaries for similar cases. When the speaker mispronounced a word and immediately corrected it, the two spellings were maintained in the transcription of the text. But if the speaker misspelled a word and went on in his speech without any correction, the standard spelling of the word was kept in the transcription. Paralinguistic forms and onomatopoeia not registered in the reference dictionaries were transcribed to represent, as much as possible, the sound produced.

These comparable corpora are the first step towards the development of linguistic studies of the Portuguese Varieties of African countries where Portuguese is the official language and is taught, according to the EP variety, as second or foreign language (Bacelar do Nascimento, 2006; Bacelar do Nascimento *et alii*, 2006 e 2007).

2. Extraction of lexical information

The first studies undertaken based on the five comparable corpora are centered on the contrastive properties of each variety's lexicon: contrastive lexicons of the main POS categories, nuclear *vs*. peripheral vocabulary and divergent derivational processes.

2.1 Corpus annotation

In order to achieve these studies, the five comparable corpora have been automatically annotated with POS and lemma information using Eric Brill's tagger (Brill, 1993), previously trained over a written and spoken Portuguese corpus of 250.000 words, morphosyntactically annotated and manually revised.

The initial tag set for the morphosyntactic annotation of the written corpus covered the main POS categories (Noun, Verb, Adjective, etc.) and secondary ones (tense, conjunction type, proper noun and common noun, variable *vs.* invariable pronouns, auxiliary *vs.* main verbs, etc.), but person, gender and number categories were not included.

2.2 Corpus-extracted contrastive lexicons

Five lexicons had been extracted from the corpora, one per each variety, comprising lexical items from the main categories of Common Name, Adjective and Verb, as well as a category for Foreign Words. For each lexical item, the following information is given: POS, lemma and index of frequency of occurrence in the corpus. A total number of 25.523 lemmas have been described: 14.666 (57%) nouns, 6.268 (25%) adjectives, 4.292 (17%) verbs and 297 (1%) foreign words.

The lexicons of the different varieties have been compared and treated statistically, in the form of contrastive lists, with data of frequency and distribution, and are also available at CLUL's webpage for online query.

2.3 Nuclear *vs*. peripheral vocabulary

One of the most important aspects of the contrastive studies on corpora of varieties of a given language, especially languages such as Portuguese, English, Spanish or French, which are spoken in a great diversity of countries, is to establish the grammatical and vocabulary nucleus to all the varieties. This cohesion will assure the understanding among the speakers of these varieties.

In what concerns English, Quirk *et al.*(1985) agree that: «A common core or nucleus is present in all varieties so that, however esoteric a variety may be, it has running through it a set of grammatical and other characteristics that are present in all the others. It is this fact that justifies the application of the name "English" to all the varieties.» (Quirk et *al.*, 1985, *apud* Nelson, 2006, p. 115).

Using the terminology in Nelson (2006), we have extracted the core vocabulary or nucleus of the five corpora (i.e., the common lexicon to all five varieties), as well as the peripheral vocabulary (i.e., that area of the lexicon where, in the corpus, overlapping between varieties do not occur). The common core data are completely reliable, but even in corpora with bigger dimensions, as the International Corpus of English where each variety is 1M words, it is difficult to consider non-overlapping lexical items as definitively specific of one variety, since many situational and contextual factors may determine the occurrence, or not, of lexical items in one subcorpus. Nevertheless, the results of the peripheral vocabulary must be taken into consideration as being an important contribution to our lexical knowledge of AVP, even though they ought to be validated in corpora of bigger dimensions.

Lexical indexes gave us information on the lemmas that constitute the common nucleus of the five subcorpora and on those that had occurred in four, three, two or only one of the subcorpora. We present in Table 2 the quantitative results, in percentile terms, of these occurrences. As we can see, the percentage of common lemmas to the five corpora is lower than the lemmas that have occurred in only one of the subcorpora.

That common nucleus contains the lemmas with bigger frequency of occurrence in the corpus and it can be considered the Basic Vocabulary of the Africa Corpus.

Core lexicon	Common to 5 varieties	26%
Lexicon From core to periphery	Common to 4 varieties	11%
	Common to 3 varieties	11%
	Common to 2 varieties	15%
Peripheral lexicon	Specific to 1 variety	37%

Table 2. Core and peripheral vocabulary in AVP

This common vocabulary to the five corpora (26% of the lemmas) corresponds to 91.75% of occurrences in the corpus. The lemmas that occurred in just one of the corpora present low frequencies or are hapax legomena and are, in fact, more representative cases of lexical change, or africanization, of the lexicon of the Portuguese language.

2.4 Divergent derivational processes

The neologisms presented in Table 3 were collected in the peripheral zones of the vocabulary and are the result of processes of lexical formation with radicals and affixes available in the European variety. This makes possible morphologic structures that derive from the standards of EP and that, therefore, are predictable and of easy interpretation (Rio-Torto, 2007). We only marked as neologisms lexical items that were not present in the exclusion corpus that we first established (i.e., all the lexical items included in two dictionaries of reference: *Vocabulário da Língua Portuguesa* from Rebelo Gonçalves and *Grande Dicionário da Língua Portuguesa*, Porto Editora) or that were labelled as *africanism*. Of course, this does not mean that some of these neologisms cannot occur in spoken or written productions of EP. We present in Table 3 an example of the lexical productivity encountered in the Africa Corpus, with cases of nouns, verbs and adjectives formed with the prefix *des-* 'un-'.

	Angola	**Cape Verde**	**Guinea-Bissau**	**Mozambique**	**S. Tome and Principe**
Nouns	desatracção desinteriorização	desaculturação descrucificação descravização	desfeitura	descamponês desemergência destriunfo	desarrazoável
Verbs	desconseguir desestrelar	desbaralhar		desconseguir desconter desinventar destrabalhar descosturar desimperializar	
Adjectives	descrispado	desapontador desmamentado		desapetitoso	

Table 3. Neologisms with prefix *des* 'un' in AVP

3. Verb complementation

Verb complementation, at the lexicon-syntax interface, is one of the aspects where AVP are diverging from EP. Based on our preliminary analysis of the five corpora, each AVP shows, in fact, an important internal variation regarding verb complementation (and other properties), either converging or diverging from EP patterns. However, data from the five comparable corpora point to several general tendencies.

3.1 Diverging linguistic properties

First, cases where direct objects in AVP (corpus examples (1a) and (2a)) occur as indirect or prepositional objects in EP ((1b) and (2b)):

(1)a. "Pediram o Ministério da Educação Nacional para assumir a sua responsabilidade" G(W)[1]
'[They] asked the Ministry of National Education-dirOBJ to assume its responsibility'

b. Pediram ao Ministério da Educação Nacional que assumisse a sua responsabilidade (EP)
'[They] asked to the Ministry of National Education-indirOBJ that it assume its responsibility-dirOBJ

(2)a. "tinha que ir a escola" M(S)
'[I] had to go the school-dirOBJ

b. tinha que ir à escola
'[I] had to go to the school-prepOBJ

Second, the opposite situation where indirect or prepositional objects in AVP ((3)-(4)) correspond in EP to direct objects:

(3)a. "o Adolfo pegou então a doença que lhe foi matar" A(W)
'Adolfo got then the disease that him-indOBJ killed (killed him)

b. o Adolfo apanhou então a doença que o foi

[1] The codes following the corpus examples indicate their origin: A – Angola; CV – Cape Verde; G – Guinea-Bissau; M – Mozambique; ST – Sao Tome and Principe; S – Spoken; W – Written. The same codes are used for countries in Tables 4-7.

matar (EP)
'Adolfo got then the disease that him-dirOBJ killed

(4)a. "para combater com a delinquência" G(S)
'to fight with the delinquency-prepOBJ'

b. para combater a deliquência (EP)
'to fight the delinquency-dirOBJ'

And third, the fact that complements are frequently introduced in AVP by a different preposition than the one occurring in EP, like in (5). It seems that in AVP the range of prepositions tends to be more limited and some prepositions are extensively used in contexts which would show in EP a large variation, like the case of preposition *em* 'in' (see example (5)) which covers different semantic values.

(5)a. "Menino esperto, você precisa ir na escola" A(W)
'Smart boy, you need to go in-the school

b. Menino esperto, você precisa de ir à escola (EP)
'Smart boy, you need to go to-the school

Moreover, pronominal verbs in EP, either intrinsic pronominal verbs or verbs intrinsically pronominal in a specific meaning, do occur very frequently in AVP as non-pronominal:

(6)a. "O Partido da Renovação Social (PRS) congratulou ontem com a nomeação de Aristides Gomes" G(W)
'The Party of the Social Renovation (PRS) congratulated [himself] yesterday with the nomination of Aristides Gomes'

b. O Partido da Renovação Social (PRS) congratulou-se ontem com a nomeação de Aristides Gomes (EP)
'The Party of the Social Renovation (PRS) congratulated-clitic=himself) yesterday with the nomination of Aristides Gomes'

This affects furthermore inchoative alternations, typically pronominal with certain lexical verb classes in EP, and frequently lexicalized as non-pronominal in AVP, like in (7), a behaviour that requires an in-depth contrastive analysis of lexical verb classes in AVP so as to capture relevant insights on the meaning-syntax relationship.

(7)a. "O seu partido não preocupa com o abandono ou não de Tagme Na Waye" G(W)
'His party does not worry with the abandonment or not of Tagme Na Waye'

b. O seu partido não se preocupa com o abandono ou não de Tagme Na Waye (EP)
'His party does not se-clitic=himself worry with the abandonment or not of Tagme Na Waye'

The differences in verb complementation presented above imply significant changes in the syntax of AVP. For example, the lexicalization of indirect objects as direct objects leads to a structure with double objects, a possibility excluded in EP (see example (8)).

(8)a. "perguntas uma pessoa 'o que é que tu queres fazer?" G(S)
'[you] ask a person-dirOBJ 'what do you want to do'-dirOBJ'

b. "perguntas a uma pessoa 'o que é que tu queres fazer?" (EP)
'[you] ask to a person-IndirOBJ 'what do you want to do'-dirOBJ'

The general tendency to transform indirect and prepositional objects into direct objects leads to the possibility of forming structures like passive and certain inchoative alternations with verbs which do not allow for those constructions in EP (see passive construction in (9)):

(9)a. "É-lhe informado que a chegada do barco seria no dia seguinte" CV(W)
It was-to him-indOBJ informed that the arrival of the boat would be on the next day-SUBJ
'He was informed that the arrival of the boat would be on the next day'

b. Ele é informado de que a chegada do barco seria no dia seguinte (EP)
He-SUBJ was informed of that the arrival of the boat would be on the next day

The passive construction reveals other important aspects, namely, the fact that passives are encountered in AVP with verb classes which do not allow passivization in EP. It is the case of the verb *nascer* 'to be born' (10), an unaccusative verb, i.e. a verb with a subject argument that presents certain syntactic and semantic properties that differ from typical intransitive verbs.

(10)a. "em sessenta e seis fui nascido" A(S)
'in sixty six [I] was born'

b. em sessenta e seis nasci (EP)
in sixty six [I] born

3.2 Analysis of specific lemmas

Although the properties sketched above are pervasive in the five AVP corpora, we wanted to observe possible differences in verb complementation in each variety. In order to do so, we first started by analysing, in the five comparable corpora, concordances of verb lemmas with different syntactic structures, belonging to different lexical classes and in some cases, having a pronominal construction or participating in a pronominal inchoative alternation: *matar* 'to kill', *responsabilizar* 'to hold responsible / to assume responsability', *informar* 'to inform', *combater* 'to fight', *perguntar* 'to ask a question', *pedir* 'to ask for', *habituar* 'to make/get used to', *chegar* 'to arrive', *voltar* 'to return', *precisar* 'to need', *congratular* 'to congratulate', and *preocupar* 'to worry'.

The general results are presented in Table 4, with information, for each variety, on the type of linguistic phenomena, the number of contexts showing divergence from EP (DF), the total frequency of the lemmas considered for this study (TF) and the percentage of diverging contexts from EP. The general phenomena considered are the passage of direct objects to indirect objects (DO > IO) or to prepositional objects (DO > PP), the opposite, namely the passage of indirect or prepositional objects to direct objects (IO/PP > DO), the use of a different preposition introducing a prepositional object, the absence of preposition *de* 'of' introducing noun phrases or clauses and the use of pronominal constructions (either reflexive, inherent or anticausative) as non pronominal ones.

The first important information to draw from these results is the fact that contexts diverging from EP in the corpora are not largely extensive and that the lemmas behave in most cases according to the European norm.

Variety ►	**A**			**CV**			**G**			**M**			**ST**		
Linguistic phenomena ▼	**DF**	**TF**	**%**	**DF**	**TF**	**%**	**DF**	**TF**	**%**	**DF**	**TF**	**%**	**DF**	**TF**	**%**
DO > IO	4	250	**1,6**	-	185	**0,00**	12	225	**5,33**	8	376	**2,13**	1	250	**0,40**
DO > PP	-	32	**0,00**	-	35	**0,00**	1	102	**0,98**	0	51	**0,00**	-	33	**0,00**
IO / PP > DO	1	349	**0,29**	1	408	**0,25**	11	426	**2,58**	3	354	**0,85**	3	216	**1,39**
different preposition	26	1163	**2,24**	2	946	**0,21**	5	1700	**0,29**	6	872	**0,69**	8	644	**1,24**
no preposition DE	19	197	**9,64**	54	246	**21,95**	21	270	**7,78**	23	181	**12,71**	21	174	**12,07**
pron. > non pron.	1	137	**0,73**	-	116	**0,00**	13	170	**7,65**	1	102	**0,98**	3	92	**3,26**
Total	51			57			63			41			36		

Table 4. Diverging syntactic behaviour regarding EP of selected lemmas in the 5 AVP

The second aspect is that these data confirm our first impression when confronted with the five AVP corpora, namely the fact that the Portuguese variety of Guinea-Bissau is the one presenting more diverging patterns regarding EP, although, of course, results are still preliminary. In fact, in Table 4, the Portuguese of Guinea-Bissau presents the most diverging numbers in comparison with EP in what concerns direct objects (DO) becoming indirect objects (IO) or prepositional objects (PP), indirect objects or prepositional objects realized as direct objects, and, also, the use of pronominal verbs as non-pronominal (pron. > non pron.). The opposite general tendency occurs with the variety of Cape Verde, with almost no verbal contexts differing from EP. But surprisingly, the Cape Verde variety shows an extremely high number of occurrences where preposition *de* is omitted. Although the absence of preposition *de* introducing noun phrases or clauses is also a general tendency in EP, the large majority of the cases encountered in AVP differs from the ones found in EP. Table 4 points to several linguistic phenomena where all five varieties diverge from the pattern of EP (the transitivization of verbs, the change of preposition introducing verb complements and the absence of preposition *de* introducing object noun phrases and clauses) while two other patterns occur in 4 varieties (direct objects as indirect ones and the realization of pronominal constructions as non pronominal). These data point to the fact that at least four AVP share a general tendency towards changes in verb complementation, leaving the Cape Verde variety as a special case where a specific tendency is uncovered.

	A			**CV**			**G**			**M**			**ST**		
	DF	**TF**	**%**	**DF**	**TF**	**%**	**DF**	**TF**	**%**	**DF**	**TF**	**%**	**DF**	**TF**	**%**
Different preposition	**26**	**1163**	**2,24**	**2**	**946**	**0,21**	**5**	**1700**	**0,29**	**6**	**872**	**0,69**	**8**	**644**	**1,24**
Perguntar *OI > PP*	8	161	**4,97**	-	177	**-**	-	147	**-**	-	149	**-**	-	37	**-**
responsabilizar *por* 'by' > *de* 'of'	1	15	**6,67**	-	13	**-**	4	34	**11,76**	-	19	**-**	-	18	**-**
Responsabilizar *por* 'by' > *em* 'in'	-	15	**-**	-	13	**-**	1	34	**0,68**	-	19	**-**	1	18	**2,70**
chegar *a* 'to' > *em* 'in'	12	441	**2,72**	2	449	**0,45**	1	458	**0,22**	5	376	**1,33**	7	317	**2,21**
voltar *a* / *para* 'to' > *em* 'in'	5	523	**0,96**	-	279	**-**	-	1028	**-**	-	319	**-**	-	256	**-**
habituar *a* 'to' > *com* 'with'	-	23	**-**	-	28	**-**	-	33	**-**	-	9	**-**	1	16	**6,25**

Table 5. Detailed syntactic behaviour of a specific linguistic property in AVP

Although these conclusions are certainly accurate in what regards the corpus data which was analysed, when observing the results in more detail, one is confronted with a larger amount of variation for the set of lemmas under study among the five varieties. In Table 5, one of the patterns mentioned in Table 4, namely the tendency for the change of preposition introducing object noun phrases and clauses, is furthermore detailed into the lemmas presenting contexts diverging from EP. Although change in preposition is a property very generally attributed to AVP and although this is a property pervasive of all five varieties, we see that frequencies are in fact low, with some more cases in Angola, but with the rest of the AVP presenting some lemmas with just one diverging context. So, the rising percentages correspond in many cases to non significant context frequencies. When observing each lemma independently in each variety, we see that they do behave in very different ways. While *chegar em* 'to arrive in' is a systematic verb complementation change in all AVP, *perguntar em* 'to ask in' is a somehow frequent pattern only in Angola. The general patterns of change in verb complementation only arise in Table 5 and in the full results when grouping contexts from different lemmas.

These data strongly confirm the need to treat each AVP as an independent system showing specific tendencies, even if the five properties do share more general principles of change. Considering our objective of preparing materials for the teaching and learning of Portuguese in those five African countries, we soon understood the need to establish different manuals for each variety, focusing on the diverging phenomena regarding EP, uncovered by corpus analysis. For example, in the case of Guinea-Bissau, all the linguistic aspects analysed in this corpus would have to be covered, but taking into attention specific lemmas which proved to present more diverging patterns from the EP norm, like the verbs *informar*, *precisar* and *preocupar*. While the Portuguese variety of Angola will need more focus on the change of preposition with verb *perguntar* and *responsabilizar* and the Cape Verde variety will essentially need some attention to the absence of preposition *de*, especially with verb *precisar* when followed by an infinitive verb.

The choice of four lemmas (verbs *preocupar*, *responsabilizar*, *congratular* and *habituar*) pointed to the fact that pronominal constructions tend to be used as non pronominal, and the global frequencies were presented in Table 4. But a closer look to other verbs (non pronominal in EP) show contexts where those are used pronominally, with a clitic pronoun which does not correspond to a verb complement, but rather to a particle with different functionalities. This raises the question of whether we are facing a tendency towards a non pronominal use of pronominal verbs or a mixed tendency, towards both insertion and loss of clitic pronoun. We then analysed all the contexts where the two patterns were found in the five spoken subcorpora, for all lemmas. The results are presented in Table 6, with information, for each variety, on the total frequency of contexts showing insertion or absence of a clitic pronoun and thus diverging from EP.

Pronoun	**A**	**CV**	**G**	**M**	**ST**
Insertion	4	5	11	9	5
Absence	23	29	63	42	75

Table 6. Differences in pronominal constructions in AVP compared to EP

Indeed, numbers are clearly more significant in the case of non pronominal uses of verbs which would be pronominal in EP, and this tendency is shared by all AVP. The insertion of a clitic pronoun in contexts of non pronominal constructions could be seen as a consequence of the tendency to omit the pronoun, since this would inevitably generate some confusion on the use of pronominal constructions and also a tendency towards overcorrection.

It is important to take into account that we are facing varieties of Portuguese that have not yet reached a stable point of evolution, so that only a detailed contrastive analysis of more data concerning verb complementation could point us towards the path of understanding the current ongoing changes.

4. Degree of variation regarding European Portuguese

As already mentioned, most of the verb lemmas that we analysed in the corpus had low frequencies and especially low frequencies of contexts diverging from EP. Another question was the strong variation found among the five AVP regarding those lemmas, which seemed to make it difficult to assess the degree by which each AVP differ from the European norm. However, when observing texts from the different varieties, the higher or lower degree of variation was more evident. This lead us to search for the totality of diverging contexts regarding some linguistic phenomena instead of focusing on some lemmas. Of course, this objective was not doable over the whole corpus and we decided to limit this study to the spoken subcorpus, since it does present more AVP-specific patterns than the written one.

Four important linguistic phenomena showing divergence from the EP standard were selected for a comparison between the five spoken corpora: the position of the clitic pronoun regarding the verb, the concordance inside the nominal phrase and between subject and predicate, verbal conjugation and insertion or absence of clitic pronouns (already discussed in Table 6). The results are presented in Table 7.

	A	**CV**	**G**	**M**	**ST**
Position of clitic pronouns	**40**	**23**	**46**	**68**	**20**
Concordance: total	**136**	**57**	**241**	**116**	**64**
Nominal concordance					
Determiner-noun: gender	6	5	47	12	4
Determiner-noun: number	38	5	39	27	11
Noun-adjective: gender	2	5	22	2	2
Noun-adjective: number	13	6	21	11	1
Subject-predicative noun: gender	13	-	16	6	5
Subject-predicative noun: number	13	-	2	4	3
Verbal concordance					
Person	25	10	39	8	5
Number	26	26	55	46	33
Conjugation	**59**	**29**	**96**	**78**	**43**
Mood	26	19	69	52	20
Time	33	10	27	26	23
Pronominal constructions	**27**	**34**	**74**	**51**	**80**
Insertion of clitic pronoun	4	5	11	9	5
Absence of clitic pronoun	23	29	63	42	75
Total	**262**	**143**	**457**	**313**	**207**

Table 7. Patterns of variation in the five AVP

Table 7 includes the number of diverging occurrences in each variety and for each property, together with the final number of diverging contexts for each country. These contexts were found over a total number of around 25-27 thousand words, the number of words of each spoken subcorpus.

When looking globally at different linguistic aspects in the whole subcorpora and comparing the data in Table 7, the variety of Guinea-Bissau emerges as the most diverging one regarding EP. It comes as a confirmation of the conclusion already attained with the study of verb complementation (which seemed however limited by low frequencies). Although the Mozambican variety shows higher results in what concerns the position of clitic pronouns and although the variety of Cape Verde shows slightly higher results in what concerns pronominal constructions, the contexts regarding concordance and verb conjugation and the global results isolate Guinea-Bissau. The other varieties follow gradually, Mozambique, Angola, Sao Tome and Principe, in that order, and, finally, Cape Verde, which is the less diverging variety of the five, as the results of verb complementation already showed.

4. Conclusions

If, on the one hand, the adequate description of the linguistic properties of AVP requires the use of balanced corpora, it is also true that, on the other hand, the evaluation of the degree by which they diverge from the EP norm as well as the contrastive study the five AVP will only be possible through the access to comparable corpora of those varieties.

The recently compiled comparable corpora of the AVP are the first step towards a better understanding of the similarities and differences encountered among them and EP and between each variety. The first linguistic results based on these corpora have been a contrastive lexicon which establishes a common core vocabulary, as well as peripheral lexical sets for each variety. Two strongly diverging phenomena are under contrastive study, morphological and lexical analysis of derivational processes in AVP, as well as verb complementation.

However important the results may be, in order to reach confident observations regarding the evolution of AVP and their relationship with EP, it is necessary to ensure the enlargement of the existing five comparable corpora, with special attention to the spoken subpart, since most AVP-specific properties only show in the spoken register. In fact, since most of the properties where AVP differs from the EP norm are still emergent and show strong variation inside each variety, it is essential to rely on comparable corpora which are balanced. Only then will we be able to establish more stable tendencies of linguistic change across the varieties and inside each variety.

This will be important to give teachers more knowledge about the linguistic properties that are characteristic of the African varieties of Portuguese as well as teaching and learning materials that could point to the process of identification and understanding of the unity and diversity factors that are at stake between these varieties and between those and European Portuguese.

5. Aknowledgements

The Corpus Africa was compiled and explored in the scope of the projects "Linguistic Resources for the Study of the African Varieties of the Portuguese", of the Center of Linguistics of the University of Lisbon (CLUL) and of the Center of Theoretical and Computational Physics of the University of Lisbon, and "Properties of the African Varieties compared with the European Portuguese". The first project had been financed by the Foundation for Science and the Technology under the cover of the Lusitânia Program and for the Foundation Calouste Gulbenkian and the second had been financed only by this Foundation. The first project was supported by Center de Information and Documentation Amílcar Cabral (CIDAC), the Open University of Lisbon, the Camões Institute, the Embassy of the Democratic Republic of S. Tome and Principe and the Group of Natural Language of the Department of Computer Science of the College of Sciences of the University of Lisbon. The consultant of the first project was Perpétua Gonçalves and, of the second, Perpétua Gonçalves and Amália Mendes (For more information on the projects, see the webpage http://www.clul.ul.pt/sectores/linguistica_de_corpus/linguistica_de_corpus.php.

