# Workshop Programme

**Saturday, May 22, 2010**

9:00-9:15 Opening Remarks

**Invited talk**

9:15 *Comparable Corpora Within and Across Languages, Word Frequency Lists and the KELLY Project*
Adam Kilgarriff

10:30 Break

**Session 1: Building Comparable Corpora**

11:00 *Analysis and Evaluation of Comparable Corpora for Under Resourced Areas of Machine Translation*
Inguna Skadiņa, Andrejs Vasiļjevs, Raivis Skadiņš, Robert Gaizauskas, Dan Tufiş and Tatiana Gornostay

11:30 *Statistical Corpus and Language Comparison Using Comparable Corpora*
Thomas Eckart and Uwe Quasthoff

12:00 *Wikipedia as Multilingual Source of Comparable Corpora*
Pablo Gamallo and Isaac González López

12:30 *Trillions of Comparable Documents*
Pascale Fung, Emmanuel Prochasson and Simon Shi

13:00 Lunch break

**Saturday, May 22, 2010 (continued)**

**Session 2: Parallel and Comparable Corpora for Machine Translation**

14:30 *Improving Machine Translation Performance Using Comparable Corpora*
Andreas Eisele and Jia Xu

15:00 *Building a Large English-Chinese Parallel Corpus from Comparable Patents and its Experimental Application to SMT*
Bin Lu, Tao Jiang, Kapo Chow and Benjamin K. Tsou

15:30 *Automatic Terminologically-Rich Parallel Corpora Construction*
José João Almeida and Alberto Simões

16:00 Break

**Session 3: Contrastive Analysis**

16:30 *Foreign Language Examination Corpus for L2-Learning Studies*
Piotr Bański and Romuald Gozdawa-Gołębiowski

17:00 *Lexical Analysis of Pre and Post Revolution Discourse in Portugal*
Michel Généreux, Amália Mendes, L. Alice Santos Pereira and M. Fernanda Bacelar do Nascimento

17:30 *From Language to Culture and Beyond: Building and Exploring Comparable Web Corpora*
Maristella Gatto

**Panel Session**

18:00 A Roadmap for Comparable Corpora

19:00 End of Workshop

**Organisers:**

Reinhard Rapp (University of Tarragona, Spain)
Pierre Zweigenbaum (LIMSI-CNRS, France)
Serge Sharoff (University of Leeds, UK)

**Invited Speaker:**

Adam Kilgarriff (Lexical Computing Ltd, UK)

**Program Committee:**

Srinivas Bangalore (AT&T Labs, USA)
Caroline Barrière (National Research Council Canada)
Chris Biemann (Microsoft / Powerset, San Francisco, USA)
Lynne Bowker (University of Ottawa, Canada)
Hervé Déjean (Xerox Research Centre Europe, Grenoble, France)
Kurt Eberle (Lingenio, Heidelberg, Germany)
Andreas Eisele (DFKI, Saarbrücken, Germany)
Pascale Fung (Hong Kong University of Science & Technology, China)
Éric Gaussier (Université Joseph Fourier, Grenoble, France)
Gregory Grefenstette (Exalead, Paris, France)
Silvia Hansen-Schirra (University of Mainz, Germany)
Hitoshi Isahara (NICT, Tokyo, Japan)
Kyo Kageura (University of Tokyo, Japan)
Min-Yen Kan (National University of Singapore)
Adam Kilgarriff (Lexical Computing Ltd, UK)
Natalie Kübler (Université Paris Diderot, France)
Philippe Langlais (Université de Montréal, Canada)
Tony McEnery (Lancaster University, UK)
Emmanuel Morin (Université de Nantes, France)
Dragos Stefan Munteanu (Language Weaver Inc., USA)
Carol Peters (ISTI-CNR, Pisa, Italy)
Emmanuel Prochasson (Hong Kong University of Science & Technology, China)
Reinhard Rapp (University of Tarragona, Spain)
Sujith Ravi (ISI, University of Southern California, USA)
Serge Sharoff (University of Leeds, UK)
Michel Simard (National Research Council Canada)
Richard Sproat (OGI School of Science and Technology, USA)
Michael Zock (LIF, CNRS Marseille, France)
Pierre Zweigenbaum (LIMSI-CNRS, Orsay, France)

Proceedings assembled with LaTeX using the ACLPUB package.

# Preface

Comparable corpora are collections of documents that are comparable in content and form in various degrees and dimensions. This definition includes many types of parallel and non-parallel multilingual corpora, but also sets of monolingual corpora that are used for comparative purposes. Research on comparable corpora is active but used to be scattered among many workshops and conferences. The workshop series on "Building and Using Comparable Corpora" (BUCC) aims at promoting progress in this exciting emerging field by bundling its research, thereby making it more visible and giving it a better platform.

Following the two previous editions of the workshop which took place at LREC 2008 in Marrakech and at ACL-IJCNLP 2009 in Singapore, this year the workshop was co-located with LREC 2010 in Malta. With the workshop's theme being "Applications of Parallel and Comparable Corpora in Natural Language Engineering and the Humanities" the focus was on bringing together researchers from different disciplines, thereby giving an indication of the breadth of research taking place in this field.

We would like to thank all people who in one way or another helped in making this workshop a success. Our special thanks go to Adam Kilgarriff for accepting to give the invited presentation, to the participants of the panel discussion, to the members of the program committee who did an excellent job in reviewing the submitted papers, and to the LREC organizers. Last but not least we would like to thank our authors and the participants of the workshop.

Reinhard Rapp, Pierre Zweigenbaum, Serge Sharoff

# Table of Contents

# Author Index

# Comparable Corpora Within and Across Languages, Word Frequency Lists and the KELLY Project

**Adam Kilgarriff**

Lexical Computing Ltd
Brighton, UK
adam@lexmasterclass.com

**Abstract**

Word frequency lists play a pivotal role as we explore and exploit comparable corpora. They form a compact summary of what is in a corpus. They also make it possible to assess how similar two corpora are, and how they contrast with each other. They are also widely used by educators, psychologists and publishers in their own right. In the recently-started EU project KELLY, we are exploring these issues across nine languages, including starting from loosely comparable corpora across languages. The paper describes how word frequency lists can be developed from corpora, and how they might be used, complete with plans and experiences from Kelly.

## 1. Measuring comparability

What makes comparable corpora 'comparable'? They should have roughly the same text type(s), covering the same subject matter, in the same proportions. Given that definition, comparable corpora may be of the same or different languages.

In 2003 Maia could not help but conclude that "comparability is in the eye of the beholder" (Maia, 2003). This is not a satisfactory state of affairs: we do not want the sampling for the datasets underlying our scientific endeavour to be subjective. We could avoid subjectivity if we could make measurements. We would like to be able to measure how comparable, or similar, two corpora are.

Then it becomes useful that the definition of comparability (or, hereafter, similarity) relates equally to same-language and different-language corpora. It gives us a reference point: any corpus is entirely similar to itself. It also gives us some history in quantitative comparison of same-language corpora.

(Biber, 1988) opened the field, showing how corpus counts could be used for the systematic study of contrasts between language varieties. For him the object of the study was the differences between the text types, rather than calibrating differences between corpora. I explored the calibration question in (Kilgarriff, 2001). At the time the issue was of largely theoretical interest as corpus users tended to be beggars not choosers: most corpus users were using any corpus of approximately the right type that they could lay their hands on, options being few and far between.

Since then we have had BootCaT and the 'web as corpus' strategy, making it possible to quickly and cheaply build a corpus to a specification (Baroni and Bernardini, 2006). In that model, once you have built a corpus the overriding questions are "is it what I wanted? What kind of a corpus (in terms of text types, subject matter, proportions) is it?" The collection strategy may have been more, or less, succesful in gathering what was wanted, and will probably have picked up some things that were not wanted along the way, so the builder wants to evaluate the corpus.

The simplest place to start is a word frequency list.

## 2. Word Frequency Lists

Word frequency lists can be seen from several perspectives. For computational linguistics or information theory, they are also called unigram lists and can be seen as a compact representation of a corpus, lacking much of the information in the corpus but small and easily tractable.

Psychologists exploring language production, understanding, and acquisition are interested in word frequency, as a word's frequency is related to the speed with which it is understood or learned so frequency needs to be allowed for in choosing words to use in psycholinguistic experiments. Educationalists are interested too, so frequency can guide the curriculum for learning to read and similar. To these ends, Thorndike and Lorge prepared *The Teacher's WordBook of 30,000 words* in 1944 by counting words in a corpus, creating reference set used for many studies for many years (Thorndike and Lorge, 1944). It made its way into English Language Teaching via West's General Service List (West, 1953) which was a key resource for choosing which words to use in the English Language Teaching curriculum until the British National Corpus[1] replaced it in the 1990s.

In language teaching word frequency lists are used for:

- defining a syllabus
- deciding which words are used in
  - learning-to-read books for children
  - textbooks for non-native learners
  - dictionaries
  - language tests for non-native learners.

### 2.1. Creating word frequency lists

There are three ways to get a word list: copy, guess, or count.

[1] http://natcorp.ox.ac.uk



*Proceedings of the 3rd Workshop on Building and Using Comparable Corpora, LREC 2010*, pages 1–5
Malta, 22 May 2010

Most word lists for most languages have used the first and the second. Where there are no corpora available, this is forgiveable. In 2010, this is no longer an excuse for any medium-sized or larger language. Principled word lists must be based on corpora.

Following on from Thorndike and Lorge, in the 1960s Kučera and Francis developed the Brown Corpus, a carefully compiled selection of current American English of a million words drawn from a wide variety of sources (Francis and Kučera, 1982). They undertook a number of analyses of it, touching on linguistics, psychology, statistics, and sociology. The corpus has been very widely used in all of these fields. The Brown Corpus is the first modern English-language corpus, and a useful reference as a starting-point for the sub-discipline of corpus linguistics (from an English-language perspective).

While the Brown Corpus was being prepared in the USA, in London the Survey of English Usage was under way, collecting and transcribing conversations as well as gathering written material. It was used in the research for the Quirk *et al* Grammar of Contemporary English (Quirk et al., 1972), and was eventually published in the 1980s as the London-Lund Corpus, an early example of a spoken corpus.

My personal involvement in word lists came about when, in 1994 and 1995 I counted the words in the (then new) British National Corpus, the first time for inclusion in LDOCE3 (LDOCE3, 1995; Kilgarriff, 1997), and the second, for the world at large, putting them on the web. The web version has been used and used, for example as the source of the JCET 8000 which defines the English syllabus in Japan. So people have come to think of me as an expert on word lists. The work described below is an attempt to live up to that cheaply-earned reputation.

There are various steps in getting from corpus to high-quality word list, as spelt out below.

### 2.2. Core and sublanguage

A language consists of core vocabulary and sublanguages. Core vocabulary is used across the board, sublanguage vocabulary changes according to what is being talked about (and in what genre) so will be different from corpus to corpus. My suspicion is that the core is quite small. When preparing word frequency lists, one strategy is to, firstly, identify the core, and secondly, decide which sublanguages are privileged, in the context of, for example, language learners: perhaps sublanguages like family relationships (*brother, sister, uncle aunt* etc) and body parts (*eye ear nose throat wrist shoulder* etc).

### 2.3. What is a word

In English, a (textual) word is, to a first approximation, an item found between spaces comprising a-z characters. English is a particularly easy language here. Chinese does not have spaces between the words at all, Arabic (and, to a lesser extent Italian) often incorporates pronouns, articles and other grammatical items into the same space-delimited object. Swedish, Norwegian, German and Dutch have compounding and separable verbs.

### 2.4. Words and lemmas

In texts we find word forms (*invade invading invades invaded*) whereas in dictionaries we find lemmas, also called dictionary headwords: just *invade*. Word lists for educators should be lists of lemmas. To get from word forms to lemmas is the process of lemmatisation: not needed at all for Chinese (which has no inflections), simple for English, middling for Italian, Greek, Norwegian and Swedish, and very complex for Russian, Polish and Arabic.

### 2.5. Grammatical classes

English *brush* can be a noun or a verb. Should the noun and the verb be counted as separate for purposes of the word list, or as a single item? Some dictionaries treat them as separate headwords, others as the same. Languages also vary: Chinese has a weak sense of word class so for Chinese, giving different noun and verb entries is less appealing as it may force decisions as to whether a word is a noun, a verb, or both. English has a lot of freedom for using nouns as verbs and *vice versa*, but, in context, there is usually a right answer as to whether a word is being used as a noun or verb (or adjective; for *-ed* and *-ing* forms this becomes difficult).

If the word list is to distinguish different word classes, we shall need a taxonomy of word classes for the language. It is desirable that this is the same for each language except where there is a good linguistic reason why it cannot be. The work done in EAGLES and associated projects presents an approach for this task (EAGLES, 1996).

### 2.6. Non-central word types

There are various marginal classes of word:

- numbers, ordinals, fractions
- names (of people, places of various kinds, organisations)
- countries, currencies, nationalities, languages, ethnic groups, religions and philosophies and their adherents (nouns and adjectives)
- days of week, months, decades, festivals
- abbreviations, initials, acronyms
- informal, slang, offensive language
- dialect words, regional variants

Decisions will be required on what to include.

### 2.7. Multiwords

English *according to* is, from a linguistic point of view, a word, but is written with a space. Let us call all such items multiwords. (This does not relate to Chinese or Japanese as they are not written with spaces between words at all.) Big classes of multiwords for English are phrasal verbs, compound prepositions and compound nominals. Linguistically, word lists should contain multiwords but, unlike simple words, we cannot easily count them. If we count

| ANW | | | NlWaC | | |
|---|---|---|---|---|---|
| **Theme** | **Word** | **English gloss** | **Theme** | **Word** | **English gloss** |
| Belgian | Brussel | (city) | Religion | God | |
| | Belgische | Belgian | | Jezus | |
| | Vlaamse | Flemish | | Christus | |
| Fiction | Keek | Looked/watched | | Gods | |
| Newspapers | vorig | previous | Web | http | |
| | kreek | watched/looked | | Geplaatst | posted |
| | procent | Percent | | Nl | (Web domain) |
| | miljoen | million | | Bewerk | edited |
| | miljard | billion | | Reacties | Replies |
| | frank | (Belgian) Franc | | www | |
| | Zei | said | English | And | In book/film/song |
| | aldus | thus | | The | titles, names etc |
| | Meppel | City with local newsp | History | Arbeiders | workers |
| | gisteren | yesterday | | Dus | thus |
| | Foto | Photo | | Macht | power |
| | Auteur | Author | | Oorlog | war |
| | Van | (in names) | | Volk | people |
| Pronouns | Hij | Him/he | Pronouns | We | we |
| | haar | She/her(/hair) | | Ons | us |
| | Ze | (They/them) | | Jullie | you |

Table 1: Keywords in ANW and NlWaC

all two-word strings in an English corpus the commonest is *of the* but no-one wants that in their wordlist. Very many common two word strings are not multiwords. So, if we use a direct strategy for including multiwords in a wordlist, we are back to copying or guessing.

### 2.8. Homonymy

The English noun *bank* can be the side of a river or a financial institution. Should these count as two separate items in a frequency list?

Every different dictionary makes different decisions about what is to count as a separate meaning so if we try to build homonymy into word lists, we shall introduce some arbitrariness.

## 3. Contrasting corpora

Word frequency lists as compact representations of corpora, and word lists for use by educators may seem very different things, but if the latter do not in some way come from the former we are either copying or guessing. A word frequency list is only of value for educators if it is based on 'the right corpus', which throws us back on the question of how we might assess corpora.

We assess a corpus by comparing its word frequency list with the list from another corpus. While other approaches are possible (for example, measuring cross-entropy between the corpora) it is harder to interpret their outcomes.

The simplest strategy is to compare the top ten, or top twenty, words in the two lists. Often, many of them are the same, and it is not clear whether there are interesting differences between the words that are in a different position in the two lists.

A better method is to identify the words that are most different in their frequencies between the two corpora: the keywords of each with respect to the other. To do this we

- normalise frequencies to per-million
- for each word, calculate the ratio between normalised frequencies in the two corpora
- sort by ratios
- the top and bottom items are the keywords (of the first corpus versus the second, and vice versa).

We can make the scheme more flexible, and address the fact that we cannot compute a ratio against zero, by adding a constant to all normalised counts before computing ratios. The higher the constant, the more the frequency list will focus on higher-frequency items, as shown in (Kilgarriff, 2009).

Provided the lists are prepared in uniform ways in relation to tokenization, lemmatisation etc., an examination of the keywords will allow us to rapidly identify the main contrasts between two corpora. We used this method to compare a Dutch web corpus, NlWaC, with the ANW corpus, a balanced corpus of 100 million words built to support the lexicography for the ANW, a major new dictionary of Dutch. It comprises: present-day literary texts (20%), texts containing neologisms (5%), texts of various domains in the Netherlands and Flanders (32%) and newspaper texts (40%).

The twenty highest-scoring (ANW) keywords and the twenty lowest-scoring (NlWaC) keywords, with English glosses and clustered by themes, are given in *Table* 1.

The classification into themes was undertaken by checking where and how the words were being used. The analysis shows that these two large, general corpora of Dutch have different strengths and weaknesses, and different areas that might be interpreted as over-representation or under-representation. The ANW has a much stronger representation of Flemish (the variety of Dutch spoken in Belgium). It has 20% fiction: *keek* (looked, watched) is used almost exclusively in fiction. It is 40% newspaper and newspapers talk at length about money (which also interacts with time and place: franks were the Belgian currency until 1999; also the units were small so sums in franks were often in millions or even billions). There is a particularly large chunk from the Meppel local newspaper. Most occurrences of *foto* were in "Photo by" or "Photo from" and of *auteur*, in newspaper by-lines, which might ideally have been filtered out. Daily newspapers habitually talk about what happened the day before, hence *gisteren*.

NlWaC has a large contingent of religious texts. It is based on Web texts, some of which could have been more rigorously cleaned to remove non-continuous-text and other non-words like URL components *www, http, nl*. The English might appear to be because we had gathered mixed-language or English pages but when we investigated, we found most of the instances of *and* and *the* were in titles and names, for example "The Good, the Bad and the Ugly", where the film was being discussed in Dutch but with the title left in English.

This analysis (also in (Kilgarriff et al., 2010)) is presented here to illustrate how we can assess how 'comparable' same-language corpora are.[2]

## 4. The KELLY Project

KELLY is an EU Lifelong Learning Project with the goal of developing language-learning cards, with a word in one language on one side and its translation on the other. The languages involved are Arabic, Chinese, English, Greek, Italian, Norwegian, Polish and Swedish. In the past, tools of this kind have rarely been corpus-based, or even corpus-informed. In Kelly we hope to be able to prepare high-quality lists which are fully corpus-based.

The method is as follows ('lempos' is shorthand for lemma plus part of speech; our lists will be lists of lemposes):

- prepare (tokenised, lemmatised, POS-tagged) corpora
- Generate lempos-lists (call these M1 lists, for monolingual first-stage lists)
- Study keywords lists from different corpora; review and fix anomalies to give M2 lists
- Translate into all eight other languages, to give T1 (first Translated) lists
- Review candidate additions to lists
- Review and finalise monolingual lists and bilingual lists for word cards (M3, T2 lists)

We hope that omissions and failings of the M2 list for a language might be rectified by the set of translations of lists from eight other languages into that language. In particular, although the M2 list will not include multiwords, multiwords are, by definition, akin to a single word linguistically so one can expect them to have single-word equivalents in other languages, so they are likely to feature as translations. We expect to acquire many items to add to M2 lists in this way, to give M3 lists.

At time of writing M2 lists are being finalised.

We wished to use comparable corpora for each language for preparing M1 lists. The only type of large, general corpus that we could obtain for all languages was a BootCaT-style web corpus. (For Swedish, where we did not know of any such corpus, we prepared one (Kilgarriff et al., 2010).)

To get from M1 lists to M2 lists, which can reasonably be presented to translators, a gamut of issues have been encountered. Junk needed deleting. POS-taggers and lemmatisers made many errors. The most heated debates at our initial project meeting related to multiwords and homonymy, with the one argument being that lists including multiwords and homonymy decisions would include a large dose of arbitrariness, and the counter-argument being that the eight translators-out-of-English needed guidance, to know, for example, that the English noun *mean* occurred in the M2 list because of its occurrence in *by means of*. For homonyms, how were the eight translators to know whether to translate *money bank* or *river bank*? Consortium members for different languages have adopted slightly different strategies on these issues, each according to their own perspective.

A further problem relates simply to the text type mix of web corpora. A recent email thread was titled *alphabet, orange, banana and elbow*: Swedish equivalents of these words were not in the top-6000 list, yet they were basic vocabulary. Responses have included:

- for English and Norwegian, corpora of conversational speech were available and have been used as comparison corpora, so words such as these have entered the M2 lists via that route if not otherwise
- if the words are there in the M2 list for one language, it is likely they will percolate across to all languages
- we may do further checking of lists against textbook vocabularies
- we may allow addition of items simply because the person preparing the list knows they should be there!
- I am not sure that *elbow* is such a common word in any text type, but there is nonetheless an argument for including it as a body-part term: as noted above, some domains are privileged from a language-learning perspective. (It is part of the project's agenda to relate word cards to the language levels as defined in the Common European Framework (CEF, 2010). The

[2] See (Kilgarriff, 2001) for global figures of how similar two corpora are; a drawback of these figures is that they can only be used to compare similarity scores between two or more pairs of corpora, and cannot be interpreted in isolation.

CEF makes explicit reference to some thematic areas including food and drink, and health and body care.)

The list of words added to the English M2 list from the English conversational speech corpus started with *yeah mum dad okay sorry hello dear*. We fear we underestimated the mismatch between web corpus frequencies and frequencies from everyday language use, as required by a learner.

### 4.1. Frequencies and points

Where the person looking at keywords lists decides that a word needs adding to the list, or a word has too high or too low a score, how should they implement it? The word cards are, in due course, to be divided into six levels (of 1500 words each) so we need to retain order information. The list is, initially, a frequency list, so should the person make up a frequency that puts it in a position that they judge to be appropriate?

Making up frequencies feels monstrous. We have a slightly less bad variant: first translate frequencies into points, then promote or demote words by adding or subtracting points.

The initial list is of 6000 items: the top 500 get twelve points, the next 500, eleven, the next 500, ten, and so on. When a word is introduced into the list from the top of a spoken conversation list, it will be introduced with twelve points; ones introduced from lower down the spoken list may be introduced with a smaller number of points. Words found to be entirely absent in the spoken corpus can be demoted, say, four points.

We begin with each band containing 500 items, but that will not stay true. If, at some point, we need to specify the top 1500 items, we can use frequency in the web corpus as a second level of sorting for words with the same number of points.

While the strategy makes no claim to objectivity, it provides a framework for systematic amendment of a starter list.

### 4.2. The Translations database

All translations will be entered in a database. With translations of 6000 items for each of nine languages into all eight other languages, it will be a large and rich resource.

There are just 6000 words in M2 lists as against 9000 word cards for each language pair eventually required. We anticipate making up the difference from "back translations": words and multiwords which were not in M2 but do occur as translations from other languages. In addition to the 6000 M2 words for a language, there will be up to $6000 \times 8 = 48,000$ additional items: most will overlap with the M2 list and each other, and it remains to be seen how many are useful. We envisage adding items according to rules such as:

> if a multiword or word not in M2 occurs more than once as a translation (either as the translation of equivalent terms from two different other languages, or otherwise) then it is a candidate for inclusion.

## 5. Summary

Word frequency lists play a pivotal role as we explore and exploit comparable corpora. They form a compact summary of what is in a corpus, and make it possible to assess where two corpora of the same language are comparable, and how they contrast with each other. They are also widely used by educators, psychologists and publishers in their own right. In the recently-started EU project KELLY, we are exploring the preparation of word lists from corpora across nine languages, including starting from loosely comparable corpora across languages and the large-scale translation of lists. We hope to shed light on how we might measure comparability between corpora across, as well as within, languages in due course.

## Acknowledgments

This work was supported by the EU Lifelong Learning Project KELLY. Much of the work discussed has been undertaken together with members of the KELLY team.

## 6. References

Marco Baroni and Silvia Bernardini, editors. 2006. *Wacky! Working Papers on the Web as Corpus*. Gedit, Bologna.

Douglas Biber. 1988. *Variation across speech and writing*. Cambridge University Press.

CEF. 2010. Common european framework of reference for languages. Technical report, Council of Europe.

EAGLES. 1996. Synopsis and comparison of morphosyntactic phenomena encoded in lexicons and corpora. a common proposal and applications to european languages. Technical Report EAG-CLWG-Morphsyn/R, ILC-CNR, Pisa.

N. Francis and H. Kučera. 1982. *Frequency Analysis of English Usage*. Houghton Mifflin, Boston.

Adam Kilgarriff, Siva Reddy, Jan Pomikálek, and Avinesh PVS. 2010. A corpus factory for many languages. In *Proc. LREC*, Malta.

Adam Kilgarriff. 1997. Putting frequencies in the dictionary. *Int Jnl Lexicography*, 10(2):135–155.

Adam Kilgarriff. 2001. Comparing corpora. *Int Jnl Corpus Linguistics*, 6(1):1–37.

Adam Kilgarriff. 2009. Simple maths for keywords. In *Proc. Corpus Linguistics*, Liverpool, UK.

LDOCE3. 1995. *Longman Dictionary of Contemporary English*. Longman, 3rd edition.