6. References

Bacelar do Nascimento, M. F. , Bettencourt Gonçalves, J. (2006). The Role of Spoken Corpora in Teaching/Learning Portuguese as a Foreign Language - The Case of Adjectives of Intensification. In Kawaguchi, Y., Zaima, S., Takagaki T. (eds.) *Spoken Language Corpus and Linguistic Informatics*. Amsterdam: Jonh Benjamins, Coll. Usage-Based Linguistic Informatics, vol.V, pp. 219-226.

Bacelar do Nascimento, M. F., Pereira, L. A. S., Estrela , A., Bettencourt Gonçalves, J., Oliveira, S. M., Santos, R. (2006). The African Varieties of Portuguese: Compiling Comparable Corpora and Analyzing Data-derived Lexicon". In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC2006)*, 24-26 May, Genoa, Italy, pp. 1791-1794.

Bacelar do Nascimento, M. F., Pereira, L. A. S., Estrela, A., Bettencourt Gonçalves, J., Oliveira, S. M., Santos, R. (2007). As variedades africanas do português: um corpus comparável". *X Simposio Internacional de Comunicación Social*, Ministerio de Ciencia, Tecnología y Medio Ambiente, v. I, Janeiro, Santiago de Cuba.

Bacelar do Nascimento, M. F., Pereira, L. A. S., Estrela, A., Bettencourt Gonçalves, J., Oliveira, S. M., Santos, R. (2007). Especificidades das Variedades Africanas do Português na Formação dos Professores de Português". *Saber Ouvir, Saber Falar, 7º Encontro Nacional da Associação de Professores de Português* (CD-ROM).

Gonçalves, P., Stroud, C. (Orgs.) (1998). *Panorama do Português Oral de Maputo*, Vol. 3 – *Estruturas Gramaticais do Português: Problemas e Exercícios*, Cadernos de Pesquisa nº 27. Maputo, INDE.

Gonçalves, P., Stroud, C. (Orgs.) (2000). *Panorama do Português Oral de Maputo*, Vol. 4 – *Vocabulário Básico do Português (espaço, tempo e quantidade) Contextos e Prática Pedagógica*, Cadernos de Pesquisa nº 36. Maputo, INDE.

Gonçalves, P., Stroud, C. (Orgs.) (2002). *Panorama do Português Oral de Maputo*, Vol. 5 – *Vocabulário Básico do Português, Dicionário de Regências*, Cadernos de Pesquisa nº 41. Maputo, INDE.

Gonçalves, P. (1997). Tipologia de Erros. In Stroud, C., Gonçalves, P. (Orgs.) (1997b).

Gonçalves, R. (1966). *Vocabulário da Língua Portuguesa*. Coimbra: Coimbra Editora.

Grande Dicionário da Língua Portuguesa (2004). Porto, Porto Editora.

Greenbaum, S. (1996). *Comparing English Worldwide: The International Corpus of English*. Oxford, Clarendon Press.

Nelson, G. (2006). The core and the periphery of world Englishes: a corpus-based exploration. *World Englishes*, 25:1, pp. 115-129.

Quirk, R., Greenbaum, S., Leech, G., Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. London, Longman.

Rio-Torto, G. (2007). Caminhos de Renovação Lexical: Fronteiras do Possível. In Isquerdo, A. N., Ieda, M. A. (Orgs.), *As Ciências do Léxico, Lexicologia, Lexicografia, Terminologia*, Vol. 3. Campo Grande, MS, Ed. UFMS; São Paulo, Humanitas.

Stroud, C., Gonçalves, P. (Orgs.) (1997a). *Panorama do Português Oral de Maputo*, Vol. 1 – *Objectivos e Métodos*, Cadernos de Pesquisa nº 22. Maputo, INDE.

Stroud, C., Gonçalves, P. (Orgs.) (1997b). *Panorama do Português Oral de Maputo*, Vol. 2 – *A Construção de um Banco de " Erros"*, Cadernos de Pesquisa nº 24. Maputo, INDE.

Empirical Studies on Language Contrast Using the English-German Comparable and Parallel CroCo Corpus

Oliver Čulo[1], Silvia Hansen-Schirra[1], Stella Neumann[2], Mihaela Vela[3]
[1]Applied Linguistics Johannes Gutenberg University Mainz, Germersheim, Germany
[2]Applied Linguistics, Translation and Interpreting Saarland University, Saarbrücken, Germany
[3]German Research Center for Artificial Intelligence, Saarbrücken, Germany
culo@uni-mainz.de, hansenss@uni-mainz.de, st.neumann@mx.uni-saarland.de, Mihaela.Vela@dfki.de

Abstract

This paper presents results from empirical studies on language contrasts, translation shifts and translation strategies gained by exploiting the CroCo Corpus. The aim of this paper is to show that the insights from investigating the comparable parts of the corpus can be complemented by additionally exploiting the parallel parts of the corpus using the examples of word order peculiarities and diverging part-of-speech frequencies in English and German. The exploitation of the corpus proceeds in two steps. First, contrastive differences are identified in the comparable parts of the corpus. In the second step, the solutions chosen by human translators to deal with the contrastive differences are identified. These can be used to decide between different possible translation strategies and can serve as templates for translation strategies to be adopted in the development of MT systems.

1. Multilingual Corpora in Translation

The creation of linguistic corpora in the past decades has made possible new ways of researching linguistic phenomena and refining methods of processing language with the computer. In the field of translation, corpora are making inroads as well. Corpus-based translation studies are steadily gaining interest thus potentially serving as an input to research in the field of machine translation as well.

The aspects we can study from comparable and parallel corpora differ. However, the decision is not necessarily between creating either a comparable or a parallel corpus. One outcome of the CroCo project[1] is a corpus that contains both parts.

This paper demonstrates how the CroCo corpus (Neumann & Hansen-Schirra, 2005) can be used both as comparable as well as parallel corpus and what kind of insights we can gain for each of the fields mentioned above. It also shows how techniques from both worlds can complement each other.

The paper is organized as follows. In section 2 we shortly introduce the topics of language contrasts, translation shifts, translation strategies and information structure. In section 3 we will present the design and representation of our English-German corpus of originals and translations as well as its exploitation. Section 4 discusses the findings from the corpus exploitation. Section 5 gives an overview of our conclusions and offers an outlook on computational applications of our findings.

2. Strategies for Handling Language Contrasts

Language contrasts can be studied by investigating corpora of the languages involved using multilingually comparable techniques (Granger et al., 2003). Contrasts become visible at all levels of language, in graphology (in written mode), morphology, syntax and on text level and can be investigated empirically with the help of comparable corpora. For instance, the claim that English has a more rigid word order than German with the subject mostly in sentence-initial position can easily be tested on a corpus like the annotated CroCo corpus by simply querying the number of sentences where the subject is in sentence-initial position in both languages (see section 4.1). Examples from the corpus may be helpful to understand how German word order relates to English in terms of rigidness.

When comparing source texts and their translations in another language (parallel techniques), translation shifts become apparent. Translation shifts have been discussed in translation studies since the 1950s (Vinay & Darbelnet, 1958; Catford, 1965; Newmark, 1988; van Leuven-Zwart, 1989). The accounts are similar in that they categorize lexical, grammatical, and semantic shifts. On the level of lexis, the focus is on strategies for gaps or *lacunae*, i.e. lexical items that do not exist in the target language. Grammatical shifts are often called *transpositions* and refer to changing tense, number, person, part-of-speech. They function in the target text without changing the meaning. A special case is what Catford (1965) calls level shifts where the shift involves both lexis and grammar, because a given grammatical construction is not available in the target language and has to be replaced by an alternative lexical item reflecting the meaning of the construction. In semantic shifts, or *modulations* (Vinay & Darbelnet, 1958), a change of perspective occurs between source and target text. This may involve concretion, explication, negation of the opposite, (de-) passivization, etc.

In computational linguistics, translation shifts of all types are a crucial issue for the development of MT sys-

[1] http://fr46.uni-saarland.de/croco, funded by the German Research Foundation as project no. STE 840/5-2 and HA 5457/1-2

tems. Identification, classification and formalization of translation shifts have received considerable attention in the MT community (e.g. in the Eurotra project, Copeland et al., 1991). Within this context, Barnett et al. (1991) introduce a rough distinction between translation *divergences* for mere structural differences and *mismatches* for changes which also comprise shifts in meaning. Under the umbrella term *complex transfer*, Lindop & Tsujii (1991) present a comprehensive discussion of examples that appear to be problematic for MT. On this basis, Kinoshita et al. (1992) classify these divergence problems into four categories: argument-switching, head-switching, decomposition and raising. Dorr (1994) proposes a more fine-grained categorization of MT divergences. She distinguishes between thematic, promotional, demotional, structural, conflational, categorical and lexical divergences, thus using linguistic categories. Additionally, she presents a formal description of these divergences and an interlingua approach to a systematic dealing with divergences.

In more recent studies, multiply annotated parallel corpora are used to develop interlingual representations (Farwell et al., 2004) or to learn transfer rules (Čmejrek et al., 2004; Hinrichs et al., 2000). These approaches implicitly include translation shifts in MT procedures and could benefit from input from translation studies. Cyrus (2006) combines the two perspectives, but her focus on the predicate argument structure restricts the findings to semantic shifts. A further limitation of the study results from the direct annotation of translation shifts. A theory-neutral annotation and alignment on different levels like the one proposed here offers the opportunity to query the corpus for different purposes.

On sentential and textual level, the translator is faced with an information structure which, due to grammatical, lexical and other differences cannot always be directly reproduced thus entailing modulation (see section 4.1). The *translation strategies* used to map information structures from one language onto another result in shifts that may occur on all linguistic levels and are due to the translator s understanding as well as idiosyncratic preferences during the translation process, to contrastive differences between the languages involved or to different register characteristics.

The present paper presents a linguistically founded approach to detecting translation shifts and studying language contrasts and translation strategies in a multiply annotated and aligned comparable and parallel corpus.

3 . Corpus Design, Representation and Exploitation

The CroCo corpus was built to investigate contrastive commonalitites and differences between the two languages involved as well as peculiarities in translations. It consists of English originals (EO), their German translations (GTrans) as well as German originals (GO) and their English translations (ETrans). Both translation directions are represented in eight registers, with at least 10 texts totalling 31,250 words per register. Altogether the CroCo Corpus comprises approximately one million words. Additionally, register-neutral reference corpora are included for German and English including 2,000 word samples from 17 registers.

The corpus thus consists of both, comparable and parallel, parts. The registers are political essays (ESSAY), fictional texts (FICTION), instruction manuals (INSTR), popular-scientific texts (POPSCI), corporate communication (SHARE), prepared speeches (SPEECH), tourism leaflets (TOU) as well as websites (WEB) and were selected because of their relevance for the investigation of translation properties in the language pair English-German. All texts are annotated with

- meta information including a brief register analysis that allows additional filter options following the TEI standard (Sperberg-McQueen & Burnard, 1994),
- part-of-speech information using the TnT tagger (Brants, 2000) with the STTS tag set for German (Schiller et al., 1999) and the Susanne tag set for English (Sampson, 1995),
- morphology using MPRO (Maas, 1998) which operates on both languages,
- phrase structure again using MPRO and
- grammatical functions of the highest nodes in the sentence, manually annotated with MMAX2 (Müller & Strube, 2006).

Furthermore, all texts are aligned on

- word level using GIZA++ (Och & Ney, 2003),
- chunk level indirectly by mapping the grammatical functions onto each other,
- clause level manually again using MMAX2,
- sentence level using the WinAlign component of the Trados Translator s Workbench (Heyn, 1996) with additional manual correction.

For an effective exploitation of the annotated data, the annotation and alignment is converted into a MySQL database. The information on token level, such as tokenization, part-of-speech, lemmatization and word alignment, is written into tables in the database. The tokens in one language are indexed, each index referring to a string, a lemma, a part-of-speech tag and an index for its alignment in the other language. At chunk level, the tables are filled with information about chunk type and the grammatical function it fulfills. The tables for chunks are connected to the information at token level. Analogously, the clause and sentence segmentations as well as the corresponding alignments are transformed into tables connected to the token tables in the MySQL database. This type of storage offers an easy and fast method to query the corpus. Additionally, a query interface with a menu-like, predefined set of queries can be connected to the database, also allowing non-experts to query the corpus.

4 . Findings

4.1 Information Structure in German and English-German Translations

The CroCo corpus is used to study and compare linguistic phenomena both from a cross-lingual and a monolingual perspective using original and translated texts. This has been done for grammatical functions in theme (i.e. sentence-initial) position as table 1 illustrates. The figures have been computed for the register SHARE.

	subj	obj	compl	adv	verb	other
EO_SHARE	63.43	0.15	0.15	27.14	0.80	8.35
ETRANS_SHARE	64.20	0.19	0.45	27.13	0.25	7.77
GTRANS_SHARE	55.47	2.42	0.22	36.08	0.51	5.29
GO_SHARE	50.25	8.46	1.70	31.00	1.20	7.39

Table 1: Grammatical functions in theme position (in percent).

Focussing on the grammatical functions subject (abbreviated subj) and adverbials (adv), the quantitative-figures confirm the widespread assumption that English has a stronger tendency than German to put the subject in theme position. The proportion of subjects in sentence-initial position in EO is more than 13 percentage points higher than in GO. The figures suggest a general tendency in German SHARE texts to vary the function located in sentence-initial position. This can be attributed to language-typological peculiarities of mapping the grammatical functions on semantic roles in the two languages involved (Hawkins, 1986). English is more restricted as to the location of the subject, but the subject can accommodate various semantic roles more easily than German. Conversely, German is more flexible as to which element goes first in the sentence, but requires different grammatical functions to reflect the various semantic roles.

Both, the human translator and the MT system, have to accommodate these differences in the translation. There are two possible solutions for cases, where a one-to-one translation is not possible. Either (1) the order of the grammatical functions remains constant and the semantic content of the original is moved to a different grammatical function or (2) the linear precedence of the semantic content is kept and the order of grammatical functions is changed.

To retrieve the strategy preferred by human translators, we query the source sentence subject chunk in combination with the word alignment. Where the semantic content is not part of the target sentence subject chunk, the word alignment points to a different grammatical function. At present, the results have a low precision and recall rate and can therefore only be seen as a first indication.

Two findings (cf. Kast, 2007) seem particularly interesting: In the translation direction German-English, the lexical content of subjects is often shifted to direct objects.

> **GO**: Auch im Berichtsjahr setzte [die SAP] ihre bewährte Politik des offenen und intensiven Meinungs- und Informationsaustausches fort.
> **ETrans**: [1994] saw SAP continue to pursue its proven policy of open and intensive exchange of information and values.

Here, the translator has chosen solution 2: The subject in GO (in squared brackets) is located after the verb, a position that is not easily accessible to the English subject. Consequently, the perspective is changed in the translation with the temporal information now in the subject and the former agent SAP now a direct object (underlined), thus leading to a modulation (see section 2.1).

The translation direction English-German highlights a shift from subject in EO to adverbial in GTrans.

> **EO**: [Day 2] covered new thinking in Globalization, Six Sigma and Product Services.
> **GTrans**: Am zweiten Tag widmete [man] sich dem Gedankenaustausch und neuen Ideen zu den Themen Globalisierung, Six Sigma und produktbezogene Dienstleistungen.

Again, solution 2 seems to be the preferred one: Rather than changing the precedence of the semantic content, the translator chose to map the content on another function that is more amenable to temporal information in German, namely an adverbial.

These initial findings point to a preference in human translation to preserve information sequencing while varying the mapping of grammatical functions, thus accepting a change in perspective. This result can be used in the development of MT systems when aiming at producing a more natural output.

4.2 Part-of-speech Distributions and Shifts

As on the level of chunks, parts-of-speech reflect clear differences between the two languages as can be seen from the comparable corpora displayed in table 2. Both, the reference corpora (ER and GR) and the register-controlled corpora (EO_ and GO_SHARE) show divergences that require handling during translation. The interpretation of these divergences, however, is not always straightforward.

	noun	adj	verb	adv
ER	24.60	6.24	15.72	4.63
GR	22.93	9.20	13.04	5.02
EO_SHARE	29.14	6.97	13.83	3.15
GO_SHARE	25.30	10.69	11.64	4.30

Table 2: Part-of-speech statistics in %

Interestingly, we find a higher percentage of nouns in English than in German. One reason for the former observation is a clearly technical one. German compounds are written in one word (e.g. Gerichtsentscheidung), where as the parts of English compounds are mostly separated (e.g. court decision). The POS tagger does not decompose compounds, so where a compound containing two or more nouns is only counted once for German, each part is counted separately in English.

Furthermore, the proportion of verbs seems to be higher in English originals than in the German comparable texts. This divergence can be observed in the contrastive reference corpora as well as in the register-controlled corpora. Rather than for technical reasons, this seems to be a genuine contrastive difference between the two languages, that can be expected to have an effect on translation in the form of transpositions (see section 2). Transpositions can be retrieved from the corpus by querying for an aligned word pair with different part-of-speech tags. Table 3 illustrates the frequency of the different transpositions for both translation directions, taken from SHARE.[2]

Type of shift	English-German	German-English
verb-noun	24.31	16.98
verb-adjective	11.69	2.80
verb-adverb	6.95	0.25
adjective-noun	17.43	9.48
adjective-verb	1.84	9.92
adjective-adverb	1.42	11.58
noun-adjective	13.89	21.63
noun-verb	5.74	16.98
noun-adverb	3.40	1.08
adverb-adjective	10.06	1.34
adverb-noun	3.05	1.59
adverb-verb	0.21	6.36

Table 3: Frequencies of transpositions in %

For this sub-corpus, we have a total of 40,090 English-German aligned lexical word pairs, among which 1,411 (3.52%) shifts are found, and 37,694 German-English aligned word pairs with 1,572 (4.17%) shifts. Comparing the types of shifts, we can generalize that we find more verb to x alignments for English-German, but fewer x to noun alignments and more noun to x alignments for German-English. This means that English translations are less nominal than their German originals. The following excerpt is taken from the English-German verb-noun list and displayed as follows: original pos ### translation pos.

do	-	vd0	###	Handeln	-	nn
play	-	vv0	###	Spielen	-	nn
work	-	vv0	###	Arbeiten	-	nn
programming	-	vvg	###	Programme	-	nn
communicate	-	vv0	###	Kommunikation	-	nn
believe	-	vv0	###	Auffassung	-	nn
computing	-	vvg	###	Computers	-	nn
compared	-	vvn	###	Vergleich	-	nn
learn	-	vv0	###	Lernen	-	nn
enters	-	vvz	###	Schwelle	-	nn
integrate	-	vv0	###	Integration	-	nn
develop	-	vv0	###	Entwicklung	-	nn
browsing	-	vvg	###	Browsen	-	nn
manage	-	vv0	###	Verwalten	-	nn
connect	-	vv0	###	Verbindung	-	nn
control	-	vv0	###	Kontrolle	-	nn

The following example illustrates these English-German transpositions, which result in nominalizations in the German translation.

EO: Whether you want to communicate, learn, work, or play, the PC can enrich and improve the experience.
GTrans: Ganz gleich, ob Sie ein Hilfsmittel zur Kommunikation oder zum Lernen, Arbeiten oder Spielen benötigen, der PC kann diese Erfahrung eindringlicher und besser gestalten.

The solutions found by the human translators might be mistaken as deficient. In fact, the above example shows an appropriate solution to the difference between English and German in terms of the frequency of infinite constructions. Transpositions can therefore serve as a basis to develop transfer rules in MT systems that handle contrastive differences between the languages involved.

5 . Conclusions and Outlook

The paper has presented findings from empirical studies in a German-English comparable and parallel corpus. It has shown that techniques applied for either comparable or parallel corpora can complement each other, providing explanations from the latter for observations made using the former corpus. The findings from the analyses presented here demonstrate that solutions chosen be human translators which appear to deviate from the source text must not necessarily be defective. They can rather be viewed as a valuable resource for creating a more natural output of MT systems taking into account contrastive differences in language use. These can only be identified with the help of a comparable corpus. The findings encourage further investigation into language contrasts, translations shifts and translation strategies.

For the application in MT, the use of corpora - and thus empirical resources for language contrasts, translation shifts and translation strategies - is expected to be more dynamic than rule-based approaches. The combination of linguistic corpus enrichment and the extracted translation shifts allows compiling a comprehensive set of transfer rules for MT systems, ideally evaluated on the basis of translation models from translation studies. With this approach, existing translations serve as a basis for solving translation problems thus making the MT output more similar to human translation.

6 . References

M. Baker. (1993). Corpus linguistics and translation studies: implications and applications. M. Baker, G. Francis & E. Tognini-Bonelli (Eds.), *Text and technology: in honour of John Sinclair*. Amsterdam & Philadelphia: John Benjamins, pp. 233-250.

[2] The error rate for the part-of-speech tagger is 3.07% for the German subcorpora and 5.09% for the English subcorpora. Tested on a small sample from SHARE, the word aligner reaches 78.1% precision and 62.8% recall. Other influences on precision and recall include problems of mapping the contrastive tag sets. However, these are difficult to quantify.

J. Barnett, I. Mani, P. Martin & E. Rich. (1991). Reversible machine translation: What to do when the languages don t lineup. *Proceedings of the Workshop on Reversible Grammars in NLP*. ACL-91, Berkeley, pp. 61-70.

T. Brants. (2000). TnT - A Statistical Part-of-Speech Tagger. *Proceedings of ANLP-2000,* Seattle.

J.C. Catford. (1965). *A Linguistic Theory of Translation*. Oxford University Press, Oxford.

M. Čmejrek, J. Cuřín, J. Havelka, J. Hajič & V. Kuboň. (2004). Prague Czech-English Dependecy Treebank: Syntactically Annotated Resources for Machine Translation. *Proceedings of LREC 04,* Lisbon.

C. Copeland, J. Durand, S. Krauwer & B. Maegaard. (1991). The Eurotra linguistic specifications. *Studies in MT and NLP* Vol. 1, Brussels.

L. Cyrus. (2006). Building a resource for studying translation shifts. *Proceedings of LREC 06*, Genoa, pp. 1240-1245.

B. J. Dorr. (1994). Machine Translation Divergences: A Formal Description and Proposed Solution. *Computational Linguistics* 20:4, pp. 597-633.

D. Farwell, S. Helmreich, B. J. Dorr, N. Habash, F. Reeder, K. Miller, L. Levin, T. Mitamura, E. Hovy, O. Rambow & A. Siddharthan. (2004). Interlingual Annotation of Multilingual Text Corpora. *Proceedings of the HLT-NAACL Workshop on Frontiers in Corpus Annotation*, Boston, pp. 55 62.

S. Granger, J. Lerot & Stephanie Petch-Tyson (eds.). (2003). *Corpus-based approaches to contrastive linguistics and translation studies.* Amsterdam & New York: Rodopi.

J. A. Hawkins. (1986). *A Comparative Typology of English and German: Unifying the Contrasts.* London: Croom Helm.

M. Heyn. (1996). Integrating machine translation into translation memory systems. *European Association for MT, Workshop Proceedings*, ISSCO, Geneva, pp. 111-123.

E. W. Hinrichs, J. Bartels, Y. Kawata, V. Kordoni & H. Telljohann. (2000). The VERBMOBIL Treebanks. *Proceedings of KONVENS 2000 Sprachkommunikation*, ITG-Fachbericht 161, VDE-Verlag, pp. 107-112.

M. Kast. (2007). *Variation innerhalb der grammatischen Funktion "Subjekt" bei Übersetzungen Englisch-Deutsch und Deutsch-Englisch.* Unpublished diploma thesis. Saarbrücken: Applied Linguistics, Translation and Interpreting, Universität des Saarlandes.

S. Kinoshita, J. Phillips & J. Tsujii. (1992). Interaction between structural changes in machine translation. *Proceedings of COLING 92,* Nantes, pp. 679-685.

J. Lindop & J. Tsujii. (1991). *Complex transfer in MT: A survey of examples*. Technical report, CCL/UMIST 91/5, Manchester.

H. D. Maas. (1998). Multilinguale Textverarbeitung mit MPRO. *Europäische Kommunikationskybernetik heute und morgen 98,* Paderborn.

C. Müller & M. Strube (2006). Multi-Level Annotation of Linguistic Data with MMAX2. S. Braun, K. Kohn, J. Mukherjee (eds.) *Corpus Technology and Language Pedagogy. New Resources, New Tools, New Methods*. Peter Lang, Frankfurt. pp. 197-214.

S. Neumann & S. Hansen-Schirra. (2005). The CroCo Project. Cross-linguistic corpora for the investigation of explicitation in translations. *Proceedings of the Corpus Linguistics Conference Series* Vol. 1 no. 1.

P. Newmark. (1988). *A Textbook of Translation*. Prentice Hall, New York.

F. J. Och & H. Ney. (2003). A Systematic Comparison of Various Statistical Alignment Models. *Journal of Computational Linguistics* Nr.1, vol. 29, pp. 19-51.

G. Sampson. (1995). *English for the Computer. The Susanne Corpus and Analytic Scheme*. Clarendon Press, Oxford.

A. Schiller, S. Teufel & C. Stöckert. (1999). Guidelines für das Tagging deutscher Textkorpora mit STTS. *Technical report, IMS, University of Stuttgart, Seminar für Sprachwissenschaft,* University of Tübingen.

C. M. Sperberg-McQueen & L. Burnard (eds.). (1994). *Guidelines for Electronic Text Encoding and Interchange (TEI P3)*. Text Encoding Initiative, Chicago and Oxford.

K. van Leuven-Zwart. (1989). Translation and Original. Similarities and Dissimilarities, *Target* 1(2), pp. 151-181.

J.-P. Vinay & J. Darbelnet. (1958). *Stylistique comparée du français et de l anglais* Les éditions Didier, Paris

Poster presentations

Heuristics for Term Extraction from Parallel and Comparable Text: The KB-N Legacy

Magnar Brekke
Norwegian School of Economics and Business Administration
Helleveien 30
N-5045 Bergen
Norway
Magnar.Brekke@nhh.no

Abstract

KB-N (KnowledgeBank of Norway) is the result of a 3-year project to establish a concept-oriented knowledge-bank for economic-administrative domains. Our point of departure is the basic assumption that special subdomain knowledge is embedded in text produced by domain experts and expressed through terminological units (specialist vocabulary) reflecting the fundamental concepts of the subdomain. The current state of NLP research reflects major strides in the automatic processing of large text corpora for the purpose of mining such knowledge residing in professional text. Semi-automatic bilingual terminology extraction based on parallel text processing is largely in place, and the KB-N team has developed software and methods for approaching Norwegian along these lines, despite the limited supply of parallel text especially when Norwegian is the other member of a language pair. The possibility for transferring and extending the approach to also handle comparable corpora, where the text supply is virtually unlimited (but subject to significant constraints, to be discussed shortly) is much less certain. This paper reports on an experiment to explore these possibilities by exploiting the notion of "weirdness" or significance/salience in evaluating the statistical likelihood of a given vocabulary unit turning up in a specific text. Preliminary results are encouraging.

1. Background

KB-N has two major components. One is a comprehensive corpus of specialized domain text representing relevant document types and text genres. Initial focus has been on capturing parallel texts in English and Norwegian where one is a certified translation of the other. Using Stuttgart's CWB and Oracle as a platform each text has been XML-coded and POS-tagged (using the Oslo-Bergen Tagger (Hagen et al., 2000)). The parallel text versions have then been aligned via Hofland's lexical anchor method (Hofland, 1996) and made available for automatic as well as user-initiated bilingual parallel concordancing, an essential feature of automatic term extraction (to be described below). The current holdings comprise about 800,000 words of strictly parallel text (English/Norwegian 50/50). The plans are to enrich and extend this initial parallel corpus with a considerably larger collection of comparable text, i.e. original either English or Norwegian monolingual texts representing the same domains and communication types.

The other major component is a concept-oriented bilingual terminological database, a repository of domain-specific knowledge extracts from the corpus supplemented by relevant items not represented in the text samples (as determined by domain expert) in order to attain a significant domain coverage. Each term record accommodates term equivalents, synonyms, acronyms etc. in the respective languages (p.t. English and Norwegian) and links them to their common concept. The pivotal role played by the concept facilitates future inclusion of other languages in the term bank. We are currently experimenting with representing conceptual structures graphically in a separate ontology window, where structures or elements can be established, accessed, inspected and manipulated. For each concept its relative position in the concept structure (whether hierarchic, cognitive or otherwise) is indicated.

The emergence of computerized corpus-based methods has of course had an enormous impact on terminology research but without entirely displacing the time-honored technique of excerpting by hand. In fact the general problems of "silence" and "noise" in terminology extraction (well described in Castellvi et al. (2001)) provide ample justification for viewing the two approaches as complementary, by allowing relevant items not represented in the text samples to be supplied by a subdomain expert. For each term one or more characteristic authentic usage contexts are given, and to aid (automatic) word sense disambiguation (essential for MT) a set of domain-specific collocations are listed whenever identifiable. The link between term and concept must be established by a domain expert, whose tacit knowledge is also required for the identification of "missing" concepts based on the systematization of conceptual structures. Other than such input the remaining knowledge represented in the term record is either extracted from or in the main based on the corpus text samples.

The basic mechanism linking the text base and the term base is the semi-automatic extraction of term candidates from parallel text, to which we now turn.

2. Term Extraction from parallel text

Most of the specific techniques proposed for automatic term extraction are strongly sensitive to typological differences between languages. Thus the strategies available for English differ markedly from those relevant for Romance languages, or for those of Germanic stock, including Norwegian. Øvsthus (2005) is the first published work indicating that this language has been tackled with reasonable success. Our heuristic approach is three-pronged, exploiting linguistic, lexical, as well as statistical techniques applied to a very large corpus, see Table 1.

1. *Linguistic filter*:
a) regular expressions
(adj. in positive form)* + noun (minus genitive form),
adj +”og/eller” + adj + noun,
noun + ”-” + ”og/eller” + noun

b) general vocabulary trap (cumulative stop-list of non-focal adj)

2. *Named Entity Recognizer*:
Evaluates output of linguistic filter according to specific criteria

3. *Statistical Significance (”Weirdness”) ratio*:
Text occurrence ratio of given text checked against occurrence ratio in major general language corpus:

Frq. of W in t
No. of tokens in t
divided by
Frq. of W in C
No. of tokens in C
(If occurrence of W in C is zero formula returns “Infinite”)

Table 1: Norwegian Term Candidate Extraction

Table 1 points to a fairly straightforward identification of complex NPs followed by cumulative list of stopwords which filters out the more obvious general language NPs, while an algorithm for named entity recognition (from an independent project called Nomen Nescio) prevents the elimination of desirable items. At this point the lexical properties of a massive and independently existing language resource is brought to bear on the task at hand: a list of word occurrence ratios drawn from Hofland’s cumulative corpus of general Norwegian newspaper text (currently at 600m words; see Hofland n.d.) is accessed and compared with the ratios generated from the new text, and a salience ratio is calculated for items exceeding a set threshold level (currently set at 60).
The samples from the FAS109 “weirdness” list given in Table 2 (after function words have been removed) are very small and carry no statistical significance in isolation; they do, however, provide interesting glimpses of the sort of lexical substance that turns up. We will now take a look at four “critical” points in the word list based on “weirdness” values:

Freq	Match	SLR	SL/GLR
160	deductible	0,0034	inf!
117	carryforward	0,0025	inf!
105	carryforwards	0,0022	inf!
46	pretax	0,0010	inf!
36	carryback	0,0008	inf!

Table 2a: Vocabulary at top of list

Table 2a are the items with the highest “weirdness” score, they are also absent from the general word list which constitutes the standard of comparison.

Freq	Match	SLR	SL/GLR
28	expenses	0,0006	60,4255
20	illustrates	0,0004	60,3073
3	amends	0,0001	60,0331
86	benefits	0,0018	58,2475
3	lenders	0,0001	55,0304
2	allocations	0,0000	55,0304

Table 2b: Vocabulary near cut-off point (60)

2b shows the items straddling the weirdness score of 60, understood as occurring in the text being examined about 60 times over their expected frequency in a general language text.

Freq	Match	SLR	SL/GLR
6	meet	0,0001	1,0136
3	aspects	0,0001	1,0067
1	enable	0,0000	1,0006
2	access	0,0000	0,9983
2	august	0,0000	0,9938
2	selection	0,0000	0,9938

Table 2c: Vocabulary near “general word”

Table 2c shows items with a score near 1, i.e. occurring with the same frequency in the test object as their expected frequency in a general language text, in other words, “everyday” items making no claim to specialist content.

Freq	Match	SLR	SL/GLR
1	great	0,0000	0,0322
1	again	0,0000	0,0266
1	come	0,0000	0,0263
1	own	0,0000	0,0252
1	know	0,0000	0,0188
1	way	0,0000	0,0180

Table 2d: Vocabulary at bottom of list

2d is the lower end of this list (of 1810 items), items occurring much less frequently here than they do in general language.

Again the material exhibited here can only point in the general direction of a conclusion, but it is clearly consistent with the view that the top of a “weirdness” list (unlike the top of a frequency list) gives a good indication of the characteristic terminology of the subdomain that the text is drawn from. Using 60 as a cutoff-point (chosen for practical reasons) seems quite arbitrary (see 2b) and should be examined critically, while 2c and 2d represent lexical units of no obvious interest to a terminologist or accounting professional. Quite coincidentally the average word length for the items listed in Table 2a, 2b, 2c and 2d is 10, 8, 6.3 and 4, respectively, a nice confirmation of the general (and superficial) observation that word length is related to degree of-specialization.

The end result of this filtering process is a list of “recommended” term candidates presented to the human expert for validation before final inclusion into the term bank. Critics may object that this human intervention constitutes a significant bottleneck which slows down processing and reduces the efficiency of the system by several orders of magnitude. The objection is valid – but such efficiency would come at a considerable cost, since fully automatic entry of term candidates into the term base would quickly clog the system with linguistic and conceptual debris which could only be eliminated through hugely complicated and expensive procedures (if at all).

It has been our philosophy that the superior capabilities of computer-based language processing are best exploited when harnessed to the professional knowledge of a human expert, such that each partner handles the tasks it is best equipped for handling: the computer performing speed-of--light indexing, harvesting and sophisticated pattern matching, the person exercising professional skills and competence in refining raw data into validated knowledge prior to the inclusion of the output in the term bank. In the suite of computer-assisted working procedures developed through the KB-N project we consider our semi-automatic bilingual terminology matching technique for term capture the jewel in the crown, and a brief description here will bring out the essential features and procedures of KB-N's semi-automatic method for term extraction from parallel texts.

The actual text under scrutiny is that of *International Accounting Standard (IAS) No. 12: Income Tax*, previously translated into its official and legally binding Norwegian equivalent entitled *Internasjonal regnskapsstandard (IAS) 12: Inntektsskatt*, for all intents and purposes a near perfect pair of parallel texts, each roughly 12,000 words long. In the (previous) text window the file IAS-12 (Norwegian version) has already been selected and the button for starting term extraction pressed. Depending on the size of the file the result will quickly turn up in the term window, which has three frames. The leftmost frame lists all term candidates identified by the algorithm i this text, the most salient ones at the top ("inf!" indicating a form not found in the reference corpus), and the terminologist/domain expert selects one Norwegian candidate.

This action produces the contents of the middle frame. The term candidate here appears at the top, accompanied by a set of possible English translation candidates for this term which already have been picked out from the corresponding location in the carefully aligned English text. At the same time the system presents the authentic contexts for each term candidate for both language variants. The terminologist/domain expert evaluates the alternatives, indicates preferences by ticking the corresponding boxes in the left-hand margin, and presses the "Register" button

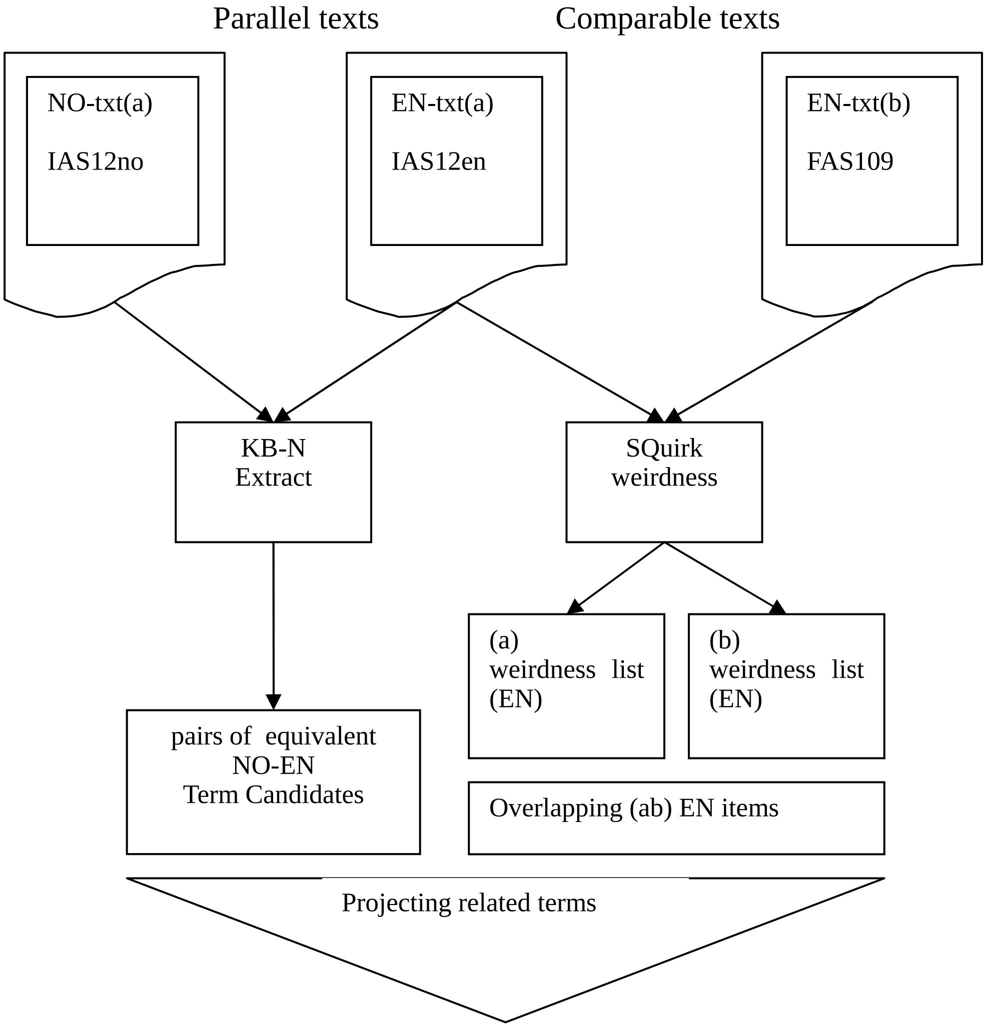


Figure 2: Structure of experimental setup

This action in turn produces the filled-in term record in the rightmost frame, which can be edited and supplemented before being entered in the term base by terminologist making the appropriate choice under "Edit". Getting from raw output to finished term record has taken six mouseclicks.

Such is the bare outline of KB-N's approach to term capture based on parallel text mining. Can this knowledge engineering platform be extended and exploited into the much larger domain of comparable text, with reasonable gains in productivity and quality of results?

3. Term Extraction from comparable text

To investigate these possibilities we selected *Financial Accounting Standard (FAS) No. 109*: *Accounting for Income Taxes*, which is the American equivalent to the British/European IAS-12. FAS109 is a much longer text (about 47,000 words) but in terms of genre, text type and target audience fully comparable with IAS-12.

As a consequence of the increasing globalization of financial markets and even more so the major accounting fraud scandals following in its wake (suffice it to mention the fall of Enron and the demise of Anderson), great efforts go into the harmonization ofgeneral accounting regulations. The convergence seen so far makes it a reasonable hypothesis that the terminology used in these documents will show significant overlap. Can our methodological approach corroborate this?.

Figure 2 above indicates the overall layout for testing the general hypothesis that texts sharing subdomain, genre, text type and audience orientation will have a significant terminological overlap.

The left-hand side represents the procedure for bilingual term capture from aligned parallel texts (described in section 2), while the right-hand side indicates the stages of attempted monolingual term capture utilizing salience statistics for identifying term candidates in comparable texts. The KB-N term extraction algorithms (see Table 1) has been specifically targeted at Norwegian and does not work for English, which can be handled by a wide variety of available extraction strategies and algorithms. One of the more interesting suites for computerized term capture is System Quirk, whose concept of "weirdness" was part of the initial inspiration for KB-N's project. In the case of System Quirk the lists representing general language vocabulary have been compiled on the basis of a commercial general language dictionary, which implies that the occurrence figures calculated for a given item will not be strictly comparable with KB-N data. In particular this is likely to affect the number of items designated as "inf!", since "non-occurrence" is less likely in a comprehensive dictionary than in a random text corpus however large.

4. Preliminary results

We have earlier had a brief look at some critical statistical points in one of the lists will serve to illustrate how the "weirdness" figures are reflected in defining lexical units along the special/general continuum; table 2 has four snippets from the FAS109 "weirdness" list.

The samples from the FAS109 "weirdness" list given in Table 2 (after function words have been removed) are very small and carry no statistical significance in isolation; they do, however, provide interesting glimpses of the sort of lexical substance that turns up. Table 2a are the items with the highest "weirdness" score, they are also absent from the general word list which constitutes the standard of comparison. 2b shows the items straddling the weirdness score of 60, understood as occurring in the text being examined about 60 times over their expected frequency in a general language text. Table 2c shows items with a score near 1, i.e. occurring with the same frequency in the test object as their expected frequency in a general language text, in other words, "everyday" items making no claim to specialist content. 2d is the lower end of this list (of 1810 items), items occurring much less frequently here than they do in general language.

Again the material exhibited here can only point in the general direction of a conclusion, but it is clearly consistent with the view that the top of a "weirdness" list (unlike the top of a frequency list) gives a good indication of the characteristic terminology of the subdomain that the text is drawn from. Using 60 as a cutoff-point (chosen for practical reasons) seems quite arbitrary (see 2b) and should be examined critically, while 2c and 2d represent lexical units of no obvious interest to a terminologist or accounting professional. Quite coincidentally the average word length for the items listed in Table 2a, 2b, 2c and 2d is 10, 8, 6.3 and 4, respectively, a nice confirmation of the general (and superficial) observation that "specialist words" tend to be long, everyday words tend to be short.

5. General conclusion

The KB-N KnowledgeBank in its current phase is somewhat lacking in volume of captured and coded text from the relevant domains, since we have given quality priority over quantity. Consideration of text for inclusion in the corpus requires careful scrutiny of stylistic quality, lexical representativity as well as conceptual substance, and, in the case of parallel texts, the professional quality and equivalence of the target text must be assessed.

In addition a serious hindrance to the efficient establishment of a domain-specific digital archive has emerged from the legal agencies guarding the intellectual property rights of major publishing houses. Communication regarding permissions has proved extremely laborious, not just because some copyright holders refuse to respond to our formal requests but because those that do, often fail to comprehend the nature of our interest in their texts. For many languages this problem constitutes a major obstacle in most efforts to establish representative public language bank, archives of suitable and necessary linguistic resources for documentation and research.

The volume of term records currently held in the KB-N KnowledgeBank (about 8,500), on the other hand, is quite respectable in four subdomains of major importance (Accounting being one of these, with about 1,200 records), a situation attributable not only to the continuous development and refinement of automatic term extraction, but in large measure also to the efficiency of the custom made tools for human-machine communication which have been created to speed up the work flow in the KB-N project (cf. section 2).

KB-N is designed as a web-enabled resource available for systematic terminology look-up and usage documentation. Search for a given term into the term bank will retrieve essential registration data relating to meaning, cross-language equivalence or usage etc. For any given special domain included KB-N will constitute a conceptual/terminological clearing house, a precise control system for the conceptual underpinnings of the domain knowledge and their terminological and textual manifestations, in one or more languages.

We consider automatic term extraction the computationally most interesting achievement of the KB-N project so far, in exploiting the empirical value of linguistic resources (acquired for quite different purposes) in developing precise algorithms for automatically generated term candidate lists. This operation constitutes a significant link between the text bank and the term bank and exploits human-machine interaction to combine text-embedded domain knowledge with human expertise in a form which can be utilized in e.g. e-learning, machine translation, human translation, and knowledge management.

Nevertheless, the lack of major headway being made in applying automatic term extraction beyond parallel corpora, and the so far largely untapped resources encapsulated in huge repositories of comparable text, constitute a major challenge for the natural language processing community. The tentative results hinted at in this preliminary version of the paper (to be expanded and substantiated in the final version) are taken as encouragements in our continued efforts to refine the use of statistical salience in achieving significant term capture on the basis of comparable corpora.

KB-N has been developed in the context of KUNSTI funded by the Norwegian Research Council (Brekke 2004).

References:

Ahmad K. & M. Rogers. (2001). Corpus linguistics and terminology extraction. In: Ahmad K, Wright SE, Rogers M, Budin G, editors. Handbook of terminology management, vol. 2. Philadelphia: Benjamins;. p. 725-60.

Brekke, M (2004). ”KB-N: Computerized extraction, representation and dissemination of special terminology”. Costa, R. et al. (eds). Workshop on Computational and Computer-assisted Terminology, IV LREC 2004, Lisbon.

Brekke, M et al. (2006). Automatic Term Extraction from Knowledge Bank of Economics, V LREC 2006, Genoa, pp. 1912-1915.

Cabré, M.T. et al. (2004). The GENOMA-KB project: towards the integration of concepts, terms, textual corpora and entities. Lisbon: IV LREC 2004. p. 87-90.

Castellvi M.T.C., Bagot R.E, J. Vivaldi Palatresi (2001). Automatic term detection: a review of current systems. In: Bourigault D. et al. (eds). Recent advances in computational terminology. Philadelphia: Benjamins;. p. 53-87.

Hagen K., J. Bondi Johannessen & A. Nøklestad (2000). "A Constraint-Based Tagger for Norwegian". In Lindberg, C.-E. og S. Nordahl Lund (eds). 17th Scandinavian Conference of Linguistics, Odense Working Papers in Language and Communication, No. 19, vol I.

Hofland, K. (1996). "A program for aligning English and Norwegian sentences." In S. Hockey et al. (eds), Research in Humanities Computing. Oxford: OUP. 165-178.

Hofland, K. (n.d.) Norwegian Newspaper Corpus at http://helmer.aksis.uib.no/aviskorpus/..

Ide, N. & J. Veronis (1998). Word Sense Disambiguation: the state of the art. Computational Linguistics, 24(1), pp. 1-41.

Kristiansen, M. (2005). Disciplinary autonomy and concept relations in electronic knowledge bases. A theoretical approach to KB-N - a knowledge base for economic-administrative domains.. SYNAPS - Fagspråk. Kommunikasjon. Kulturkunnskap, 17(2005). Bergen: NHH, pp.1-7.

Maia, B & Sarmento, L. (2006). The Corpografo – an Experiment in Designing a Research and Study Environment for Comparable Corpora Compilation and Terminology Extraction. V LREC 2006, Genova.

Sarmento, L. et al. (2006). Corpografo V3. From Terminological Aid to Semi-automatic Knowledge Engineering. V LREC 2006, Genova.