Belinda Maia. 2003. What are comparable corpora? In *Multilingual Corpora: Linguistic Requirements and Technical Perspectives. A Workshop on the Corpus Linguistics Conference*, Lancaster, UK.

Randolph Quirk, Sydney Greenbaum, Geoff Leech, and Jan Svartvik. 1972. *A Grammar of Contemporary English*. Longman.

E. L. Thorndike and I. Lorge. 1944. *The Teachers Word-Book of 30,000 words*.

Michael West. 1953. *A General Service List of English Words*. Longman.

# Analysis and Evaluation of Comparable Corpora for Under Resourced Areas of Machine Translation

**Inguna Skadiņa*, Andrejs Vasiļjevs*, Raivis Skadiņš*, Robert Gaizauskas†, Dan Tufiş‡ and Tatiana Gornostay***

* Tilde, Riga, Latvia
† Department of Computer Science, University of Sheffield, Sheffield, UK
‡ Research Institute for Artificial Intelligence, Romanian Academy Bucharest, Romania
inguna.skadina@tilde.lv, andrejs@tilde.lv, raivis.skadins@tilde.lv, r.gaizauskas@sheffield.ac.uk,
dan_tufis2006@yahoo.com, tatiana.gornostay@tilde.lv

**Abstract**

Lack of sufficient linguistic resources and parallel corpora for many languages and domains currently is one of the major obstacles to further advancement of automated translation. The solution proposed in this paper is to exploit the fact that non-parallel bi- or multilingual text resources are much more widely available than parallel translation data. This position paper presents previous research in this field and research plans of the ACCURAT project. Its goal is to find, analyze and evaluate novel methods that exploit comparable corpora in order to compensate for the shortage of linguistic resources, and ultimately to significantly improve MT quality for under-resourced languages and narrow domains.

## 1. Introduction

In recent decades data-driven approaches have significantly advanced the development of machine translation (MT). However, the applicability of current data-driven methods directly depends on the availability of very large quantities of parallel corpus data. For this reason the translation quality of current data-driven MT systems varies dramatically from being quite good for language pairs with large corpora available (e.g. English and French) to being barely usable for under-resourced languages and domains (e.g. Latvian and Croatian).

The problem of availability of linguistic resources is especially relevant for "smaller" or under-resourced languages. For example, one of the few parallel corpora of reasonable size for Latvian is the JRC Acquis corpus (Steinberger et al, 2006) which contains EU legislation texts. SMT trained on this corpora performs well on EU legislation documents (Koehn et al, 2009; Skadiņa and Brālītis, 2009), but it has unacceptable results for other domains.

The solution proposed in ACCURAT project and presented in this paper is to exploit the fact that comparable corpora, i.e., non-parallel bi- or multilingual text resources are much more widely available than parallel translation data.

Comparable corpora have several obvious advantages over parallel corpora – they can draw on much richer, more available and more diverse sources which are produced every day (e.g. multilingual news feeds) and are available on the Web in large quantities for many languages and domains. Although the majority of these texts are not direct translations, they share a lot of common paragraphs, sentences, phrases, terms and named entities in different languages. Expansion of Web content and massive library digitization initiatives make comparable corpora much more available than parallel corpora

The FP7 ACCURAT (Analysis and Evaluation of Comparable Corpora for Under Resourced Areas of Machine Translation) project has started on January 1, 2010. The main goal of this 2.5 year project is to find, analyze and evaluate novel methods that exploit comparable corpora in order to compensate for the shortage of linguistic resources, and ultimately to significantly improve MT quality for under-resourced languages and narrow domains.

A typical statistical MT system is based on a language model trained on monolingual target language corpus, and a translation model. Methods for the creation of translation models from parallel corpora are well studied and there are several techniques developed and widely available.

However, similar methods and techniques for non-parallel, or comparable corpora, have not been worked out thoroughly and there has been relatively little research on this subject.

This position paper presents research plans of the ACCURAT project to create a methodology and fully functional model for exploiting comparable corpora in MT, including corpus acquisition from the Web and other sources, analysis and metrics of comparability, multi-level alignment and extraction of lexical data and techniques for applying aligned text and extracted lexical data to increase the translation quality of existing MT systems.

The paper describes the state of the art in research related to use of comparable corpora for MT, presents related work regarding MT strategies and corpora use in MT, and describes the ACCURAT project goals and planned innovation.

## 2. Related Work in Corpus Use in MT

### 2.1 MT Strategies

Several approaches are used in the development of translation technologies: rule-based, statistical and example-based approaches. Cost-effectiveness is one of



*Proceedings of the 3rd Workshop on Building and Using Comparable Corpora, LREC 2010*, pages 6–14
Malta, 22 May 2010

the key reasons that the statistical paradigm has come to be the dominant current framework for MT theory and practice, as it has proven to be the most effective solution both from the point of view of time and labor resources and translation output quality.

Statistical Machine Translation (SMT) started with word-based models, but significant advances have been achieved with the introduction of phrase-based models (Koehn et al., 2003). Currently the most competitive SMT systems use phrase translation, such as the ATS (Och and Ney, 2004), CMU (Vogel et al., 2003) and IBM (Tillmann, 2003) systems. Recent work has also incorporated syntax or quasi-syntactic structures (Chiang, 2007). There are efforts to integrate in SMT systems linguistic annotation either at the word-level with factored translation models (Koehn and Hoang, 2007) or using tree-based models (Yamada and Knight, 2001, 2002; May and Knight, 2007). The proposed methods improve MT performance especially for languages with rich morphology and free word order, and help to solve such problems as long distance reordering and sentence-level grammatical coherence.

Until recently SMT research has been mainly focused on widely used languages, such as English, German, French, Arabic, and Chinese. For "smaller" languages MT solutions, as well as language technologies in general, are not as well developed due to the lack of linguistic resources and technological approaches that enable MT solutions for new language pairs to be developed cost effectively. This has resulted in a technological gap between these two groups of languages.

Although in the past few years translation services like Google Translate have started to broaden the set of translation language pairs, incorporating, e.g. the Baltic languages, translation quality lags behind significantly compared to major language pairs.

Also the EuroMatrix project[1] represents a major push in MT technology, applying the latest MT technologies systematically to all pairs of EU languages. The EuroMatrixPlus project[2] is continuing the rapid advance of MT technology, creating sample systems for every official EU language. Still these services and projects rely on available parallel corpus data.

### 2.2. Corpora Use in MT

In the area of rule-based MT systems, approaches towards using corpus-based technology for bilingual term extraction, and importing such terms into the dictionary of a rule-based system have been researched (Eisele et al., 2008).

Changes in the MT engine's process of data-driven term selection in the transfer component show that disambiguation of transfer alternatives can be significantly improved using the corpus-based data-driven techniques (Thurmair, 2006).

While SMT techniques are language independent, they require very large parallel corpora for training translation models. Translation systems trained on data from a particular domain, e.g. parliamentary proceedings, will perform poorly when used to translate texts from a different domain, e.g. news articles (Munteanu et al., 2004).

Parallel corpora remain a scarce resource covering few language pairs with too little data in only a few domains. For smaller languages parallel corpora are very limited in quantity, genre and language coverage. This remains true despite the creation of automated methods to collect parallel texts from the Web (Goutte et al., 2009; Hewavitharana and Vogel, 2008; Maia and Matos, 2008; Alegria et al., 2008; Munteanu, 2006; Munteanu and Marcu, 2005; Resnik and Smith, 2003).

The ACCURAT project goal is to overcome the bottleneck of insufficient parallel corpora for less widely used languages by extracting linguistic data from comparable corpora. Such corpora can be obtained by taking advantage of existing methods for mining the Web for similar documents or by other methods that will be explored in the project, such as mining Wikipedia.

## 3. Comparable Corpora

A comparable corpus is a relatively recent concept in MT, corpus linguistics and NLP in general. In contrast to the notion of a parallel corpus, a comparable corpus can be defined as collection of similar documents that are collected according to a set of criteria, e.g. the same proportions of texts of the same genre in the same domain from the same period (McEnery and Xiao, 2007) in more than one language or variety of languages (EAGLES, 1996) that contain overlapping information (Munteanu and Marcu, 2005; Hewavitharana and Vogel, 2008).

Examples of comparable corpora are:

- Comparable multilingual Document Collection in the Multilingual Corpora for Cooperation includes financial newspaper articles from the early 1990s in six European languages: Dutch (8.5 million words), English (30 million words), French (10 million words), German (33 million words), Italian (1.88 million words), and Spanish (10 million words).
- Bulgarian-Croatian comparable corpus (Bekavac et al. 2004) in news domain: 3,500,000 tokens (393 Kw Bulgarian; 3.1 Mw Croatian) was built from subsets of two larger newspaper corpora of respective languages from the texts selected using the same criteria (e.g., identical year, same domain etc.);
- English-Finnish-Swedish comparable corpus in news domain (University of Tampere);
- English-French-Norwegian comparable corpus in science domain (academic prose), 450 reviewed scientific papers; 3,2 million words;
- project INTERA and its four parallel sub-corpora.

These comparable corpora cannot be readily used for MT and are restricted to particular languages and certain domains. The degree of comparability of these corpora varies significantly, since texts were selected on the basis of one criterion only – topic.

[1] http://www.euromatrix.net

[2] http://www.euromatrixplus.net

Research in comparable corpora started about 15 years ago with the first works on general lexica (Rapp, 1995) and named entity translation derived from *noisy parallel corpora* (Fung, 1995). The authors supposed that the quantity of training data has an impact on the performance of statistical machine translation and a comparable corpus can compensate for the shortage of parallel corpora. This has been confirmed by other recent experiments (Munteanu and Marcu, 2005; Maia and Matos, 2008; Hewavitharana and Vogel, 2008; Goutte et al., 2009)

The latest research has also shown that adding extracted aligned parallel lexical data (additional phrase tables and their combination) from comparable corpora to the training data of an SMT system improves the system's performance in view of un-translated word coverage (Hewavitharana and Vogel, 2008). It has been also demonstrated that language pairs with little parallel data are likely to benefit the most from exploitation of comparable corpora. Munteanu (2006) achieved performance improvements of more than 50% from comparable corpora of BBC news feeds for English, Arabic and Chinese over a baseline MT system, trained only on existing available parallel data. The authors stated that the impact of comparable corpora on SMT performance is "comparable to that of human-translated data of similar size and domain".

One of the most challenging tasks is to perform alignment of comparable corpora for extraction of necessary translation data. Zhao and Vogel (2002) and Utiyama et al. (2003) extended algorithms designed to perform sentence alignment of parallel texts to apply them for comparable corpora. They started by attempting to identify similar article pairs from the two corpora. Then they treated each of those pairs as parallel texts and aligned their sentences by defining a sentence pair similarity score and use dynamic programming to find the least-cost alignment over the whole document pair. The performance of these approaches depends heavily on the ability to reliably find similar document pairs. Moreover, comparable article pairs, even those similar in content, may exhibit great differences at the sentence level (reordering, additions, etc). Therefore, they pose hard problems for the dynamic programming alignment approach.

The STRAND Web-mining system by Resnik and Smith (2003) can identify translational pairs. However, STRAND focuses on extracting pairs of parallel Web pages rather than sentences.

Muntenau and Marcu (2005) proposed a maximum entropy classifier that, given a pair of sentences, can determine whether or not they are translations of each other. This approach supposedly overcomes some of the limitations of previous approaches. Their experiments were carried out on Chinese, Arabic, and English non-parallel newspaper corpora.

ACCURAT will investigate previous multi-level alignment methods and will work on a complex approach to extract maximum linguistic data from comparable corpora for a number of under resourced languages (Croatian, Estonian, Greek, Latvian, Lithuanian and Romanian) and narrow domains. In such a way we will continue from a point reached by previous research.

# 4. ACCURAT Project

The main goal of the ACCURAT research is to find, analyze and evaluate novel methods how comparable corpora can compensate for this shortage of linguistic resources to improve MT quality significantly for under-resourced languages and narrow domains. Thus the project has the following key objectives:

- To create comparability metrics, i.e., to develop the methodology and determine criteria to measure the comparability of source and target language documents in comparable corpora.
- To develop, analyze and evaluate methods for automatic acquisition of comparable corpora from the Web.
- To elaborate advanced techniques for extraction of lexical, terminological and other linguistic data (e.g., named entities) from comparable corpora to provide training and customization data for MT.
- To measure improvements from applying acquired data against baseline results from SMT and RBMT systems.
- To evaluate and validate the ACCURAT project results in practical applications.

We will use the latest state-of-the-art in SMT and rule-based MT systems as a baseline and will provide novel methods to achieve much better results by extending these systems through the use of comparable corpora. Initial research demonstrates promising results from the use of comparable corpora in SMT (Munteanu and Marcu, 2005) and RBMT (Thurmair, 2006) and this makes us confident of the feasibility of the proposed approach.

The ACCURAT target is to achieve strong improvement in translation quality for a number of new EU official languages and languages of associated countries (Croatian, Estonian, Greek, Latvian, Lithuanian and Romanian), and propose novel approaches for adapting existing MT technologies to specific narrow domains, significantly increasing language and domain coverage of automated translation.

## 4.1. Comparability Metrics

The issue of comparability of corpora can be traced back to the origin of large-scale corpus research, when the aim was to balance the composition of a corpus to achieve representativeness (Sinclair, 1987). However we still lack definite methods to determine the criteria of comparability and comparability metrics to evaluate corpus usability for different tasks, such as machine translation, information extraction, cross-language information retrieval.

### 4.1.1. Criteria of Comparability and Parallelism

Comparability and parallelism is a complex issue, which can be applied to different levels, such as

- document collections,

- individual documents,
- paragraphs or sentences of documents.

Until now there has been no agreement on the degree of similarity that documents in comparable corpora should have, or even agreement about the criteria for measuring parallelism and comparability. There are only a few publications discussing the characteristics of comparable corpora (Maia, 2003). There have been some attempts to determine different kinds of document parallelism in comparable corpora, such as complete parallelism, noisy parallelism and complete non-parallelism, and define criteria of parallelism of similar documents in comparable corpora, such as similar number of sentences, sharing sufficiently many links (up to 30%), and monotony of links (up to 90% of links do not cross each other) (Munteanu, 2006). In addition to these criteria there have been some attempts to measure the degree of comparability according to *distribution of topics* and *publication dates* of documents in comparable corpora to estimate the *global comparability of the corpora* (Saralegi et al., 2008). ACCURAT will research criteria of comparability for different document groups with different types of parallelism, e.g., translated texts, texts on the same topic, texts on comparable topics, etc.

As we will focus on under-resourced languages and domains, some of the existing methods for detecting parallel sentences are not always applicable due to the lack of initial resources. For example, a simple word-overlap filter for comparable corpora needs sufficient parallel resources and a number of lexical resources specific to under-resourced languages and narrow domains (e.g., bilingual dictionaries, semantic lexica). ACCURAT research results could be portable to other comparable corpora in under-resourced areas resulting in a language- and domain-independent methodology.

Parallelism on the level of individual sentences will be studied in cases of rough translation equivalents, e.g., when the same event is reported in two different languages, as well as in cases of structural equivalents, e.g., when two conceptually similar events are discussed involving different entities in each language, such as names of organizations, persons, quantities or dates.

#### 4.1.2. Metrics of Comparability and Parallelism

Using defined criteria for parallelism, we would like to develop formal automated metrics for determining the degree of comparability.

Recent studies (Kilgarriff, 2001; Rayson and Garside, 2000) have added a quantitative dimension to the issue of comparability by studying objective measures for detecting how similar (or different) two corpora are in terms of their lexical content. Further studies (Sharoff, 2007) investigated automatic ways for assessing the composition of web corpora in terms of domains and genres. We will study and investigate existing measures and metrics for assessing corpus comparability and document parallelism. Different existing measurement techniques, such as counting word overlap, vector space models (including both bag of words and document structure sensitive approaches), cosine similarity, classification scores, etc. will be explored and combined. The methods of detecting similar documents and sentences in a comparable corpus will be evaluated for precision and recall.

### 4.2. Methods and Techniques for Building a Comparable Corpus from the Web

Although there are many more potential data sources for comparable corpora than there are for parallel texts, and they are easily accessible via the web, the problem of how to collect these data automatically for under resourced languages and for narrow domains poses a significant technical challenge.

We will begin with building general, i.e. non-domain specific, corpora for under-resourced languages by exploring the limits of techniques that have been developed for extracting parallel corpora for well-resourced languages – for example those exploiting URL and HTML structure, document and text chunk length and basic content matching (Resnik and Smith, 2003; Zhang et al., 2006; Shi et al., 2006). Based on preliminary investigations for the ACCURAT languages, the volume of parallel pages obtainable in this way is too low to yield satisfactory statistical MT models or to extract satisfactory lexical resources on their own. Still such pages, when they exist, are useful for seeding or supplementing lexical resources for use in searching/assembling comparable corpora. Hence we will start by building tools based on existing techniques for automatically building parallel corpora from the web for application to under resourced languages.

Given the paucity of web page pairs that are actual translations for under-resourced languages, we seek pairs of web documents that contain individual sentences which are translations or, weaker still, sentence or phrasal near equivalents. One likely source of such documents is news web sites where one news provider provides news in multiple languages (e.g. Agence France Presse, Xinhua News, Reuters, CNN, BBC). Stories on such sites may not be direct translations, but are likely to share considerable content. Munteanu and Marcu (2005) build their approach to extracting parallel sentences from comparable corpora around such sites, exploiting the LDC gigaword corpora for Chinese, Arabic and English drawn from Agence France Presse and Xinhua news. Unfortunately, none of these major news providers offer services in Croatian, Estonian, Greek, Latvian, Lithuanian, Slovenian or Romanian. However, ACCURAT will explore the underlying idea that contemporaneous news stories in multiple languages will be topically similar by crawling major national monolingual news providers and building comparable corpora of news documents. This will be done by

- restricting the news categories crawled to categories likely to contain stories shared between language communities, e.g. international news, international sporting events (Bekavac et al., 2004);

- restricting date ranges so that documents are likely to be reporting the same events (Munteanu and Marcu, 2005);
- exploring content checking during corpus collection to increase the likelihood that stories are on the same topic, e.g. presence of multiple common named entities.

One question that needs investigation is whether it is better to assemble monolingual corpora independently in multiple languages using the same constraints for each language, or, whether constraints should be established for one language and from the crawled documents meeting these constraints generate queries for other languages using cross-language IR techniques.

Aside from contemporaneous news reports in different languages, another under-exploited source of comparable texts is Wikipedia. There are now a substantial number of Wikipedia articles in each of the under-resourced languages ACCURAT aims to address (Croatian - 65 466, Estonian - 66 308, Greek - 44 173, Latvian 23 058, Lithuanian - 91 315, Slovenian – 79 289, Romanian - 414 091 articles on 24.08.2009). Many of the articles are linked to articles on the same topic in other languages.

ACCURAT will explore the selection of similar documents in multiple languages from Wikipedia. A primary approach is to crawl Wikipedia for comparable articles. The terms and multi-word units that are gathered in this crawl will then be used to seed comparable corpora searches (Bekavac and Tadić, 2008). Here we will be issuing multi-language queries to web search engines to locate such corpora. The similarity of different languages will be tested on different levels, starting from the level of headwords to the level of HTML links that form the structure of relations to other concepts worded as single-word units or multi-word units.

Another approach to generating effective searches will be to issue structured searches, looking for pages written in one language, which are linked to pages written in another language. Also learning the typical tags and text found in links between comparable articles will be examined.

In sum, we propose to explore three classes of techniques to address the problem of automatically assembling comparable corpora for under-resourced languages:

- techniques based on URL and HTML structure, geared at finding web pages which are translations of each other on multilingual sites or which point to related material in other languages
- techniques based on exploiting genre, topicality and shallow content matching to find comparable texts, e.g. news texts in the same category on the same date mentioning the same named entities are likely to report the same events
- techniques based on exploiting cross-language linkages between articles in Wikipedia both to extract comparable corpora directly from Wikipedia and as sources of terms to seed web searches to expand such corpora.

## 4.3. Techniques for Extraction of Lexical, Terminological and Other Linguistic Data from Comparable Corpora

Multi-level alignment of documents, paragraphs, sentences, phrasal units, named entities and terms for comparable corpora is much more challenging than for parallel corpora.

In parallel corpora, a source language text is translated into one or more sentences in the corresponding target language text and the order of sentences in the two texts tends to be more or less the same. Relatively simple sentence alignment algorithms (e.g., Gale and Church 1991) have proven quite successful at this task and the resulting sentence-aligned texts may then be directly exploited by statistical MT systems.

For comparable corpora the situation is much less straightforward, since, depending on the nature of the comparable corpus, only some or perhaps none of the sentences in any pair of texts from the two languages will be translations of each other. Thus, non-alignment of sentences may well be the norm, and even in cases where two texts communicate information on the same topic (e.g. the same news story), the ordering of information, distribution of information over sentences and the inclusion or exclusion of additional information makes the alignment task extremely challenging.

ACCURAT will address this challenge by investigating a number of multi-level alignment methods for comparable corpora. While our focus and novel contributions are on the alignment of, and acquisition of bilingual lexical resources from comparable corpora, we do not exclude the use of existing parallel corpora. On the contrary, starting from whatever parallel resources are available, we will extract at least seed lexical knowledge to be used in, and enhanced by, the process of aligning comparable corpora.

### 4.3.1. Selection of Similar Documents from a Comparable Corpus

Given a comparable corpus consisting of documents in two languages, L1 and L2, the first step is to find similar documents in L1 and L2.

Typical approaches involve treating a document in the L1 collection as a query and then using cross-language information retrieval (CLIR) techniques to retrieve the top n documents from the L2 collection (Muntenu and Marcu, 2005, Quirk et al., 2007). This approach requires some sort of bilingual dictionary for use in query translation.

One innovation will be the exploration of bootstrapped bilingual lexical resources: initial bilingual lexicons used for text and sentence alignment will lead to new lexical translation mappings and those with the most confidence will be added to the bilingual lexicons for use in subsequent iterations of text and sentence alignment.

### 4.3.2. Phrasal Alignment

After similar documents are selected, similar text fragments need to be identified. These fragments may be

sentences or possibly only phrases.

Recent research results have shown that in most cases methods designed for parallel texts perform poorly for comparable corpora. For example, most standard sentence aligners exploit the monotonic increase of the sentence positions in a parallel corpus, which is not observed in comparable corpora.

ACCURAT will investigate how successful the reified sentence aligner (Ceauşu et al. 2006) is in aligning similar sentences in comparable corpora. This reified sentence aligner, based on SVM technology, builds feature structures characterizing a pair of sentences considered for alignment (number of translation equivalents, ratio between their lengths, number of non-lexical tokens, such as dates, numbers, abbreviations, etc., word frequency correlations). These feature structures are afterwards classified as describing GOOD or BAD sentence alignments with respect to experimentally determined thresholds. This aligner has been evaluated and has an excellent F-measure score on parallel corpora, being able to align N-M sentences. It is much better than Vanilla aligner[3] and slightly better than HunAlign[4]. The state-of the-art sentence aligner is Moore's (2002), but this aligner produces only 1-1 alignments (almost perfect), loosing N-M alignments (which downgrades its F-measure score). As comparable corpora do not exhibit the monotonic increase of aligned sentence positions, we anticipate that many of the alignments will be of the type 0-M, N-0 and N-M sentences, thus this alignment ability is a must. The SVM approach to sentence alignment has the advantage that it is fully trainable, the statistical parameters being learnt from the training examples (both positive and negative ones).

Another promising method for identifying similar sentence pairs within comparable corpora, proposed by Munteanu and Marcu (2005), will be also investigated. To select candidate sentences for alignment, they propose a word-overlap filter (half the words of the source language sentence have a translation in the target language sentence) together with a constraint on the ratio of lengths of the two sentences. Given two sentences that meet these criteria, the final determination of whether they are or are not parallel sentences is made by a Maximum Entropy classifier trained over a small parallel corpus, using such features as percentage of words with translations (according to the dictionary), length of sentences, longest connected and unconnected substrings. We will expand this method to sentences / paragraphs which are only to some extent translations of each other, thus adapting the proposed method to comparable corpora.

A challenging research avenue for detecting meaning-equivalent sentence pairs within comparable corpora is using cross-lingual Q&A techniques. The main idea is to exploit dependency linking (Ion and Tufis, 2007) and the concepts of superlinks and chained links (Irimia, 2009) for determining the most relevant search criteria. The keywords, extracted from the dependency linking of a source paragraph/sentence, will be translated (using whatever bilingual resources available, e.g. aligned wordnets, terminology resources or bilingual lexicons - where available, seed translation-pair lists extracted from existing parallel corpora) into a target language and available search engines will look for the most relevant candidate paragraph/sentences. The possible pairs of translation equivalent textual units will be scored by a reified sentence aligner and will be accepted or rejected based on previously determined thresholds.

### 4.3.3. Named Entity and Terminology Alignment and Extraction

Finding common named entities (NEs) or technical terms in phrases from texts in different languages is a powerful indicator that the phrases may be translation equivalents, and their absence almost certainly suggests that the phrases are not equivalents (modulo anaphora).