System Quirk:
http://www.computing.surrey.ac.uk/AI/SystemQ/

Øvsthus, K. et al. (2005). Developing automatic term extraction. Automatic domain specific term extraction for Norwegian. Proceedings of Terminology and Knowledge Engineering, Copenhagen: CBS

Using a comparable corpus to investigate lexical patterning in English abstracts written by non-native speakers

Carmen Dayrell[1] and **Sandra Aluísio**[2]
University of São Paulo (USP) - Brazil
[1] Department of English
Rua Prof. Luciano Gualberto, 403/sala 14 - Cidade Universitária CEP: 05508-900 - São Paulo
[2] Centre of Computational Linguistics (NILC)/ Department of Computer Sciences
Caixa Postal: 668 - CEP: 13560-970 - São Carlos
E-mails: dayrellc@gmail.com, sandra@icmc.usp.br

Abstract

This study is set in a broad context of development of courses and computational tools to aid Brazilian graduate students in writing scientific papers in English. The main focus is on experimental research papers from the disciplines of physics, pharmaceutical and computer sciences. One of our primary objectives is to give students feedback and raise their awareness of the most typical lexical patterns used by their academic discourse community while at the same time draw their attention to the various available alternatives. Errors related to lexical use are by far the most frequent errors made by Brazilian graduate students when writing academic English. The aim of this paper is to carry-out a corpus-based study to investigate potential differences in the collocational patterns of *work* in abstracts written by Brazilian graduate students as opposed to abstracts collected from published papers of the same discipline. Relevant differences were found between the two subcorpora. The results were validated by examining the identified lexical patterns in a reference corpus of English abstracts. We also identified various items other than *work* which may be used to refer to the study described in the abstract as well as other lexical variations within the lexical patterns analysed.

1. Introduction

Scientific writing poses considerable challenges for non-native speakers of English. In addition to complying with the conventions and norms adopted by their academic discourse community, they also have to deal with the various difficulties involved in the complex process of writing in a foreign language. These problems are even more acute if the writer is an inexperienced researcher and he/she does not have full command of English grammar and usage at the sentence level.

Within the specific context of English Language Teaching (ELT), much effort has been spent on producing material to aid non-native speakers in overcoming the various problems which they may face when writing research papers. Swales (1990, 2004), Swales & Feak (2000), Weissberg & Buker (1990) are good examples of pedagogically useful studies which focus on the description of phrases and lexical patterns which are frequently used in academic discourse. It is also worth mentioning that several websites[1] are now available to provide users with practical guidelines when producing academic English.

Another important contribution is offered by studies based on learner corpora of academic English (among others, Thompson, 2001, 2006; Lee and Swales, 2006; Hyland, 2008a, 2008b). These studies opened up new perspectives and provided useful insights which enhanced our understanding of underlying regularities in the language produced by students.

Benefits were also gained with computer-aided writing tools such as the *English Grammar and Spelling Software - Advanced Solutions for Your Writing*[2] which, in addition to a grammar and spell checker, also includes a dictionary, a thesaurus and a list of the most relevant collocates. Further achievements came from computational tools which take a step further and provide users with extracts from authentic research papers retrieved from a reference corpus of the discipline in question. This is the case of two writing tools developed at the University of Sao Paulo, namely AMADEUS and *Scipo-Farmácia*[3]. The former focuses on the disciplines of physics and computer science (Aluísio & Oliveira, 1995; Aluísio & Gantenbein, 1997; Aluísio *et al.*, 2001) while the latter deals with pharmaceutical sciences (Aluísio *et al.*, 2005; Genoves Jr. *et al.*, 2007). Similarly, Narita *et al.* (2003) and Anthony (2006) focus on English texts by Japanese speakers and developed computational tools to help user structure the text and produce adequate sentences in English.

Needless to say, various courses on English for Academic Purposes are offered worldwide. Most relevant to this study are the courses on academic writing offered annually by the University of São Paulo to graduate students. As we shall see in the next section, these courses have provided the data which is analysed in this paper. For the time being, what is important to mention is that the present study is set in a broad context, which includes a joined effort by various departments at the University of São Paulo to develop courses and computational tools to

[1] To mention but a few: *A Guide to Grammar and Writing*: http://grammar.ccc.commnet.edu/grammar/, *A Guide for Writing Research Papers Based on Modern Language Association (MLA) Documentation*: http://www.ccc.commnet.edu/mla/index.shtml, *Common Errors in English*: http://www.wsu.edu/~brians/errors/

[2] http://www.whitesmoke.com/

[3] *Scipo-Farmácia* can be accessed at http://www.nilc.icmc.usp.br/scipo-farmacia/

aid Brazilian graduate students in writing scientific papers in English. The main focus is on experimental research papers from the disciplines of physics, pharmaceutical and computer sciences. Our long-term objective is two-fold: to improve course materials and resources for academic English and to provide computer-aided writing tools with linguistic knowledge so as to enable the automatic identification and correction of errors at the lexical, syntactical and rhetorical levels.

This paper is part of a larger project which aims to investigate errors made by Brazilian graduate students in academic writing. Our primary aim is to carry out a corpus-based study on the collocational behaviour of lexical items which frequently pose a challenge for Brazilian writers. The focus is on errors related to lexical use which is by far the most frequent error made by Brazilian students when writing scientific papers in English (Genoves Jr. *et al.*, 2007). These refer to misuse of a word to express a particular meaning. They may refer to direct translations of a Portuguese item into a false cognate in English (*pretend* for *intend*) or errors made in a common idiom (*as* for *such as*) or common collocation (*do contributions* for *make contributions*). There are other cases which are related to naturalness, that is, the writer's lexical choice is not most frequently used in that particular context, although its semantic meaning is fairly appropriate.

This paper focuses on the lexical patterning of the item *work*. A pilot study is carried out to investigate potential differences in the collocational patterns of *work* in abstracts written by Brazilian graduate students in comparison with abstracts collected from published papers of the same discipline. This idea relies on Sinclair's (1991:6,108 and 2003:3) suggestion that words do not occur randomly in a text but are instead closely associated with their surrounding context. According to Sinclair's (1991:6), the use of a given lexical item is related to specific lexical and grammatical patterns. Thus, by identifying differences in the lexico-grammatical patterning of abstracts written by students and published abstracts, we hope to be able to raise students' awareness of their most frequent errors as well as to draw their attention to the use of chunks which are regularly used within their academic discourse community. The results are validated by examining the identified lexical patterns in a reference corpus of English abstracts which were collected from papers published in reference journals of the disciplines under analysis. All procedures described below are carried out by means of the software package *WordSmith Tools*, version 4.0 (Scott, 2004).

2. The Comparable Corpus of English Abstracts (CCEA)

The data analysed in this paper is drawn from a monolingual comparable corpus of English abstracts which consists of two separate subcorpora: one made up of abstracts written by Brazilian graduate students and the other consisting of abstracts from published papers.

The subcorpus of English abstracts written by students (hereafter EA-STS) contains 84 abstracts which were collected in four courses on academic writing offered to graduate students from the disciplines of pharmaceutical sciences (20 abstracts), biology/genetics (11), physics (27) and computer science (26) at two universities in Brazil between 2004 and 2006. Here, we examine the first version of the abstracts, that is to say, abstracts handed in before the course started.

The subcorpus of English abstracts extracted from published papers (hereafter EA-PUB) was designed to match the specifications of the EA-STS subcorpus so that the two collections could be made comparable. Thus, EA-PUB also includes 84 abstracts from the same four fields of research, paying special attention to the number of abstracts in each. The abstracts were randomly selected from of various academic journals, using the WebBootCat tool[4] as a starting point to select websites to be consulted. WebBootCat is a tool designed to help users to quickly produce corpora from any domain or subject (Baroni *et al.*, 2006). In an attempt to diversify the selection of journals as much as possible, no more than 3 abstracts were selected from each journal. In terms of number of words (tokens), the EA-STS and the EA-PUB subcorpora contain 18,004 and 21,061 words respectively.

An important methodological point to make here is that by published abstracts we do not mean that they are all written by native speakers of English. What is assumed is that they are of acceptable quality because they have been published by recognised bodies of a given discipline. Thus, published abstracts are presumably more likely to comply with the pre-established conventions adopted by the discourse community in question. Another difference between the two subcorpora is that most abstracts included in the EA-PUB come from papers by more than one author. This may mean that they has been more carefully revised and edited and hence less likely to contain deviations.

3. Data Analysis

This paper focuses on the lexical item *work*. The main rationale behind this choice is the fact that *work* is one of the most frequent lexical items in the EA-STS (9th position), with 38 instances, and it occurs only nine times in the EA-PUB. This difference in the number of instances may be interpreted an indication that the item behaves differently in the two subcorpora.

Work is used as a noun in the vast majority of instances of both subcorpora (89%): 34 occurrences in the EA-STS and eight occurrences in the EA-PUB. These are therefore the focus of the study and all instances of *work* as a verb or an adjective are discarded.

In both collections, *work* tends to be part of recurring lexico-grammatical patterns and refer to the study described in the abstract. Two instances in the EA-STS and three instances in the EA-PUB are exceptions. In these cases, *work* either refers to someone else's work or it

[4] Further details of this web service can be found at http://www.sketchengine.co.uk/auth/wbc/mycorp.cgi.

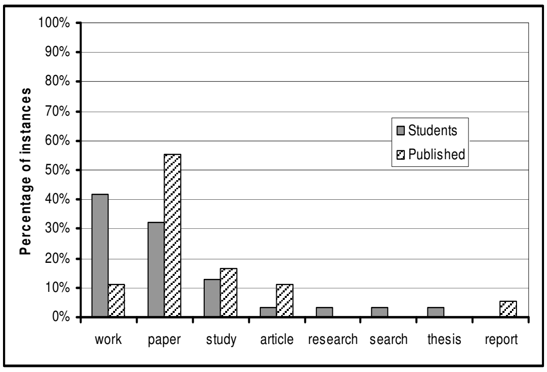
is related to the effort required to do a given task. Table 1 summarises these findings. VERB refers to any verb which appeared in the first position on the right of the *work* such as *aims*, *presents*, *shows*, etc. In pattern (iii), NOUN is used to indicate the various nouns which appeared in that particular position such as *aim* and *objective*. Optional items are indicated between brackets.

	Patterns	Students' Abstracts	Published Abstracts
i.	in this/my/ the present work	13	2
ii.	This/Our work VERB	13	2
iii.	the (main) NOUN of this/ the present work	6	1
iv.	*work* does not refer to the study described in the abstract	2	3
TOTAL		**34**	**8**

Table 1: Lexical patterns identified in the CCEA

Taking into consideration the low number of times *work* occurs in the EA-PUB subcorpus, our next step is to identify word(s) other than *work* which occur within these specific recurring lexical patterns in published abstracts. Here, we are interested in items which may also be used to refer to the study in question, even though they may not be exactly synonymous. Once these items are identified, we go back to the EA-STS subcorpus and examine whether these same items are used by students in these specific contexts.

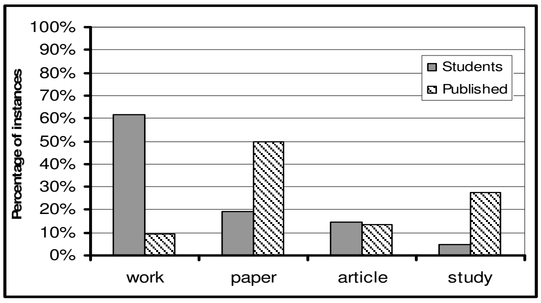
For instance, pattern (i) refers to the sequences *in this/the present/my* ***. Eight different lexical items are used: *work*, *paper*, *study*, *article*, *report*, *thesis*, *research* and *search*. *Search* occurs once in the EA-STS subcorpus; it is most probably to be a mistranslation and meant to be *research*. Also, the pattern *in my work* occurs once in the EA-STS subcorpus. Pattern (i) occurs 31 times in the EA-STS and 18 times in the EA-PUB.



Graph 1: Pattern (i) - *in this/the present/my* ***

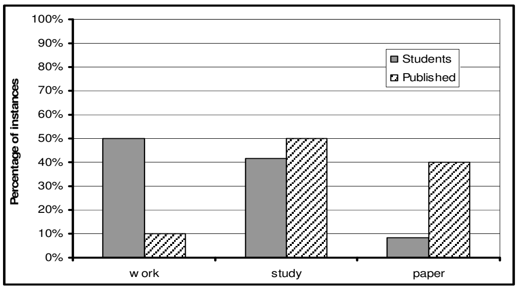
Work is the most frequent item in the EA-STS (42%) and *paper* is the most frequent in the EA-PUB (56%). However, it is interesting to notice that *paper* appears in as much as 32% of instances in the EA-STS whereas *work* occurs only twice (11%) in the EA-PUB. *Study* is the only item which is used in similar percentage in the two subcorpora. *Article* appears in a higher proportion in the EA-STS than in the EA-PUB (11% compared to 3%). *Thesis*, *research* and *search* occur once in the EA-STS and *report* appears once in the EA-PUB.

Pattern (ii) refers to the sequences *this/the present/the/our* *** VERB in the beginning a clause. In addition to *work*, three other items are used in this context: *paper*, *article* and *study*, yielding chunks such as *this paper examines* and *this study presents*. The pattern occurs 21 times in the EA-STS and 22 times in the EA-PUB. In the EA-STS, *work* accounts for the vast majority of instances (62%) and *paper* is the second most frequent item with 19% of occurrences. By contrast, *paper* occurs in 50% of the cases in the EA-PUB whereas *work* is only used twice (9%). *Study* is also more frequent in the EA-PUB (27%) in comparison with the EA-STS (5%). *Article* is the only item which is used in similar percentage in both subcorpora (14%).



Graph 2: Pattern (ii) – *this/the present/our* *** VERB

By examining the lexical variations within pattern (iii) – *the* (ADJ) NOUN *of this/the present* *** –, we find that, in addition to *work*, *paper* and *study* are also used.



Graph 3: Pattern (iii): *the* (ADJ) NOUN *of this/the present* ***

As for the position of NOUN, for the overwhelming majority of instances in both subcorpora (83% in the EA-STS and 70% in the EA-PUB), it is related to 'aim' (*aim*, *objective*(s) and *purpose*). We also find *result* and *contributions(s)* in the EA-PUB and *expectancy* in the

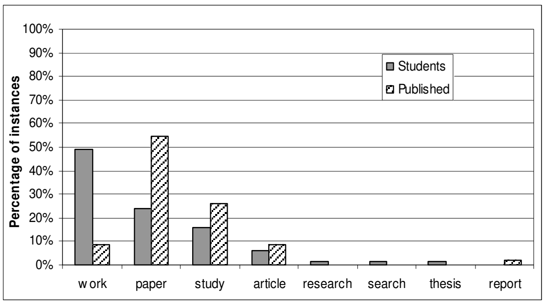
EA-STS, which is most probably a mistranslation of the Portuguese item *expectativa* (expectation). One instance in the EA-STS shows *object of this work*. Adjectives (ADJ) such as *main*, *primary* and *key* may occur before the noun. *Study* shows a high frequency of occurrence in both subcorpora: 42% of instances in the EA-STS and 50% of instances in the EA-PUB. However, w*ork* accounts for 50% of instances in the EA-STS and paper represents 40% of instances in the EA-PUB.

Taking into consideration the overall number of recurring patterns in the corpus, no striking difference is found between the percentages of each across the two subcorpora (Table 2). Patterns (i) and (ii) are used in fairly similar proportion in both collections; pattern (ii) is slightly more frequent in the EA-STS.

	Patterns	**Students' Abstracts**	**Published Abstracts**
i	*in this/the present/my ****	31 (48%)	18 (36%)
ii	*this/the present/the/our* *** VERB	21 (33%)	22 (44%)
iii	*the* (ADJ) NOUN *of this/the present ****	12 (19%)	10 (20%)
	TOTAL	**64 (100%)**	**50 (100%)**

Table 2: Lexical patterns identified in the CCEA

By examining the overall number of times each lexical item is used in each subcorpus, we find that *work* occurs in 51% of the instances in the EA-STS compared to 11% in the EA-PUB subcorpus. *Paper* is the most frequent item in the EA-PUB, representing 54% of all occurrences. It is the second most frequent item in the EA-STS (24%). Study shows a higher percentage of instances in the EA-PUB component (30%) in comparison with the EA-STS subcorpus (16%).



Graph 4: Overall number of times each lexical item is used in the EA-STS and EA-PUB subcorpora

4. Discussion

The figures above indicate a clear tendency of students to use the item *work* when referring to the study described in the abstract whereas published abstracts show a marked preference for the word *paper*. However, one cannot afford to ignore that the EA-PUB subcorpus is very limited in size and hence it does not allow the researchers to make generalizations on the collocational behaviour of these two items in scientific abstracts.

An important point to make here is that our findings are very much consistent with the results revealed by Orasan (2001) on the use of the word *paper* in scientific abstracts. Orasan (ibid.) examines 917 abstracts (146,489 words) from the disciplines of artificial intelligence, computer science, biology, linguistics, chemistry and anthropology. *Paper* is frequently used as the subject of verbs such as *present* (62 times), *describe* (50), *be* (45), *introduce* (15) and tends to yield patterns like *this paper presents* (44) or *this paper describes* (39). Another clear pattern is the sequence *in this paper*, which occurs143 times in the corpus. Taking into consideration the items used to refer to the study described in the abstract, Orasan (ibid.) shows that *paper* is the most frequent option (53% of the instances) in relation to other items such as *study*, *research* and *work* (Table 3).

Item	**Number of instances**	**% of instances**
paper	499	53%
study	170	18%
research	154	17%
work	111	12%
TOTAL	**934**	**100%**

Table 3: Number and percentage of instances of the items used to refer to the study in question (Orasan, 2001)

Similar to our study, Orasan (2001) also concludes that the high frequency of these patterns in abstracts is not by chance but instead that it is a strong indication that they are frequently used in this specific context of abstracts.

However, it is important to bear in mind that Orasan (ibid.) uses a corpus which includes abstracts from various disciplines and does not focus on the specific recurring lexical patterns that we are interested here. Thus, in order to validate our findings and be able to obtain a clearer picture of how *work* and *paper* are used by the academic discourse communities in question, we necessarily need access to a reference corpus of abstracts which matches the specifications of the texts included in the CCEA. Here, we use a corpus consisting of 723 scientific abstracts from the disciplines of physics (369) and pharmaceutical sciences (354) (Genoves et al., 2007). All abstracts were collected from reference journals of these two disciplines such as *Physical Review Letters (A-D), Science, Nature* and *Biotechnology Progress*. The overall size of the corpus is 115,913 words (tokens).

We first focus on the three recurring patterns discussed above and look at number of instances in which *work* and *paper* are used. The analysis is extended to include other items which may also be used to refer to the study described in the abstract.

Unlike the results discussed above, we find that, in the reference corpus, *study* is by far the most frequent item, with 51% of the instances (Table 4). *Paper* is the second most frequent item, accounting for 21% of the occurrences. *Work* is used in 17% of the instances. The reference corpus also shows that the item *review* can also be used in these specific contexts; however, it does not occur in the CCEA.

	Item	**Number of instances**	**% of instances**
1.	study	76	51%
2.	paper	31	21%
3.	work	26	17%
4.	review	8	5%
5.	article	4	3%
6.	report	3	2%
7.	research	1	1%
	TOTAL	**149**	**100%**

Table 4: Number and percentage of instances for the lexical items in the reference corpus

The reference corpus also reveals several variations of the patterns under analysis. For instance, pattern (i) is the most frequent pattern in the reference corpus with 101 instances. In addition to *in this* and *in the present*, which account for the vast majority if instances – 85 and 10 respectively, we also find *in our* (3), *in the current* (2) and *in the* (1). Seven lexical items are used to refer to the study in question: *study*, *paper, work*, *article*, *review, report* and *research* (Table 5). *Study* is the most frequent item, accounting for 53% of the instances. *Paper* (22%) is slightly more frequent than *work* (16%).

	Item	**Number of instances**	**% of instances**
1.	study	54	53%
2.	paper	22	22%
3.	work	16	16%
4.	article	4	4%
5.	review	3	3%
6.	report	1	1%
7.	research	1	1%
TOTAL		**101**	**100%**

Table 5: Items within pattern (i) in the reference corpus

Pattern (ii) – *This* *** VERB – occurs 38 times in the corpus (Table 6). In addition to *this*, which appears in 23 instances (61%), the following appears before our search item, mentioned in order of frequency of occurrence: *the present* (5), *our* (5), *the current* (3) and *the performed* (1). One instance shows *the study presented here* VERB. All these occurrences have been considered as variants of pattern (ii). Five different lexical items are used to refer to the study in question, in order of frequency: *study*, *work*, *paper, review* and *report*. *Study* is again the most frequent item in pattern (ii), accounting for 40% of the instances.

	Item	**Number of instances**	**% of instances**
1.	study	15	40%
2.	work	9	24%
3.	paper	7	18%
4.	review	5	13%
5.	report	2	5%
TOTAL		**38**	**100%**

Table 6: Items within pattern (ii) in the reference corpus

Pattern (iii) occurs 10 times in the corpus (Table 7). Only three items appears within this pattern. Here again. *study* is the most frequent item, representing 70% of the instances (Table 6).

	Item	**Number of instances**	**% of instances**
1.	study	7	70%
2.	paper	2	20%
3.	work	1	10%
TOTAL		**10**	**100%**

Table 7: Items within pattern (iii) in the reference corpus

As can be seen, in the reference corpus, *study* is the most frequent item within the three patterns. *Paper* is the second most frequent item in patterns (i) and (iii) and *work* comes third in the frequency ranking. For pattern (ii), it is interesting to notice that *work* is more frequent than *paper* (24% compared to 18%). By contrast, in the EA-STS, *work* is the most frequent item in the three patterns whereas *study* shows a high percentage of instances for pattern (iii) only. For patterns (i) and (ii), *paper* is the second most frequent item.

In terms of percentage of instances for each pattern, we notice that the reference corpus shows a strong preference for pattern (i) (68%, Table 8). This same tendency is seen in EA-STS abstracts, although not as marked (Table 2).

	Patterns	**Reference Corpus**
i	*in this/the present/current/ our/the ****	101 (68%)
ii	*this/the present/current /performed/ our **** VERB	38 (25%)
iii	*the* (ADJ) NOUN *of this/the present ****	7 (10%)
TOTAL		**149 (100%)**

Table 8: Number and percentage of instances for each pattern in the reference corpus

5. Final Remarks

This paper has examined the collocational behaviour of item *work* in abstracts written by Brazilian graduate students as opposed to abstracts collected from published papers of the same discipline. Relevant differences were found between the two subcorpora. Taking into consideration the same specific contexts, the former displayed a strong preference for the item *work* whereas the latter showed a clear tendency to use *paper*. The results were validated by examining the identified lexical patterns in a reference corpus of English abstracts. *Study* was by far the most frequent item in the reference corpus. *Paper* came second, showing a slightly higher proportion than *work*.

Given that our long-term objective is to provide support to the development of course materials and computer-aided writing tools to aid Brazilian graduate students in writing scientific papers in English, this study took a step further and searched for items other than *work*, *paper* and *study* which may also be used to refer to the study described in the abstract. We also looked at instances which could be regarded as variants of the identified lexical patterns.

Thus, in addition to contrasting collocational patterns of *work* in abstracts written by students and published abstracts, this study has identified various lexical items used in specific lexical patterns as well as described their usage according to frequency. These findings can be incorporated into course materials and computational resources. This would enable us to raise students' awareness of the most typical lexical patterns used by their academic discourse community while, at the same time, it allows us to draw students' attention to the various other alternatives available to them when writing academic English.

6. Acknowledgements

We would like to thank our Brazilian sponsors FAPESP and CNPq (Brazil) for providing the financial support needed in this research.

7. References

Aluisio, S.M., Barcelos, I., Sampaio, J., Oliveira Jr., O. (2001). How to Learn the Many Unwritten 'Rules of the Game' of the Academic Discourse: A Hybrid Approach Based on Critiques and Cases. In *Proceedings of the IEEE International Conference on Advanced Learning Technologies*. Madison, Wisconsin. Los Alamitos, CA: *IEEE Computer Society*, 1, pp. 257--260.

Aluísio, S.M., Gantenbein, R.E. (1997). Educational Tools for Writing Scientific Papers. In *VIII Simpósio Brasileiro de Informática na Educação*, ITA, pp. 239--253.

Aluísio, S.M, Oliveira Jr., O. (1995). A Case-Based Approach for Developing Writing Tools Aimed at Non-native English Users. In *Lecture Notes in Artificial Intelligence* 1010, pp. 121--132.

Aluísio, S.M, Schuster, E., Feltrim, V.D., Pessoa Jr., A., Oliveira Jr., O.N. (2005). Evaluating Scientific Abstracts with a Genre-specific Rubric. In *The 12th International Conference on Artificial Intelligence in Education - AIED*, 18-22 July, Amsterdam.

Anthony, L. (2006). Developing a Freeware, Multiplatform Corpus Analysis Toolkit for the Technical Writing Classroom. In *IEEE Transactions on Professional Communication*, Vol. 49, No. 3, pp. 275--286.

Baroni, M., Kilgarriff, A., Pomikálek, J., Rychlý, P. (2006). WebBootCaT: instant domain-specific corpora to support human translators. In *Proceedings of EAMT 2006*. Oslo, pp. 247--252.

Genoves Jr., L., Lizotte, R., Schuster, E., Dayrell, C., Aluísio, S. (2007). A two-tiered approach to detecting English article usage: an application in scientific paper writing tools. In *Proceedings of the International Conference RANLP´2007*. Borovetz, Bulgaria, Sep 26, 2007, pp. 225--239.

Hyland, K. (2008a). Academic Clusters: Text Patterning in Published and Postgraduate Writing. In *International Journal of Applied Linguistics*, 18(1), pp. 41--61.

______ (2008b). As Can Be Seen: Lexical Bundles and Disciplinary Variation. In *English for Specific Purposes*, 27, pp. 4--21.

Lee, D. and Swales, J. (2006). A Corpus-Based EAP course for NNS doctoral students: Moving from Available Specialized Corpora to Self-compiled Corpora. In *English for Specific Purposes*, 25, pp. 56--75.

Narita, M., Kurokawa K., Utsuro, T. (2003). Case Study on the Development of a Computer- Based Support Tool for Assisting Japanese Software Engineers with their English Writing Needs. In *IEEE Transactions on Professional Communication*, Vol. 49 (3), pp. 194--209.

Orasan, C. (2001). Patterns in scientific abstracts. In *Proceedings of Corpus Linguistics 2001 Conference*. Lancaster University, Lancaster, UK, pp. 433--443

Scott, M. (2004). *WordSmith Tools* version 4. Oxford: Oxford University Press.

Sinclair, J. (1991). *Corpus Concordance and Collocation*. Oxford: Oxford University Press.

______ (2003) *Reading Concordances*, London: Pearson Education Ltd., Longman.

Swales, J.M. (1990). *Genre Analysis: English in Academic and Research Settings*. Cambridge: Cambridge University Press.

______ (2004) *Research Genres. Explorations and Applications*. Cambridge: Cambridge University Press.

Swales, J.M., Feak, C.B. (2000). *English in Today's Research World: A Writing Guide*. Michigan: The University of Michigan Press.

Thompson, P. (2001). Looking at Citations: Using Corpora in English for Academic Purposes. In *Language Learning and Technology* 5(3), pp. 91--105.

______ (2006). Assessing the Contribution of Corpora to EAP Practice. In Z. Kantaridou, I. Papadopoulou & I. Mahili (Eds.), *Motivation in Learning Language for Specific and Academic Purposes*. Macedonia: University of Macedonia [CD-ROM].

Weissberg, R., Buker, S. (1990). Writing up Research: Experimental Research Report Writing for Students of English. Englewood Cliffs (NJ): Prentice Hall Regents.

A Comparative Approach to Diachronic Comparable Corpus Investigation

Meng Ji
Humanities, Imperial College London
85 Sterling Place, South Ealing, London W5 4RB
E-mail: m.ji05@imperial.ac.uk

Abstract

The present paper sets out to investigate the evolving nature of Mandarin Chinese through a comparative study of the distribution of Chinese idioms in two large-scale modern Chinese monolingual corpora, i.e. Lancaster Corpus of Mandarin Chinese (LCMC) (1990s) and the UCLA Chinese Corpus (early 2000s). The very interesting results presented here show that idioms, which represent the most conventionalized part of Chinese, seem to have undergone a considerable change in the last decade of the twentieth century, for when compared to the LCMC, many of the text types or genres have witnessed a noticeable decrease in the occurrence of idioms in the UCLA corpus.

The present paper sets out to investigate the evolving nature of Mandarin Chinese through a comparative study of the distribution of Chinese idioms in two large-scale modern Chinese monolingual corpora, i.e. Lancaster Corpus of Mandarin Chinese, also known as LCMC (1990s) and the UCLA Chinese Corpus (early 2000s). The two corpora have been constructed by following the same sampling framework as that of the Brown or LOB corpus, and are thus essentially comparable. The very interesting results presented here show that idioms, which represent the most conventionalized part of Chinese, seem to have undergone a considerable change in the last decade of the twentieth century, for when compared to the LCMC, many of the text types or genres have witnessed a noticeable decrease in the occurrence of idioms in the UCLA corpus.

As two widely distributed monolingual corpora of modern Chinese, both LCMC and UCLA Corpus have been built to address the increasing need for large-scale comparable corpora to do contrastive language studies, usually in combination with purposely-built specific corpora of much smaller size. However, the present paper hopes to show that a quantitative study of the two diachronically successive corpora, which seems to have been less discussed in the past, may also bring us valuable insights into the changing nature of Chinese, as being focused upon at a particular historical point. The linguistic phenomenon under investigation is the distribution of Chinese idioms, as a central lexicographical component of the language, among the various text types included in the two corpora, which add up to some fifteen categories.

Code	Text Type	Raw Frequency (LCMC)	Raw Frequency (UCLA)
AD	Adventure/Martial Arts Fiction	338	300
ES	Essays and Biographies	931	363
GF	General Fiction	290	223
HU	Humor	108	76
MY	Mystery/ Detective Fiction	266	493
NED	News Editorials	369	111
NREP	News Reportage	484	236
NREV	News Reviews	249	117
PL	Popular Lore	501	171
RE	Religion	112	7
REP	Reports/Official Documents	108	36
RO	Romantic Fiction	378	263
SC	Science (Academic Prose)	344	51
SF	Science Fiction	45	255
SK	Skills/Trades/Hobbies	244	9
Total	Total	4767	2711

Table I Distribution of idioms in LCMC versus UCLA Chinese corpus[1]

Table I exhibits the raw frequency of idioms in different text genres, which is an initial comparison of the two monolingual corpora. However, it should be noted that the

[1] Last access to LCMC and UCLA corpus was on February 8, 2008

first impression that we may have of such comparison may turn out to be misleading, due to the different size of the two corpora: while the LCMC contains one million tokens[2], the current version of the UCLA corpus holds 687, 634 running words in its collection[3]. As a result, it would be rather difficult to tell from the outset whether the two corpora genuinely differ from each other with regards to the distribution of idioms across the fifteen text types. To overcome this technical problem, the statistical procedure Pearson's moment-product correlation test has been employed, which yields the important statistical result shown in Table II.

Pearson's correlation test is widely used in corpus linguistics to test the strength of association between different corpus texts. It does not assume any causal relationship between the variables under test and may only deal with continuous data. It expresses the strength of correlation numerically through the correlation coefficient, R, which varies from menus one to one as the maximum values at two extremes. Table II shows that firstly, the mean frequency of idioms in the LCMC is as high as 317.8, which is almost twice that of the UCLA corpus. The computed coefficient of the correlation model is approximately 0.435, whose further interpretation requires the consultation of the index of the Pearson's coefficient critical values set at different significant levels[4].

STATISTIC	Variable X (LCMC)	Variable Y (UCLA)
Mean	317.8	180.73
Biased Variance	44198.69	18634.86
Biased Standard Deviation	210.23	136.51
Covariance	13364.37	
Correlation	0.43	
Determination	0.19	
Degrees of Freedom	13	
Number of Observations	15	
Critical value for Pearson's test (two tailed at 5% level)	0.514	
Significance (Y/N) (two tailed at 5% level)	No (no significant correlation between the two corpora)	

Table II Summary of Pearson's correlation test

As a normal practice in corpus linguistics, we opt for the five per cent as the threshold level to measure the strength of correlation between the two Chinese corpora. Given that we do not an obvious reason to assume or hypothesize the existence of a strong relationship between the two corpora in advance, we shall check the computed coefficient value with the critical value at the two-tailed non-directional category, which is always more prudent than using the one-tailed directional value.

The mechanism of the Pearson's correlation test is that we start the statistical procedure by assuming a null hypothesis which treats the two corpora as having no relationship at all; and in order to subvert the default hypothesis, the computed coefficient must be equal or greater than the critical value. However, as Table II shows, the coefficient r obtained from the two Chinese corpora, which is as low as 0. 435, is definitely below the threshold value at the critical five per cent, which is 0.514.

The result suggests that despite the many similarities shared by the two corpora, such as the same sampling framework, the same language type, standard Mandarin Chinese, they indeed differ from each other in terms of the frequency of occurrence and distribution of idioms. This may well turn out to a surprising outcome to many people who sustain the idea that given the high conventionality of idiom in Chinese, it would hardly allow such a rapid change to take place within a rather limited period of time[5]. To have a clear view of the considerable decrease of idioms in the UCLA corpus, as well as the general patterns of distribution of idioms in each corpus, the raw frequencies summarized in Table I have been normalized into frequencies per 10k words (see Table III).

Code	Frequency per 10k words (LCMC)	Frequency per 10k words (UCLA)
REP	18	5.45
SC	21.5	18
SK	32	10
RE	32.9	11.7
SF	37.5	42.5
GF	50	54.4
NREP	55	28.10
MY	55.4	58
PL	56.9	68.4
AD	58.3	54.5
HU	60	23
ES	60.5	51.1
RO	65.2	39.3
NED	68.3	44.4
NREV	73.2	36.6
Total	47.7	39.5

Table III Comparison of normalized frequencies between LCMC and UCLA corpus

[2] See http://bowland-files.lancs.ac.uk/corplang/lcmc/

[3] See http://bowland-files.lancs.ac.uk/corplang/ucla/

[4] See Appendix 8 The Pearson's product-moment correlation coefficient, in *Statistics in Corpus Linguistics*, Oakes, M (1998), Edinburgh: Edinburgh University Press, p. 267

[5] Xiang, G (1979) "Relationships between Chinese Idioms, Natural Environment, Cultural Traditions, and Linguistic Characteristics", in *Chinese Language*, vol. 2, pp. 112-121

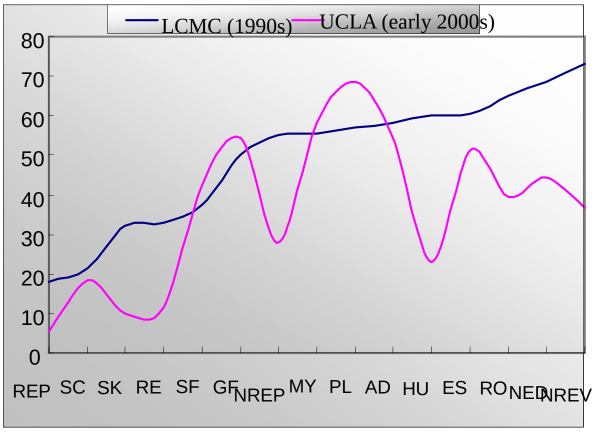


Diagram I Comparison of normalized frequencies between LCMC and UCLA corpus

To allow us to have an easier access to the numerical information provided in Table III, the results have been used to draw a histogram in which the two coloured curves represent the distribution of idioms across different text types in the LCMC (blue line) and the UCLA corpus (pink line), respectively. As may be seen from the graph, several important patterns regarding the evolving nature of Chinese idioms in written texts[6] seem to emerge.

Firstly, the blue line which embodies the LCMC shows a general trend to run above the pink line representing the UCLA corpus. This fits in well the descriptive statistics shown in Table II, where the mean frequency of the LCMC is twice that of the UCLA corpus. This seems to suggest that at an overall level, the language recorded in the LCMC is more idiomatic than the material compiled in the UCLA, which was constructed some ten years later. However, the term of idiomaticity is a very complex notion (Nunberg, et al. 1994), which may well have different connotations in different text types or genres.

In Nunberg et al, the notion of idiomaticity is broken down into six dimensions of English idioms which range from central to more peripheral properties. They are conventionality, inflexibility, figuration, proverbiality; informality and affect. Though it is arguably true that idiomaticity exists universally and to a large extent, shares fundamental features in all human languages, it has been noticed that the statements made in Nunberg et al. have limited applications in the study of Chinese idioms. In a previous study based on the evidence collected from large-scale modern Chinese corpora, i.e. the Modern Chinese Corpus[7], it has been pointed out that the three defining features of Chinese idioms, or Cheng Yu as we say in Chinese, are conventionality, figuration or archaism and potential structural flexibility (Ji, 2007).

There are five perceivable low ebbs along the pink curve line, which take place (1) in the first genre, REP (reports or official documents); (2) between the SK (skills/trades or hobbies) and RE (religion) categories; (3) in one of the middle text type, NREP (news reportage); (4) HU (humour) and (5) in the last category, NREV (news review). Among these six categories, which have been highlighted due to the detected sharp decrease in the use of idioms in the relevant text genres, we can see there are invariably non-fictional Chinese text types.

As mentioned above, the pragmatic functions of Chinese idioms may be approached either by its figurative or archaic attributes. Then, in light of the distinctive discursive features or writing conventions of these highlighted Chinese text genres, we could assume that with regards to Chinese text types which would use idioms as an important rhetoric device to enhance the formality of the language style, such as NREP (news reportage), NREV (news reviews), RE (religious), REP (reports and official documents), the language used in these text genres has been evolving quite visibly towards an informal style.

On the other hand, in text genres where idioms may be explored as figurative tropes such as the case of HU (humour), the significant drop in the use of idioms in a time span of ten years seems to suggest that the semantic transparency of such texts has been enhanced considerably. At the same time, the two peaks featured along the pink curve seem to suggest that the rhetorical or aesthetic value of idioms in Chinese fictional or popular writings has been steadily enhanced, which is well represented by the two small peaks along the pink line as above its blue counterpart SF (science fiction), GF (general fiction) and PL (popular lore). Such interesting findings uncovered through a comparable study of diachronic Chinese comparable corpora would require further explanations within a broader sociolinguistic framework to allow us a fuller understanding of the evolving nature of Chinese language in the last decade of the twentieth century.

Acknowledgement

I would like to thank my first reviewer of the paper, whose kind suggestions have been taken with gratitude in the revision of my submission to the workshop and thus have helped to improve its quality visibly.

References

Ji, M (2007) “What is the starting point? In search of a working definition of Chinese idioms”, in *Asian and African Studies*, vol. 6 1-2, Netherlands: Brill, pp. 1-11

Li, Yu Ling (2003) “Linguistic inheritance of idioms from ancient Chinese”, in *Journal of Kai Feng University*, vol.

[6] Both the LCMC and the UCLA corpus have been constructed with material collected from sources of written texts, such as online electronic libraries, or electronic texts posted on the web.

[7] The Modern Chinese Corpus may be accessed online at http://ccl.pku.edu.cn:8080/ccl_corpus/index.jsp?dir=xiandai

17, no.4, pp. 44-47

Nunberg, G. et al. (1994) “Idioms”, in *Language 70*, pp. 491–538

Oakes, M (1998) *Statistics in Corpus Linguistics*, Edinburgh: Edinburgh University Press, p. 267

The LCMC corpora may be accessed online at http://bowland-files.lancs.ac.uk/corplang/lcmc/

The UCLA corpus may be accessed online at http://bowland-files.lancs.ac.uk/corplang/ucla/

The Modern Chinese Corpus may be accessed online at http://ccl.pku.edu.cn:8080/ccl_corpus/index.jsp?dir=xiandai

Xiang, G (1979) “Relationships between Chinese Idioms, Natural Environment, Cultural Traditions, and Linguistic Characteristics”, in *Chinese Language*, vol. 2, pp. 112-121

A comparable Learner Translator Corpus: creation and use

Natalie Kübler
CLILLAC, University Paris Diderot
5, rue Thomas Mann, case 7016
F-75205 Paris cedex 13
nkubler@eila.univ-paris-diderot.fr

Abstract

In this paper, we summarise the development and present the use of a multilingual annotated Learner Translator Corpus (hereafter LTC) -a corpus whose core is composed of specialised translations produced by trainee translators and whose primary purpose is to provide insights into the most significant characteristics of such texts in order to inform translation pedagogy. The LTC has been developed in the frame of a European project, which ended in the Fall 2007. The corpus consists of a series of specialised texts which have been translated into several language pairs in both directions; translations have been annotated with PoS tagging in the different languages; all translations have meta-data concerning the trainee translator's background and the situation in which the translation was made. A specific error typology has been designed, in order to error-tag translations. As those were made from and into different languages, different tagsets had to be used; in order to harmonise all types of information and meta-data, it was decided to use XML stand-off annotations. An on-line query tool was developed to allow easy access to the LTC whose sentences are fuzzily aligned to show all possible translations of a given sentence; users can then see how each trainee has translated the sentence, and can also decide on which type of texts they want to see, depending on the choice of the trainee type and the translation context. . We show examples of the different errors annotated in the corpus and explain how the LTC can be used to inform translation teaching. The way translation teaching material can be customised by using the LTC results is illustrated with different types of pedagogical uses. We also make suggestions about how the corpus can be used by professional translators to improves their translation strategies.

1. Introduction

When dealing with corpora in reference with translation training or translation studies, aligned translation corpora (called hereafter *parallel corpora*), as well as bilingual or multilingual comparable corpora play a very important role. While parallel corpora can be used to study possible translations or study the way target text can be influenced by the source text for example, comparable corpora lead to the possibility of finding collocations and phraseological units that resort to the idiom principle (Sinclair 1991). Parallel corpora and comparable corpora in languages for specific purposes (LSPs) are also of great interest to terminologists and specialised translator, since they can be used to extract bilingual terminology and bilingual specialised phraseological units. From the more theoretical point of view, parallel and comparable corpora can give new insights into the languages that are studied, as a contrastive approach underlines features that may not be usually studied in a monolingual approach.

On the other hand, multilingual comparable corpora also have also tackled the question of second language from a varied range of theoretical and applied points of view.

Learner corpora have allowed linguists to question a L2 learner's interlanguage (coined by Selinker 1972), to question language transfer (Odlin T. 1989, Granger 1988, Granger et al. 2002) between the mother tongue (and all other languages the learner already knows) and the second language, or to develop applications that are either corpus-based or corpus-driven, to teach a second language (Granger 2003) or correct errors non native speakers make in a second language (Cornu *et al.*1996). The corpus we present here is concerned with those two types of corpora, as it is a multilingual parallel corpus (source text aligned with several translations in different source and target languages) and learner corpus (translations made by trainee translators). As the creation and development of the LTC has been widely described in Castagnoli *et al.* (to appear) this paper will thus briefly summarise the creation and content of the Learner translation Corpus (hereafter LTC) in order to focus on the use and applications that can be derived from the corpus as it is.

The LTC was developed in the frame of the MeLLANGE project, a European-funded project, which lasted three years and comprised ten partners. The MeLLANGE project aimed at devising a methodology for the collaborative creation of eLearning teaching content in the fields of translation and translation technology, producing corpus-based teaching materials and, more ambitiously, establishing a framework for a European Master's in Translation Technologies[1] .

2. Related work

Whereas learner corpora have known a huge development, work on learner translation corpora has been sparse. Castagnoli et al. (to appear) describe pioneer work at the end of the nineties and compare four corpora that differ in source and target languages and that are mainly aimed at error detection: the *Student translation Archive* Bowker & Bennison 2003), the PELCRA project (Uzar &Walinski 2001), the ENTRAD corpus (Floren 2006, to appear, and the *Russian Translation Learner Corpus* (Sosnina 2006).

3. Aims and design of the LTC

The initial aim for designing the LTC consisted in providing translation trainers with corpus-based pedagogic material and researchers with a corpus that would allow comparative observations, such as the influence of the source language on the native speakers' language of translation students, the influence of mastering more than one second language on translation into the native language, differences in translation error types taking into account different genres of texts or different specialised domains, and other related observations. However, an application such as providing pedagogical material means that not only the errors will be studied, but also, the different strategies for correctly translating a source text. This leads to the study of the specific interlanguage translation trainees show in translating into their own native language.

Five different texts, available in the different languages of the partnership (i.e. CA, DE, EN, ES, FR, IT) and considered as source texts, were chosen to be translated. The source texts were translated by students in the different language pairs of the partnership. They belong to different domains, namely administrative, technical, law, and journalistic. As the aim is to not only provide users with pedagogic material, but also to allow them to query the corpus, it was also necessary for it to carry specific metadata and linguistic information.

3.1 Corpus collection and metadata

The corpus to collect consisted of translations in the different language pairs of the partnership made by translation students. Some were also made by professional translators, in order to have correct reference translations, as part of the corpus as well. The collection process was made on-line on a collaborative platform, which is still available today (http://mellange.upf.es). Students and professional are able to download archives containing the source text, the extract to be translated and a brief giving all the information about the text, the source language, the domain, etc. Contributors can translate the text into the target language and then upload the translation onto the collaborative platform. They have to create an account the first time they access the platform, which makes it

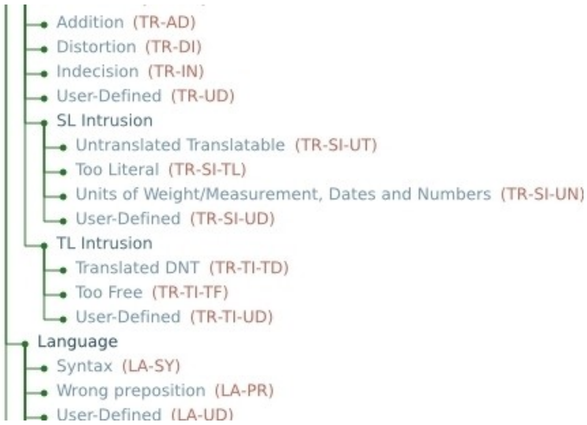
[1] Further information can be found at the following URL: http://mellange.eila.univ-paris-diderot.fr

possible to collect personal metadata about the translator. Personal metadata consist of information about language (mother tongue, second language, other second language), and curriculum (language studies, translation studies, first-year student, Master's student). Those metadata are stored once with the account of the contributor; then each time a translation is submitted, other questions are asked about the conditions in which the translation was made (in the classroom, at home, marked or not, constrained time, dictionaries allowed, with internet access). Each translation is then anonymised and stored in the database with all the metadata attached to it.

3.2. PoS and error tagging

The LTC is PoS tagged with different tagsets. As each language differs from others on PoS, different taggers and tagsets were used depending on the language. Instead of using only one approach for all languages, it was decided to keep different tagsets and harmonise them. An example of those different PoS tagging can be consulted in Castagnoli et al. (to appear).

Translations made by translation trainees contain different types of errors which represent the most interesting part of the corpus. In order to allow users to find out about errors, and thus specific difficulties in translating different texts in different language pairs, the partnership decided to annotate all errors in the corpus. A survey was made by Secara (2005), in order to produce a first version of an error typology, based on existing error categories used both by academia and professional translators. The typology was then tested on a sample of translations and enhanced, according to the observations made on those actual translations. Error types are divided into two sections: content transfer errors and language errors. Each section is then subdivided into subtypes. As each language and culture have their own specific way of tagging errors, a free error-category was left for each error subtype, namely a *user-defined* category. Therefore, each translation also carries error-tagging , which had to be done manually. However, in order to facilitate the error-tagging process, an error-tagging programme, based on MMAX2, originally developed by Mark-Christoph Müller at the European Media Laboratory, Heidelberg (2006), was customised to enable-error-tagging of the LTC. Figure 1 below shows a sample of the error-typology. We will show and explain examples of errors in the application section of this paper. Working on manual error-tagging at different sites and in various language pairs revealed the differences in considering the definition of a translation error. However, as we had to tag a corpus that would allow automatic handling, we had to all agree on errors. The main point of this error typology is that it does not show the reason why the translator made a mistake, but only the result of the mistake, which can be used as a marker for specific difficulties in the translation process.



- Addition (TR-AD)
- Distortion (TR-DI)
- Indecision (TR-IN)
- User-Defined (TR-UD)
- SL Intrusion
  - Untranslated Translatable (TR-SI-UT)
  - Too Literal (TR-SI-TL)
  - Units of Weight/Measurement, Dates and Numbers (TR-SI-UN)
  - User-Defined (TR-SI-UD)
- TL Intrusion
  - Translated DNT (TR-TI-TD)
  - Too Free (TR-TI-TF)
  - User-Defined (TR-TI-UD)

Language
- Syntax (LA-SY)
- Wrong preposition (LA-PR)
- User-Defined (LA-UD)

Figure 1: Sample of Mellange error typology

3.3. Corpus-query tool

As the partnership chose an XML stand-off annotation scheme, the corpus has a modular structure that allowed the development of a complex query tool, which is available at http://corpus.leeds.ac.uk/mellange. This query tool also gives access to another corpus, the eCoLoRe (http://ecolore.leeds.ac.uk/) corpus built in a previous project. The tool enables users to search the corpus for information on errors, alternative translations, reference translation, as well as various types of metadata. Each error is presented in the context of a whole sentence and is aligned, not with the sentence of the source text, but with the approximate context of the it. Then, the target text is also aligned with the reference translation and all other translations made of the same context. This is very useful in a pedagogic perspective and will be explained in the next section.