Named entities and many technical terms are typically not found in general purpose lexicons and so their mapping must be established in other ways. Such multi-word expressions typically fall into two types: those which are more or less phonetically equivalent in two languages (e.g. person names like "Barack Obama" – "Barack/Baraks Obama" in Croatian/Latvian and biological terms like "photosynthesis" – "fotosintēze" in Latvian) and those some or all of whose component words are translated individually (e.g. "Black Sea" – "Melnā jūra" in Latvian). In cases where the NEs or terms are not phonologically related, i.e. contain component words that are translations of each other, entity type equivalence together with dictionary matching on component words may be used to align them. In cases where they are phonologically related, however, a process of matching based on transliteration similarity may be used. It is well known that even NE's that are phonologically equivalent across languages are frequently not orthographically equivalent thus to perform named entity matching requires transliteration from the writing system of one language to that of another.

Transliteration can be performed either orthographically or phonetically. In orthographic approaches (e.g. Aswani and Gaizauskas, 2005) possible cross-language n-gram character mappings observed in training data can be recorded and then, for test names, candidate sequence transliterations can be proposed and scored against the candidate name equivalent using string similarity measures, such as edit distance. In phonetic-based approaches (e.g. Kondrak, 2000; Mani et al., 2008), names are transduced into a phonetic representation and then candidate matches are determined using edit distance measures with learned thresholds.

In the ACCURAT project we will explore both approaches, developing adaptive HMM and/or CRF-based techniques (e.g. Zhou et al., 2008) trained on name pairs gathered initially from parallel training data and then bootstrapped using lexicons derived in the project.

[3] http://nl.ijs.si/telri/Vanilla

[4] http://mokk.bme.hu/resources/hunalign

We will also exploit new advances in adaptive, semi-supervised NE recognition (e.g. Nadeau, 2007) that allow powerful NERC systems to be built for a wide range of entity types from only a handful of examples in each entity class together with suitable corpora. These techniques have not been extensively explored for languages other than English.

Since terminology is of utmost importance in the translation of technical documents, automated terminology extraction will be a basic facility for development of MT systems for narrow domains. Our work will be focused on exploring the use of existing term extraction techniques for terms within the narrow domains. Various techniques exist for identifying terms within a domain-specific (monolingual) corpus and we will build on these. One of techniques is supervised and weakly supervised for semantic labeling of terms within specialist domains in English (e.g. biomedicine; Roberts et al, 2008) which should be relatively portable across languages. Other techniques do not attempt semantic labeling but just attempt to recognize multiword units that are domain specific terms (Bourigault et al, 2000).

Research on bilingual terminology extraction has started recently and relies on assumption that words with same meaning in different languages tend to appear in the same context (Rapp, 1995). The most common approach is to use context vectors and evaluate candidate translations. On single words this approach demonstrated good results (e.g. Chiao and Zweigenbaum, 2002). Recently Daille and Morin (2008) adapted this direct context vector approach for single and multi-word terms and added compositional translation methods for French-Japanese languages. This method increases by 10% the results of Morin et al. (2007), however they are still rather low for multi-word terms.

#### 4.3.4. Relation Extraction for Phrasal Alignment

Another novel way information extraction techniques can assist in aligning comparable corpora is through the identification of cross-language mappings between relation-expressing contexts. Hasegawa et al. (2004) propose a technique for unsupervised relation discovery in texts, whereby contexts surrounding pairs of NEs of given types are extracted and then clustered, the clusters correspond to particular relations (e.g. the relation "company X ACQUIRES company Y" may be expressed as "X's purchase of Y", "X has agreed to buy Y"). This technique achieves impressive results and could be used to align relation expressing contexts as follows. First relation clusters could be established monolingually, given NERC tools in each language. These clusters could then be aligned cross-lingually, using aligned sentence pairs containing NE pairs found in the clusters, aligned sentences coming either from a small amount of parallel data or from high confidence alignments in the comparable corpus. Once relation clusters are aligned cross-lingually, presence of a pair of NEs from an aligned relation cluster in an L1 and L2 sentence pair would constitute evidence that sentences should be aligned.

### 4.4. Comparable Corpora in Machine Translation Systems

To evaluate the efficiency and usability of the approach proposed in the ACCURAT project for under-resourced areas of MT, we will integrate research results into SMT and rule-based systems. We will measure improvements from applying acquired data against baseline results from SMT and RBMT systems and will evaluate the ACCURAT project results in practical applications. The ways how comparable corpora will be integrated and evaluated in MT are described in Eisele and Xu (2010).

## 5. Conclusions

The ACCURAT project has the ambitious goal of developing solutions for the application of comparable corpora in machine translation. Previous research and the planned approach described in this paper allow us to expect promising results from this research.

## 6. Acknowledgements

The project has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under Grant Agreement n° 248347. Part of this work was funded by European Social Fund.

Many thanks for project preparation to colleagues in ACCURAT partner organizations: Andreas Eisele from DFKI (Germany), Serge Sharoff from University of Leeds (UK), Gregor Thurmair from Linguatec (Germany), Nikos Glaros from Institute for Language and Speech Processing (Greece), Marko Tadić from University of Zagreb (Croatia), Boštjan Špetič from Zemanta (Slovenia).

## 7. References

Alegria, I., Ezeiza, N., Fernandez, I. (2008). Translating Named Entities using Comparable Corpora. In *Proceedings of the Workshop on Comparable Corpora, LREC'08*, pp. 11-17.

Aswani, N., Gaizauskas, R. (2005). Aligning Words in English-Hindi Parallel Corpora. In *Proceedings of the ACL 2005 Workshop on Building and Using Parallel Texts: Data-driven Machine Translation and Beyond*, pp. 115-118.

Bekavac, B., Osenova, P., Simov, K., Tadić, M. (2004). Making Monolingual Corpora Comparable: a Case Study of Bulgarian and Croatian. In *Proceedings of the 4th Language Resources and Evaluation Conference: LREC04*, Lisbon, pp. 1187-1190.

Bekavac B. and Tadić M. (2008). A Generic Method for Multi Word Extraction from Wikipedia. In *Proceedings of ITI2008 Conference*, SRCE, Zagreb, pp. 663-667. Chiang D. (2007) Hierarchical Phrase-Based Translation. In *Computational Linguistics* 33(2): 201-228.

Bourigault, D., Jacquemin, C. and L'Homme, M. (eds.). (2002), *Recent Advances in Computational Terminology*, CNRS, ERSS, Université Toulouse-le-Mirail / CNRS-LIMSI, Orsay, France /

Université de Montréal.

Ceauşu A., Ştefănescu D., Tufiş D. (2006). Acquis Communautaire Sentence Alignment using Support Vector Machines. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006), pp. 2134-2137.*

Chiang, D. (2007) Hierarchical Phrase-Based Translation. In *Computational Linguistics* 33(2): 201-228.

Chiao, Y., Zweigenbaum, P. (2002) Looking for Candidate Translational Equivalents in Specialized, Comparable Corpora. In : *COLING 2002*, pp. 1208-1212, Tapei, Taiwan.

Daille, B. and Morin E. (2008) An Effective Compositional Model for Lexical Alignment. In : *Proceedings of IJCNLP-08*, pp. 95-102.

Daille, B., Morin, E. (2005) French-English Terminology Extraction from Comparable Corpora. In : *IJCNLP 2005*, pp. 707-718

EAGLES. (1996). Preliminary recommendations on corpus typology. Electronic resource: http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html.

Eisele, A., Federmann, C., Uszkoreit, H., Saint-Amand, H., Kay, M., Jellinghaus, M., Hunsicker, S., Herrmann, T., Yu Chen (2008) Hybrid Machine Translation Architectures within and beyond the EuroMatrix project. In *Proceedings of EAMT*, Hamburg.

Eisele, A. and Xu, J.(2010). Improving machine translation performance using comparable corpora. In: *Proceedings of 3rd Workshop on Building and Using Comparable Corpora* , Malta.

Gale, W. and Church, K. (1991). A Program for Aligning Sentences in Bilingual Corpora. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pp. 177-184.

Goutte, C., Cancedda, N., Dymetman, M., Foster, G. (eds.). (2009). *Learning Machine Translation*. The MIT Press. Cambridge, Massachusetts, London, England.

Hasegawa, T., Sekine, S., and Grishman, R. 2004. Discovering relations among named entities from large corpora. In *ACL '04*.

Hewavitharana, S. and Vogel, S. (2008). Enhancing a Statistical Machine Translation System by using an Automatically Extracted Parallel Corpus from Comparable Sources. In *Proceedings of the Workshop on Comparable Corpora, LREC'08*, pp. 7-10.

Fung, P. (1995). A Pattern Matching Method for Finding Noun and Proper Noun Translations from Noisy Parallel Corpora. In *Proceedings of the Association for Computational Linguistics*, pp. 236-243.

Ion, R. and Tufiş, D. (2007). RACAI: Meaning Affinity Models. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007*), pp. 282-287, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

Irimia, E. (2009). Metode de traducere automată prin analogie. Aplicaţii pentru limbile română şi engleză. (Methods for Analogy-based Machine Trannslation. Aplications for Romanian and English). PhD thesis, March 2009.

Kilgarriff, A. (2001). Comparing corpora. *International Journal of Corpus Linguistics*, 6(1):1–37.

Koehn, P., Birch, A., Steinberger, R. (2009).462 machine translation systems for Europe. In *MT Summit XII: proceedings of the twelfth Machine Translation Summit*, August 26-30, 2009, Ottawa, Ontario, Canada; pp. 65-72.

Koehn, P. and Hoang, H. (2007). Factored Translation Models. In *Proceedings of EMNLP'07*.

Koehn, P., Och, F. J., Marcu, D. (2003). Statistical phrase based translation. In *Proceedings of the Joint Conference on Human Language Technologies and the Annual Meeting of the North American Chapter of the Association of Computational Linguistics (HLT/NAACL)*.

Kondrak, G. (2000). A New Algorithm for the Alignment of Phonetic Sequences. In *Proceedings of the First Meeting of the North American Chapter of the Association for Computational Linguistics ANLP-NAACL'00*, pp. 288-295.

Mani, I., Yeh, A., Condon, S. (2008). Learning to Match Names Across Languages. In *Proceedings of the COLING 2008 Workshop on Multi-source Multilingual Information Extraction and Summarization MMIES'08*, pp 2-9.

McEnery, A.M. and Xiao, R.Z. (2007). Parallel and comparable corpora: What are they up to? In *Incorporating Corpora: Translation and the Linguist.* Translating Europe. Multilingual Matters, Clevedon, UK.

Maia, B. (2003). What are Comparable Corpora? *Electronic resource*: http://web.letras.up.pt/bhsmaia/belinda/pubs/CL2003%20workshop.doc

Maia, B. and Matos, S. (2008). Corpógrafo V.4 – Tools for Researchers and Teachers Using Comparable Corpora. In *Proceedings of the Workshop on Comparable Corpora, LREC'08*, pp. 79-82.

May, J. and Knight, K. (2007). Syntactic Re-Alignment Models for Machine Translation. In *Proceedings of EMNLP-CoNLL'07*.

Moore, R. C. (2002). Fast and Accurate Sentence Alignment of Bilingual Corpora. In *Machine Translation: From Research to Real Users* (Proceedings, 5th Conference of the Association for Machine Translation in the Americas, Tiburon, California), Springer-Verlag, Heidelberg, Germany, pp. 135—244.

Morin, E., Daille, B., Takeuchi, K., Kageura, K. (2007) Bilingual Terminology Mining - Using Brain, not brawn comparable corpora. In: *ACL 2007*

Munteanu, D. (2006). Exploiting Comparable Corpora (for automatic creation of parallel corpora). *Online presentation. Electronic resource*: http://content.digitalwell.washington.edu/msr/external_release_talks_12_05_2005/14008/lecture.htm

Munteanu, D. and Marcu, D. (2005). Improving Machine Translation Performance by Exploiting Non-Parallel

Corpora. *Computational Linguistics*, 31(4): 477-504.

Munteanu, D., Fraser, A., Marcu, D. (2004). Improved Machine Translation Performance via Parallel Sentence Extraction from Comparable Corpora. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, HLT / NAACL'04.*

Nadeau, D. (2007). Semi-Supervised Named Entity Recognition: Learning to Recognize 100 Entity Types with Little Supervision. PhD Thesis, University of Ottawa, 2007.

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

Quirk, C., Udupa, R., Menezes, A. (2007). Generative Models of Noisy Translations with Applications to Parallel Fragment Extraction. In *Proceedings of MT Summit XI, European Association for Machine Translation.*

Rapp, R. (1995). Identifying Word Translations in Non-Parallel Texts. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pp. 320-322.

Rayson, P. and Garside, R. (2000). Comparing corpora using frequency profiling. In *Proceedings of the Comparing Corpora Workshop at ACL'00*, pp. 1-6.

Resnik, P. and Smith, N. (2003). The Web as a Parallel Corpus. *Computational Linguistics*, 29(3) pp. 349-380.

Roberts, A., Gaizauskas, R., Hepple, M., Guo, Y. (2008). Combining terminology resources and statistical methods for entity recognition: an evaluation. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation* (LREC 2008).

Saralegi, X., San Vicente, I., Gurrutxaga, A. (2008). Automatic extraction of bilingual terms from comparable corpora in a popular science domain. In *Proceedings of the Workshop on Comparable Corpora*, *LREC'08*.

Sharoff, S. (2007). Classifying Web corpora into domain and genre using automatic feature identification. In *Proceedings of Web as Corpus Workshop*. Louvain-la-Neuve.

Shi, L., Nie, C., Zhou, M., Gao, J. (2006). A dom tree alignment model for mining parallel data from the web. In *Joint Proceedings of the Association for Computational Linguistics and the International Conference on Computational Linguistics*, Sydney, Australia.

Sinclair J. (1987) (eds.) Looking up: an account of the COBUILD Project in lexical computing. Collins, London and Glasgow.

Skadiņa, I., Brālītis, E. (2009). English-Latvian SMT: knowledge or data? In: *Proceedings of the 17th Nordic Conference on Computational Linguistics NODALIDA.* NEALT Proceedings Series, Vol. 4, pp. 242–245.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D., Varga, D. (2006) The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the 5th International Conference on Language Resources and Evaluation: LREC'06*.

Thurmair, G. (2006). Using Corpus Information to Improve MT Quality. In *Proceedings of the Workshop LR4Trans-III, LREC*, Genova.

Tillmann, C. (2003). A projection extension algorithm for statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural language processing*.

Tufiş, D., Ion, R., Ceauşu, A., Ştefănescu, D. (2006). Improved Lexical Alignment by Combining Multiple Reified Alignments. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL2006)*, pp. 153-160, Trento, Italy, April 2006.

Utiyama, M., Isahara, H. (2003). Reliable Measures for Aligning Japanese-English News Articles and Sentences. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pp. 7-12 July 2003, Sapporo, Japan.

Vogel, S., Zhang, Y., Huang, F., Tribble, A., Venugopal, A., Zhao, B., Waibel, A. (2003). The CMU Statistical Machine Translation System. In *Proceedings of MT-Summit IX*.

Yamada, K. and Knight, K. (2001). A Syntax-Based Statistical Translation Model. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 523–530, Toulouse, France, July.

Yamada, K. and Knight, K. (2002). A Decoder for Syntax-Based Statistical MT. In *Proceedings of the Conference of the Association for Computational Linguistics, ACL'02*.

Zhang, Y., Wu, K., Gao, J., Vines, P. (2006). Automatic Acquisition of Chinese-English Parallel Corpus from the Web. In *Proceedings of 28th European Conference on Information Retrieval*.

Zhao, B., Vogel, S. (2002). Adaptive Parallel Sentences Mining from Web Bilingual News Collection. In *Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM'02)*, p.745, December 09-12, 2002.

Zhou, Y., Huang, F., Chen, H. (2008). Combining probability models and web mining models: a framework for proper name transliteration. In *Information Technology and Management* 9(2).

# Statistical Corpus and Language Comparison using Comparable Corpora

**Thomas Eckart, Uwe Quasthoff**
NLP Group, University of Leipzig
Johannisgasse 26, 04103 Leipzig, Germany
E-mail: {teckart, quasthoff}@informatik.uni-leipzig.de

## Abstract

Corpora of different languages but similar genre allow language comparison. Applying the same methods to corpora of the same language but of different genre or origin results in corpus comparison. Having many corpora in identical formats, these statistical methods will generate various data for manual or automatic analysis. The introduced system reports more than 150 results per corpus, for approximately 150 corpora right now. The results are presented on more than 22,000 pages which are generated automatically. Intelligent Browsing allows contrasting of different corpora with respect to different questions, languages, text genres and varying corpus size. As a side effect, shortcomings in the corpus preprocessing usually produce statistical anomalies that are easily noticeable and lead to an improved processing chain.

## 1. The Leipzig Corpora Collection

Basis for all further considerations are the corpora of the Leipzig Corpora Collection. For about fifteen years corpora are created by using text material of all kind, focusing on the Internet as text resource. By using the Web text material in more than 50 languages and in partially enormous sizes were gathered from various sources.

By now hundreds of corpora were created, which can be classified in three dimensions: language (including dialects), genre (currently: news texts, random web texts, governmental and Wikipedia texts) and size (measured in number of sentences). For easy corpus comparisons, subcorpora of normed sizes (containing 10,000, 30,000, …, 3 million sentences), are created.

All texts are segmented into sentences and words and all relevant data is stored in a relational database (cf. Quasthoff et al., 2006), containing information like word frequencies and word co-occurrences. To ensure comparability, the corpus preprocessing was standardized as much as possible (cf. Quasthoff & Eckart, 2009). Currently, corpora in 15 languages are made freely available, an extensive expansion of the download portal is planned for the near future[1].

## 2. Analysis Procedure

With a standardized creation process and a uniform data schema on the one hand and a fast growing amount of different corpora on the other, it became obvious that there was a lack of analysis tools to evaluate existing data and to ensure corpus quality without extensive manual work. As a result, existing tools (mostly Python and Perl scripts of different complexity) were replaced by a new tool with the intention to separate the knowledge- and labor intensive creation of an evaluation task from the execution of this task on a specific corpus.

Therefore every evaluation is encapsulated in a single script, that holds all necessary information and that validates against a proprietary XML schema. In general, one script consists of a set of SQL statements that are executed on a database, specified by the user. Each result set can be processed further by the scripting languages Perl or PHP, including: merging of data, reformatting of result sets or computing interesting values that couldn't be provided by the database management system itself. These data are sufficient for many problems of corpora analysis. To offer more intuitive ways, especially in the field of statistical evaluation, a graphical component is needed. Hence, the plotting tool Gnuplot[2] was integrated, that offers various possibilities of graphical presentation.

To ensure platform independence only software was used that is provided for different platforms and systems, namely Java, PHP, Perl and Gnuplot. Additionally an

1 *http://corpora.informatik.uni-leipzig.de/download.html*

2 *http://www.gnuplot.info*



*Proceedings of the 3rd Workshop on Building and Using Comparable Corpora, LREC 2010*, pages 15–20
Malta, 22 May 2010

easy-to-use Graphical User Interface was developed, and the possibility of executing a set of evaluation scripts in a batch mode.

## 3. Analysis Types

To cover as many fields of interest as possible, more than 150 different evaluation scripts were created and classified in six sections of analysis:

**Corpus Meta Information**

Information regarding the corpus and its creation: size, versions of preprocessing tools, duration of the processing tool chain etc.

**Characters and Character N-Grams**

Information regarding the distribution of characters, especially on word beginnings or endings, character successor rates, character transition probabilities etc.

**Words and Multi-words**

Information regarding words (including multi-words if existing): length distribution, text coverage, samples, several variants of Zipf's law (cf. Zipf, 1949), word transition probabilities, word similarity using Levenshtein distance, average word length, longest words in different frequency ranges etc.

**Sentences**

Information regarding distribution of sentence lengths measured in words or characters, typical sentence beginnings or endings, similar sentences, sentences containing only words of either high or low frequency etc.

**Word Co-occurrences**

Samples for typical word (sentence / neighbour-) co-occurrences (cf. Dunning, 1994), visualization of Zipf's law for co-occurrences, semantic word similarity using joint co-occurrences, small world parameters for the co-occurrence graph etc.

**Sources**

Information regarding sources like: number of used sources, typical size of each source, differences between various sources measured in parameters as above, etc.

These fields are steadily extended and will be developed further. The focus here is especially on customization and extension of existing scripts to character sets and syntactic structures that haven't been dealt with yet.

## 4. Language and Corpora Comparison

### 4.1 General Structure

An analysis script as described above usually generates three different types of output:

- A table containing the measured data, together with a Gnuplot diagram
- One or two parameters (like the slope for Zipf's law) to approximate the function plotted above
- Example corpus data for extreme data points (for Zipf's law: the most frequent words)

These three distinct output types can be used for different purposes: a plotted diagram is fine for manual inspection and manual corpora comparison. Numeric parameters are more interesting for automatic comparisons: the parameters of different analysis can be considered as components of a feature vector for a corpus. Clustering techniques can then be used to identify families of similar corpora or languages.

Sample words or sentences with extreme parameters are of interest due to their specific linguistic properties or may help to find corpus preprocessing problems, as will be shown below.

### 4.2 Intra-language and Inter-language Comparisons

While language dependent parameters are expected to vary for different languages, their behavior for different genres within one language is difficult to predict. The following table compares three parameters first for different text genres of German, and then the same parameters for newspaper corpora for different languages. The intra-language variation may help to decide whether differences between languages can be considered as significant. Moreover, for corpora of mixed or unknown genre such data help to decide whether more detailed information about the genres are necessary.

| | **Text coverage (20 top words)** | **Avg. word length** | **Avg. sentence length** |
|---|---|---|---|
| **News** | 22.10% | 13.59 | 16.19 |
| **Web** | 21.57% | 14.06 | 16.03 |
| **Wikipedia** | 23.10% | 12.57 | 16.71 |
| **Movie Subtitles** | 21.20% | 10.42 | 6.57 |

Table 1: Intra-language comparison

Table 1 and 2 show the text coverage for the 20 most frequent words, the average word length in characters (without multiplicity) and the average sentence length in words.

| | Text coverage (20 top words) | Avg. word length | Avg. sentence length |
|---|---|---|---|
| **German** | 22.10% | 13.59 | 16.19 |
| **English** | 26.23% | 10.62 | 19.46 |
| **Czech** | 16.78% | 8.65 | 14.95 |
| **Vietnamese** | 12.44% | 4.97 | 23.64 |
| **Finnish** | 12.37% | 12.28 | 11.50 |

Table 2: Inter-language comparison

### 4.3 Insights into Language Structure

The following example counts the number of letter n-grams as a measure for character successor variability. Because rare words (especially when containing spelling errors) will contain nearly any n-gram, only the $N=10^k$ most frequent words (for k=2, 3, 4, …) are used.

Table 3 shows the number of different letter n-grams at word beginnings, taken from a newspaper corpus in Finnish.

| N | # of bigrams | # of 3-grams | # of 4-grams | # of 5-grams |
|---|---|---|---|---|
| **100** | 51 | 82 | 95 | 99 |
| **1000** | 211 | 449 | 654 | 822 |
| **10000** | 577 | 1821 | 3256 | 4829 |
| **100000** | 1391 | 6852 | 16804 | 28622 |
| **1000000** | 2512 | 14910 | 44494 | 86492 |

Table 3: Finnish n-grams at word beginnings

In figure 1, the values of table 3 are plotted with logarithmic scale. The nearly straight lines suggest a power law.
Similar results are true for counting letter n-grams at word endings or counting letter n-grams regardless of their position. The same is true for many other languages. Of course, the slope varies for the different n-gram types and languages.

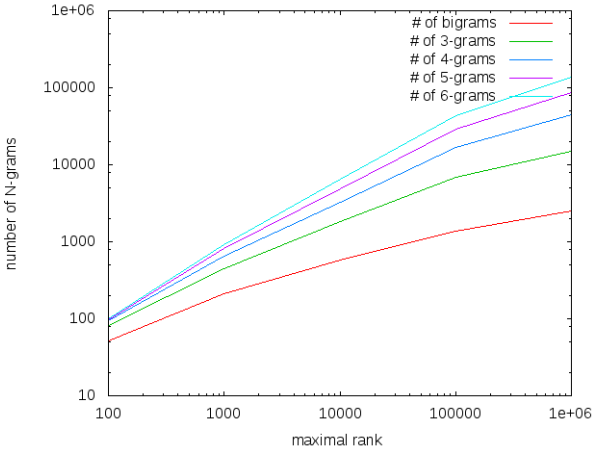



Figure 1: Letter n-grams of Finnish word beginnings

### 4.4 Non-linear Growth Rates

The non-linear growth of certain parameters gives rise to new difficulties when comparing different corpora or languages. For such comparisons we can use the corpora of normed size as explained in section 1. Figure 2 shows the number of distinct word forms, the number of sentence based word co-occurrences and the number of next neighbor co-occurrences. These numbers are taken for corpora of 100.000, 300.000, 1 million and 3 million sentences. Again, the nearly straight lines imply a power law. A more detailed inspection using different languages still shows nearly straight lines, but with slightly different parameters.

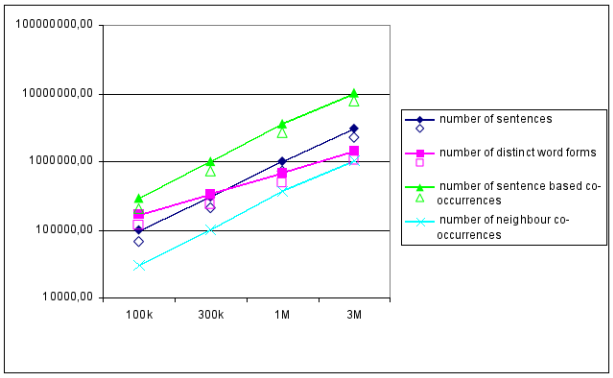



Figure 2: Non-linear growth

## 5. Quality Assurance

In many of the above mentioned analysis types, a special value is measured for many objects, like sentence length in characters for every sentence of a corpus. Looking at objects with extreme values (i.e. very small or very large), we often find effects of errors in the input material or poor preprocessing (cf. Eskin, 2000; Dickinson & Meurers, 2005).

In the case of very short sentences, we may find broken sentences. Moreover, sentences containing many very low frequent words are usually not well-formed. Table 4 shows sentences of an English Web corpus that consists of words that have a very low average frequency. Apparently there were encoding problems in the input material and the language identification failed in rejecting some non-English sentences.

| **Avg. word rank** | **Sentence** |
|---|---|
| 35844 | Bidh an luchd-aithisg a' gabhail notaichean tron choinneimh agus 's dÃ²cha […] briathran air an togail. |
| 31711 | "HCPT - The Piligrimage Trust" jest organizacja charytatywna zalozona w Wielkiej Brytanii. |
| 28524 | Gjelder dette for barn og unge mennesker under 18 Ã¥r? |

Table 4: Examples of sentences that consist of words with low average frequency

Another hint for problems in the corpus generation process is looking at extreme points of the distribution of specific characters.

| **# of semicolons** | **# of sentences** |
|---|---|
| 2 | 183 |
| 3 | 28 |
| 4 | 16 |
| 5 | 4 |
| 6 | 1 |
| 12 | 1 |

Table 5: Part of a semicolon distribution in Ukrainian sentences

Table 5 shows an excerpt of the distribution of semicolons in a 100,000 sentences Ukrainian corpus.

These sentences that were segmented by the sentence boundary detection and accepted by the following quality assurance procedures include "*Територією області течуть річки Ілі з притоками Чарин, Чілія, Текес, Курти; Каратал із притокою Коксу; Аксу; Лепси; Аягуз; Тентек; Кеген.*" (Ukrainian), "*Machiaj: Divizia Make-up DUMAREX Parteneri media: EVENIMENTUL ZILEI; ZIUA; JURNALUL NATIONAL; CAPITALA; COTIDIANUL; METROBUS; AZI; BURDA ROMANIA; ANTENA 1 - Doina Levintza, "'Neata"; PRIMA TV - "Clubul de Duminica", "Stil".*" (Romanian) or "*Siippainen kirjoitti lehtijuttujaan eri nimimerkeillä kuten Iloinen, Petteri; Kaaleppi; Karho, Otto; Kimpinen; Kimpinen, Kalle; Mäikiä, Urmas; O. S.; O. S-nen; Robin Hood; Saarto, Olavi; Svejk; Uolevi.*" (Finnish).

This information provides a fast feedback and leads to more accurate data resources in the future. Statistical values that may indicate problems with input selection, inaccurate preprocessing tools or other issues are widely spread, ranging from character analysis to show character set problems to automated rating of the corpora sources based on their homogeneity of various statistical values. This is still to be evaluated.

## 6. Presentation of the Results

Central goal for the presentation of the created result pages was a web portal that should allow both researchers in the fields of natural language processing and linguistics an easy access and overview of existing corpora and a starting point for evaluating linguistic phenomena in the field of corpus, genre and language comparison.

Each question, answered for a certain corpus, produces an HTML page containing the results. As described above, these result pages consist of a plot or of a (set of) table(s), or both. For comparisons, all corpora are assigned to three different categorization dimensions: language, text genre and corpus size. The Corpora and Language Statistics Website presented at *www.cls.informatik.uni-leipzig.de* supports this complex navigation. To achieve an easy access, despite the thousands of pages strongly related to each other, the ISO standard Topic Maps was used as underlying technology. Based on JRuby Topic Maps (cf. Bleier et al., 2009) and tinyTIM, all existing resources were merged while allowing extensions to new fields and dimensions in the future.

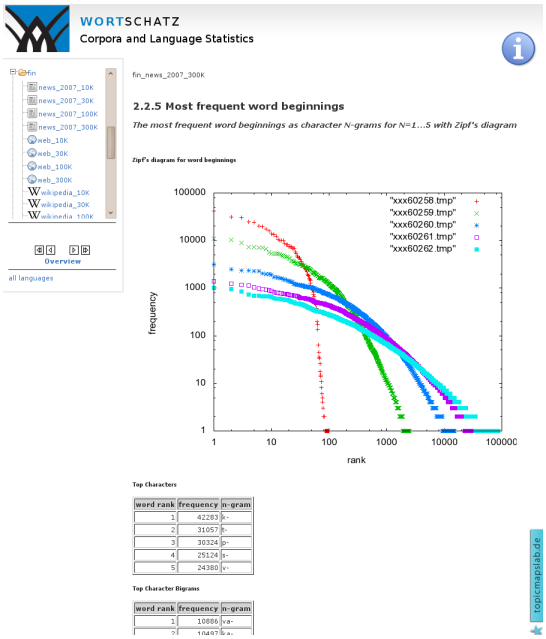



Figure 3: Sample HTML page

The user interface is designed to lower the entry barrier for the (possibly inexperienced) user: on the left side one can select between other languages, genres and corpus sizes. These links will show the corresponding page for the same question, but another corpus. The arrows allow linear scrolling through the different questions for one corpus.

An additional help screen gives detailed information about the data shown and the intentional background of the question. A (possibly slightly simplified) select-statement is provided. This can be used or modified for similar questions asked by the user. Some open problems and cross references complete this help screen.

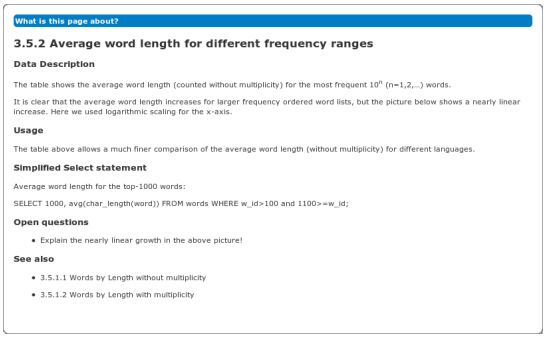



Figure 4: Sample help screen

# 7. Experimental Setup

## 7.1 Configuration of a Single Analysis

To allow contributions by different kind of persons including undergraduate students of different disciplines the underlying XML schema was designed in an uncomplex way that is nonetheless powerful through its universality.

There already exists a huge amount of scripts in different analysis domains. Therefore the standard procedure to extend the stock of evaluations is the modification of a template or an already used script and the adaption to the new problem.

Every task (as requests to the database management system, further processing like linking of temporary results or defining the specific visual output) is a single working step. As most new scripts try to examine an already considered field in more detail most parts of an existing script are still valid and can be adopted (especially simple post processing or output definitions). Therefore the effort of further extension is quite low. As a consequence whole ranges of new scripts could be generated by very simple replacements in already used SQL statements and explanatory text strings.

Listing 1 shows an excerpt of a simple evaluation script with all changes highlighted that are necessary to adapt the script to a new character.

```
<title>Distribution of Letter F</title>
<description>Number of sentences containing a fixed
        number of occurrences of this
        character</description>

<step descriptor="0">
        <sql-step>
                <statement>select
        round(char_length(sentence)-
        char_length(replace(lower(sentence),"f",""))),
        as freq, count(*), sentence from
        BASEDB.sentences group by freq order by
        freq</statement>
         </sql-step>
 </step>
```

Listing 1: Excerpt of an evaluation script

## 7.2 Reproducibility of the Results

To compare results of the CLS Website with similar results on other corpora it is essential to have free access

to the corpora used here. Moreover, it must be transparent how the measurement was performed. The first condition is fulfilled by the availability of the Leipzig corpora collection, the second by the detailed description given in the help screens.

## 8. Further Work

At present not all existing corpora are already evaluated, many are still to be processed. To enhance usability and to achieve an easier access to the evaluation data it is intended to offer more interactive ways in the future. These will allow the user to compare values across self-chosen corpora and to inspect the data in more detail. Another aim is the adoption of the created tools and structures to other domains. As an example in eAQUA (cf. Heyer &. Schubert, 2008), a co-operational project of researchers of Computer Science and Ancient Science, a similar approach is used to give both sides a fast comparison of existing data resources and helps finding problems in the complex (pre-)processing of ancient texts.

## 9. References


Bleier, A.; Bock, B., Schulze, U., Maicher, L. (2009): *JRuby Topic Maps*. In *Proceedings of the Fifth International Conference on Topic Maps Research and Applications* (TMRA 2009). Leipzig, Germany.

Dickinson, M. and Meurers, D. (2005): *Detecting Annotation Errors in Spoken Language Corpora.* In: *Proceedings of the Special Session on Treebanks for Spoken Language and Discourse at the 15th Nordic Conference of Computational Linguistic* (NODALIDA-05), Joensuu, Finland.

Dunning, T. (1993): *Accurate methods for the statistics of surprise and coincidence*. Computational Linguistics, Volume 19, number 1.

Eskin, E. (2000): *Automatic Corpus Correction with Anomaly Detection*. In *Proceedings of the First Conference of the North American Chapter of the Association for Computational Linguistics* (NAACL-00). Seattle, Washington, USA.

Heyer, G. and Schubert, C. (2008): *eAQUA - Extraktion von strukturiertem Wissen aus Antiken Quellen für die Altertumswissenschaft*. Word Wide Web electronic publication, http://www.eaqua.net.

Quasthoff, U.; Richter, M.; Biemann, C. (2006): *Corpus Portal for Search in Monolingual Corpora.* In: *Proceedings of the fifth international conference on Language Resources and Evaluation*, LREC 2006, Genoa, Italy.

Quasthoff, U. and Eckart, T. (2009): *Corpus build process of the project 'Deutscher Wortschatz'.* Workshop "Linguistic Processing Pipelines", GSCL Conference 2009, Potsdam, Germany.

Zipf, G.K. (1949): *Human behavior and the principle of least effort : an introduction to human ecology*. Hafner reprint, New York, 1972, 1st ed. (Addison-Wesley, Cambridge, MA, 1949).

# Wikipedia as Multilingual Source of Comparable Corpora

**Pablo Gamallo Otero, Isaac González López**

University of Santiago de Compostela
Galiza, Spain
pablo.gamallo@usc.es, isaacjgonzalez@gmail.com

**Abstract**

This article describes an automatic method to build comparable corpora from Wikipedia using *Categories* as topic restrictions. Our strategy relies of the fact Wikipedia is a multilingual encyclopedia containing semi-structured information. Given two languages and a particular topic, our strategy builds a corpus with texts in the two selected languages, whose content is focused on the selected topic. Tools and corpora will be distributed under free linceses (General Public License and Creative Commons).

## 1. Introduction

Wikipedia is a free, multilingual, and collaborative encyclopedia containing entries (called "articles") for more than 300 languages. English is the more representative one with almost 3 million articles. As table 1 shows, the number of entries/articles for the most used languages in Wikipedia is so high that it could be considered a reliable multilingual resource. However, Wikipedia is not a parallel corpus as their articles are not translations from one language into another. Rather, Wikipedia articles in different languages are independently created by different users.

In accordance with fast growth of Wikipedia, many works have been published in the last years focused on its use and exploitation for multilingual tasks in natural language processing: extraction of bilingual dictionaries (Yu and Tsujii, 2009; Tyers and Pieanaar, 2008), alignment and machine translation (Adafre and de Rijke, 2006; Toms et al., 2001), multilingual retrieval information (Pottast et al., 2008). In addition, there exists theoretical work on the degree of comparability among the different multilingual versions of an entry/article in Wikipedia (Filatova, 2009). In particular, the author analyzes symmetries and asymmetries in multiple descriptions of multilingual entries.

In this paper, our main concern is the use of Wikipedia as a source of comparable corpora. The EAGLES - Expert Advisory Group on Language Engineering Standards Guidelines (see `\http://www.ilc.pi.cnr.it/EAGLES96/browse.html`) gives us the following definition for "comparable corpora":

> *A comparable corpus is one which selects similar texts in more than one language or variety.*

One of the main advantages of comparable corpora is their versatility to be used in many linguistic fields (Maia, 2003), like terminology extraction, Information Retrieval, and Knowledge Engineering. In addition, they can also be used as training corpus to improve statistic machine learning systems, in particular when parallel corpora are scarce for a given pair of languages. Another advantage concerns their availability. In contrast with parallel corpora, which require (not always available) translated texts, comparable corpora are easily retrieved from the web. It is much easier to find original texts on a particular subject than to find a pair consisting of the original and a good translation. Among the different web sources of comparable corpora, Wikipedia is likely the largest repository of similar texts in many languages. We only require the apropriate computational tools to make them comparable.

| Languages | number of articles |
|---|---|
| English | 2,826,000 |
| German | 888,000 |
| French | 786,000 |
| Polish | 593,000 |
| Italian | 576,000 |
| Japonese | 556,000 |
| Dutch | 528,000 |
| Portuguese | 470,000 |
| Spanish | 460,000 |
| Rusian | 376,000 |

Table 1: The top ten languages in Wikipedia ranked by number of articles (April 2009)

By taking into account multilingual potentialities of Wikipedia, the goal (and main contribution) of this paper is to describe a method to extract comparable corpora from this freely available encyclopedia, according with two parameters of variation: languages and topic. More precisely, given two languages and a particular topic, our strategy builds a corpus with texts in the selected languages, whose content is focused on the selected topic. Both the generated corpora and the tools used to generate them will be available under Creative Commons license in `http://gramatica.usc.es/pln`. Experiments will be performed with articles in English, Spanish, and Portuguese. As Table 1 shows, Spanish is the ninth most used language in Wikipedia, with 460 thousand articles, very close to Portuguese, which reaches 470 thousands.

This paper is organized as follows. Section (2.) describes how we convert the original Wikipedia into a new codified corpus, called "CorpusPedia". Section (3.) introduces different strategies to build comparable corpora from CorpusPedia. In Section (4.), we give some empirical data of CorpusPedia, as well as the results of some experiments per-

This work has been supported by the Galician Government, within the projects PGIDIT07PXIB204015PR and 2008/101.



*Proceedings of the 3rd Workshop on Building and Using Comparable Corpora, LREC 2010*, pages 21–25
Malta, 22 May 2010

```
<page>
<title>Arqueoloxía</title>
<id>3</id>
<revision>
  <id>1310468</id>
  <timestamp>2009-10-06T02:42:14Z</timestamp>
  <contributor>
    <username>SieBot</username>
    <id>2109</id>
  </contributor>
  <minor />
  <comment>bot Engadido: [[ku:Arkeolojî]]</comment>
  <text xml:space="preserve">{{Historia en progreso}}

A '''arqueoloxía''' é a [[ciencia]] que estuda as [[arte|artes]],
[[monumento|monumentos]] e [[obxecto]]s da
[[antigüidade|antigüidade]], especialmente a través dos
seus restos. O nome ven do [[lingua grega|grego]]
"archaios", &quot;vello&quot; ou &quot;antigo&quot;, e
"logos", &quot;ciencia&quot;, &quot;saber&quot;.

[...]

[[zh: 考古学]]
[[zh-yue: 考古]]</text>
  </revision>
</page>
```

Figure 1: XML example of Wikipedia: excerpt of Galician entry "Arqueoloxía" (Archaeology)

formed using the strategies defined in (3.). The last section discusses future tasks we intend to implement in order to extend and improve our tools.

## 2. CorpusPedia

The first step of our method consists in converting the source files of Wikipedia to a set of files with a more friendly and easy-to-use XML structure: CorpusPedia. For this purpose, we developed tools aimed to automatically download Wikipedia in the required languages and then to apply the process of transforming the downloaded XML file into the new XML files of CorpusPedia. In the following, we will compare the structure of those two formats.

### 2.1. Format of Wikipedia

The whole Wikipedia is downloadable in XML files containing a great variety of metadata. Figure 1 shows an example of an article codified in this way. An entry/article is identified by the tag *page*, which contains a title, a data, an author, and the text of the article. The difference with regard to the most usual web markup languages (html or xhtml) is that the text of all articles is codified in *wiki format*, as the the tag *text* in Figure 1 illustrates. One of the main taks of CorpusPedia is to build a plain text version from that wiki text.

### 2.2. Format of CorpusPedia

The format of CorpusPedia also consists, essentially, in both the title and the text of each entry/article. Besides, further information is also provided using semi-structured data of Wikipedia and some conventions among editors (Clark et al., 2009). Figure 2 depicts the XML code used to markup the new format generated from Wikipedia. Tags *title*, *category*, *plaintext*, and *translations* are those required to generate comparable corpora.

The tag *category* is used to identify all the topics classifying the text content. In Wikipedia, each article is explicitly assigned to one or more categories representing different topics. This tag will allow us to extract those articles classified with similar categories and, then, with high degree of comparability. The tag *wikitext* contains the original format of Wikipedia. We keep this format since it can be useful for further extractions. based on semi-structured content. The tag *plaintext* (i.e., text without any codification) is generated from *wikitext* by applying a *wiki2plaintext* parser we developed for this purpose. Unlike other *wiki2plaintext* converters, we took into account specific semantic features of Wikipedia. The tag *translations* codifies a list of interlanguage links (i.e., links to the same articles in other languages). As we will explain in the next section, these links are useful to align article-by-article a comparable corpus if it is required by the user. The list of interlanguage links is always ranked in the same way (gl pt es en fr ca eu al it cs bg el). Besides, if there is no a specific interlanguage link, the symbol "#" is used to explicitly mark that the translation is not available. Other languages can be easily added to the list if they are required.

The remaining tags of CorpusPedia provide further useful relations with other articles in Wikipedia. This way, the tag *related* adds those articles that are somehow related to the current one and which have been explicitly marked in Wikipedia. Finally, *links* introduce the set of links to other

<article>
<title>Arqueoloxía</title>
<category>Arqueoloxía</category>
<related>Antropoloxía, Arqueoloxía industrial, Arqueoloxía submarina</related>
<links>ciencia, arte|artes, monumento|monumentos, obxecto, antigüidade|antigüidade, lingua grega|grego, cultura, estudo, psicolóxico, condutistas, antropoloxía, idade de pedra, Idade Media, Arqueoloxía industrial, Antropoloxía, Arqueoloxía industrial, Arqueoloxía submarina</links>
<translations># Arqueologia Arqueología Archaeology Archéologie Arqueologia Arkeologia # Archeologia Archeologie Археология Αρχαιολογία</translations>
<plaintext>A arqueoloxía é a ciencia que estuda as artes, monumentos e obxectos da antigüidade, [...] o que se coñece como Arqueoloxía industrial.</plaintext>
<wikitext>{{Historia en progreso}}

A '''arqueoloxía''' é a [[ciencia]] que estuda as [[arte|artes]], [[monumento|monumentos]] e [[obxecto]]s da [[antigüidade|antigüidade]], [...]
[...]
[[yi:ארכעאלאגיע]]
[[zh: 考古学]]
[[zh-yue: 考古]]</wikitext>
</article>

Figure 2: XML example of CorpusPedia: excerpt of Galician entry "Arqueoloxía" (Archaeology)

articles that were explicitly mentioned within the text (also called "interlinks").

## 3. Strategies to Elaborate Wikipedia-Based Comparable Corpora

Given the information structure of CorpusPedia, it is possible, not only to easily collect articles about the same topic in the same language, but also to put them in relation with articles about the same topic in other languages. It means the structure of CorpusPedia enables to easily build comparable corpora. For this purpose, we developed three tools aimed to extract corpora with different degrees of comparability. These tools, which correspond to three strategies, are described in the following subsections.

### 3.1. Not-Aligned Comparable Corpora

This strategy extracts those articles in two languages having in common the same topic, where the topic is represented by a category and its translation (for instance, the english-spanish pair "Archaeology-Arqueología"). The algorithm used to extract not-aligned comparable corpora from CorpusPedia is the following:

```
Given two languages, L1 and L2, and
two bilingual categories, C1 and C2,
where C2 is the translation of C1 in
L2:
    (1) extract those articles in L1
    containing C1 within the section
    <category> ;
    (2) Repeat the same process in L2,
    using C2.
```

It results in a not-aligned comparable corpora, consisting of texts in two languages ($L1$ and $L2$) sharing the same topic: $C1$-$C2$. We called it "not-aligned" because the version of an article in one language may have not its corresponding version in the other language. In technical terms, it means articles extracted from L1 will contain both empty and not empty interlanguage links to articles in L2.

### 3.2. Strong Alignment

The corpus resulting of the previous process can be considered as being too heterogeneous, since it may contain articles in one language that have not their corresponding versions in the other one. For instance, we can find an English article with the title "Australian archaeology" that has not any interlanguage link in Spanish, i.e., that has not a Spanish version with the title "Arqueología australiana" in the Wikipedia. To build an aligned corpus at the level of articles, we define a strategy to extract only those articles that have interlanguage links to the target language. The algorithm of this strategy is the following:

```
Given two languages, L1 and L2, and
two bilingual categories, C1 and C2,
where C2 is the translation of C1 in
L2:
    (1) extract those articles in L1
    with the following properties:
        - C1 is within the section
        <category>
        - there is a interlanguage
        link to an article in L2
        containing C2 in the section
        <category>
```

```
    (2) Repeat the same process from
    L2 and remove inconsistencies.
```

We obtain a comparable corpus constituted by the same articles in both languages. The strategy used to align article by article is very restrictive and then has very low coverage. In fact, no only each article in one language must have its corresponding article in the other one, but also both articles must share the same categorial restriction. Let's note that we have to automatically remove inconsistencies in annotations, such as for instance ill-defined interlangue links. These annotations problems were inherited from the source file.

### 3.3. Soft Alignment

The strong alignment algorithm is not able to extract some relevant articles, in particular those that, having interlanguage links to the target languange, do not fill the categorial restriction. For instance, there may be articles categorized in the English Wikipedia by means of the term "Archaeology"[1], which have not been categorized in the Spanish Wikipedia with the corresponding term "Arqueología". However, these Spanish articles can be considered as being indirectly classified by the English category. In fact, Spanish Wikipidia is less categorized as the English one (Spanish editors tend to use fewer categories by article). Similarly, the Portuguese Wikipedia is still less categorized as the Spanish one. This categorial asymmetry is responsible for the low coverage reached by the previous strategy (strict alignment). To solve this problem, we propose a less rigid alignment. The goal is to extract pairs of bilingual articles related by interlanguage links if, at least, one of both contains the required category. The algorithm is the following:

```
Given two languages, L1 and L2, and
two bilingual categories, C1 and C2,
where C2 is the translation of C1 in
L2:
    (1) extract those articles in L1
    with the following properties:
        - C1 is within the section
        <category>
        - there is an interlanguage
        link to an article in L2
    (2) extract those articles in L2
    with an interlanguage link to
    the articles in L1 which have
    been already extracted, and
    remove inconsistencies.
```

It results in a corpus that has also been alingned article by article, but using a technique not so restrictive as in the previous method.

## 4. Experiments and Results

### 4.1. Size of CorpusPedia

In the last version of CorpusPedia, the plaintext in English contains about 1,2 billion token words, 180 million in Spanish, and 120 million in Portuguese. Notice that the Spanish version contains more words than the Portuguese one. However, the Portuguese Wikipedia contains a larger number of articles, as is shown in Table 1. It follows the plaintext content of Portuguese articles tends to be smaller than that of the Spanish version.

[1]The structured list of categories in the English Wikipedia avoids language variation. In this case, the normalized term is the British Archaelogy instead of Archeology.

| strategy | size (in words) | number of articles |
|---|---|---|
| en/es not-aligned | 738,000 / 344,000 | 1120 / 462 |
| en/pt not-aligned | 738,000 / 64,000 | 1120 / 100 |
| es/pt not-aligned | 344,000 / 64,000 | 462 / 100 |
| en/es strong-align | 34,000 / 23,000 | 34 / 34 |
| en/pt strong-align | 29,000 / 11,000 | 16 / 16 |
| es/pt strong-align | 27,000 / 11,000 | 19 / 19 |
| en/es soft-align | 220,000 / 134,000 | 191 / 191 |
| en/pt soft-align | 161,000 / 60,000 | 124 / 124 |
| es/pt soft-align | 132,000 / 64,000 | 119 / 119 |

Table 2: Comparable corpora in english-spanish, english-portuguese, and spanish-portuguese. They were obtained using category "Archaelogy-Arqueología-Arqueologia" and three strategies.

### 4.2. Size of Comparable Corpora Generated with the Three Strategies

Taking CorpusPedia as input source, we performed several experiments to build comparable corpora (english-spanish, spanish-portuguese, and english-portuguese) containing texts on the same topic, namely Archaeology. We used the three strategies described in the previous section. The specific topic in both Spanish and Portuguese was selected with the corresponding translations of "Archaeology", that is: "Arqueología" in Spanish and "Arqueologia" in Portuguese. Table 2 summarizes the quantitative description of all generated corpora.