4. Using the LTC

As said above, the LTC is part of a bigger project aiming at providing translator trainers with pedagogic material based on authentic translation difficulties. As is the general case with corpus use in teaching, the first step consists in browsing the corpus to get more familiar with the types of difficulties that are encountered. This approach can then be taught to the trainee translator, knowing that any future (and actual) translator has become familiar with the specific domain in which the translation has to be made. The second step, for trainers, consists then into exploiting the corpus as pedagogic material. Exploiting a learner corpus tends to lead the trainer towards focussing on errors, which represents the first obvious use of such a corpus. We will illustrate however that another approach can be applied, which is based on the analysis of various possible solutions for translating a text.

4.1. Error-oriented approach

The first obvious approach consists in working with a sample of errors. Students can be presented with sentences containing a specific error-type and be asked to understand why there is an error, and what could be considered a correct translation. The reference translation can then be used to compare their suggestions with a professional translation.

EN	He may be excluded from the management of bodies under public law and from the exercise of an office under public law.
FR	Il peut être exclu de la gestion d'organismes et être démis de ses fonctions **au nom du droit public**. **TR-DI**
REF	il peut être exclu de la participation à la gestion d'organismes de droit public et de **l'exercice d'une fonction de droit public**.

Table 1: An example of a distortion error; all other error types have been removed in the French sentence.

EN	For the first time this huge country - which is the world's tenth-ranked industrial nation, with a population of 170m - is about to have a democratic government under a leader with roots in the radical left who rejects liberal globalisation.
FR	Le dixième pays industriel avec une population de 170 millions d'habitants connait pour la première fois de son histoire un gouvernement démocratique dirigé par un politique provenant de la gauche radicale qui rejette la mondialisation libérale.
FR tagged	[Le dixième pays industriel avec une population de 170 millions d'habitants]**LA-ST-AW** [connait]**LA-ST-AW** [pour la première fois de son histoire]**TR-UD**]**TR-SI-TL**[un gouvernement démocratique]**LA-ST-AW** dirigé par un [politique provenant]**LA-ST-AW** de la gauche radicale [qui rejette]**LA-IA-TA** la mondialisation libérale.
REF	Pour la première fois, l'immense Brésil - 170 millions d'habitants, dixième puissance industrielle du monde - s'apprête à être gouverné, dans des conditions démocratiques, par un leader issu de la gauche radicale qui rejette la mondialisation libérale.

Table 2: The second row shows the target text without error-tagging, the third row shows the same text with error tags, and the last row shows the reference translation.

Table 1 shows an example of a distortion error in a translation from English into French. The type of the text is administrative and deals with the rights of European workers in the European Union.

In this example, the prepositional phrase *under public law* is attached to the noun phrase *an office*. In the French translation, the PP *au nom du droit public* has been attached to the verb *être démis* which results in completely modifying the meaning of the source text. Here one can see that the error-tagging does not rely on any explanatory approach of the error: the segment concerned with the error is highlighted, but the interpretation is left to the student. Students have to think of the meaning of the sentence to understand the error and find a correct translation. The advantage of the corpus lies in the fact that students can be presented with dozens of errors of the same type, but occurring in different contexts. It helps them practice this particular translation skill which consists not only in understanding the meaning of the source text, but also in reformulating it into a target text that both clearly renders the meaning of the source text and sounds idiomatic to native speakers.

Another type of practice consists in collecting errors of different types and asking students to locate the error and find out what it is about. We show an example of this type of work in table 2.

4.2. Strategy-oriented approach

Another type of pedagogic approach with translation students consists in triggering a reflection and a discussion in the classroom about different translation strategies. The LTC allows this approach as each target sentence containing error is aligned with all the other translations of the same context. Among those translations some of those do not contain errors and can thus be used to work on strategies. We show in table 3 an example of variation in translation strategies for the same source text sample as in table 2.

5. Conclusion

We have tried to show here that comparable corpora could be coupled with a parallel approach, as the corpus presented consists of a comparable corpus containing parallel corpora. The second point we tried to make here was that learner corpora do not only deal with the process of learning a language, but can also focus on the process of learning a *process*, as the LTC which can be used to help translation trainees learn how to translate.

T	Pour la première fois, ce pays gigantesque – qui est le dixième pays industriel, avec une population de 170 millions d'habitants – est sur le point d'avoir à sa tête un gouvernement démocratique, avec un dirigeant dont les racines se trouvent dans la gauche radicale, qui rejette la mondialisation.
TR2	Pour la première fois, ce pays immense, qui est l'une des dix premières puissances mondiales, avec une population de 170 millions d'habitants, est sur le point d'avoir un gouvernement d'extrême gauche qui est contre la mondialisation libérale.

Table 3: Two different correct translations

6. Acknowledgements

The author would like to thank Susan Armstrong, Guy Aston, Toni Badia, Silvia Bernardini, Sarah Castagnoli, Dragos Ciobanu, Carme Colominas, Beto Garrido Boulosa, Tony Hartley, Kerstin Kunz, Stella Neumann, Andrei Popescu-Belis, Alina Secara, and Alexandra Volanschi , as the MeLLANGE LTC is the result of the collaborative work of all.

7. References

Bowker, L. & Bennison, P. (2003) Student Translation Archive: Design, development and application. In Zanettin F., Bernardini S., & Stewart D. *Corpora in Translator Education* Manchester: St Jerome, 103-117

Castagnoli, S., Ciobanu, D., Kunz, K., Volanschi, A., Kübler, N. (to appear), Designing a Learner Translator Corpus for Training Purposes, *Proceedings of TALC2006.*Amsterdam: Rodopi

Cornu, E., Kübler, N., Bodmer, F., Grosjean, F., Grosjean, L., Léwy, N., Tschichold, C. & Tschumi, C. (1996). Prototype of a second language writing tool for French speakers writing in English. *Natural Language Engineering*, 2 (3), 211-228.

Floren C. (to appear) ENTRAD, an English Spanish Parallel Corpus Created for the Teaching of

Translation. *Proceedings of TALC2006*. Amsterdam: Rodopi

Granger, S. (1988): *Learner English on Computer*. London : Longman.

Granger, S. (2002): A Bird's-eye view of learner corpus research. In: Sylviane Granger, Joseph Hung, Stephanie Petch-Tyson (Eds.), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam, Philadelphia: John Benjamins. P. 3~~3~~3.

Granger S. (2003). Error-tagged learner corpora and CALL: a promising synergy" In *CALICO* (special issue on Error Analysis and Error Correction in Computer-Assisted Language Learning) 20 (3): 465-480.

Müller C. (2006) Representing and Accessing Multi-Level Annotations in MMAX2. *Proceedings of the workshop on Multi-dimensional Markup in Natural Language Processing (NLPXML-2006)*, EACL, Trento, Italy, April 2006.

Odlin,T.(1989) *Language Transfer: crosslinguistic influence in language learning*. Cambridge: Cambridge University Press.

Secară A. (2005) Translation Evaluation - a State of the Art Survey. *eCoLoRe/MeLLANGE Workshop Proceedings*, 39-44. (http://www.leeds.ac.uk/cts/research/publications/leeds-cts-2005-03-secara.pdf)

Selinker, L. (1972) Interlanguage. In *IRAL*, 10/3. pp. 31-54.

Sinclair, J.M. (1991) Corpus, Concordance, Collocation, Oxford: Oxford University Press

Sosnina E. P. (2005) Development and Application of Russian Translation Learner Corpus. Presented at *Corpus Linguistics – 2006* ST Petersburg, Russia, 10-14 Octoebr 2006

Uzar R. & Walinski J. (2001) Analysing the Fluency of Translators. *International Journal of Corpus Linguistics*, Vol. 6 (Special Issue), 155-166.

Corpógrafo V.4 – Tools for Researchers and Teachers Using Comparable Corpora

Belinda Maia, Sérgio Matos
Linguateca – PoloCLUP
Faculdade de Letras da Universidade do Porto
Via Panorâmica s/n
4150-563 Porto
Portugal
E-mail: bmaia@mail.telepac.pt, sgmatos@letras.up.pt

Abstract

This paper describes the response by the NLP project Linguateca to the needs of researchers, teachers and students in the areas of terminology, translation, contrastive linguistics, and related areas, for user-friendly tools for building and using comparable corpora. It will present the latest developments of the Corpógrafo, a suite of freely available and fully integrated online tools that allow for individuals or small groups to do linguistic research, or simply study the implications of corpus and terminology research for translators. The new developments include considerable improvements to the previous corpus and terminology database tools, a parallel corpus aligner, an aligner of parallel segments in comparable corpora, the integration of the NooJ engine with dictionaries in English, French and Portuguese, and a lexical / phrasal database structure designed for both normal lexicography and for the storage and analysis of the multi-word expressions of interest to those researching genre, text or discourse analysis. The results of this research will, in turn, contribute to the enrichment of the Corpógrafo tools.

1. Introduction

The reasons for building comparable corpora vary considerably, but the call for papers for this workshop focuses on several of the computational interests involved. We shall begin by referring briefly to the way the Corpógrafo functioned in the past and describe some of the improvements made for finding terms in special domain comparable corpora. We shall then concentrate on the possibilities of our new tools for collecting and analyzing phrases in comparable corpora. It is hoped that the data thus acquired can, in turn, be used to enrich and develop the tools themselves. Our approach is based on our own experience of the symbiosis needed between developing such tools, finding a practical research usage for them, and improving them using feedback from users.

The tools are not particularly new individually, but as an integrated suite they are useful. Before we discuss the computational tools for linguistic analysis of comparable corpora, however, we shall begin by reflecting briefly on the nature of comparable corpora and how they can serve as a basis for a wide variety of research projects for which computational tools offer possibilities.

2. Reasons for Building Comparable Corpora

What is a comparable corpus? It is not that easy to either define a comparable corpus or, having done so, to find suitable texts with which to build one. However, it is clear that, more often than not, comparable corpora are seen as domain or subject specific, such as texts about composite materials, fire hazards, or pet cats. Once the domain has been chosen it is also normal to restrict the genre so that, for example, scientific texts and publicity texts are paired separately. Besides this, it is often assumed that comparable corpora are bi- or multilingual.

As has been said in the call for papers, comparable corpora are of increasing interest because of the scarcity of reliable parallel corpora. Most of the workshop topics contemplate comparable corpora which are bi or multilingual, and presume that one will build a corpus of this kind for mining information of various kinds. As comparable corpora also have the advantage that most specialized texts will have been written by domain experts, they will therefore be more reliable for terminology extraction than translations that, despite all the recommendations of the European Norm EN 15038, may not have been revised by an expert.

Another advantage is that texts in comparable corpora are usually written by native speakers and should be better examples of the language or languages being studied. This means that they can also serve for various kinds of genre, text and discourse analysis.

There are also several reasons for creating monolingual comparable corpora. Someone may wish to discover why one text is more successful with its audience than another as, for example, in publicity texts. Others may want to study different authors, in the attempt to find out who influenced whom, and this has applications for discovering plagiarism and for forensic linguistics. Yet others may wish to create a corpus of exemplary texts in different domains and genres in English and extract phrases that would be useful for the growing number of non-native English speakers who feel obliged to write directly in English.

3. Linguistic v. Computational Approaches

It should be clear by now that our approach will necessarily have to combine computational tools with ‘manual’ intervention by linguists, and we believe that it is essential to unite the two skills for better research. Computational approaches tend to favour acquiring large quantities of text in the hope that the number of examples of the required information, terms, or phrases will prove significant enough to allow one to safely ignore anything that appears infrequently or not at all. There are a variety of computational methods for finding texts in certain domains, an example of which is BooTCaT (Baroni & Bernardini, 2004). However, one of the problems of dredging the internet for such texts is that a lot of repeated material and noise come back with whatever it is we are looking for. Internet mirror pages and plagiarism are responsible for much of this.

On the other hand, corpora consisting of texts that have been carefully chosen by a linguist may not need to be enormous in order to provide useful information. For several years now, translation teachers have encouraged students to create small corpora for specific assignments, called 'do-it-yourself' corpora (Maia, 1997) or 'disposable' corpora (Varantola, 2003), and they have proved pedagogically useful, despite their limitations for NLP research.

The design of the Corpógrafo was based on the assumption that individuals would invest time in finding texts that suited their research needs, but needed help in converting them into plain text and combining them selectively into searchable corpora. Choosing the texts is in itself part of the pedagogical process. Cleaning up a large automatically extracted corpus may actually take much longer and the process is hardly educational for the trainee translator, terminologist or linguist. Now that the Corpógrafo is being extended to more general language analysis, the need to create carefully chosen corpora continues to be relevant.

4. Building Comparable Corpora and Related Databases

The Corpógrafo was originally designed for the building of comparable corpora in special domains for the extraction of terminology, but the tools can be used for any kind of corpus. It offers a complete framework for working with text, from extracting text from different types of files, to editing and cleaning the texts, to grouping the files selectively into separate monolingual corpora, and using simple concordance tools for studying these corpora.

When the corpora have been created it allows users to create related multilingual databases in an efficient manner, by using the system's semi-automatic methods for registering metadata on the corpora, extracting lexical and phrasal items, as well as term candidates, using n-gram tools with or without filters, and finding definitions and semantic relations between lexical items or terms using underlying list of lexical patterns (Sarmento et al., 2006). Once the initial texts, monolingual corpora and related multilingual databases are operable, statistics on the frequency of lexical items or terms and the way they occur in the texts in a corpus can be generated automatically.

A new feature is a tool to bootstrap information from the internet directly into Corpógrafo's file preparation system using a starting list of seed expressions from this statistical information. This feature follows the same idea as implemented by the BooTCaT toolkit (Baroni & Bernardini, 2004), but allows the researcher to select and process relevant texts as needed.

This general workflow in Corpógrafo and an overview of the system's architecture are illustrated in Figure 1. All data added by the users and associated metadata, are kept on the user's working area. Operations on these data are managed by Corpógrafo, and are available to the users through graphical interfaces to the system's functions.

5. Genre Specific Comparable Corpora

One of our earliest tools was a simple n-gram tool which served to help find the lists of expressions used to find definitions and semantic relations in the Corpógrafo. It also drew our attention to what people call 'lexical bundles', 'multi-word units/expressions', 'paraphrases', and similar phenomena (Maia et al., forthcoming). Silva (2006) used the tool to search for discourse phrases in information on art exhibitions in English and Portuguese and was able to show the differences in the text conventions for this genre in the two languages/cultures. He first searched his corpora using the n-gram tool, selected expressions that could be considered discourse connectors, like *in order to, at the same time, for example* and then classified these expressions in terms of discourse markers, such as 'purpose', 'inclusion' and 'exemplification', respectively, He then analysed the examples in comparable corpora of about 128,000 words for each language, quantified the results and drew certain conclusions about the cultural differences between English and Portuguese conventions when writing on the subject of art exhibitions.

This experiment led us to create the possibility of creating multilingual lexical and phrasal databases with appropriate classifications for lexical and syntactic information, as well as for lexical and semantic conceptual relations, similar to those used in the terminology databases. This will allow us to develop Silva's methodology and apply it to further research.

The new lexical/phrasal database structure also offers the possibility of classifying the word or phrase for the effect of discourse analysis. The choices of classification offered are derived from the Rhetorical Structure Theory discourse relations developed by Maite Taboada (see: http://www.sfu.ca/rst/index.html) and adapted for Portuguese by Rui Silva. It is also possible to create one's own classifications, if one wishes.

The objective here is to develop lists of expressions that will semi-automatically retrieve the discourse elements according to this classification. Since there is a growing interest at both a research and pedagogical level in raising awareness of the conventions of different genres and text types and comparing these conventions in different social and cultural situations, this development offers new opportunities for this type of analysis.

Another development is the use of the NooJ engine (see: http://www.nooj4nlp.net) to query the corpora for phrasal units, using regular expressions and grammatical (part-of-speech) tags. This now works in French, English and Portuguese. In the future, we plan to allow users to save and edit the NooJ annotation so that it becomes possible to correct the results and even add – semi-automatically - tags related to one's own discourse analysis or similar

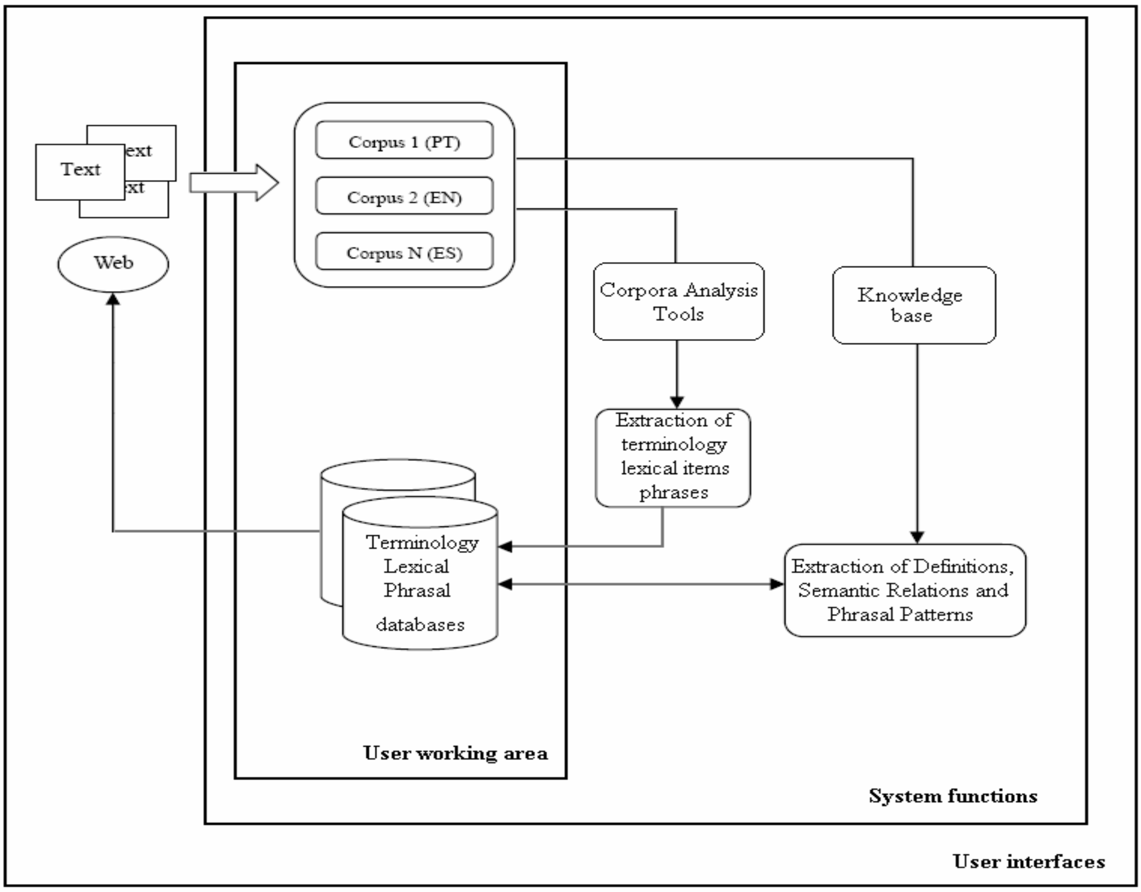

Text
ext
xt
Web
Corpus 1 (PT)
Corpus 2 (EN)
Corpus N (ES)
Corpora Analysis Tools
Knowledge base
Extraction of terminology lexical items phrases
Terminology Lexical Phrasal databases
Extraction of Definitions, Semantic Relations and Phrasal Patterns
User working area
System functions
User interfaces

Figure 1: General workflow and architecture of Corpógrafo

research. This will help with the identification and correlation between languages of phrasal, syntactic or discourse patterns, and once these patterns are entered in the multilingual databases, they can be observed using the concordancing tools for parallel and comparable corpora described below.

6. Aligning Parallel and Comparable Corpora

To address the need of some of our users, we have integrated a sentence alignment tool (from IMS-CWB) in the Corpógrafo environment. The alignment is performed without the user's interaction, allowing users to create their own parallel corpora, without the need of any knowledge of programming or on how to configure the aligner. The alignment results are presented on screen in a tabular form, each row representing an alignment unit, and can be easily edited to correct any alignment errors.

We are at present working on a tool which offers dual concordances from two monolingual comparable corpora of sentences that include the segments that the researcher has marked as equivalents in the multilingual databases of terms, lexical items or phrases. The objective will be to verify if the information is correct and to see if the apparent equivalents do actually function in the same collocational or textual circumstances as the researcher originally supposed.

7. Research and Teaching Applications

The Corpógrafo has been used for a variety of teaching and research applications for some time. Although it was originally designed for use by individuals, it is now possible for groups of people to work on the same area and distinguish the work done by the different contributors. It is available online to whoever asks for a username and password. We use it for teaching purposes and several of our masters' and doctoral dissertations depend on the system for their research. There are also many users from all over the world, particularly from Brazil, who use if for pedagogical and research work.

So far, most of our research has been in the areas of terminology and lexical analysis, and is becoming increasingly sophisticated now that the tools have been improved. However, the new tools allow for much more.

These tools can now be used to search corpora for various forms of multi-word expressions using n-grams, normal lexical concordances and concordancing using the NooJ POS analysis. Parallel texts can be aligned, and data extracted from comparable corpora can be concordanced in two languages simultaneously. The resulting databases can be used to store and categorize lexical, syntactic and textual information that can be exported for a variety of uses.

Apart from the more obvious applications to research projects, there are several ways in which practical results can be obtained for translators and others. For example, in order to facilitate the organization and translation into English - or even the writing of the original in English - of the programmes of our university courses, we are at present using the tools to find and store in our databases, useful phrases in comparable corpora built from texts from English speaking university sites on-line. This is being done using an n-gram tool and/or the NooJ POS analysis using patterns typically associated with the type of text under analysis. The results are being used to create a list of English and Portuguese "useful phrases", available on the university intranet or on a special area of our translator's page at http://web.letras.up.pt/traducao/TRAD/trad.htm The same idea can be applied to a variety of similar uses, and provide useful pedagogical tools for teaching levels of language from lexicography to text analysis.

8. Final Remarks

The Corpógrafo has always been driven by the needs of researchers in linguistics who want to take advantage of user friendly language technology. It is also useful for teachers who want to train their students to understand the possibilities of these technologies without necessarily having to beg their universities for constant upgrades of very expensive commercial translation software with which to do so.

In other words, we have always tried to foresee a use for the tools rather than simply create tools that may or may not be wanted. The tools themselves are not a novelty, but the combination and integration of several tools into one integrated system is less usual.

We must emphasize the fact that the tools have been conceived to encourage the general linguist to use and understand the possibilities of NLP tools. This means that the tools should provide the general linguist with the possibility of collecting, observing and validating data and inserting it into the Corpografo in their personal area. The results can then be used to integrate information in the Corpógrafo tools as, for example, when lists of expressions to retrieve definitions and semantic relations were retrieved for terminology processing.

The latest developments will allow us to create lists of discourse markers, lexical bundles and other linguistic phenomena that can be used in both monolingual and multilingual comparable corpora. The work-in-progress is at the level of research and individual project work being done by post-graduates in translation, terminology and general or contrastive linguistics.

Acknowledgements

This work was done in the scope of the Linguateca, contract nº339/1.3/C/NAC, project jointly funded by the Portuguese Government and the European Union. We should also like to thank the colleagues and students who have contributed with ideas and opinions on how best to develop new tools.

References

Baroni, M. and Bernardini, S. (2004). BootCaT: Bootstrapping corpora and terms from the web. In M.T. Lino, M.F. Xavier, F. Ferreira, R. Costa & R. Silva (eds.), *Proceedings of LREC 2004* (Lisboa, Portugal, 26-28 May 2004), pp. 1313-1316.