The table shows there are significant differences in size among the three language. As it was expected, the baseline strategy without alignment yields an English corpus with 730 thousand words in contrast to only 64 thousand in Portuguese. However the difference between Spanish and Portuguese (344 against 64 thousand) is less expected since Wikipedia contains more articles in Portuguese. Two reasons explains such a difference. First, the system found 420 Spanish articles sharing the category "Arqueología" against only 100 in Portuguese. This is in accordance with the fact that Portuguese articles tend to contain fewer categories than the Spanish ones. Second, the plaintext size of Spanish articles is larger than in Portuguese. This is easily confirmed by the results obtained using alignment techniques: given the same number of extracted articles (19 with strong alignment and 119 with soft alignment), the size of the Spanish corpus is about twice larger than in Portuguese. The same tendency is verified between English and Spanish. So, it follows the size of English articles is almost three times larger than that found in Portuguese. Yet, this is true only concerning aligned articles. Those English articles that were not aligned (i.e., without their corresponding versions in Spanish or Portuguese) are much smaller than those aligned with their Spanish and Portuguese versions. All those significant asymmetries should be taken

| English articles | Spanish Articles |
|---|---|
| Adena culture | Cultura Adena |
| Afanasevo culture | Cultura Afanasevo |
| Alalakh | Alalakh |
| Alexandria National Museum | Museo Nacional de Alejandría |
| Amazons | Amazona (mitología) |
| Ancient footprints of Acahualinca | Huellas de Acahualinca |
| Antiguo Oriente | Antiguo Oriente |
| Antikenmuseum Basel und Sammlung Ludwig | Museo de arte antiguo de Basilea y colección Ludwig |
| Apadana | Apadana |
| Archaeological Museum of Asturias | Museo Arqueológico de Asturias |
| Archaeological Museum of Granada | Museo Arqueológico y Etnológico de Granada |
| Archaeological Survey of India | Servicio arqueológico de la India |
| Archaeology | Arqueología |
| Archaeology of the Americas | Prehistoria de América |
| Archaic period in the Americas | Periodo arcaico de América |

Table 3: Sample of titles extracted from en/es soft-alignment.

into account to build, not only homogeneous corpora according to a specific topic, but also balanced resources with regard to corpus size.

Finally, Table 3 shows a sample of bilingual titles english-spanish representing some of the articles extracted using the soft alignment strategy. Lists of bilingual pairs as those depicted in 3 allow us to observe the degree of comparability between texts in both languages. A large-scale automatic evaluation of quantitative features will be the goal of further experiments.

## 5. Conclusions and Future Work

The emergence of multilingual resources, such a Wikipedia, make it possible to design new methods and strategies to compile corpus from the web, methods that are more efficient and powerful than the traditional ones. In particular, the semi-structured information underlying Wikipedia turns out to be very useful to build comparable corpora. On the one hand, editors classify articles with categories corresponding to topics or genders and, on the other, a network of interlanguage links enables to create bilingual relations between articles.

Our current research is focused on how to improve the strategies by extending coverage (more articles) without losing accuracy (the same topic). For this purpose, we are testing and evaluating two techniques to expand categories using a list of similar terms: those tagged as *related* in CorpusPedia and those identified as hyponyms or co-hyponyms of the source category. In order to find hyponyms and co-hyponyms of a term, it is required to make use of an ontology well suited to encyclopedic knowledge. One of our current tasks is to build an ontology of categories using the semi-structured information of Wikipedia (Chernov et al., 2006).

Finally, in future work, we will define an evaluation protocol to measure the degree of comparability between texts. For this purpose, we will make use of techniques described in (Saralegui and Alegria, 2007).

## 6. References


S.F. Adafre and M. de Rijke. 2006. Finding similar sentences acorss multiple languages in wikipedia. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 62–69.

Sergey Chernov, Tereza Iofciu, Wolfgang Nejdl, and Xuan Zhou. 2006. Extracting semantic relationships between wikipedia categories. In *SemWiki2006 - From Wiki to Semantics*, Budva, Montenegro.

M. Clark, Ian Ruthven, and Patrick O'Brian Holt. 2009. The Evolution of Genre in Wikipedia. In *Proceedings of JLCL 2009*, volume 24, pages 1–22.

Elena Filatova. 2009. Directions for Exploiting Asymmetries in Multilingual Wikipedia. In *CLEAWS3*, pages 30–37, Colorado.

Belinda Maia. 2003. What Are Comparable Corpora. In *Workshop on Multilingual Corpora: Linguistic Requirements and Technical Perspectives*, pages 27–34, Lancaster, UK.

M. Pottast, B. Stein, and M. Anderka. 2008. A wikipeda-based multilingual retrieval model. In *Advances in Information Retrieval*, pages 522–530.

X. Saralegui and I. Alegria. 2007. Similitud entre documentos multilnges de carcter cientfico-tcnico en un entorno Web. In *Procesamiento del Lenguaje Natural*, page 39.

J. Toms, J. Bataller, and F. Casacuberta. 2001. Mining Wikipedia as a Parallel and Comparable Corpus. In *Language Forum*, volume 1, page 34.

M.F. Tyers and J.A. Pieanaar. 2008. Extracting Bilingual Word Pairs from Wikipedia. In *LREC 2008, SALTMIL Workshop*, Marrakech, Marroco.

Kun Yu and Junichi Tsujii. 2009. Bilingual dictionary extraction from wikipedia. In *MT Summit XII*, Ottawa, Canada.

# Trillions of Comparable Documents

**Pascale Fung, Emmanuel Prochasson and Simon Shi**

Human Language Technology Center
Hong Kong University of Science & Technology (HKUST)
Clear Water Bay, Hong Kong
{pascale,eemmanuel,eesys}@ust.hk

**Abstract**

We propose a novel multilingual Web crawler and sentence mining system to continuously mine and extract parallel sentences from trillions of websites, unconstrained by domain or url structures, or publication dates. The system is divided into three main modules, namely Web crawler, comparable and parallel website matching and parallel sentence extraction. Previous methods in mining parallel sentences from the Web focus on specific websites, such as newspaper agencies, or sites sharing the same URL parents. The output of these previous systems are limited in scope and static in nature. As the Web is boundless and growing, we propose to continuously crawl the Web and update the pool of parallel sentences extracted. One main objective of our work is to improve statistical machine translation systems. Another objective is to take advantage of the heterogeneous website documents to discover parallel sentences in henceforth undiscovered domains and genres, such as user generated content. We investigate a host of recall-oriented vs precision-oriented algorithms for comparable and parallel document matching, as well as parallel sentence extraction. In the future, this system can be extended to mine other monolingual or bilingual linguistic resources from the Web.

## 1. Introduction

As statistical approaches become the dominant paradigm in natural language processing, there is an increasing demand for data, more data, and yet more data. Just little more than a decade ago, "large corpora" used to mean a collection of user manuals, or 5 years of newspaper articles. The first statistical machine translation (SMT) system using the IBM model (Brown et al., 1990) was trained on a parallel corpus of Canadian parliamentary transcriptions in English and French - the Hansard, which amounted at the time to 117,000 sentence pairs. Fast forward to 2010, state-of-the-art SMT systems are trained on tens of millions of sentence pairs consisting of hundreds of millions of words. Much of the parallel data used to train SMT systems are manually translated by professional translators. The standard rate for such an effort is about US$0.15 per word, making good SMT systems extremely expensive to build. Organizations such as the Linguistic Data Consortium have been distributing some large corpora of translated texts for research and development at a lower cost to the user than directly commissioning translators. However, as SMT systems typically perform better on texts within the same genre as its training data, general purpose, open-domain SMT systems are only attainable if the developers of such systems have access to the world's data.

In today's world, only the most powerful search companies are privy to such information. One organization with such access - Google, the world's top search engine company, whose mission is to "organize all the world's information", has access to trillions of websites, billions of email content, videos, images, speech files, and other user generated content. As of March 2009, the (indexable) Web contains at least 25.21 billion pages (World Wide Web Size, 2009). Google search had discovered one trillion unique URLs. And its translation system is statistically trained from all the data that is within its grasp. Google, while having this access, does not distribute the result of its mining to the public, except through its services. Yet, as the Web founder Tim Berners-Lee famously put it, "*The power of the Web is in its universality. Access by everyone regardless of disability is an essential aspect.*"

In this paper, we address the "disability" of statistical natural language research in general, and SMT systems in particular, to access the information on the Web as a training corpus, and propose a multilingual Web crawling and mining system as a tool to facilitate our community to mine the Web for more linguistic resources.

The World Wide Web is a "*boundless world of information interconnected by hypertext links*". We argue that the Web is a virtually infinite and continuously growing corpus for natural language processing. Rather than taking a snapshot of it at one moment, and use the result as a static corpus, we propose to continuously crawl the Web for new, comparable data for mining parallel sentences. Rather than focusing on a single domain such as news, or on translated parallel sites with matching structures, we propose to look for sites that are comparable in content, HTML structure, link structure, URL as well as in temporal distance as they potentially contain parallel sentences.

Much effort has been made in the past to try to automatically extract parallel resources from comparable corpora on one hand, and to use the Web as a corpus on the other. Both approaches (often combined) allow more diversity in the data harvested. (Resnik and Smith, 2003) directly extracted parallel texts from the Web, relying mostly on URL names. Some work has been done to extract parallel resource (sentences, sub-sentential fragments, lexicon) from comparable data. (Munteanu and Marcu, 2005) showed they can extract relevant parallel sentences using a supervised approach on newspaper corpora, although their main goal was to show how they manage to use such resources to improve Statistical Machine Translation. (Fung and Cheung, 2004; Wu and Fung, 2005) extracted parallel sentence from quasi-comparable corpora, that is corpora containing



*Proceedings of the 3rd Workshop on Building and Using Comparable Corpora, LREC 2010*, pages 26–34
Malta, 22 May 2010

documents from the same domains as well as documents of different domains.
We need to be able to combine advanced IR/Web crawling techniques with advanced NLP methods in order to obtain large and high quality sets of parallel sentences. From this point of view, we do not want to focus on one particular domain (such as newspaper, as it is often the case in related works). Of course, we are aware and will keep in mind than better results can be obtained from certain kind of documents (for example, Wikipedia constitutes a large source of very comparable, easy to harvest and well structured documents), but propose a general approach for mining from any website, in any dominant Web language. We strive to reduce the language dependency and domain dependency to a minimum.
This is work in progress and this paper is intended as a position paper to present our objectives and arguments to the community of NLP researchers. In the next section, we take a look at the challenges that we encounter and how we plan to solve them, step by step. Section 3 describes the experimental setup and preliminary results of our experiments. We then conclude in Section 4 and discuss future directions in Section 5.

## 2. Challenges

Existing tools (Munteanu and Marcu, 2006; Resnik and Smith, 2003; Ma and Liberman, 1999) mine parallel sentences from a pre-defined set of archival data, with temporal and domain constraints. Some of these tools do not crawl the Web but rather, they try to mine parallel texts (Resnik and Smith, 2003) or parallel sentences (Munteanu and Marcu, 2006) from a pre-existing archive. (Ma and Liberman, 1999; Chen and Nie, 2000) developed tools that dynamically mine parallel sentences from a subset of the Web. However, these tools have become obsolete over time and the Web has since grown tremendously in the last decade. Most other methods of mining parallel sentences from comparable or parallel corpora require training from existing parallel corpora and therefore, are often only applicable to a single domain or genre. Many issues related to the challenge of mining parallel sentences from the Web has been studied and some interesting achievements have been made.
Two strategies can be adopted when mining parallel sentences: favoring recall or precision. Favoring recall will provide many pairs of sentence, but the quality of those pairs (the parallelness) is likely to be low. However parallel sub-sentential fragments (Munteanu and Marcu, 2006) can still be of great value, especially if they can be post-processed to filter out the non-parallel segments (Abdul-Rauf and Schwenk, 2009). On the other hand, favoring precision yields high quality parallel sentences (moreover, reliable alignment of sentences) at the cost of probably missing many valuable information. We focus on both approaches. For the purpose of improving statistical machine translation systems, we need to mine parallel sentences with high precision, measurable by SMT performance, not just human judgment. At the mean time, as "more data is better data" for statistical MT systems, we will also strive to improve the recall rate, while maintaining precision. We are also interested in obtaining large amounts of data quickly.
Last but not the least, even though our current objective is to mine parallel sentences from the Web, it is potentially useful to crawl the Web for other language resources, such as translation lexicons, or monolingual resources. Since the Web crawling and indexing task is non-trivial and time consuming, we need to design the system so that useful information are retained for future processing, without having to recrawl the Web for the same pages.
To summarize, we need to meet the following challenges for our task of mining parallel sentences from the Web:

1. Recall - include as many websites as possible that might contain parallel sentences
2. Precision - to be able to find high quality parallel sentences that can improve SMT performance
3. Domain and topic - to be able to find parallel sentences in as many domains/topics as possible
4. Language - to be able to find parallel sentences in different language pairs
5. Heterogenous - the system must find websites that are not just translations of each other but also others that have similar content
6. Up-to-date and always available - the system needs to crawl the Web continuously for new additional document resources
7. Query-driven - the system can accept queries to crawl and search for specific websites
8. Scalability - the system needs to be scalable to run on multiple nodes of servers in parallel.
9. Speed - fast algorithms are needed to enable us to crawl the Web efficiently for the mining task.
10. Extendable - the system needs to be modular and extendable to other mining tasks, in addition to parallel sentence mining.

The whole process is described in figure 1 and the different modules are described in the following sections.

### 2.1. Crawling the Web

A Web crawler is a program that automatically downloads pages from the Web. To mine parallel sentences from the entire World Wide Web continuously and automatically, a main component of our tool is a Web crawler that collects as many documents from the Web in a given language pair continuously and indexes each page for comparable document searching. The Web crawler indexes Web pages on the Web to enable them to be searchable. The main function of our system currently is to act as an comparable document search engine which discovers articles in another language that are comparable or parallel to any input text. So in the first stage, we need to crawl and index both the English Web (i.e. all English websites) and the Chinese Web. We build an index including all English pages like a search engine. When the index has reach a certain size, say 1M pages, we

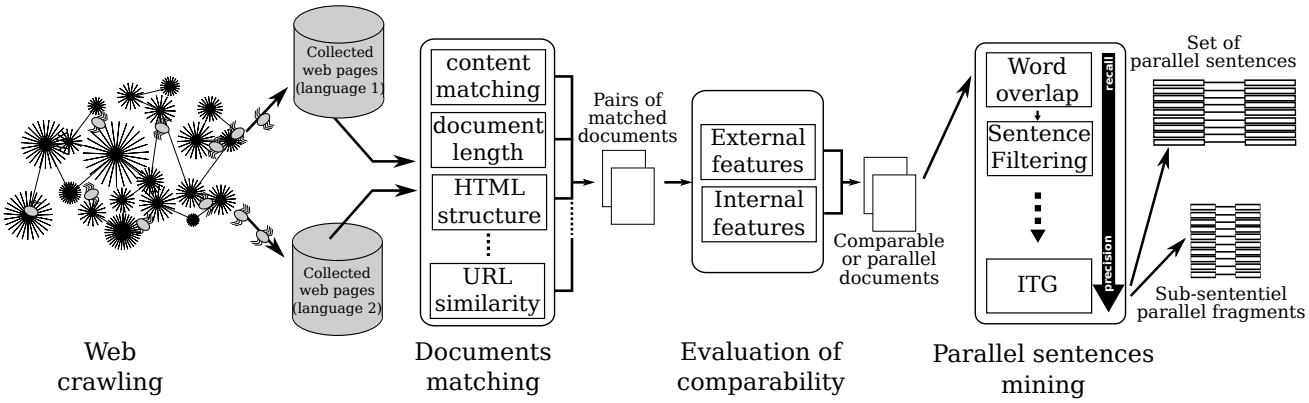



Figure 1: Overview of the sentence extraction system

will process each Chinese page to find its comparable English page in the index.

Queries to conventional search engines normally contain one or more distinct keywords. However, the query to our system at this stage is a document which may contain hundred of words. The tool searches the index and finds documents in another language that are comparable to the input. It is a high dimensional search problem with time complexity of $O(n \times m)$ where $n$ and $m$ are number of websites in the two languages (i.e. English and Chinese respectively). (Gionis et al., 1999) introduced a hashing method for high dimensional similarity search which can be used to reduce computation time. For our purposes, we suggest that some kind of topic or genre clustering can be carried out first to reduce the search dimension. Methods for topic classification, taking into consideration content and other information, can be used to speed up the search as well.

After we indexed a significant amount of Web pages, say 1 million pages, we start to use the search engine to get comparable documents. For each Chinese document, we first translate it into English by an MT system, such as Google Translate, or simply convert it to the word index in a bilingual lexicon. Then by searching the index we can obtain a ranked list of English texts, in terms of comparability. Those document pairs are returned as the output of the search engine. We assume that for each Chinese documents, there will be some comparable documents in English.

Simple bag-of-words comparison cannot tell us whether two pages are actually comparable, noisy parallel or parallel. So we will need other measures, described in the next section, to achieve our mining objective. In consideration of such measures, we must first index the websites accordingly. In our system, the following features are considered in the indexing step:

- Page content in terms of words
- Position of words in the document
- URL structure
- HTML structure
- Link structure
- Image file names
- Time of creation if relevant

During indexing, unlike conventional Web crawlers, we must convert all information above into index numbers. Word IDs, for example, must correspond to those in a bilingual lexicon for our source and target languages. Multiple translations of the same word can be considered. Word features such as tf/idf, frequency rank within the same page, word positions, etc. should be indexed.

In addition, the Web crawler is configured to collect different types of documents by various regular update intervals. A stochastic model for crawl target selection (Akamine et al., 2009) is implemented to control the revisit time of the crawler in order to keep the document up-to-date. For news websites, Web pages can be collected daily by the crawler while the visit frequency of other websites can be much longer.

Previously, (Chen and Nie, 2000; Yang and Li, 2004; Gleim et al., 2006) developed a parallel text mining system on bilingual websites sharing the same root URL. (Munteanu and Marcu, 2005) focused on some news websites only. They tried to extract parallel sentences from given sets of known websites without crawling the Web. Whereas the result of such work has shown to improve SMT performance, many parallel sentences exist on other websites and the sentence pairs reside on different hosts are never discovered by their more limited and static approach. (Chen and Nie, 2000) developed a tool PTMiner which mines parallel sentences under the same hostname. The Web crawler of PTMiner performs breadth first search on the same host only. In our case, we must crawl and index boundless number of websites (hostnames) continuously, rather than search for and download a part of the Web only like these previous work.

The Web crawling speed is mainly constrained by connection bandwidth. In the initial testing, we crawl the Web using 10 spiders over Ethernet, reaching the speed of one page per second. For indexing each page, a single PC with Core Duo processor at 2.0GHz is able to index 50 pages per minute. With very limited optimization, a PC running as the database server takes 10 seconds to process each Chinese document when there are 10,000 pages in the database.

We use MySQL as the central database server which is scalable to run on clusters. The Web crawlers work independently. It is possible to have several groups of spiders to crawl the Web and index pages.
We also use a black list to avoid crawling sites containing mostly non-textual material, such as YouTube, Picasa, Flicker, etc.

### 2.2. Matching comparable and parallel documents

To improve the recall of mining parallel sentences, we need to be able to measure and classify document pairs into not comparable, quasi-comparable, comparable, noisy parallel and parallel in order to match them better. As mentioned above, using quantitative measures, we will select documents that are comparable and noisy parallel (including parallel). According to (Fung and Cheung, 2004), quasi-comparable and comparable documents are those that were written independently but on more or less the same topic. In such cases, structural features are not useful. Noisy-parallel documents refers to a pair of source and translated document, that were either adapted or evolved in different ways. For example, Wikipedia article that was once the translation of another Wikipedia page, but evolved in time due to different contributors can be either noisy parallel or comparable to the source article.
In order to improve recall of parallel sentences between two texts, it is important to select very comparable documents but not be restricted to translated, parallel documents only. The notion of comparability is hazy and is still an open question. Practically, it depends on the expected usage of the documents. The comparability is generally evaluated on both internal and external criterion. External criterion are qualitative features, such as the topic, the domain, the time of publishing or the discourse, whereas internal criterion are quantitative features, such as the quantity of common vocabulary.
(Kilgarriff, 2001) tried to answer a related question by measuring the similarity of two corpora. He observed that such a measure is not trivial since corpora are complex and multidimensional objects. Two corpora can be close for one dimension and distant for another. In this context, the notion of similarity is connected to the notion of homogeneity in one corpus. A homogeneous corpus contains the same kind of document (Biber, 1989), that is, where some particular linguistic distinctiveness can be found. We focus on comparable documents rather than a collection of corpora. The question of homogeneity is in our case not really relevant. We therefore focus on different features, external and internal. (Fung and Lo, 1998; Fung and Cheung, 2004; Carpuat et al., 2006) previously proposed to compare the frequency rank of seed words in documents to be matched. Similar documents should have a similar representation of the common vocabulary. Such comparison can be visually evaluated, see Figure 2. Identical documents should rise a perfect diagonal, unrelated documents should show no such tendency. To quantify the similarity of documents, we also use a regression score which evaluate the dispersion of the data from the diagonal.
This score works well for documents containing a significant number of content words, but is brittle on smaller documents. If few seed words are found between two documents, the dispersion will be small, whereas documents with many common seed words might be seen more similar, since more dots will be compared. Therefore, we need to weight the raw score to get more significant information. An example is given in Figure 2: less than 50 words are common to both texts, which is too sparse for our measure. We then need to rely on other features to evaluate comparability or to be more precise, to evaluate whether two documents might contain translated sentences.

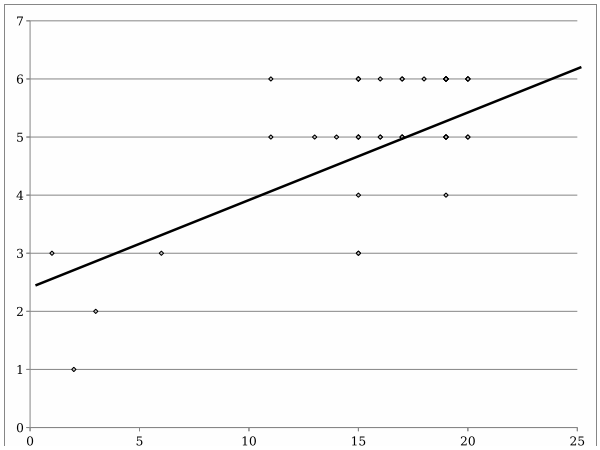



Figure 2: $R^2$ computation on two parallel documents about Lamma Island.

(Resnik and Smith, 2003) looked for pairs of document in translation by searching for specific link in a parent page (with links to several version of one document, in many languages) or in sibling pages (with link such as "this document in English"). We suggest that external features can be used, such as URL structure, document length, html structure, link structure, or image file names.

### 2.3. Mining parallel sentences

Mining parallel resources from comparable corpora has been done in several studies. (Munteanu and Marcu, 2005) proposed an approach to mine parallel sentences from selected comparable documents using a supervised Maximum Entropy classifier. One goal of their work was to rely on large amount of out-of-domain parallel data and small amount of in-domain parallel data to complete in-domain knowledge for MT. The initial parallel data are used to train the EM classifier, which will determine which sentences are good translation candidates (based on many features, starting with word overlap and length ratio of pairs of sentences). They work on newspaper data in English, Chinese and Arabic. (Fung and Cheung, 2004) looked for parallel sentences and bilingual lexicon from very non-parallel corpora, defined as collection of document on the same topic (in-topic) or not (off-topic). Rather than relying on the "*find-topic-extract-sentence*" principle (e.g. find in-domain documents, then look for translations), they proposed to "*find-one-get-more*". In other words, if parallel sentences have been found between two documents, they are likely to share more parallel sentences. They used a cosine sim-

ilarity measure to compare pairs of sentence and raised pairs above a given threshold for English/Chinese alignment. This approach raised interesting parallel resources, but they were shown to be quite scarce among unrelated documents. Furthermore, this approach applies on large amount of data.

For texts that are translations but contains a lot of noise, such as one-to-many translations, or inserted examples and graphs, or even occasional segments that are not translations of each other, we propose to adapt the DK-vec algorithm (Fung, 1998; Fung and McKeown, 1997; Fung, 1995) which use an iterative Dynamic Time Warping method to match a bilingual lexicon, used later as anchor points to align sentences. This method is interesting for it is totally unsupervised and language independant: the bilingual resources can be boostrapped from the document. Furthermore, this approach has been shown to be efficient for document without strict sentence boundary information. It was designed for noisy-parallel corpora, basically yielding a path of lexicon alignment that is not necessarily the diagonal if there is noise. DK-vec is also unique in that it uses the position feature and the (sentence) length feature implicitly in the dual objective of alignment and bilingual lexicon extraction. Other methods either use an existing lexicon and position feature to perform alignment, or use the length feature for alignment.

Finally, the results provided by high-recall method can be filtered, for example using Inversion Transduction Grammar (Wu and Fung, 2005). When using word overlap methods (or cosine similarity), sentences that share a common vocabulary but do not have the same meaning are likely to yield a high score. As an example, this pair of sentence, extracted from French newspaper *Le Figaro* and English *New York Times* obtain a high score when using word overlap:

> **En:** "National Highway Traffic Safety Administration has received about 100 complaints involving the brakes of the Prius new model."
>
> **Fr:** « Aux Etats-Unis, une centaine de plaintes ont été déposées auprès de l' administration de sécurité routière américaine pour des difficultés de freinage avec la Prius. »
>
> **Trans:** *"In the United States, about one hundred complaints have been submitted to the american administration of traffic safety for difficulties when braking with the Prius"*

Even though both sentences have roughly the same meaning, they cannot be considered parallel. ITG can then be used to take a closer look at the sentence constituent structures (predicate argument dependencies) and will eventually allow us to filter out this candidate pair, to only keep strictly parallel candidates. ITG has been shown to be efficient for this particular task and are language independant.

All in all, the overall process, from crawling the Web to parallel sentence extraction can be seen as refining a raw material (the Web) to obtain golden resources, each of the step attempting to filter out irrelevant data.

## 3. Preliminary Experiments

We ran an experiment to roughly evaluate the feasibility of our task by trying to extract parallel sentences from a subset of French and English Wikipedia. It is hard to precisely estimate the amount of parallel sentences available from the Web, for several reasons:

- the availability and density of parallel sentences is highly related to the type of document processed; the Web is a heterogeneous resource. It is not possible to infer an accurate estimation from a small subset evaluation.
- assuming we already had a high-recall and -precision tool to mine parallel sentence from the Web, we can not ensure we have found them all (recall estimation is, in that case, impossible). We can estimate the precision on a small subset, but the precision is also related to recall.

It would be presumptuous therefore to claim anything regarding the density of parallel sentences from the Web, however we might still want to have a look, at least to confirm that there are some, and that they can be extracted automatically.

### 3.1. Experimental setup

We randomly extracted 1,000 pairs of articles from French and English Wikipedia by considering articles with the exact same title (at the time we write this paper, there were 548,900 pairs of articles available). Most of these articles refer to proper names (e.g. biography of a famous figure, book titles, other works) and few of them are animal species. No distinction was made for articles that are translations or just comparable. We tried to mine parallel sentences in pairs of documents only, using a simple word-overlap measure and a French-English bilingual dictionary. The word-overlap score is evaluated based on the number of common words between two sentences, penalized by the number of words whose translation is in the dictionary and that can not be found in the other sentence. The word-overlap score is detailed in equation 1.

$$wo(S_1, S_2) = \frac{|S_1 \cap S_2|}{|S_1 \cup S_2| + |S_1 - S_2| + |S_2 - S_1|} \quad (1)$$

In equation 1, intersections/disjunctions of set is computed only on known elements, no penalty is imposed on unknown words. The sentences are cleaned to filter functional words using a list of stop words in English and French. We used a threshold to keep interesting candidates ($> 0.2$). This threshold is arbitrary and can be increased to maximize precision, but will allow us to observe the translation candidates.

### 3.2. Results

Using this experimental setup, we extracted 1,233 candidate translations. The top-ranked ones happen to be correct but are mostly useless, as they concerns short titles or structure information (typically, we obtained 29 occurrences of the correct translation *Reference*/*Références*, and

37 occurrences of *See Also*/*Voir Aussi*). We also obtained many alignment of dates or proper nouns, or alignment of irrelevant data. Moreover, due to the type of documents found in Wikipedia (especially given the constraints of selection we used), we found many "identical matches", such as *Ravat-Malvern Star*/*Ravat-Malvern Star*. This kind of alignment accounts for more than 85% of the sentences extracted. This latter observation shows that cleaning documents from the web is an issue that should not be underestimated. We need to ensure to process *content* of webpages, and make sure to get rid of useless information such as menus or advertising.
Apart from short sentence alignment, we can classify other candidates into three groups:

- Exact parallel sentences. Same meaning, same organisation of the sentence, same amount of information.
- Partially parallel sentences. One sentence is likely to contain more information, or they are organised differently. Those can still be of interest if they can be post-processed.
- False Positives. Sentences that were matched but don't share a common meaning.

Surprisingly, we obtain very few false positives with a score higher than $0.25$. One example is given below:

> **English:** The DC2 Type R was the only Type R ever sold in North America (With the Acura badge)
>
> **French:** Les Honda Type R sont les modèles sportifs les plus performants du constructeur Honda automobile.
>
> **Translation:** *The Type R Honda are the most performant race models from the Honda motor company.*

Overall, we found about 12 true wrongly aligned sentences only. This is a very interesting result, since it shows that, as long as we ensured that i) the documents we are processing are strongly comparable and ii) we found some translation candidates with a high enough score, those candidates are reasonably reliable. We found about 150 parallel or partially parallel sentences. They were manually classified and some examples are given in Tables 1 and 2.
These results are interesting because they show a reasonable amount of parallel sentences can be found. However, as we emphasized previously, these results can not help us evaluate precision/recall or ratios of parallel sentences among documents from the Web. Our ultimate goal is not to harvest parallel sentences from Wikipedia, in French and English. Some effort will be necessary to obtain more interesting results from the rest of the Web.

## 4. Conclusion

We argue that it is possible to mine a heterogenous corpus of parallel sentences in the dominant Web languages, in any domain and any topic, from the Web. We propose to combine sophisticated information retrieval methods with statistical natural language processing methods to better harvest the material from the Web. Many assumptions made by previous work do not hold as we move from mining from limited domain, and limited genre websites to the entire Web. We suggest that an optimal combination of recall-oriented algorithms and precision-oriented ones will enable us to mine the gold nuggets - linguistic resources - in the information ocean that is the World Wide Web. The Web is boundless and amorphous. The innovation of our proposed work lies in our consideration of the Web as a dynamic, time-variant corpus, rather than a static archive. We propose a combination of content, structural, and temporal features to crawl the Web with the objective of continuously mining useful multilingual linguistic resources such as comparable or parallel corpus. We suggest to investigate a host of recall vs precision-oriented methods to mine parallel sentences from comparable websites returned by our Web crawler. Some initial experimental results have been shown as the existence proof of parallel sentence pairs in non-parallel websites, such as the Wikipedia.

## 5. Discussion, future work

This project is large and ambitious, and each step will require extensive study of state-of-the art approaches, and hopefully improvement of previous approaches. As we mentioned, many of the assumptions made previously might or might not hold for a project at this scale. For example, relevant documents that are useful for cross-lingual retrieval, based on page-ranked search results, might not contain any parallel sentences. Websites that are not translations of each other, might still contain parallel segments. Extraction systems using classifiers and rankers trained from an in-domain corpus are not applicable to our system as we do not focus on any specific domains. Nevertheless, it can be useful to classify the final extracted sentences into different domains for training domain-specific SMT systems.
With the rising popularity of Web 2.0 and Web 3.0 websites, there are more and more user generated content on the Web and many of them relate to each other in very interesting ways, such as user feedback on the latest Apple products, fan club discussion on the latest gossip of a celebrity. Such topics are temporal in nature - and available in multiple languages. Our system downloads and compares these websites as part of its output. We would like to analyze the results and see whether such data can be used to improve an SMT system on user generated content.
The Semantic Web is another effort by the W3C community to improve upon the current HTML annotation of Web pages to include the "meaning" of Web content for Web browsers and search engines to better "understand" and satisfy user queries. When mature, the new semantic annotation scheme can potentially provide a new feature, the semantic feature, to our system in mining and comparing websites.
A problem that remains to be addressed by our system is that there are many more parallel (and other) data available on the Web than those indexed by a search engine or by our system - there are compressed files of translated texts, such as the United Nations Parallel Corpus, or image files of scanned documents, such as books in translated into multiple languages, contents of tables, subscription-based

| French | English |
|---|---|
| Histone H4, un composant de la structure de plus haut niveau de l'ADN des cellules eucaryotes | Histone H4, a component of DNA higher structure in eukaryotic cells |
| L'important engagement d'Henry Ford à réduire les coûts aboutit à de nombreuses innovations techniques et commerciales, notamment un système de franchise qui installe une concession dans toutes les villes en Amérique du Nord et dans les grandes villes, sur les six continents. | Henry Ford's intense commitment to lowering costs resulted in many technical and business innovations, including a franchise system that put a dealership in every city in North America, and in major cities on six continents. |
| Dans celui-ci les angles sont confinés à un plan ; donc l'étape suivante devrait être une algèbre quadruple quand l'axe du plan devient variable. | In it the angles are confined to one plane ; hence the next stage will be a quadruple algebra, when the axis of the plane is made variable. |
| Le segment six a un motif semblable mais avec moins de bleu et le segment sept est presque entièrement noir, avec seulement une fine bande bleue à la base. | Segment six has a similar pattern but with more restricted blue and a broader area of black, and segment seven is mostly black, with just a narrow blue area at the base. |
| Swami Shivananda Saraswati ( 8 septembre 1887 - 14 juillet 1963 ) est un maître spirituel hindou très réputé et un promoteur du Yoga et du Vedanta. | Swami Sivananda Saraswati ( September 8, 1887 July 14, 1963 ) was a Hindu spiritual teacher and a well known proponent of Sivananda Yoga and Vedanta. |

Table 1: Sample of parallel sentences extracted.

| French | English |
|---|---|
| **L'album est sorti le 18 novembre 2009 sous le label Regain Records.** | **The album was** officially **released on November 18, 2009 via Regain Records.** |
| **De 1977 à 1981, il travaille dans l'équipe la Commission des vétérans à la Chambre des représentants.** | **From 1977 to 1981,** Webb **worked on the staff of the House Committee on Veterans Affairs.** |
| Elle **donne** à toute personne recevant le logiciel **le droit** illimité **de l'utiliser, le copier, le modifier, le fusionner, le publier, le distribuer, le vendre et de changer sa licence.** | The MIT License **states** more explicitly **the rights** given to the end-user, including the right **to use, copy, modify, merge, publish, distribute, sublicense, and/or sell the software.** |

Table 2: Sample of partially parallel sentences extracted.

websites etc. This Deep Web (or Hidden Web) is orders of magnitudes larger than the visible Web. The current Web reachable by search engines is about 167 terabytes whereas the Deep Web is estimated to be 91,000 terabytes. Whereas developing a comprehensive tool to crawl the Deep Web is perhaps beyond the scope of our proposed work, for a specific natural language task, such as SMT, we might want to dig deeper into a specific genre of data.

One of the most interesting part, and a cornerstone of this work is the ability to evaluate comparability. This is a particularly tricky question, since the comparability concept itself is hazy. Some assume than noisy-parallel corpora are comparable, some assume that document in each languages has to be written independently while others claim there is a continuum from non-related to parallel corpora. Quantitatively and qualitatively evaluating the comparability might bring to light a more precise definition of comparability and comparable corpora. For websites, structural comparability does not necessarily lead to content comparability. Given the large amount of websites, should we first constrain our search with URL structural matching as in (Resnik and Smith, 2003)? Or should we start with the least stringent criteria for recall? We argue for the latter. All Wikipedia articles have similar URL names and HTML structures, but with very different content. For example, Chinese Wikipedia is clearly not a translation of the English Wikipedia.

As we mentioned that our system aims to help users mine multilingual resources from the Web for more than one applications. As an example, one of the main interest in comparable corpora concerns bilingual lexicon extraction, which is generally performed on large corpora (millions words (Fung, 1995; Rapp, 1995)) following *the more data is better data* principle, or relying on smaller but more constrained, specialized corpora (Daille and Morin, 2005; Chiao and Zweigenbaum, 2002) to focus on terminology. Both approaches fail to find relevant translations for rare words, for two reasons: (1) Even in large corpora, there is no guarantee that a source word will occur in the target corpus (Zip's law); (2) these approaches mostly rely on context-based comparison - a word and its translation are likely to have similar contexts, just as a word and its synonyms share the same context (following the Firthian principle that "you shall know a word by the company it keeps" (Firth, 1957)). Rare words by definition do not occur frequently enough to create a meaningful context and cannot be compared efficiently. (Pekar et al., 2006) tried to circumvent this issue by smoothing the context of rare words using the context of their k-nearest neighbors. They obtained a significant improvement in the quality of the lexicon alignment, by lowering the rank of correct translation candidates.

This raises another interesting question: Are there rare words in the Web? Does the notion of hapaxes still exist? There is a direct answer: yes, of course. First, one can invent a word that could not be found anywhere else, but this is a trivial case. Rare words occur in languages that are scarcely represented on the Web. Apart from these cases, can we assume than all the words and terms of the world's top Web languages can be found on the Web? A related experiment done by our group found that all Chinese named entities in the Wikipedia pages are translated into English somewhere on Chinese websites. A simple regular expression search can return the translation results.

Finally, rather than relying on large quantities or highly constrained corpora, we believe we can take advantage of the diversity and availability of comparable documents (and typically, take advantage of the availability of comparable documents in many languages, to perform multi-source alignment). A lexicon acquired in such a way can be used as feedback to the whole sentence alignment process, to increase the quality of word overlap estimation and comparability evaluation, raising better matched documents and higher quality parallel sentences.

## 6. References


Sadaf Abdul-Rauf and Holger Schwenk. 2009. On the use of comparable corpora to improve SMT performance. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL'06)*, pages 16–23.

Susumu Akamine, Yoshikiyo Kato, Daisuke Kawahara, Keiji Shinzato, Kentaro Inui, Sadao Kurohashi, and Yutaka Kidawara. 2009. Development of a large-scale web crawler and search engine infrastructure. In *Proceedings of the 3rd international Universal Communication Symposium (IUCS'09)*, pages 126–131.

Douglas Biber. 1989. A typology of english texts. *Linguistics*, 27:3–43.

Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. 1990. A statistical approach to machine translation. *Computational Linguistic*, 16(2):79–85.

Marine Carpuat, Pascale Fung, and Grace Ngai. 2006. Aligning word senses using bilingual corpora. *ACM Transactions on Asian Language and Information Processing*, 5(2):89–120.

Jiang Chen and Jian-Yun Nie. 2000. Parallel web text mining for cross-language information retrieval. In *Recherche d'Informations Assistée par Ordinateur (RIAO)*, pages 62–77.

Yun-Chuang Chiao and Pierre Zweigenbaum. 2002. Looking for candidate translational equivalents in specialized, comparable corpora. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING'02)*, pages 1208–1212.

Béatrice Daille and Emmanuel Morin. 2005. French-English Terminology Extraction from Comparable Corpora. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing (IJCLNP'05)*, pages 707–718.

John Firth. 1957. *A synopsis of linguistic theory 1930-1955*. Studies in Linguistic Analysis, Philological. Longman.

Pascale Fung and Percy Cheung. 2004. Mining Very-Non-Parallel Corpora: Parallel Sentence and Lexicon Extraction via Bootstrapping and EM. In Dekang Lin and Dekai Wu, editors, *Proceedings of Empirical Methods on Natural Language Processing (EMNLP'04)*, pages 57–63, Barcelona, Spain.

Pascale Fung and Yuen Yee Lo. 1998. An IR approach for translating new words from nonparallel, comparable texts. In *Proceedings of COLING-ACL98*, pages 414–420.

Pascale Fung and Kathleen McKeown. 1997. A technical word- and term-translation aid using noisy parallel corpora across language groups. *Machine Translation*, 12(1/2):53–87.

Pascale Fung. 1995. Compiling bilingual lexicon entries from a non-parallel English-Chinese corpus. In David Yarovsky and Kenneth Church, editors, *Proceedings of the 3rd Workshop on Very Large Corpora (VLC'95)*, pages 173–183.

Pascale Fung. 1998. A Statistical View on Bilingual Lexicon Extraction: From Parallel Corpora to Non-parallel Corpora. In David Farwell, Laurie Gerber, and Eduard Hovy, editors, *Proceedings of the 3rd Conference of the Association for Machine Translation in the Americas (AMTA'98)*, pages 1–16.

Aristides Gionis, Piotr Indyk, and Rajeev Motwani. 1999. Similarity search in high dimensions via hashing. In *VLDB '99: Proceedings of the 25th International Conference on Very Large Data Bases*, pages 518–529, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Rüdiger Gleim, Alexander Mehler, and Matthias Dehmer. 2006. Web corpus mining by instance of wikipedia. In *WAC '06: Proceedings of the 2nd International Workshop on Web as Corpus*, pages 67–74, Morristown, NJ, USA. Association for Computational Linguistics.

Adam Kilgarriff. 2001. Comparing corpora. *International Journal of Corpus Linguistics*, 6(1):1–37.

Xiaoyi Ma and Mark Liberman. 1999. Bits: A method for bilingual text search over the web. In *Proceedings of Machine Translation Summit VII*, page 6.

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-parallelcorpora. In *21st International Conference on Computational Linguistics (ACL'05)*.

Viktor Pekar, Ruslan Mitkov, Dimitar Blagoev, and Andrea Mulloni. 2006. Finding translations for low-frequency words in comparable corpora. *Machine Translation*, 20(4):247–266.

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.

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

World Wide Web Size. 2009. `http://www.worldwidewebsize.com/`.

Dekai Wu and Pascale Fung. 2005. Inversion transduction grammar constraints for mining parallel sentences from quasi-comparable corpora. In *IJCNLP*, pages 257–268.

Christopher C. Yang and Kar Wing Li. 2004. Building parallel corpora by automatic title alignment using length-based and text-based approaches. *Information Processing & Management*, 40(6):939 – 955.

# Improving Machine Translation Performance Using Comparable Corpora

**Andreas Eisele, Jia Xu**

{Andreas.Eisele,Jia.Xu}@dfki.de

DFKI GmbH, Language Technology Lab
Stuhlsatzenhausweg 3
D-66123 Saarbrücken Germany

**Abstract**

The overwhelming majority of the languages in the world are spoken by less than 50 million native speakers, and automatic translation of many of these languages is less investigated due to the lack of linguistic resources such as parallel corpora. In the ACCURAT project we will work on novel methods how comparable corpora can compensate for this shortage and improve machine translation systems of under-resourced languages. Translation systems on eighteen European language pairs will be investigated and methodologies in corpus linguistics will be greatly advanced. We will explore the use of preliminary SMT models to identify the parallel parts within comparable corpora, which will allow us to derive better SMT models via a bootstrapping loop.

## 1. Introduction

State-of-the-art machine translation based on the statistical approach is a data-driven process. The quality and quantity of the training data is crucial for the performance of a translation system. However, the increasing amount of training corpora can still not meet the demand of automatic translation on different language pairs and in various domains. Rich data are mostly available for few languages and only certain domains. There are still a great number of under-resourced languages. Thousands of languages are spoken by less than 50 million native speakers, with a big group of more than 200 languages that have between 1 and 50 million native speakers. Most of these languages are lacking sufficient linguistic resources. This brings difficulties to improve the translation qualities on these languages.

For instance, the majority of the European languages are under-resourced and lack both parallel corpora and language technologies for MT. The project ACCURAT (Analysis and Evaluation of Comparable Corpora for Under-Resourced Areas of Machine Translation) will focus on developing and evaluating language pairs of English-Latvian, English-Lithuanian, English-Estonian, English-Greek, English-Croatian, Croatian-English, English-Romanian, English-Slovenian, Slovenian-English, English-German, German-English, German-Romanian, Romanian-German, Greek-Romanian, Lithuanian-Romanian, Romanian-Greek, Romanian-English and Latvian-Lithuanian. We also work on the language pair of German and English which is well investigated previously. This can help us find the impact of comparable corpora on translations between language pairs with both rich and poor resources. More details can be found in (Skadina et al., 2010). The participants include organizations of Tilde, USFD, CTS, LISP, FFZG, DFKI, RACAI, Linguatec and Zemanta.

The main goal of the ACCURAT research is to find, analyze and evaluate novel methods how comparable corpora can compensate for this shortage of linguistic resources to improve MT quality for under-resourced languages and narrow domains. The work will be carried out on the listed European language pairs and adapted to narrow domains, e.g. automotive engineering. We expect an enhancement of language and domain coverage in MT.

The ACCURAT project will provide novel methodologies and models that exploit comparable corpora to enhance the translation quality of current MT systems, which are universal and can be used to new language pairs and domains. We will define criteria to measure the comparability of texts in comparable corpora. Methods for automatic acquisition of a comparable corpus from the Web will be analyzed and evaluated. Advanced techniques of obtaining parallel sentences and phrases from comparable corpora will be applied and extended to provide training and customization data for MT. Domain dependent MT will be exploited by automatic clustering of training data into genres according to their contents. Given limited amounts of available in-domain data, we will also perform the adaptation of domain specific translation systems to enhance the system performance in specific domains. Improvements from applying acquired data will be measured against baseline results from MT systems and validated in practical applications. As a summary, the most important results of ACCURAT will be

- Criteria and metrics of comparability
- Tools for building comparable corpora
- Tools for multi-level alignment and information ex-



*Proceedings of the 3rd Workshop on Building and Using Comparable Corpora, LREC 2010*, pages 35–41
Malta, 22 May 2010

traction from comparable corpora

- Multilingual comparable corpora for under-resourced languages and narrow domains
- Improved baseline translation systems for under-resourced European language pairs using data extracted from comparable corpora
- Report on requirements, implementation and evaluation of usability in applications for specialists in narrow domain and specific languages

## 2. State of the Art

Machine translation, in particular the statistical approach to it, has undergone significant improvements in recent years. However SMT research has been mainly focused on widely used languages, such as English, French, Arabic, Chinese, Spanish, and German. Languages with less native speakers such as Romanian are not as well developed due to the lack of linguistic resources. This results in a technical gap between the translation on widely spoken languages and on other languages.

Building statistical machine translation system requires a great amount of parallel corpora for model training. Good results can be easily achieved when the domain of the training corpus is closer to that of the test data. Rule-based machine translation can also profit from the data-driven technique: a MT system can have better translation quality, when bilingual lexical data has been extracted from parallel resources and imported into an RBMT system dictionary (Eisele et al., 2008). Nowadays parallel corpora are still limited in quantity, genre and language coverage.

There have been many investigations to exploit comparable corpora. Whereas early work on alignment such as the sentence aligners described in (Gale and Church, 1993) and (Brown et al., 1991) assumed parallel corpora, models that incorporated lexical information to increase performance on noisy data were investigated early after, e.g. in (Chen, 1993; Fung and McKeown, 1994; Jones and Somers, 1995; Fung, 1995; Rapp, 1995). In (Zhao and Vogel, 2002), sentence length models and lexicon-based models are combined under a maximum likelihood criterion. Specific models are proposed to handle insertions and deletions that are frequent in bilingual data collected from the web. Using the mined data, word-to-word alignment accuracy machine translation modeling is improved as shown in the experiments. In (Utiyama and Isahara, 2003), language information retrieval and dynamic programming methods are applied to align the Japanese and English articles and sentences. In (Munteanu and Marcu, 2005) the parallel sentences are discovered using a maximum entropy classifier, where similar sentence pairs are analyzed using a signal processing-inspired approach. The extracted data have been shown to improve the performance of a state-of-the-art translation system. In (Shi et al., 2006), a new web mining scheme for parallel data acquisition is presented based on the document object model. A comparison of different alignment methods and more approaches considering non-monotone sentence alignments are described in (Khadivi, 2008) and (Xu et al., 2006).

One very promising approach for the iterative bootstrapping of improved translation models from comparable corpora is given in (Rauf and Schwenk, 2009) for the case of English and French. We will apply these methods for all the 18 language pairs investigated in the project and report on the question how well the methods generalize to language pairs from different families.

Also, a number of techniques have been developed for automatically assembling domain specific corpora from the web, e.g. BootCaT in (Baroni and Bernardini, 2004), Corpógrafo in (Maia and Matos, 2008). However, state-of-the-art fully automatic extraction results in noisy output and requires human processing. To select similar documents from comparable corpora, CLIR techniques are applied in selection process for widely used languages, e.g. (Quirk et al., 2007) and (Munteanu and Marcu, 2005).

Furthermore, several phrasal alignment methods have been researched for parallel corpora: IBM Models 1-6 (Brown et al., 1993); applying lexico-syntactic categories for word tagging and the identification of semantically equivalent expressions (Aswani and Gaizauskas, 2005); Phrase-based joint probability model (Marcu and Wong, 2002); factored phrase-based alignments (Koehn and Hoang, 2007).

There are only a few parallel corpora publicly available for the languages we work on. The JRC-Acquis is a huge collection of European Union legislative documents translated into more than twenty official European languages (Steinberger et al., 2006) including under-resourced languages such as Latvian, Lithuanian, Estonian, Greek, Croatian and Romanian. The European Parliament Proceedings Parallel Corpus (Europarl corpus) was extracted from the proceedings of the European Parliament (1996-today) and has included versions in 11 European languages: French, Italian, Spanish, Portuguese, English, Dutch, German, Danish, Swedish, Greek and Finnish (Koehn, 2005). The Europarl corpus was aligned at the sentence level using a tool based on the Church and Gale algorithm (Gale and Church, 1991). Other available multilingual parallel corpus are developed in the framework of projects of Multilingual Corpora for Cooperation (MLCC), the Integrated European language data Repository Area (INTERA2) eContent, SEEERAnet and so on. Very interesting corpora are contained in the OPUS collection described in (Tiedemann, 2009).

## 3. Domain Adaptation

Here we will focus on methods of sentence, paragraph and phrasal alignment and domain adaptation. The discussion on comparability metrics and building comparable corpora is described in (Skadina et al., 2010).

To select similar documents from a comparable or parallel corpus and to find multilingual comparable corpora for certain domains, the cross language information retrieval (CLIR) techniques will be proposed. Bootstrapped bilingual lexical resources will be explored for document selection.

Given a comparable corpus consisting of documents in two languages, L1 and L2, the first step is to find similar documents in L1 and L2. Typical approaches involve treating a document in the L1 collection as a query and then using CLIR techniques to retrieve the top n documents from the L2 collection as described in (Munteanu et al., 2004) and (Quirk et al., 2007). This approach requires some sort of bilingual dictionary in query translation.