Maia, B. (1997). Do-it-yourself corpora ... with a little bit of help from your friends! in B. Lewandowska-Tomaszczyk and P. J. Melia (eds) *PALC '97 Practical Applications in Language Corpora.* Lodz: Lodz University Press. 403-410.

Maia, B., Sarmento, L., Santos, D., Cabral, L., & Pinto, A.S. (2005). CORPÓGRAFO - an online suite of tools for the construction and analysis of corpora, semi-automatic extraction of terminology and the construction of conceptual databases. *Proceedings from the Corpus Linguistics 2005 Conference Series* (Birmingham, UK, 14-17 July 2005), s/pp.

Maia, B. Silva, R., Barreiro, A., & Frois, C. (forthcoming). N-grams in search of theories, in B. Lewandowska-Tomaszczyk *PALC 2007 Practical Applications in Language Corpora.*

Oliveira, D., Sarmento, L., Maia, B., & Santos, D. (2005). Corpus analysis for indexing: when corpus-based terminology makes a difference. In P. Danielsson & M. Wagenmakers (eds.), *Proceedings from the Corpus Linguistics 2005 Conference Series* (Birmingham, UK, 14-17 July 2005), s/pp.

Sarmento, L., Maia, B., & Santos, D. (2004). The Corpógrafo - a Web-based environment for corpora research. In M.T. Lino, M.F. Xavier, F. Ferreira, R. Costa & R. Silva (eds.), *Proceedings of LREC 2004* (Lisboa, Portugal, 26-28 May 2004), pp. 449-452.

Sarmento, L., Maia, B., Santos, D., Pinto, A. & Cabral, L. (2006). "Corpógrafo V3: From Terminological Aid to Semi-automatic Knowledge Engine". In N. Calzolari, K. Choukri, A. Gangemi, B. Maegaard, J. Mariani, J. Odjik & D. Tapias (eds.), Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006) (Genoa, Italy, 22-28 de Maio de 2006), pp. 1502-1505.

Silva, R. (2006). *Performance and Individual Act Out: The Semantics of (Re)Building and (De)Constructing in Contemporary Artistic Discourse*. Master's dissertation. Porto: FLUP.

Varantola, K. (2003). Translators and disposable corpora. in Zanettin, F., S. Bernardini and D. Stewart (eds). *Corpora in translator education*. Manchester: St Jerome. 55-70.

Looking for Transliterations in a Trilingual English, French and Japanese Specialised Comparable Corpus

Emmanuel Prochasson[∗], Kyo Kageura[†], Emmanuel Morin[∗], Akiko Aizawa[⋆]

[∗]Laboratoire d'Informatique de Nantes-Atlantique, Université de Nantes,
2 rue de la Houssinière, 44322 Nantes, France
{emmanuel.prochasson, emmanuel.morin}@univ-nantes.fr
[†]Graduate School of Education, the University of Tokyo,
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-0033, Japan
kyo@p.u-tokyo.ac.jp
[⋆]Digital Contents and Media Sciences Research Division, National Institute of Informatics,
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan
aizawa@nii.ac.jp

Abstract

Transliterations and cognates have been shown to be useful in the case of bilingual extraction from parallel corpora. Observation of transliterations in a trilingual English, French and Japanese specialised comparable corpus reveals evidences that they are likely to be used with comparable corpora too, since they are an important and relevant part of the common vocabulary, but they also yield links between Japanese and English/French corpora.

1. Introduction

Bilingual lexicon extraction from comparable corpora has received specific attention in recent years. This attention is motivated by the scarcity of parallel corpora, especially for language pairs not involving the English language. However, since comparable corpora are "sets of texts in different languages, that are not translations of each other" (Bowker and Pearson, 2002, p. 93), methods proposed for parallel corpora — that make use of fixed correlations between bilingual textual units such as word, sentence, paragraph… — are not applicable. For comparable corpora, the standard approach is based on lexical context analysis and relies on the assumption that a word and its translation tend to appear in the same lexical contexts (Rapp, 1995; Fung and McKeown, 1997; Peters and Picchi, 1998).

Although processing methods are distinct, bilingual corpora such as parallel or comparable corpora share, by essence, some transverse features such as words in one language that are orthographically or phonetically similar to a semantically related word in another language (cognates or transliterations). Cognates and transliterations yield anchor points that are useful to find extra clues for alignment of parallel texts (Simard et al., 1993). In the same way, we want to investigate the usefulness of the transliterations for the task of bilingual terminology extraction from specialised comparable corpora. We first introduce the concept of transliteration, especially concerning Japanese language and then present observations about transliterations in a trilingual English/French/Japanese specialised comparable corpus.

Note that this paper is not about automatic transliteration in comparable corpora, all transliterated units were extracted and aligned manually, as we were only concern by their prominence and relevance among specialised comparable corpora.

2. Overview of the transliteration phenomenon

In this study, we call *transliteration* the phenomenon of picking a word in one language to use it in another language, generally using different and non equivalent graphical symbols (to be accurate, a *loan word* is *transliterated* to fit a target language). This phenomenon differs from *cognates*, which are words sharing a common origin but evolved in different ways. For example, the English/Japanese pair volley-ball/バレーボール (ba-re-e-bo-o-ru – note that we will always give the Hepburn romanised version of Japanese terms introduced, each mora separated by a hyphen) is a transliteration, whereas the Spanish/Portuguese pairs estrella/estrela, meaning *star*, is a cognate.

In some cases, transliteration process is direct and the word is not changed at all (for example, *café*, *voilà*, *vis-à-vis* or *raison d'être*, which are used in French and in English, even though English language does not include any diacritical symbols in its alphabet). In other cases, however, the word need to be drastically transformed, which happen in English/French to Japanese transliterations, since Japanese does not share the same alphabet and does not include some very common English or French speech sound, such as cluster of consonants. Therefore, *hovercraft* is transformed to ホバークラフト (ho-ba-a-ku-ra-fu-to). Thus, transliterations can be seen as *the projection of a word, from a source language, into a target language*.

This phenomenon appears with many pairs of language such as western language (English, French, German…) and oriental language (Arabic, Chinese, Japanese...), in both ways. It is frequent in all languages which keep evolving, to allow a dynamic evolution of the vocabulary to fit needs of speakers. This is especially the case with technical vocabulary, which is intended to be shared by a community of experts and, at first, do not go through the regu-

	1	2	3	4	5	6	7	8
1	あ/ア/a	い/イ/i	う/ウ/u	え/エ/e	お/オ/o			
2	か/カ/ka/	き/キ/ki	く/ク/ku	け/ケ/ke	こ/コ/ko	きゃ/キャ/kya	きゅ/キュ/kyu	きょ/キョ/kyo
3	さ/サ/sa	し/シ/shi	す/ス/su	せ/セ/se	そ/ソ/so	しゃ/シャ/sha	しゅ/シュ/shu	しょ/ショ/sho
4	た/タ/ta	ち/チ/chi	つ/ツ/tsu	て/テ/te	と/ト/to	ちゃ/チャ/cha	ちゅ/チュ/chu	ちょ/チョ/cho
5	な/ナ/na	に/ニ/ni	ぬ/ヌ/nu	ね/ネ/ne	の/ノ/no	にゃ/ニャ/nya	にゅ/ニュ/nyu	にょ/ニョ/nyo
6	は/ハ/ha	ひ/ヒ/hi	ふ/フ/fu	へ/ヘ/he	ほ/ホ/ho	ひゃ/ヒャ/hya	ひゅ/ヒュ/hyu	ひょ/ヒョ/hyo
7	ま/マ/ma	み/ミ/mi	む/ム/mu	め/メ/me	も/モ/mo	みゃ/ミャ/mya	みゅ/ミュ/myu	みょ/ミョ/myo
8	ら/ラ/ra	り/リ/ri	る/ル/ru	れ/レ/re	ろ/ロ/ro	りゃ/リャ/rya	りゅ/リュ/ryu	りょ/リョ/ryo
9	や/ヤ/ya		ゆ/ユ/yu		よ/ヨ/yo			
10	わ/ワ/wa				を/ヲ/wo			
11	が/ガ/ga	ぎ/ギ/gi	ぐ/グ/gu	げ/ゲ/ge	ご/ゴ/go	ぎゃ/ギャ/gya	ぎゅ/ギュ/gyu	ぎょ/ギョ/gyo
12	ざ/ザ/za	じ/ジ/ji	ず/ズ/zu	ぜ/ゼ/ze	ぞ/ゾ/zo	じゃ/ジャ/ja	じゅ/ジュ/ju	じょ/ジョ/jo
13	だ/ダ/da	ぢ/ヂ/ji	づ/ヅ/zu	で/デ/de	ど/ド/do			
14	ば/バ/ba	び/ビ/bi	ぶ/ブ/bu	べ/ベ/be	ぼ/ボ/bo	びゃ/ビャ/bya	びゅ/ビュ/byu	びょ/ビョ/byo
15	ぱ/パ/pa	ぴ/ピ/pi	ぷ/プ/pu	ぺ/ペ/pe	ぽ/ポ/po	ぴゃ/ピャ/pya	ぴゅ/ピュ/pyu	ぴょ/ピョ/pyo
16	ん/ン/n							

Table 1: Standard Japanese mora. Column from 6 to 8 are mora composed with two symbols (note that the second one is smaller). Line from 10 to 15 are voiced sound, transformed with the ゛ and ゜ diacritical symbol (は/ha → ば/ba → ぱ/pa). There is one more mora, to be used inside words, the "small tsu", ッ/っ refers to a silent mora (romanised by repeating the following consonant).

lar process of being appropriated and integrated by regular users of a language. Numerous examples can be found in computer science technical vocabulary, being used "as-is" in French (*shell*, *login*, *OS*, *web*, *cd-rom*, *e-mail*...) even when translation can be easily found (*ligne de commande*, *enregistrement/connexion*, *SE*, *toile*, *disque compact*, *courrier électronique*...). Spotting transliterations is therefore even more interesting since it concerns a vocabulary likely to be missing in regular multilingual dictionaries.

We chose here to focus on Japanese transliterations and introduce some features of the Japanese language in the next part.

3. Characteristics of transliterations in Japanese language

3.1. Japanese writing systems

Japanese language is written using three different sets of symbols (see Kageura (2005), for complete description). Kanjis, namely Chinese symbols, are used for their meanings and can be combined to form plain words, whereas katakana and hiragana are two equivalent phonetic alphabets composed of 46 symbols each (see table 1). Hiragana are used for common words where no kanjis are available or are unknown to the writer (typically for children), for grammatical purpose and at scarce occasions to represent onomatopoeia emitted by human. Katakana is mostly used to represent transliterated terms which give us an easy way to spot them and drastically prune terms comparison process. We should also note that katakana are also frequently used for emphasise (for example, in advertising) and to represent onomatopoeia.

3.2. Origin of Japanese transliterations

Japanese language borrowed word from many languages, especially Asian languages (more often Chinese) and western languages (English, French, German…). Most of Japanese western transliterations have been borrowed to English language (even country names are for the most transliterated using the English pronunciation, for example スペイン/`su-pe-i-n`, standing for Spain). However, some transliterations are issued from other languages:

- from French, for example クロワッサン/`ku-ro-wa-s-sa-n` – *croissant* or エスカルゴ/`e-su-ka-ru-go` – *escargot*, in English *snails*, (cooked one, the name of the animal being カタツムリ/`ka-ta-tsu-mu-ri` – this last example shows that species name are often written using katakana too) ;
- from German, for example レントゲン/`re-n-to-ge-n`, corresponding to *x-rays*, from Wilhelm Röntgen who discovered them
- from other western languages, for example パン/`pa-n` from Portuguese (*bread*).

3.3. Transliteration relations with French language

Even though French to Japanese transliterations are rare, it might still be interesting to try to align them with French vocabulary (Tsuji et al., 2002). Indeed, a lot of French vocabulary is common, or very close to English vocabulary and by extend, to western languages (several terms being cognates or transliterations among those languages), especially concerning specialised technical vocabulary. Therefore, transliteration alignment between French and katakana can give interesting result due to a common *bridge word*. Table 2 shows a set of examples extracted from our corpora. Note that knowing the origin of a transliterated term is not really relevant since bridge terms and French terms are generally cognates, originally from a third common language, mostly Greek and Latin.

However, this can lead to attempt to align transliterations with *faux amis*. As an example, the Japanese term フィルム/`fi-ru-mu` is to be aligned with the English term *film*, which also exists in French although the meaning is slightly different. Whereas in French *film* is generally

Japanese / Romanised	→	Bridge term	→	French
インスリン / i-n-su-ri-n	→	insulin	→	insuline
ホルモン / ho-ru-mo-n	→	hormone	→	hormone
ミネラル / mi-ne-ra-ru	→	mineral	→	minéral
ヘモグロビン / he-mo-gu-ro-bi-n	→	hemoglobin	→	hémoglobine
ビタミン / bi-ta-mi-n	→	vitamin	→	vitamine

Table 2: Example of katakana/French indirect transliterations

used for *movie*, in English it mostly refers to *reel*, which is also the meaning of フィルム/fi-ru-mu. We therefore take cautious to talk about transliteration relation between two term only when both conditions are met: terms are phonetically related and are mutual translations.

On the next part, we will shortly present the comparable corpus and observation concerning transliterations and their importance among corpora.

4. Analysis

4.1. Point of observation

We harvested the Web in order to compile an English-French-Japanese comparable corpus. Documents selected all refer to *diabetes* and *nutrition* and are all of *scientific* discourse (“*expert addressing experts*”; (Pearson, 1998), p. 36). Documents were extracted manually, following search engine results or using PubMed[1] for the English part. Documents were finally converted from HTML or PDF to plain text. We obtained 257,000 words for the French corpus, 235,000 for the Japanese corpus and 1,877,000 words for the English corpus. The Japanese corpus is processed through the Chasen morphological analyser[2], French and English corpora are tokenised to isolate words.

The first observation concern all potential transliterations extracted from the Japanese corpus (see part 4.2.) sorted depending on language alignment possibility criteria. We then try to find corresponding source term in English and French corpora (see 4.3.) and finally take a look at a sample of the vocabulary involved in transliteration found between English and Japanese comparable corpora (see part 4.4.). Our goal here is to show the importance and the relevance of transliteration in specialised French/English and Japanese comparable corpora, in order to use them for bilingual lexicon extraction.

4.2. Starting from Japanese corpora

We extract all potential transliterations from the Japanese corpus, by isolating every sequence of katakana. We only work on Japanese single word and exclude hapax for this part, for they are likely to be unstable. 627 different terms were extracted. Note that, due to issue in PDF to text conversion, some candidates are incorrect and are therefore removed (typically single katakana). We finally obtain 493 potential transliterations (i.e. existing Japanese terms written using katakana), which stand for about 8% of the Japanese part unique vocabulary used in context vectors. We then manually translate them, in French when possible, in English if not. Table 3 summarises statistics and shows some samples concerning every sets. *French only trans.* (resp. *English only trans.*) refers to the amount of transliterations, in the Japanese corpus, that can be aligned with a French term (resp. English term — that is, phonetically related and translation of each other) but not with an English term (resp. French term). On the other hand, *French/English trans.* stands for the amount of transliterations that can be aligned with a French and an English term. Finally, *Adapted* refers to transliterations originally from any language, which can not be aligned with French or English because they have been adapted, generally shorten, such as コンビニ/ko-n-bi-ni referring to *convenient store*.

4.3. Relations with English and French corpora

We found several transliterated term in the Japanese corpus, but can we find relation with other corpora ? To answer this question, starting from the manually translated and sorted list, we seek in French and English corpora if we can find corresponding terms. There are 449 transliterations corresponding to an existing English term in the Japanese corpus (see table 3 – 228 transliterations for English only, 221 for English and French) and 225 transliteration corresponding to an existing French term (221 for English and French, 4 for French only). That means we can, at most, find 449 English terms and 225 French terms in English and French comparable corpora.

Among English corpus, **314** terms can be found (which means, they are actually 314 transliteration relations between the Japanese and English corpora on a maximum of 449 – 26 concerning hapax, 288 concerning words appearing twice or more) whereas, among French corpus, from a set of 225, **140** relations can be found (of which 16 hapax). Those results shows that, not only transliterations appears among isolated corpora, but they also cover a part of the common vocabulary we are trying to extract and provide several links between comparable corpora. Although effectiveness of transliteration in bilingual extraction is yet to be observed, these first observations reveal a good potential of incorporating transliterated elements into bilingual term extraction methods. We now have to check if those links can be useful as anchor points by observing the vocabulary involved in transliteration relations.

4.4. Transliteration vocabulary

This last observation is hard to claim without concrete experiments, however we think it is worth to introduce it. Indeed, numerous Japanese transliterations extracted refers precisely to corpora topics (*diabetes and nutrition*) or domain (medical), or related theme such as *physical activities*, *diet and recipe*, *screening and treatment*, *doctor/patient*

[1] http://www.ncbi.nlm.nih.gov/PubMed/

[2] http://chasen.naist.jp/hiki/ChaSen/

	#occ	%	Examples
French only trans.	4	0.8%	レバー/re-ba-a/*levure*, リール/ri-i-ru/*Lille*
English only trans.	228	46%	ヘルス/he-ru-su/*health*, ダイエット/da-i-e-tto/*diet*
French/English trans.	221	45%	マネジャー/ma-ne-ja-a/*magnesium*, ヒスタミン/hi-su-ta-mi-n/*histamine*
Adapted	12	2%	ビル/bi-ru/*building*, テレビ/te-re-bi/*television*
Not English, not French trans.	5	1%	カリウム/ka-ri-wa-mu/*potassium*
Not transliteration	23	5%	ムカデ/mu-ka-de/*centipede*, カキ/ka-ki/*oyster*

Table 3: Statistics concerning katakana sequences from the Japanese corpora

conversation. . . Here is a 50 words sample randomly extracted from the 314 transliteration pairs found between English and Japanese corpora.

> fair / **advice** / library / schedule / mini / **case** / **keywords** / **insulin** / follow-up / **peak** / clear / **candy** / **interferon** / score / shopping / **signal** / copy / **isotope** / map / **nano** / curriculum / **science** / hit / venture / speed / **ion** / prior / **alcohol** / **guide** / blend / **symposium** / segment / **virus** / label / **salad** / **cheese** / **energy** / **jogging** / floor / core / **beta** / later / **sausage** / wide / end / member / file / **guidance** / **fiber** / model

We emphasise all word related to the scientific discourse in reviewed papers (such as *keywords*, *signal*, *symposium*. . .), to the medical discourse (such as *advice*, *case*, *virus*. . .) or concerning *diabetes and nutrition* as previously detailed. It would be clumsy to draw a conclusion from these fuzzy data, although this is an encouraging clue to support our proposition, and we will have to check this observation through experiment.

5. Conclusion

We highlighted here different features of Japanese transliterations and their importance in specialised corpora. Indeed, we showed that it was a frequent phenomena (numerous transliteration relations between different language corpora) and that the vocabulary concerned by transliteration relation is likely to be relevant. Those observations make us think that transliteration can be efficiently used in the case of bilingual lexicon extraction from specialised comparable corpora. However, several issues need to be circumvent, the first one being the capacity to automatically extract and align transliterations pairs in corpora. Indeed, our first experiments using tools for transliterations detection (Tsuji et al., 2002) raised a lot of noisy results which are hard to integrate in the larger bilingual lexicon extraction process. On the other hand, using known transliteration relations is not straightforward. Several ways are to be explored: transliterations can be used to increase coverage of bilingual resources used in alignment, for SWT, or for compositional translation, which is particularly interesting since many MWT involve transliterations (Daille and Morin, 2008). Transliteration relations can also be used as an independent information to assist alignment of context vectors.

6. References

Lynne Bowker and Jennifer Pearson. 2002. *Working with Specialized Language: A Practical Guide to Using Corpora*. Routeledge, London/New York.

Béatrice Daille and Emmanuel Morin. 2008. Compositionality and lexical alignment of multi-word terms. In *Proceedings of the Third International Joint Conference on Natural Language Processing (IJCNLP'08)*, volume 1, pages 95–102.

Pascale Fung and Kathleen McKeown. 1997. Finding terminology translations from non-parallel corpora. In *5th Annual Workshop on Very Large Corpora*, pages 192–202.

Kyo Kageura. 2005. Character system, orthography and types of origin in japanese writing. In Reinhard Köhler, Gabriel Atmann, and Rajmund Piotrowski, editors, *Quantitative Linguistics: An International Handbook*, pages 935–946. Walter de Gruyter.

Jennifer Pearson. 1998. *Terms in Context*. John Benjamins publishing company.

Carol Peters and Eugenio Picchi. 1998. Cross-language information retrieval: A system for comparable corpus querying. In Gregory Grefenstette, editor, *Cross-language information retrieval*, chapter 7, pages 81–90. Kluwer Academic Publishers.

Reinhard Rapp. 1995. Identifying word translations in non-parallel texts. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics (ACL'95)*, pages 320–322, Morristown, NJ, USA.

Michel Simard, George F. Foster, and Pierre Isabelle. 1993. Using cognates to align sentences in bilingual corpora. In *CASCON*, pages 1071–1082.

Keita Tsuji, Béatrice Daille, and Kyo Kageura. 2002. Extracting french-japanese word pairs from bilingual corpora based on transliteration rules. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, pages 499–502.

Coarse Lexical Translation with no use of Prior Language Knowledge

Richard Rohwer, Zhiqiang (John) Wang

HNC Software LLC, Fair Isaac Corp
3661 Valley Centre Drive, San Diego, CA 92130, USA
richardrohwer@fairisaac.com, johnwang@fairisaac.com

Abstract

This paper demonstrates that language translation resources can be created in the complete absence of prior human knowledge of the languages concerned. We introduce an information-theoretic technique, *distributional factorization*, that produces crude lexical translations from comparable corpora without requiring any pre-existing language resource, such as corpora with any level of alignment information, or cross-language lexica. Terms and documents are simultaneously clustered so that the clusters are predictive of each other without being predictive of a prescribed clustering by language. This results in semantically coherent, mixed-language term clusters. We validate the method and study its properties using cross-language document retrieval experiments.

1. Introduction

This paper introduces an information-theoretic technique, *distributional factorization*, that is capable of producing crude translations between the words of different languages without demanding any human comprehension or prior knowledge of the languages at any stage of the process. All that is needed is a corpus in each language and the ability to tokenize these corpora. The method uses neither lexical resources nor any level of alignment information.

Section 2. draws on related work to develop a qualitative intuition for what *Distributional Factorization* does, and how cross-language information can be obtained from the statistics of non-aligned corpora without using any cross-language resources or language knowledge. This explanation is facilitated by summarizing a simpler but less capable algorithm, *transLign*. The *Distributional Factorization* algorithm itself is then defined in Section 3. by reviewing the simpler *co-clustering* algorithm and then introducing a succession of generalizations and variations. In Section 4. several exploratory cross-language document retrieval results are presented that show that the method does indeed produce a cross-language resource without drawing upon language knowledge. Section 5. concludes with a brief discussion of the future work required to fully understand the technique and its range of applications.

2. Background

The key to understanding how translation is possible with zero comprehension is to appreciate that there can be statistical relationships *within* any single language that are common *across* many if not all languages. This point was made by Rapp (Rapp, 1995), who compared the distribution of word pairs (how often any given pair of lexical entries occurs within a given window of tokens) derived from an English corpus with the distribution derived from a German corpus. The appearance of either distribution, which can be thought of as a 2-dimensional matrix of probability values[1], can be altered by re-arranging the order of presentation of the terms of the vocabulary; *i.e.*, by re-ordering the rows and columns of the matrix. By restricting attention to a vocabulary of 100 cleanly translatable English words and a corresponding 1-to-1 translated vocabulary of 100 German words, the German word-pair distribution was aligned by language translation with the English distribution. By introducing a measure of distributional similarity and corrupting the translation by varying degrees, Rapp showed that the two distributions had the most similar shape when aligned according to the correct translation, and proposed that in principle one could discover the correct translation by searching the space of all possible translations for the one that made these distributions appear most similar. However, this computation was considered infeasible, and in later work (Rapp, 1999), Rapp resorted to using a given lexicon of tie-word pairs (albeit a small one) to transport distributions between languages for use in cross-language meaning comparison, and to thereby expand the cross-language lexicon.

Although we formulate the problem rather differently in detail, in essence our approach is to forgo the use of any cross-language resource and instead do the “infeasible” search through the space of possible translations by brute force using simulated annealing.