After similar documents are selected, similar text fragments need to be identified. These fragments may be sentences or possibly only phrases. Recent research results have shown that in most cases methods designed for parallel texts perform poorly for comparable corpora. For example, most standard sentence aligners exploit the monotonic increase of the sentence positions in a parallel corpus, which is not observed in comparable corpora. ACCURAT will investigate how successful the sentence aligner developed at the Romanian Academy (Tufiş et al., 2006) is in aligning similar sentences in comparable corpora. This sentence aligner, based on SVM technology, builds feature structures characterizing a pair of sentences considered for alignment, including number of translation equivalents, ratio between their lengths, number of non-lexical tokens, such as dates, numbers, abbreviations, etc., and word frequency correlations. These feature structures are afterwards classified to describe how well sentence alignments corresponds to experimentally determined thresholds. This aligner has been evaluated and has an excellent F measure score on parallel corpora, being able to align N-M sentences. It is much better than Vanilla aligner and slightly better than HunAlign. A state-of the-art sentence aligner is described in (Moore, 2002), but this aligner produces only 1-1 alignments loosing N-M alignments. As comparable corpora do not exhibit the monotonic increase of aligned sentence positions, we anticipate that many of the alignments will be of the type 0-M, N-0 and N-M sentences, thus this alignment ability is a must. The SVM approach to sentence alignment has the advantage that it is fully trainable. Another promising method to identify similar sentence pairs within comparable corpora, proposed by (Munteanu et al., 2004), will be also investigated. To select candidate sentences for alignment they propose a word-overlap filter together with a constraint on the ratio of lengths of the two sentences. Given two sentences that meet these criteria, the final determination of whether they are or are not assumed to be parallel sentences is made by a maximum entropy classifier trained over a small parallel corpus, using such features as percentage of words with translations, length of sentences, longest connected and unconnected substrings. We will expand this method to paragraphs/sentences which are only to some extent translations of each other, thus adapting the proposed method to comparable corpora. A challenging research avenue for detecting meaning-equivalent sentence pairs within comparable corpora is using cross-lingual Q&A techniques. The main idea is to exploit dependency linking and the concepts of superlinks and chained links (Irimia, 2009) for determining the most relevant search criteria. The keywords extracted from the dependency linking of a source sentence/paragraph will be translated into a target language and available search engines will look for the most relevant candidate paragraphs/sentences. The possible pairs of translation equivalent textual units will be scored by a reified sentence aligner and will be accepted or rejected based on previously determined thresholds.

## 4. Sentence, Paragraph and Phrasal Alignment

We will research on multi-level alignment and information extraction methods from comparable corpora, specially building parallel sentence aligned corpora for SMT. We expect to develop pre-processing tools, a search module for detecting similar sentences/paragraphs in given collections of documents, the proper alignment tools for paragraph, sentence and phrase as well as a user-friendly alignment editor allowing the users to view and correct the wrong alignments. By promoting web service architecture, it will integrate the existing tools , especially for the required preprocessing steps such as language identification, tokenization, tagging, lemmatization, chunking etc., and it will allow for easy integrating of new tools and new languages.

Language independent methods in the spirit of those proposed in (Munteanu and Marcu, 2005) will be further investigated and elaborated for English-Latvian, English-Lithuanian, English-Estonian, English-Greek, English-Croatian, English-Romanian, English-Slovenian, German-Romanian, Lithuanian-Romanian, Romanian-Greek and Latvian-Lithuanian, allowing sentence/paragraph alignment of comparable corpora. Such methods are knowledge-poor but there is no reason for not using current language technology to embed easy to access knowledge sources. Since all partners have tools for basic preprocessing of their languages, such as tokenizers, POS-taggers, lemmatizers, the linguistic information revealed by these tools will be relied on heavily in order to decrease the danger of data sparseness and to increase the reliability of the statistical judgments.

When sentence/paragraph level alignment is established, the next step is to compute phrasal alignment, which is a central issue to exploit comparable corpora in MT applications. ACCURAT will start with the evaluation of existing methods for phrasal alignment, such as IBM Models1-6 as described in (Brown et al., 1993) and (Och and Ney, 2003),

lexico-syntactic categories for word tagging and the identification of semantically equivalent expressions (Aswani and Gaizauskas, 2005) and reified word alignment in (Tufiş et al., 2006) and (Tufiş et al., 2008) as well as their combinations. Since in many cases under-resourced languages lack linguistic resources, we will research on possibilities to extract phrasal alignments directly from similar document pairs in comparable corpora, without the use of dictionaries or pre-processing of the training data. Phrase-based joint probability model (Marcu and Wong, 2002) will be extended with the aim to overcome the sparseness of linguistic resources for under-resourced languages. We will use log-likelihood ratio statistics to assess the reliability of alignment (Kumano et al., 2007) which allows phrasal alignments to be produced just for parallel parts of the comparable corpora. To prevent alignments being produced between unrelated phrases while searching for optimal alignments, log-likelihood ratio (LLR) statistics will be applied.

Another novel way information extraction techniques can assist in aligning comparable corpora is through the identification of cross-language mappings between relation-expressing contexts. (Hasegawa et al., 2004) propose a technique for unsupervised relation discovery in texts, whereby contexts surrounding pairs of NEs of given types are extracted and then clustered, the clusters correspond to particular relations. This technique achieves impressive results and could be used to align relation expressing contexts as follows: First, relation clusters could be established monolingually given NERC tools in each language; These clusters could then be aligned cross-lingually using aligned sentence pairs containing NE pairs present found in the clusters, the aligned sentences coming either from the small amount of parallel data or from high confidence alignments in the comparable corpus; Once relation clusters where aligned cross-lingually, then presence of a pair of NEs from an aligned relation cluster in an L1 and L2 sentence pair would constitute evidence that the sentences should be aligned. ACCURAT will also investigate potential of unsupervised discovery of relations in text using NERC tools for monolingual clustering and perform cross-lingual alignment to improve fragment alignment in comparable corpora. Orthographic and phonetic-based approaches will be explored to develop adaptive HMM and/or CRF-based techniques e.g. (Zhou et al., 2008) trained on name pairs gathered initially from parallel training data and then bootstrapped using lexicons derived in the project. New advances in adaptive, semi-supervised NE recognition e.g. (Nadeau, 2007) will be explored and applied for languages other than English. Existing named entity recognition and classification systems for Croatian, English, German, Greek and Romanian will be deployed. First NERC systems for the Baltic languages will be developed, too.

Q&A techniques will be further researched and elaborated to find most relevant candidate paragraphs/sentences in comparable corpora. Cross-lingual Q&A techniques are highly relevant for this task. Queries formulated in one language and translated in another language may be used for searching the comparable corpora to find the paragraphs or sentences which are most likely to contain similar information.

## 5. Comparable Corpora for Machine Translation

The impact of comparable corpora on MT quality will be measured for seventeen language pairs, and detailed studies involving human evaluation will be carried out for six language pairs. Existing baseline SMT systems based on the Moses decoder will be coupled with data extracted from comparable corpora. Comparative evaluation will be performed to measure improvements by applying data extracted from comparable corpora. Comparable corpora will be used to update the linguistic knowledge of RBMT systems by applying terminology and named entity extraction technology.

Comparable corpora in machine translation systems will be created with the goal to evaluate results of data extracted from the comparable corpora. MT systems will be created using existing SMT techniques (Moses decoder) and existing RBMT techniques (Linguatec RBMT engine). Innovation in MT techniques will be in (1) enabling the use of additional data extracted from comparable corpora and (2) adjusting MT systems to under-resourced languages or narrow domains. To evaluate the efficiency and usability of the approach proposed in ACCURAT for under-resourced areas of MT, we will integrate research results into SMT using existing SMT techniques. In Task 4.1 baseline SMT systems will be built using traditional SMT techniques. Translation models will be trained on parallel corpora e.g. Europarl Parallel Corpus and JRC-ACQUIS multilingual Parallel Corpus. Performance of baseline SMT systems will be evaluated using automatic metrics such as BLEU and NIST as well as human metrics including fluency and adequacy. After the baseline SMT systems are built they will be improved by the integration of additional data from the comparable corpora. Data from comparable corpora will be integrated into both the translation model and the language model. Finally, SMT systems will be adjusted for a narrow domain using factored and reified models and will include domain specific knowledge such as terminology, named entities, domain specific language models, etc. Several approaches for the integration of additional data from comparable corpora into SMT will be investigated and evaluated. One option for the integration is to add extracted phrases to the training data and to retrain SMT. Another option is to use factored translation models (Koehn and Hoang, 2007) and to add data from comparable corpora as an additional phrase table.

In the ACCURAT project comparable corpora will be used instead of parallel corpora to extract bilingual lexi-

cal data for feeding rule-based machine translation systems. Comparable corpora will be used to update the linguistic knowledge of RBMT systems by applying terminology and named entity extraction technology. This is a step towards automating the current work flow in MT lexicon for RBMT production. Once these data are imported into a RBMT system, the next problem to solve is when to activate this acquired information in a given text. Automatic topic extraction would help in determining the narrow domain to which a given text belongs (Thurmair, 2006). However, many terms stay ambiguous in the selected domain, as they often have a general meaning which is also used in this narrow domain, and additional data-driven criteria will be used to further select the right translations in the narrow domain. ACCURAT will make use of techniques developed for the enrichment of a RBMT system with new lexical entries acquired automatically from parallel corpora in a specific domain in the framework of an ongoing collaboration with the European Patent Office on hybrid MT. The solution in this case was to construct a hierarchy of lexicons of increasing specificity and to traverse these lexicons from specific to more general for each ambiguous term that arises. These techniques will be generalized in case we do not have a fine-grained mark-up of the document topics but need to infer the topic via automatic classification, and in cases where the alignments are less clean because they are built from comparable instead of parallel data.

## 6. Conclusions

Lack of sufficient linguistic resources for many languages and domains is one of the major obstacle in further advancement of automated translation currently. The main goal of the ACCURAT research is to find, analyze and evaluate novel methods how comparable corpora can compensate for this shortage of linguistic resources to improve MT quality significantly for under-resourced languages and narrow domains.

The ACCURAT project will provide researchers and developers with reimplemented baseline methods such as that in (Munteanu and Marcu, 2005) along with novel methodologies to exploit comparable corpora for machine translation. We will determine criteria to measure the comparability of texts in comparable corpora. Methods for automatic acquisition of a comparable corpus from the Web will be analyzed and evaluated. Advanced techniques will be elaborated to extract lexical, terminological and other linguistic data from comparable corpora to provide training and customization data for MT. Improvements from applying acquired data will be measured against baseline results from MT systems and validated in practical applications. ACCURAT will provide novel approaches to achieve high quality MT translation for a number of under-resourced EU languages and to adapt existing MT technologies to narrow domains, significantly increasing the language and domain coverage of MT.

## 7. Acknowledgements

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no 248347. We thank our colleagues from the ACCURAT consortium for the inspiration for many of the proposed methods and for the permission to re-use parts of the project's work plan. We apologize for some overlap with the material presented in (Skadina et al., 2010).

## 8. References

N. Aswani and R. Gaizauskas. 2005. Aligning words in english-hindi parallel corpora. In *Proceedings of the ACL 2005 Workshop on Building and Using Parallel Texts: Data-driven Machine Translation and Beyond*, pages 115–118.

M. Baroni and S. Bernardini. 2004. BootCaT: Bootstrapping corpora and terms from the web. In *Proceedings of Language Resources and Evaluation Conference LREC*.

P. F. Brown, J. C. Lai, and R. L. Mercer. 1991. Aligning sentences in parallel corpora. In *Proc. of the 29th Annual Meeting of the Association for Computational Linguistics*, pages 177–184, Berkeley, California, June.

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

S. F. Chen. 1993. Aligning sentences in bilingual corpora using lexical information. In *Proc. of the 31th Annual Meeting of the Association for Computational Linguistics*, pages 9–16, Columbus, Ohio, June.

A. Eisele, C. Federmann, H. Uszkoreit, H. Saint-Amand, M. Kay, M. Jellinghaus, S. Hunsicker, T. Herrmann, and Y. Chen. 2008. Hybrid machine translation architectures within and beyond the euromatrix project. In *Proceedings of EAMT*.

P. Fung and K. McKeown. 1994. Aligning noisy parallel corpora across language groups: Word pair feature matching by dynamic warping. In *First Conf. of the Association for Machine Translation in the Americas (AMTA 94)*, pages 81–88, Columbia, MD, October.

P. Fung. 1995. A pattern matching method for finding noun and proper noun translations from noisy parallel corpora. pages 236–243.

W. Gale and K. Church. 1991. A program for aligning sentences in bilingual corpora. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*.

W. A. Gale and K. W. Church. 1993. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1):75–90.

T. Hasegawa, S. Sekine, and R. Grishman. 2004. Discovering relations among named entities from large corpora. In *Proceedings of the 42nd Meeting of the Association*

*for Computational Linguistics (ACL'04), Main Volume*, pages 415–422, Barcelona, Spain, July.

E. Irimia. 2009. *Methods for Analogy-based Machine Translation. Applications for Romanian and English*. Ph.D. thesis, March.

D. B. Jones and H. L. Somers. 1995. Automatically determining bilingual vocabulary from noisy bilingual corpora using variable bag estimation. In *Recent Advances in Natural Language Processing*, pages 81–86, September.

S. Khadivi. 2008. *Statistical Computer-Assisted Translation*. Ph.D. thesis, RWTH-Aachen University, Aachen, Germany, July.

P. Koehn and H. Hoang. 2007. Factored translation models. In *Proceedings of EMNLP*.

P. Koehn. 2005. Europarl: a parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X*.

T. Kumano, H. Tanaka, and T. Tokunaga. 2007. Extracting phrasal alignments from comparable corpora by using joint probability smt model. In *Proceedings of TMI*.

B. Maia and S. Matos. 2008. Corpógrafo v.4 – tools for researchers and teachers using comparable corpora. In *Proceedings of the Workshop on Comparable Corpora, LREC*.

D. Marcu and W. Wong. 2002. A phrase-based, joint probability model for statistical machine translation. In *Proc. Conf. on Empirical Methods for Natural Language Processing*, pages 133–139, Philadelphia, PA, July.

R. C. Moore. 2002. Fast and accurate sentence alignment of bilingual corpora. In *Proc. of the 5th Conf. of the Association for Machine Translation in the Americas*, pages 135–244, Tiburon, California, October.

D. S. Munteanu and D. Marcu. 2005. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31(4):477–504.

D. Munteanu, A. Fraser, and D. Marcu. 2004. Improved machine translation performance via parallel sentence extraction from comparable corpora. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, HLT / NAACL*.

D. Nadeau. 2007. *Semi-Supervised Named Entity Recognition: Learning to Recognize 100 Entity Types with Little Supervision*. Ph.D. thesis, DUniversity of Ottawa, Ottawa.

F. J. Och and H. Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, March.

C. Quirk, R. Udupa, and A. Menezes. 2007. Generative models of noisy translations with applications to parallel fragment extraction. In *Proceedings of MT Summit XI, European Association for Machine Translation*.

R. Rapp. 1995. Identifying word translations in non-parallel texts. In *Proc. of the 33rd Annual Conf. of the Association for Computational Linguistics*, pages 321–322.

S. A. Rauf and H. Schwenk. 2009. On the use of comparable corpora to improve SMT performance. In *EACL*, pages 16–23, April.

L. Shi, C. Niu, M. Zhou, and J. Gao. 2006. A dom tree alignment model for mining parallel data from the web. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 489–496, Morristown, NJ, USA. Association for Computational Linguistics.

I. Skadina, A. Vasiljevs, R. Skadins, R. Gaizauskas, D. Tufis, and T. Gornostay. 2010. Analysis and evaluation of comparable corpora for under resourced areas of machine translation. In *Proceedings of the International Conference on Language Resources and Evaluation: Workshop on Building and Using Comparable Corpora (This volume)*, May.

R. Steinberger, B. Pouliquen, A. Widiger, C. Ignat, T. Erjavec, D. Tufiş, and D. Varga. 2006. The jrcacquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*.

G. Thurmair. 2006. Using corpus information to improve mt quality. In *Proceedings of the Workshop LR4Trans-III, LREC*.

J. Tiedemann. 2009. News from OPUS - a collection of multilingual parallel corpora with tools and interfaces. In N. Nicolov, K. Bontcheva, G. Angelova, and R. Mitkov, editors, *Recent Advances in Natural Language Processing (vol V)*, pages 237–248, Amsterdam/Philadelphia. John Benjamins.

D. Tufiş, R. Ion, A. Ceauşu, and D. Ştefănescu. 2006. Improved lexical alignment by combining multiple reified alignments. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL2006)*, pages 153–160, April.

D. Tufiş, K. Koeva, E. Erjavec, M. Gavrilidou, and C. Krstev. 2008. Building language resources and translation models for machine translation focused on south slavic and balkan languages. in marko tadić mila dimitrova-vulchanova and svetla koeva (eds.). In *Proceedings of the Sixth International Conference Formal Approaches to South Slavic and Balkan Languages (FASSBL 2008)*, pages 145–152, September.

M. Utiyama and H. Isahara. 2003. Reliable measures for aligning japanese-english news articles and sentences. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 72–79, Morristown, NJ, USA. Association for Computational Linguistics.

J. Xu, R. Zens, and H. Ney. 2006. Partitioning parallel

documents using binary segmentation. In *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL): Proceedings of the Workshop on Statistical Machine Translation*, pages 78–85.

B. Zhao and S. Vogel. 2002. Adaptive parallel sentences mining from web bilingual news collection. In *Proceedings of the 2002 IEEE International Conference on Data Mining*, page 745, Washington, DC, USA. IEEE Computer Society.

Y. Zhou, F. Huang, and H. Chen. 2008. Combining probability models and web mining models: a framework for proper name transliteration. *Information Technology and Management*, 9(2):91–103.

# Building a Large English-Chinese Parallel Corpus from Comparable Patents and its Experimental Application to SMT

**Bin LU**
Language Information Sciences Research Centre & Department of Chinese, Language and Linguistics, City University of Hong Kong
lubin2010@gmail.com

**Tao JIANG**
ChiLin Star Corp., Southern Software Park, Zhuhai, China & Northeastern University, Shenyang, China
jiangtaoster@gmail.com

**Kapo CHOW**
Language Information Sciences Research Centre, City University of Hong Kong
kapo.chow@gmail.com

**Benjamin K. TSOU**
Language Information Sciences Research Centre & Department of Chinese, Language and Linguistics,, City University of Hong Kong
rlbtsou@gmail.com

**Abstract**

The paper provides an account on the augmentation of a Chinese-English patent parallel corpus consisting of about 160K sentence pairs, which has been enlarged by about 45 times to more than 7 million sentence pairs mostly by the means of "harvesting" comparable patents from the Web. First, based on a large corpus of English-Chinese comparable patents, more than 22 million bilingual sentence pair candidates have been mined, of which we extract more than 7 million high-quality parallel sentences, which to our best knowledge is the largest parallel sentence corpus in the patent domain. Based on 1 million parallel sentences extracted from the *abstract* and *claims* sections, some interesting preliminary SMT results are also reported here. Last by not least, the method and approach proposed here should be applicable to other languages, which shows a novel way on how to reduce the data acquisition bottleneck in multilingual language processing.

## 1. Introduction

Parallel corpora are invaluable resources for NLP applications, including machine translation, multilingual lexicography, and cross-lingual information retrieval. Many parallel corpora have been available, such as the Canadian Hansards (Gale and Church, 1991), the Arabic-English and English-Chinese parallel corpora used in the NIST Open MT Evaluation[1] and Europarl corpus (Koehn, 2005). However, large parallel corpora are still too little.

To overcome this lack of parallel corpora, comparable corpora are also used to mine parallel sentences. For instance, Zhao and Vogel (2002) investigated the mining of parallel sentences for Web bilingual news collections which may contain much noise. Resnik and Smith (2003) introduced the STRAND system for mining parallel text on the web for low-density language pairs. Munteanu and Marcu (2005) presented a method for discovering parallel sentences in large Chinese, Arabic, and English comparable, non-parallel corpora based on a maximum entropy classifier. Wu and Fung (2005) exploited Inversion Transduction Grammar to retrieve truly parallel sentence translations from large collections of highly non-parallel documents.

However, less work has been done in the patent domain, and only the following two are found. The Japanese-English patent parallel corpus (Utiyama and Isahara, 2007) contains more than 2 million parallel sentences, and was provided for the NTCIR-7 patent machine translation task (Fujii et al., 2008). The English-Chinese patent corpus (Lu et al., 2009) contains about 160K parallel sentences which were extracted from more than 6,000 English-Chinese comparable/noisy parallel patents.

In this paper, we enlarge the Chinese-English parallel corpus (Lu et al., 2009) by over 40 times to more than 7 million sentence pairs by mostly harvesting a large corpus of English-Chinese comparable patents from the Web. Compared with the one in Lu et al. (2009), this corpus is not only much larger, but also may have different characteristics because these comparable patents were first filed with English as the original language, and then translated into Chinese and filed in China. On the other hand, the patents in Lu et al. (2009) were filed in the opposite direction (i.e. first Chinese, then English).

With the large number of comparable patents harvested from the Web, we mine parallel sentences based on two publicly available sentence aligners and simple heuristic rules. Currently, more than 22 million bilingual sentence pair candidates are found, of which we extract more than 7 million high-quality parallel sentences, which is the largest parallel sentence corpus in the patent domain to our best knowledge. Based on 1 million parallel sentences extracted from the *abstract* and *claims* sections, a small part of the whole parallel corpus, some preliminary SMT experiments are also reported here. Some sampled parallel sentences are available at http://livac.org/smt/parpat.html. Since patents cover many technical domains (e.g. chemistry, vehicle, electronics, biomedicine, etc.), the large parallel corpus could be a valuable resource for many cross-lingual information access applications not only in the patent domain but also in the related technical domains mentioned above. A rough estimation on the quantity of bilingual and multilingual patents including Chinese, Japanese, Korean, German and English is made. It shows considerable potential for easing the data acquisition bottleneck for these languages in multilingual

[1] http://www.itl.nist.gov/iad/mig/tests/mt/



*Proceedings of the 3rd Workshop on Building and Using Comparable Corpora, LREC 2010*, pages 42–49
Malta, 22 May 2010

language processing.

In the next section we introduce related work, followed by the background in Section 3. Then the process of mining comparable English-Chinese patents from the Web is described in Section 4. The method of extracting parallel sentences from comparable patents and the SMT experiment are presented in Section 5, followed by discussion in Section 6, and we give conclusion and future work in Section 7.

## 2. Related work

Parallel sentences can be extracted from parallel corpora of documents or from comparable corpora. Since parallel corpora are bilingual text collections consisting of the same content in two or more different languages, it would be easier to find parallel sentences, and different approaches have been proposed: a) the sentence length in bilingual sentences (Brown et al. 1991; Gale and Church, 1991); b) lexical information in bilingual dictionaries (Ma, 2006); c) statistical translation model (Chen, 1993), or the composite of more than one approach (Simard and Plamondon, 1998; Moore, 2002). Comparable corpora raise further challenges for finding parallel sentences since the bilingual contents are not strictly parallel. Related work include Resnik and Smith (2003), Munteanu and Marcu (2005), Wu and Fung (2005), Zhao and Vogel (2002), etc.

For bilingual patent related work, Utiyama and Isahara (2007) used the "*Detailed Description of the Preferred Embodiments*" and "*Background of the Invention*" parts in the *description* section of Japanese-English comparable patents to find parallel sentences because they found these two parts have more literal translations than others. Lu et al. (2009) derives high-quality parallel sentences from English-Chinese comparable patents by aligning sentences and filtering sentence alignments with the combination of different quality measures, followed by the work in (Lu & Tsou, 2009).

The differences between this work with these two above lie in: 1) our comparable patents are mostly harvested from the Web and the parallel sentences mined are much larger compared to 2 million in the former and 160 K in the latter; 2) their comparable patents were both filed in USPTO in English by translating from the original language (namely, Japanese and Chinese) and identified by the priority information in the US patents. However, our comparable patents were first filed in English as a PCT patent, and later translated into Chinese. The different translation process may show different characteristics which will be explored in future.

For SMT, tremendous strides have been made in two decades. Brown et al. (1990; 1993) proposed the groundbreaking IBM approach, and the IBM models are word-based models. Later comes the SMT models called phrase-based models (Och and Ney, 2004; Koehn, 2004) in which translation unit may be any contiguous sequence of words. Phrase-based translation is implemented in the open-source Moses (Koehn et al., 2007), which is widely used in the SMT research community. We also use Moses for the SMT experiments in this paper. Currently, more researchers are taking advantages of syntax-based models (Chiang et al., 2005; Chiang, 2007), in which researchers attempt to incorporate syntax into phrase-based models.

For the evaluation of machine translation, NIST has been organizing MT open evaluations for several years, and the performance of the participants has been improved rapidly. The NTCIR-7 patent machine translation task (Fujii et al., 2008) has tested SMT performance on only the Japanese-English patent translation. Jiang et al. (2010) use Part-of-Speech model for the N-best list Reranking within the phrase-based SMT based on some parallel sentences extracted in this paper.