We also relax the requirement that the rows and columns of the co-occurrence matrix both correspond to words; the rows must correspond to words but the columns can correspond to any type of data that can be statistically associated with the words using only within-language resources. As a by-product, we obtain a cross-language correspondence between the data objects corresponding to the columns. In the work reported here, for example, the columns correspond to documents, and we obtain cross-language links between documents according to similarity of topic, as well as cross-language links between words according to similarity of meaning. We refer to the data type corresponding to the columns as the *context*. Even when the contexts are words, and therefore have a trivial 1-to-1 correspondence with the rows, we do not at present use this information; instead we treat the words and contexts as separate variables that just happen to possess the same set of possible values. We expect to be able improve performance eventually by making use of this correspondence.

[1] Rapp experimented with non-linear functions of these values, but this detail does not change the essence of the argument.

Rather than seeking to discover cross-language correspondences between individual words, we seek correspondences between groups of words with similar meanings. The most straightforward way to obtain these groups is through unsupervised distributional clustering within each language separately. We use co-clustering (Dhillon et al., 2003) to simultaneously obtain groups of terms and contexts, with the same number of term clusters in each language and the same number of context clusters in each language. This way, the cluster-level co-occurrence matrices have the same size in each language, so we can proceed much as Rapp suggests by permuting the rows and columns of one of these matrices to maximize the similarity in shape of the two distributions. We call this algorithm *transLign*, and it produces sensible results.

Here we focus on a successor to *transLign* we call *Distributional Factorization* that overcomes some of its defects and is applicable to a wider class of problems. The most serious limitation of *transLign* is that it assumes 1-to-1 cross-language correspondences between the word clusters and context clusters. We can expect frequent violations of this assumption, particularly as the subject matter coverage of the two corpora is made less and less comparable. In particular, we can expect some words of one corpus to possess no close translation in the other. *Distributional Factorization* does the co-clustering and cross-language alignment simultaneously, producing a single set of semantically coherent term clusters (and context clusters) wherein each cluster can (but need not) contain members from both languages.

3. The method

The *Distributional Factorization* method can be regarded as a generalization of distributional co-clustering (Dhillon et al., 2003), which is also a helpful expository starting point. Let $p(x, y)$ be the probability that a random draw of a word from a corpus selects lexical word type x from document y. This is the density of a joint random variable (X, Y) that can be estimated by counting each occurrence of word x within document y as an event. (We simply used normalized number counts, after eliminating singleton counts, though fancier Bayesian methods might fare better. We also discarded all words and documents other than the 5000 of each for which the marginal probabilities[2] $p(x) = \sum_y p(x, y)$ and $p(y) = \sum_x p(x, y)$ are greatest.)

Let ϕ be a mapping that takes unclustered term x into a cluster $\phi(x)$, and similarly let ψ map document y into document cluster $\psi(y)$.

The mappings ϕ and ψ induce a joint distribution $p(\phi, \psi)$ over the pairs of clusters in the obvious manner[3]: $p(\phi, \psi) = \sum_{xy} p(x, y)\delta_{\phi\phi(x)}\delta_{\psi\psi(y)}$, with $\delta_{ij} = 1$ for $i = j$ and 0 otherwise. In ordinary co-clustering, one varies the maps ϕ and ψ in order to maximize (as well as practicable) the mutual information

$$I_{\Phi\Psi} = \sum_{\phi\psi} p(\phi, \psi) \log \frac{p(\phi, \psi)}{p(\phi)p(\psi)} \qquad (1)$$

where $p(\phi) = \sum_\psi p(\phi, \psi)$ is the density of cluster-valued random variable Φ and $p(\psi) = \sum_\phi p(\phi, \psi)$ is the density of cluster-valued random variable Ψ. This has the effect of organizing the terms roughly according to meaning, because meaning is characterized by usage, the usage of a term x is expressed by the conditional distribution $p(y|x) = \frac{p(x,y)}{p(x)}$ (Wang et al., 2005; Freitag et al., 2005), and clustering to maximize mutual information tends to group terms with similar conditional distributions into the same cluster. To achieve this effect, it is not necessary to cluster the documents (or whatever contexts are used to express usage) as well, but this improves computational efficiency (Rohwer and Freitag, 2004).

For all the optimization problems described in this paper, we used a modification of Simulated Annealing (Mackay, 2003) that we call *Greedy Simmering*, the complete details of which are being set out in a manuscript in preparation. Briefly, the method is much like a discrete version of Gibbs sampling, with one coordinate for each unclustered data object, the possible values of which are the clusters to which it can be assigned. But rather than assign the object to a cluster according to its Boltzmann probability conditioned on all the other cluster assignments, as Gibbs sampling would do, we truncate the distribution to the two most probable clusters. This flagrantly violates the detailed balance condition that underlies much of the theory of these methods, but experimentally produces substantially better results in substantially shorter times. Using high-end PC hardware, we can usually obtain usable results in several minutes from a fast annealing schedule, and excellent results overnight from a slow schedule. That said, the optimization technique is not an important aspect of the *Distributional Factorization* method, as long as it works fairly well. What matters is the objective function introduced below.

If co-clustering is applied to a mixed-language corpus, the terms of different languages segregate into different clusters, and so do the documents. For this reason, co-clustering alone cannot produce mixed language clusters of any description, let alone semantically unified mixtures.

In order to produce mixed-language clusters that can tie two languages lexically, one must have two *independent* co-clusterings. One, which can be pre-specified in the usual case that the languages are known, maps every term x into a cluster $\lambda = \lambda(x)$ containing all the terms of its language (a value of random variable Λ), and every document y into $\upsilon = \upsilon(y)$ containing all the documents in that language (a value of random variable Υ).[4]

[2] For brevity, we use the choice of letter for the argument of a density p to label which density is meant. A more correct notation would be $p_X(x)$ or $P(X = x)$, introducing the random variable label X to make this distinction.

[3] Here we use ϕ without an argument to designate a generic cluster and $\phi(x)$ with an argument to designate the cluster to which x is assigned, and similarly for ψ.

[4] In our experiments, each document had a known language, and all the terms in a document were regarded as terms of that language. We prefixed all terms with a language-disambiguator string so that identically-spelled words from different languages were treated as completely different words. It would defeat the scientific point of the experiments to allow any cross-language information to slip in via shared vocabulary, although this would likely be a good idea from an engineering perspective.

Then one can maximize

$$I_{\Lambda\Upsilon} + (I_{\Phi\Psi} - I_{(\Lambda,\Upsilon)(\Phi,\Psi)}) \tag{2}$$

with respect to the cluster mappings ϕ and ψ. This objective function expresses the intuition illustrated in Figure 1. Maximization of $I_{\Phi\Psi}$ expresses the usual co-clustering objective of making the word clusters Φ predictive of their context clusters (*i.e.* usages; *e.g.*, the documents that contain them) Ψ, which is illustrated by the inward pointing arrows in the figure. But the term $I_{(\Lambda,\Upsilon)(\Phi,\Psi)}$ deducts from this as much mutual information as is redundant with the mutual information obtainable by clustering according to language, as illustrated by the outward pointing double arrow. Therefore clusterings of terms Φ that make language predictable from term clusters are disfavored. Clustering Φ must therefore capture language-independent semantic information; *i.e.*, it must favor semantically-coherent (for the sake of $I_{\Phi\Psi}$), mixed-language clusters. The term $I_{\Lambda\Upsilon}$, which is not varied unless one attempts to simultaneously discover the language clusters[5] is introduced here simply to show that the objective function can be taken to be symmetric in the two co-clusterings.

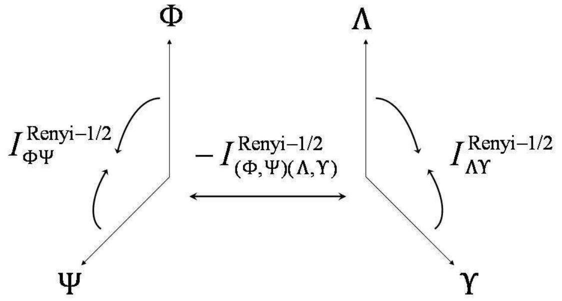

Φ
Λ
$I^{\text{Renyi-1/2}}_{\Phi\Psi}$
$-I^{\text{Renyi-1/2}}_{(\Phi,\Psi)(\Lambda,\Upsilon)}$
$I^{\text{Renyi-1/2}}_{\Lambda\Upsilon}$
Ψ
Υ

Figure 1: The *Distributional Factorization* objective function (2) (modified using Renyi information (3)) drives the term clusters Φ and document clusters Ψ to be predictive of each other, as indicated by the inward-pointing arrows, provided that this information is not also predictive of language, as indicated by the outward-pointing arrow "pushing the two co-clustering problems apart". The co-clustering of terms by language Λ and documents by language Υ would normally be specified and held fixed, although in principle it might also be learned.

Let Z be the "Cartesian product" clustering that has one cluster for every value of the pair (ϕ, λ) containing all the terms that map into both meaning category ϕ and language λ. In the case that the language assignments are to be discovered, one can think of the maximization of (2) as an effort to discover how a single random variable Z can be mapped onto the Cartesian product of a pair of random variables (Φ, Λ) that are as independent of each other as possible. That is, one discovers a "factorization" of the variable Z, which takes values over, say, nm clusters, into n clusters of one clustering plus m clusters of another, with minimal loss of information about the context (Ψ, Υ). In this sense, what *Distributional Factorization* accomplishes for discrete variables is analogous to what *Independent Component Analysis (ICA)* (Hyvarinen et al., 2001) accomplishes for continuous variables using linear mappings.

There is one final important detail. We could not achieve the desired language alignment effect by maximization of (2), despite numerous attempts. However, this problem was overcome with a simple modification. For each of the three mutual information terms in (2), we used the Renyi information of order 1/2 (Renyi, 1961) instead of the Shannon mutual information. Thus, the term $I_{\Phi\Psi}$ defined by (1) is instead defined by

$$I^{\text{Renyi-1/2}}_{\Phi\Psi} = -\log\left(\sqrt{\sum_{\phi\psi} p(\phi,\psi)p(\phi)p(\psi)}\right) \tag{3}$$

and the other two terms are modified analogously. The material difference between (1) and (3) may be that the singularity of the logarithm in (1) emphasizes differences between probability values in the numerator and denominator even when both values are quite small, whereas (3) does not. Perhaps this makes (3) more forgiving of a "loose fit" arising in language comparison.

4. Experiments

It is clear from visual inspection of the term clusters obtained that *distributional factorization* finds cross-language information, particularly for similar languages, such as English and German. Some typical example mixed-language clusters are shown in Figure 2.

To obtain a more quantitative understanding of how well the method works and how performance varies with parameters such as number of clusters, annealing rate, and languages, we conducted some cross-language document retrieval exercises. We used the Europarl JRC-Acquis 3.0 corpus [6] of European Parliament documents translated into as many as 22 European languages. We use an English document as a "query" against documents in another language, and consider the translated document to be the only correct one to retrieve. This is a relatively easy exercise due to the use of a full-length document as a query and the existence of an exact translation, yet it is adequate for our purposes, particularly for the initial exploratory experiments.

We experimented with English queries and retrieved documents in German, French and Hungarian. The JRC-Acquis corpus has 23433 corresponding documents between English and German, 23514 between English and French, and 22651 between English and Hungarian. *Distributional factorization* was performed on the most frequent 5000 words of the mixed pairwise corpora, using word-document counts and discarding all but the 5000 longest documents. Retrieval experiments were performed using

[5] This tends not to work as desired. If one really needs to discover the languages, it is better to do that first as a simple co-clustering problem. Though one normally does know the languages in advance, the technique may also be useful in problems such segmentation of a large single-language corpus into domains of an unknown nature, not necessarily related to language, and alignment of terminology between those domains.

[6] http://wt.jrc.it/lt/Acquis/

	BIN 1	(DE2 = 0.433):	
EN:licence	1.457e-05	DE:code	1.582e-05
EN:code	1.417e-05	DE:codes	1.193e-05
EN:licences	8.960e-06	DE:feld	1.144e-05
EN:codes	2.334e-06	DE:lizenz	1.017e-05
		DE:lizenzen	5.745e-06
	BIN 2	(DE2 = 0.433):	
EN:agricultural	1.305e-05	DE:lebensmittel	1.120e-05
EN:marketing	1.041e-05	DE:landwirtschaftlichen	1.079e-05
EN:crops	5.717e-06	DE:erzeugung	9.577e-06
EN:food	5.428e-06	DE:erzeuger	8.495e-06
EN:farmers	5.326e-06	DE:landwirtschaft	7.727e-06
EN:farm	4.528e-06	DE:landwirtschaftliche	6.204e-06
EN:feed	2.760e-06	DE:vermarktung	4.104e-06
EN:farming	2.629e-06	DE:landwirtschaftlicher	3.603e-06
EN:varieties	2.196e-06	DE:betriebe	3.249e-06
EN:agriculture	1.490e-06	DE:obst	2.342e-06
		DE:verwaltungsausschusses	2.771e-07
	BIN 200	(DE2 = 0.070):	
EN:accounts	1.646e-05	DE:euro	2.760e-05
EN:euro	1.606e-05	DE:transaktionen	1.219e-05
EN:transactions	1.160e-05	DE:hof	8.021e-06
EN:accounting	8.786e-06	DE:mwst	5.480e-06
EN:income	7.319e-06	DE:generaldirektion	4.276e-06
EN:cash	4.237e-06	DE:erwerb	4.115e-06
EN:vat	4.028e-06	DE:ausgewiesen	3.839e-06
EN:currency	3.449e-06	DE:garantie	3.588e-06
EN:directorate	1.463e-06	DE:buchf	3.032e-06
		DE:dollar	6.578e-07

Figure 2: The best and worst term clusters from a *Distributional Factorization* of the most frequent 5000 terms and longest 5000 documents of the combined English and German parts of the Europarl JRC-Acquis corpus into 200 clusters of each. The clusters and their members are scored and sorted according to their contribution to the objective function.

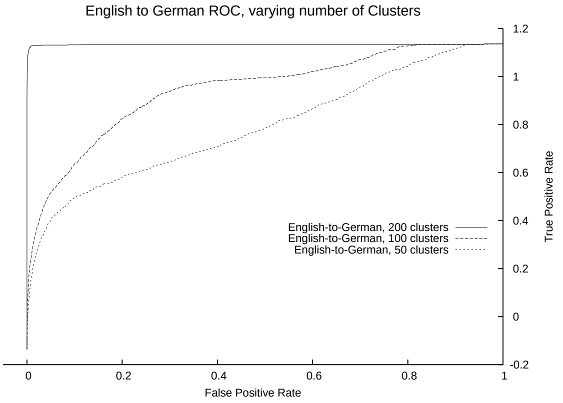


Figure 3: ROC curves for retrieval of the German translation of English documents in the Europarl-JRC corpus, using 50, 100, and 200 term clusters.

the 3000 longest documents. Each document was converted to a bag-of-mixed-language-clusters representation, and these were compared by Hellinger distance (Amari and Nagaoka, 1993):

$$D\left[p(X|\text{doc}), p(X|\text{doc}')\right] = 1 - \sum_{X} \sqrt{p(X|\text{doc})p(X|\text{doc}')}. \qquad (4)$$

Figure 3 shows the resulting ROC (receiver operating characteristic) curve for retrieving German from English using 50, 100, and 200 mixed-language clusters. Performance improves with increasing numbers of clusters, as one would expect. The F1[7] values are 0.070, 0.110, and 0.823 respectively.

Figure 4 shows the ROC curves for retrieving the 3 different languages from English, using 200 clusters. The more closely related languages to English, German and French, do best (F1=0.823 and F1=0.881 respectively) and much better than Hungarian (F1=0.146), though even for Hungarian it is obvious that cross-language information has been captured.

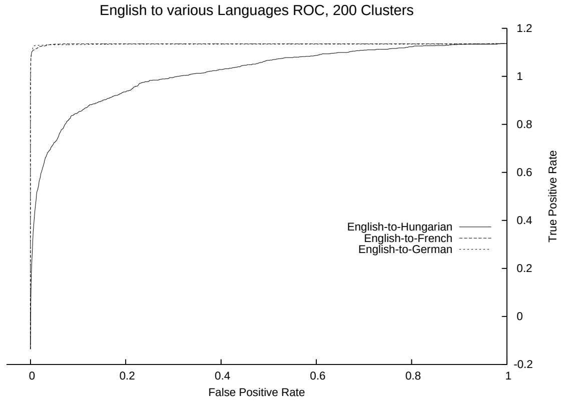


Figure 4: ROC curves for retrieval of Hungarian, French, and German translation of English documents in the Europarl-JRC corpus, using 200 clusters.

Figure 5 shows the effect of using a cluster-training corpus different from the retrieval corpus. For the most finely dotted curve at the upper left, the retrieval corpus is the JRC-Acquis corpus used throughout, and this corpus was also used to train the mixed-language clusters. Performance degenerates slightly to the dashed curve at the upper left when the retrieval corpus is changed to the European Parliament Proceedings Corpus[8] of parallel English and German articles, still using the mixed-language clusters trained from the JRC-Acquis corpus. Training the clusters on a corpus of English and German Wikipedia articles but doing cross-language retrieval in the JRC-Acquis corpus produces the lowermost dotted curve. Both curves demonstrate that it is not necessary to train the mixed-language clusters on the same corpus as is used in the retrieval experiments (although F1 drops to 0.662 and 0.054, respectively). The solid curve shows the effect of more rapid annealing (by about a factor of 10, reducing a run of around 5 hours to around 0.5 hours), which still works but drops F1 from 0.823 to 0.092.

5. Conclusions

Although this work is at an early exploratory stage, it clearly demonstrates that no input of prior human language knowledge is necessary in order to generate a cross-language resource from non-aligned, comparable corpora. The *Distributional Factorization* method introduced here accomplishes that, as is clear from cursory inspection of the mixed language term clusters it produces, such as are illustrated in Figure 2, and from the obviously better than random performance shown by the ROC curves in Figures 3,

[7] 2 (True Pos.) /(2 (True Pos.) + (True Neg.) + (False Neg.))

[8] http://www.statmt.org/europarl

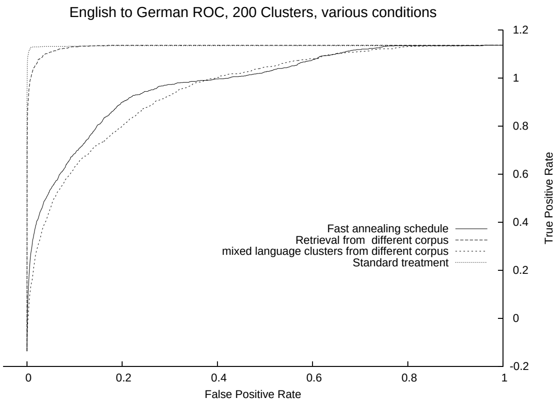


Figure 5: ROC curves for retrieval of the German translations of English documents in the Europarl-JRC or Europarl Proceedings corpora, using 200 mixed-language clusters obtained from the Europarl-JRC or Wikipedia corpora, respectively. Also shown is the retrieval from Europarl-JRC using clusters from Europarl-JRC trained with a rapid annealing schedule.

4 and 5. However, much work remains to be done in order to accurately determine the efficacy of the method for languages of interest, to explore its parameters, and to investigate the obvious variants of its objective function (2, 3).

It is not entirely clear why performance degrades as the languages become less related (Figure 4), because only term-document statistics were used in these experiments, and the documents are translations of each other. Perhaps there is more ambiguity in the possible translations. Another possibility is that increasingly innequivalent morphology is involved; we made no attempt at lemmatization, instead opting for a trivial tokenizer that does little more than case normalization and separation of strings by white space and punctuation.

One might be concerned that training on a corpus of documents that are direct translations of each other somehow trivializes the exercise (though it is not obvious how), but the result in Figure 5 with training on a corpus that is neither directly translated nor precisely topic-parallel alleviates this concern. However, much more experimentation on a wider array of problems is needed to form a clear picture of how far one can go with this zero-knowledge approach.

One obvious avenue for improvement of the method is to generalize away from hard clustering to a soft clustering approach such as mixture modeling. This would preserve more information, but with increased computational cost and complexity over what is already a computationally intensive method, for relatively small expected gains. Therefore we consider exploration of the capabilities of the current method to be a higher priority.

As noted in Section 2., it may be relatively simple to boost performance when the context data type is chosen to be the same as the word data type, simply by incorporating the constraint that the contexts be permuted in lockstep with the words rather than completely independently, as at present.

It seems plausible that this technique could seed a recursive bootstrap procedure with an initial lexical alignment that could then be employed to obtain an initial corpus alignment, then an improved lexical alignment, etc. Indeed, iterative, coordinated improvement of lexicon and corpus alignments, starting from a given seed lexicon, is already a highly developed art (Wu and Fung, 2005; Dragos Stefan Munteanu and Daniel Marcu, 2005).

The technique may also have applications within a single language, such as bridging dialects or discovering analogies and metaphors. For example, in *Structural Correspondence Learning* (Blitzer et al., 2006) part-of-speech assignments known for data in one domain of a single language are used to learn part-of-speech tags in another domain with much non-overlapping vocabulary by using carefully selected words in the common vocabulary that are referred to as “pivot features”. Although these are the same words in each domain, they function essentially as do tie-words between different languages. *Distributional Factorization* may provide an elegant way to automatically find these tie-words, and to generalize the concept so that the same pivot feature can be represented by different words in different domains.

This technique can be applied even if not a single word of one (or both!) of the languages involved is understood by anyone. This tells us, interestingly, that at least some of the meaning of the words of a language is implicit in the statistical relationships amongst the words themselves. It should not be difficult to incorporate known tie-words as constraints, thereby improving performance, but there is also risk involved relying on such *a priori* knowledge, because new meanings are often attached to old words, particularly as slang usage evolves.

It is clear that *Distributional Factorization* uncovers purely statistical structure that carries information common to multiple languages. It remains to probe the limits of this type of information, and to discover how to best use it in conjunction with other available information for language translation and more diverse applications involving any form of vocabulary alignment.

6. Acknowledgments

This work is being performed under Air Force contract FA8750-06-C-0200 with AFRL Information Directorate.

7. References

S. Amari and H. Nagaoka. 1993. *Methods of Information Geometry*. Oxford University Press and the American Mathematical Society, Providence, RI.

John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. Domain adaptation with structural correspondence learning. In *Conference on Empirical Methods in Natural Language Processing*, Sydney, Australia.

I. S. Dhillon, S. Mallela, and D. S. Modha. 2003. Information-theoretic co-clustering. In *Proc. 9th ACM SIGKDD International Conf. on Knowledge Discovery and Data Mining (KDD2003)*, pages 89–98.

Dragos Stefan Munteanu and Daniel Marcu. 2005. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31(4):477–504.

Dayne Freitag, Matthias Blume, John Byrnes, Edmond Chow, Sadik Kapadia, Richard Rohwer, and Zhiqiang Wang. 2005. New experiments in distributional representations of synonymy. In *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, pages 25–32, Ann Arbor, Michigan, June. Association for Computational Linguistics.

A. Hyvarinen, J. Karhunen, and E. Oja. 2001. *Independent Component Analysis*. Wiley.

David J. C. Mackay. 2003. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press.

Reinhard Rapp. 1995. Identifying word translations in non-parallel texts. In *Proc. of the 33rd ACL*, pages 320–322, Cambridge, MA.

Reinhard Rapp. 1999. Automatic identification of word translations from unrelated English and German corpora. In *Proc. of the 37th ACL*, pages 395–398, Maryland.

A. Renyi. 1961. On measures of information and entropy. In *Proc. 4th Berkeley Symposium on Mathematics, Statistics and Probability*, volume 1960, pages 547–561.

Richard Rohwer and Dayne Freitag. 2004. Towards full automation of lexicon construction. In *Lexical Semantics Workshop, HLT-NAACL*, Boston, MA.

Zhiqiang Wang, Ed Chow, and Richard Rohwer. 2005. Experiments with grounding spaces, pseudo-counts, and divergence formulas in association-grounded semantics. In *2005 International Conference on Intelligence Analysis*, McClean, VA.

Dekai Wu and Pascale Fung. 2005. Inversion transduction grammar constraints for mining parallel sentences from quasi-comparable corpora. In *Proc. 2nd Intl. Joint Conf. on NLP (IJCNLP)*, Jeju, Korea.