## 3. Background

A patent is a legal document representing "*an official document granting the exclusive right to make, use, and sell an invention for a limited period*" (Collins English Dictionary [2] ). Patents are important indicators of innovation. As Sun (2003) stated "a*s the economy is globalized, patenting increasingly becomes an international activity*". More firms, especially the multinational ones, are investing more and more money on intellectual property (especially patents) to protect their own technologies, and filing patents in foreign countries. There have been many legal cases involving the claims of patent infringement, such as Nokia vs Apple, Cisco vs. Huawei, Intel vs AMD, and the DVD manufacturers in China vs. the dvd6c licensing group. The companies may be interested in monitoring and analyzing the patents filed in different languages, such as English, Chinese, Japanese, Germany, etc. The traditional practice for monitoring patents filed in foreign languages is usually to involve translation companies to manually translate patents into a relevant language, which is slow, time-consuming, high-cost, and often quality-inconsistent.

Meanwhile, patent applications are increasing very quickly, especially those filed in China (Sun, 2003). The patent application numbers filed in the top leading patent offices including Japan, USA, China and Germany from 1996 to 2008 are shown in Figure 1, from which we can observe that in about 12 years, China's patent applications have increased by 10 times while USA only doubles its patent applications. The increasing trend of patent applications also impose more workload for the manual translation which demands more advanced machine translation engines and more parallel data to help us handle this problem.

Each patent application consists of different sections, namely, *bibliographical data (including title*, *abstract)*, *drawings, claims*, *description*, etc. Since we focus on the text in the patent applications, only *title, abstract, claims*

[2] Retrieved March 18, 2010, from http://www.collinslanguage.com/

*and description* are used in the experiments discussed below. From the legal perspective, the *claims* section is the most important part in one patent application, because it defines the coverage that the applicant wants to claim. The *description* section gives the technical details of the patent involved, and the descriptions of some patents have further subdivisions, such as *Field of the Invention*, *Background of the Invention*, *Objects of the Invention*, *Summary of the Invention*, etc.

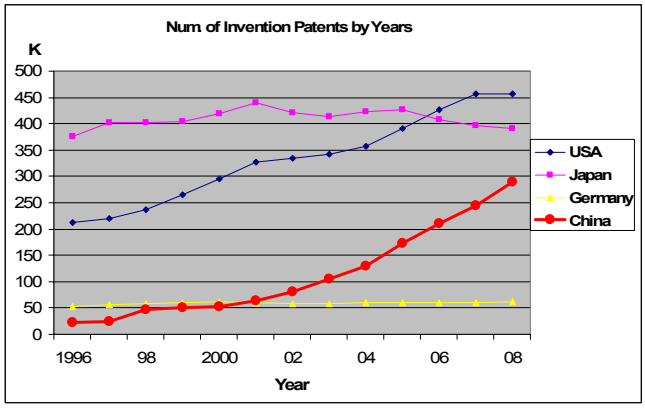



Figure 1: Patent applications by the top leading patent offices
Source[3]: WIPO: Patent Applications by Office

## 4. Mining comparable patents from the Web

The patents used in Lu et al. (2009) were first filed in China with Chinese as the original language, and we are also interested in patents which were first filed in English, and later filed in Chinese in China.

The intuition here is that from Figure 1 we can see that the number of patents filed in China was quite small in the 1990s compared to that in USA or Japan, and hence the possibility is lower for patents to be first filed in Chinese and then to be filed in English later. The opposite direction is quite different since western companies have accumulated a large amount of patents filed in other languages, and they may file Chinese patents to protect their inventions within China. Therefore, there may have many Chinese patents translated from English. The large amount of mined comparable patents which were first filed in English and later filed in Chinese prove our intuition.

### 3.1 Mining Chinese patents with English as original language

The official patent office in China is the State Intellectual Property Office (SIPO) of the People's Republic of China. SIPO was established in 1980 and began to accept patent applications since 1985. All Chinese patents are filed through SIPO. About 20 years after its creation, SIPO is regarded "*one of the more vibrant patent offices of the developing world, where an even-increasing number of domestic and non-resident applications are processed each year*" (Landry, 2008).

On the SIPO website[4], Chinese patents can be searched by many fields, such as *application number, publication number, title, International Patent Classification (IPC) code, inventor*, etc., including those patent applications which were originally filed in English with PCT publication numbers.

There were about 200 K Chinese patents both filed in China and previously filed as PCT applications in English up to early 2009. Most of the patents are invention patents. For these Chinese patents, the *bibliographical data, title, abstract* and *the major claim* were first crawled from the Web, and then *other claims* and *description* were also added. Since some contents are in the image format, the images were OCRed and manually verified. Inevitably there are errors in the data, but the quality can be generally acceptable.

### 3.2 Mining the corresponding English patents

All the PCT patent applications are filed through the World Intellectually Property Organization (WIPO). With the Chinese patents mentioned above, the corresponding English patents may be searched from the website of WIPO[5] to obtain relevant sections of the English PCT applications, including *bibliographical data, title, abstract*, *claims* and *description*. The mined English patents were automatically split into individual sections according to the respective tags inside patents.

However, not all but only about 40% out of the large number of Chinese patents had found their corresponding English ones. Some contents of the English patents were OCRed by WIPO, and hence there may be some errors in the English data.

### 3.3 Comparable patents mined

Here we give the percentage distribution of the Chinese patents in terms of their primary IPC codes. The IPC consists of 8 sections, ranging from A to H. From the category distribution in Table 1, we can see that *1) H: Electricity* and *C: Chemistry & Metallurgy* are the top two categories in terms of patent number, *2) D: Textiles & Paper* and *E: Fixed Construction* are the two categories with the smallest numbers of patents.

| | A | B | **C** | D | E | F | G | **H** | Total |
|---|---|---|---|---|---|---|---|---|---|
| **Percent (%)** | 16.6 | 11.9 | **21.7** | 1.7 | 1.7 | 4.7 | 18.0 | **23.7** | 100 |

Table 1. Percentage Distribution of Chinese Patents

Meanwhile, we obtain information on the area distribution of the patents, which shows that USA, Europe, Great Britain, Korea and Japan are the top leading areas in terms of the number of the patent priority. The distribution of publication years for the PCT patents filed in China are shown in Figure 2[6], which shows a big growth of the PCT patent applications filed in China in the 21st century.

[3] Retrieved March 20, 2010, from http://www.wipo.int/ipstats/en/statistics/patents/csv/wipo_pat_appl_from_1883_list.csv

[4] http://www.sipo.gov.cn/

[5] http://www.wipo.int/

[6] We only show the numbers within the period of 1996 to 2007, and skip the numbers for other years.

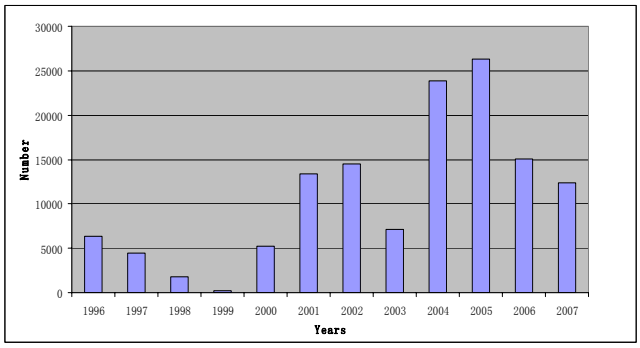



Figure 2. The distribution of publication years

The detailed statistics of each section for both Chinese and English patents are shown in Table 2.

| Sections | Chinese | | English | |
|---|---|---|---|---|
| | #Char | #Sent | #Word | #Sent |
| Title | 1.3M | 78K | 0.8M | 78K |
| Abstract | 16M | 274K | 10M | 392K |
| Claim | 183M | 3.4M | 108M | 3.7M |
| Description | 1,233M | 24.4M | 677M | 27.0M |
| **Total** | 1,435M | 28.1 M | 795M | 31.2M |
| **Avg/Patent** | 18K | 357 | 10K | 394 |

Table 2. Data Statistics of Comparable Patents

Here we consider the English-Chinese patent pairs as comparable (or noisy parallel) patents because they are not parallel in the strict sense but still closely related in terms of information conveyed. As noted in Lu et al. (2009), loose translations are very common in English-Chinese comparable patents, and the major explanations are:

1) The field of intellectual property is highly regulated in different countries, and the translation may be highly influenced by the stylistic differences in the individual countries;

2) The patent applicants may intentionally change some technical terms or the patent structure to broaden the patent coverage or to avoid potential conflict with other patents in the country when a new version is filed in another language and country;

3) Sometimes, the characteristics of different languages make it difficult to keep the original terminology/structure, and the translator may render it in a target language-specific way.

## 5. Mining parallel sentences & SMT experiments

The comparable patents are first segmented into sentences according to punctuations, and the Chinese sentences are segmented into words. The sentences in all sections of Chinese patents are aligned with those in the corresponding sections of the corresponding English patents to find parallel sentences.

### 4.1 Aligning sentences in comparable patents

To find high-quality parallel sentences in comparable patents, we combine two publicly available sentence aligners, namely Champollion (Ma, 2006) and MS aligner (Microsoft Bilingual Sentence Aligner) (Moore, 2002) with simple heuristic rules. Champollion is a sentence aligner based on bilingual dictionaries. We combine three bilingual dictionaries as the dictionary for Champollion: namely, LDC_CE_DIC2.0[7] constructed by LDC, bilingual terms in HowNet [8] and the bilingual lexicon in Champollion. The major steps for mining high-quality parallel sentences in comparable patents are as follows.

1) Champollion is used to preliminarily align the sentences in each section of the comparable patents to generate parallel sentence pair candidates. According to Lu et al. (2009), the generated candidates should have much noise and we will further explore filtering methods to remove misaligned sentences.

2) We remove sentence pairs using length filtering and ratio filtering. For length filtering, if a sentence pair has less than 100 words in the English sentence and less than 333 characters in the Chinese one, it is kept. Otherwise, it is removed. For ratio filtering, we discard the sentence pair candidates with Chinese- English length ratio outside the range of 0.8 to 1.8. The selection of the parameters here is set empirically based on the evaluation on a small sample of the large corpus.

3) MS aligner is utilized to further filter the parallel sentence candidates. MS aligner is a two-phase sentence aligner with high precision as its characteristics, and in the first pass it does alignment by using sentence length information (Gale and Church, 1991), and in the second pass it uses the sentence pairs aligned in the first pass to train an IBM Model-1 (Brown et al., 1993) and realign the sentences with the trained model.

Table 3 shows the statistics of the sentence numbers and the respective percentages of sentences kept with respect to all the sentence candidates in each step above.

| Steps | | 1. CH | 2.1 LF | 2.2 RF | 3. MS (final) |
|---|---|---|---|---|---|
| **Abstr.** | Num. | 251K | 243K | 176K | 83K |
| | Percent | 100% | 96.5% | 70% | 33% |
| **Claims** | Num. | 3.0M | 2.9M | 2.1M | 1.0M |
| | Percent | 100% | 96.5% | 72.1% | 33.4% |
| **Desc.** | Num. | 19.3M | 18.8M | 13.4M | 6.1M |
| | Percent | 100% | 97.2% | 69.4% | 31.3% |
| **Total**[9] | Num. | 22.6M | 21.9M | 15.8M | 7.2M |
| | Percent | 100% | 97.1% | 69.8% | 31.7% |
| **Average** | Num. | 286 | 277 | 200 | 91 |

Table 3. Statistics of Parallel Sentences during the Aligning Process

In the first row of Table 3, *1.CH* denotes the first step of using the Champollion to align sentences; *2.1 LF* denotes the length filter in the second step; *2.2 RF* refers to the ratio filter in the second step; *3. MS* refers to the third and

[7] http://projects.ldc.upenn.edu/Chinese/LDC_ch.htm

[8] http://www.keenage.com/html/e_index.html

[9] Here the total number does not include the number of titles. Here we did not use any method to filter the corresponding titles, and just treat them as parallel.

final step of using MS aligner to filter sentence pair candidates.

From Table 3, we can observe that 1) by using Champollion, we obtain about 22 million sentence pair candidates; 2) by filtering in step 2, the number of parallel sentences is reduced by 30%, to 16 million; 3) by using MS aligner, we final arrive at about 7 million parallel sentences.

The final parallel sentences are manually evaluated by randomly sampling 100 sentence pairs for each section of title, abstract, claims and description. The evaluation metric follows the one in Lu et al. (2009), which classifies each sentence pair into *Correct, Partially Correct* or *Wrong*[10]. The results of manual evaluation are shown in Table 4, from which we can see that the percentages of *correct* parallel sentences are quite high, and the wrong percentages are no higher than 5%. Therefore, we could conclude that the mined parallel sentences are high-quality with less than 5% wrong parallel sentences. Meanwhile, the abstract section shows the highest correct percentage, while the description section shows the lowest.

| | *Correct* | *Partially Correct* | *Wrong* |
|---|---|---|---|
| Abstr. | 97% | 2% | 1% |
| Claims | 92% | 3% | 5% |
| Desc. | 89% | 8% | 3% |

Table 4. Manual evaluation of the final corpus

One may notice that the average number of parallel sentences extracted from one comparable patent in this study is 91, while for the corpus in Lu et al. (2009), it is only about 26 (~160K/6100). Here we recomputed the average numbers of Chinese characters, English words, and Chinese and English sentences for each comparable patent in Lu et al. (2009), which are shown in Table 5.

| | **Chinese** | | **English** | |
|---|---|---|---|---|
| | **#Char** | **#Sent** | **#Word** | **#Sent** |
| **Avg/Patent** | 5.8K | 119 | 4.4K | 169 |

Table 5. Data Statistics of Comparable Patents in Lu et al.(2009)

Comparing Table 5 with Table 2, we can see that the comparable patents in Lu et al. (2009) are much smaller than those in this study in terms of numbers of Chinese characters, English words, and Chinese/English sentences. Therefore, the average number of parallel sentences extracted from the patents in this study is much bigger than that in Lu et al. (2009).

The possible explanation is that the patents in Lu et al. (2009) were first filed in China from 1996 to 2006 and later filed in USA from 1996 to 2008, and the applicants were still in their initial stage of learning how to write patent applications which may contain less content than those in this study involving patents filed by more experienced western companies.

[10] *Correct* means the English sentence is exactly the literal translation of the Chinese one, or the content overlap between them are above 80%; *partially correct* means the Chinese sentence and the English one are not the literal translation of each other, but the content of one sentence can cover more than 50% of the other; *wrong* means the contents of the Chinese sentence and the English one are not related, or more than 50% of the content of one sentence is missing in the other.

## 4.2 SMT experiments

As we have known, few SMT experiments on the English-Chinese patent translation have been reported, especially with a large scale of parallel sentences. We select 101,000 parallel sentences and divide them into three parts: 1 million sentence pairs for training, 500 sentence pairs for development and another 500 sentence pairs for testing. The statistics for the three parts are shown in Table 6.

| | **Language** | **#Sentence pairs** | **#Words** |
|---|---|---|---|
| Training | English | 1M | 33.4M |
| | Chinese | 1M | 32.1M |
| Development | English | 500 | 17.2K |
| | Chinese | 500 | 16.1K |
| Test | English | 500 | 17.2K |
| | Chinese | 500 | 16.1K |

Table 6. Data for SMT Experiments

An SMT system is setup using Moses (Koehn, 2007). We test translation in both directions (namely, Chinese to English and English to Chinese) with/without optimized parameters. The BLEU scores are as shown in Table 7. "*No MERT*" denotes the cases without optimizing parameters using minimal error-rate training (MERT) (Och, 2003) algorithm whereas "*MERT*" denotes the cases with parameter optimization of MERT on development data.

| | **Chinese->English** | | **English->Chinese** | |
|---|---|---|---|---|
| BLEU | No MERT | MERT | No MERT | MERT |
| | 0.273 | 0.274 | 0.207 | 0.240 |

Table 7. SMT experiment results

The BLEU scores here seem promising, which show that the parallel sentences extracted are of good quality for training the SMT engine. We could expect better results with more training data.

Moreover, we use the 160K parallel sentences in Lu et al. (2009) as the training data to build an SMT system, and the BLEU score for Chinese to English translation is 0.179 on the test data of 500 parallel sentences mentioned above with the MERT optimization on development data. The BLEU score of 0.274 in Table 7 based on 1 million parallel sentences shows a significant 53% relative improvement compared the BLEU score of 0.179, which demonstrates that with more training data we can get better SMT performance.

The BLEU scores for Chinese to English translation in Table 7 seem much better than those for the opposite direction. This is different from the results in NIST SMT evaluation, in which the highest BLEU scores for English to Chinese translation are usually better than those for Chinese to English translation. The possible reasons are: 1) the BLEU scores in this study are calculated without considering recasing or detokenization so we essentially ignore errors caused by them, while in NIST evaluation, recasing and detokenization are essential steps. 2) the evaluation of Chinese sentences is influenced by the boundary of Chinese words. Even when the whole sentence is correct, if the word boundaries are wrong, we

would get a low score. However, the English tokenization is much easier compared to Chinese word segmentation because there is no word boundary problem for English. 3) another relevant factor may be the translation direction of the test data, which is from English to English. Could the direction of human translation have an influence on the BLEU scores? Ozdowska & Way (2009) showed that "*data containing original French and English translated from French is optimal when building a system translating from French into English. Conversely, using data comprising exclusively French and English translated from several other languages is suboptimal regardless of the translation direction.*" Since no such observation seems have been found in Chinese-English translation, we raise the question here and are looking forward to further investigation.

Meanwhile, the MERT algorithm shows better performance on the English to Chinese translation but not on the reverse direction. One possible explanation is that MERT improves the performance with respect to the Chinese word boundary.

The server used for parallel sentence mining and SMT in this study has a 12G memory and 4 two-core 2.67GHz CPUs. Although the server is already much better than common PCs, it is still not powerful enough to do the computing-intensive SMT related tasks. Therefore, our SMT experiments only use a small part of the whole corpus, i.e. only 1 million out of more than 7 million sentence pairs.

## 6. Discussion

Here we briefly describe the efforts spent for this project. The Chinese and English websites from which the Chinese and English patents were downloaded were quite slow to access, and were occasionally down during access. Meanwhile, some patents are quite large. For example, the Chinese patent with the application number of CN200680029419.3 has 340 pages of description and 40 pages of claims, and its corresponding English patent has 396 pages of description and 46 pages of claims. These large patents would cost much time for the websites to respond and had be specifically handled. To avoid too much workload for the websites, the downloading speed had been limited. It took considerable efforts among different parties to obtain these comparable patents. By comparison, the efforts spent for parallel sentence mining and SMT experiments were much less.

According to recent investigation in 2010, the number of Chinese patent applications with English as the original language has rapidly increased, and we could expect more English-Chinese comparable patents to be filed quickly. This would allow further efforts to enlarge our corpus.

The method and approach proposed here to mine comparable patents should be also applicable to other language pairs, such as English and Japanese, English and Korean, etc. What is more, we could even build trilingual or multilingual parallel corpus by using the PCT patents and their multiple versions in different languages, such as Japanese (JP), Chinese (CH), Korean (KR), English (EN), German (DE), etc. We have searched via the website of WIPO to get an estimate on the quantity of PCT applications which were published in English and later filed in other countries in their corresponding languages, and found that the quantity of bilingual and multilingual patents for CH, KR, JP and EN seems quite considerable, which means that the multilingual patents for these languages could be harvested in remarkable quantities. For example, we have began to build a small trilingual patent corpus by leveraging the PCT patents, i.e. we search for comparable patents filed in simplified Chinese in China, filed in traditional Chinese in Taiwan, and filed in English as a PCT patent (Tsou and Lu, 2010). Although the language varieties found in mainland China and Taiwan are not two distinct languages, there are enormous differences in terms of technical terminology and even syntactic structure, this corpus is still quite useful to compare the two versions of the same PCT patent in China and Taiwan because there are linguistic convergences.

What is of special interest here is the very concept of *"parallel corpus"* in the context of translation. The commonly used BLEU and NIST scores in SMT evaluation just reduce the concept of parallelism to a rather technical mapping of language units. But it is well known that high-quality human translations often do not keep sentence units of the source language. Therefore, we may need more elaborate schemes to better evaluate the quality of machine translation, and translation studies (Munday, 2001) retain its importance.

## 7. Conclusion and future work

In this paper, we introduce our large parallel corpus which is extracted from a large corpus of English-Chinese comparable patents harvested from the Web. We first preliminarily mine parallel sentence pairs with Champollion, a publicly available sentence aligner, and then further filter the candidates with another sentence aligner, namely MS Aligner. Then, about 7 million high-quality parallel sentences out of more than 22 million bilingual sentence pair candidates are chosen as the final parallel corpus. As we know, this is the largest parallel sentence corpus in the patent domain. Based on the 1 million parallel sentences extracted from the *abstract* and *claims* section, some preliminary SMT results are also reported here.

Meanwhile, with our experimental sentence alignment efforts, only 7 million parallel sentences have been mined from 22 million sentence pair candidates. By exploring more complicated and possibly more accurate approaches such as Munteanu and Marcu (2005) or Lu et al. (2009), we could expect to find more parallel sentences from the comparable patents. More SMT experiments would be done as well since we currently only utilize 1 million parallel sentences in our SMT experiment due to limited time and computer resources.

Since different (sub-)sections (namely, title, abstract, claims, description, and subsections in the description part) in patents have their own writing styles which may influence the word choice and syntactic structure of the sentences, as well as patents cover many technical domains (such as chemistry, biomedicine, electronics, vehicle, etc.), experiments on cross-section and cross-IPC (International Patent Classification) machine translation could be enlightening for further understanding the characteristics of individual sections and technical domains. For example, *claims* have legal effect, and tend to use more relative clauses modifying head words.

Some sampled parallel sentences are available at http://livac.org/smt/parpat.html. We should be able to make some parts of our large parallel corpus available to the research community in the near future. Given the relative paucity of parallel patent data, this large parallel corpus shall be a helpful step towards MT research and other cross-lingual information access applications, in the above mentioned technical domains and especially in the patent domain. Last but not least, our method and approach should be applicable to other languages, which show a novel way on how to reduce the data acquisition bottleneck in multilingual language processing.

## 8. Acknowledgements

We acknowledge the contributions of Jacky Hui, Andy Chin and other colleagues of the Language Information Sciences Research Centre, City University of Hong Kong and of ChiLin Star Corp.

## 9. References

Brown, P.F., Cocke, J., Pietra, S.A.D., Pietra, V.J.D., Jelinek, F., Lafferty J.D., Mercer, R.L., & Roossin, P.S. (1990). A statistical approach to machine translation. *Computational Linguistics*, *16*(2), 79-85.

Brown, P.F., Lai, J.C., & Mercer, R.L. (1991). Aligning sentences in parallel corpora. In *Proceedings of the Annual Meeting of the Association of Computational Linguistics (ACL)*. pp.169-176.

Brown, P.F., Pietra, S.A.D., Pietra, V.J.D., & Mercer, R.L. (1993). Mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, *19*(2), 263-311.

Chen, S.F. (1993). Aligning sentences in bilingual corpora using lexical information. In *Proceedings of the Annual Meeting of the Association of Computational Linguistics (ACL)*. pp. 9-16.

CHIANG, D. (2005). A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the Association for Computational Linguistics (ACL)*. pp. 263-270.

Chiang, D. (2007). Hierarchical phrase-based translation. *Computational Linguistics,* 33(2), 201–228.

Fujii, A., Utiyama, M., Yamamoto, M., & Utsuro, T. (2008). Overview of the patent translation task at the NTCIR-7 workshop. In *Proceedings of the 7th NTCIR Workshop Meeting on Evaluation of Information Access Technologies (NTCIR)*. pp. 389-400. Tokyo, Japan.

Gale, W.A., & Church, K.W. (1991). A program for aligning sentences in bilingual corpora. In *Proceedings of the Annual Meeting of the Association of Computational Linguistics (ACL)*. pp.79-85.

Jiang, T., Tsou, B.K., & Lu, B. (2010). Part-of-speech model for N-best list reranking in experimental English-Chinese SMT. In *Proceedings of 1st International Workshop on Advances in Patent Information Retrieval*. Milton Keynes, UK.

Koehn, P. (2004). Pharaoh: A beam search decoder for phrase-based statistical machine translation models. In *Proceedings of the Conference of the Association for Machine Translation in the Americas (AMTA)*.

Koehn, Philipp. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MT Summit X*.

Koehn, P., Hoang H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., & Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the Annual Meeting of the Association of Computational Linguistics (ACL) Demo Session.* pp. 177-180.

Landry, P.F. (2008). How weak institutions can produce strong regimes: Patents, lawyers, and the improbable creation of an intellectual property regime in China (1985-2007). *Paper presented at Workshop on Rule of Law*, Yale University, March 28-29.

Lu, B., Tsou, B.K., Zhu,J., Jiang, T., & Kwong, O.Y. (2009). The construction of an English-Chinese patent parallel corpus. In *Proceedings of MT Summit XII 3rd Workshop on Patent Translation*. pp. Ottawa, Canada.

Lu, B., & Tsou, B.K. (2009). Towards bilingual term extraction in comparable patents. In *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation (PACLIC)*. pp. 755-762. Hong Kong. December, 2009.

Ma, X. (2006). Champollion: A robust parallel text sentence aligner. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*. Genova, Italy.

Moore, R.C. (2002). Fast and accurate sentence alignment of bilingual corpora. In *Proceedings of AMTA*. pp.135-144.

Munday, J. (2001). *Introducing translation studies: theories and applications*. Oxon, UK: Routledge.

Munteanu, D.S., & Marcu, D. (2005). Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, *31*(4), 477–504.

Och, F.J. (2003). Minimum error rate training in statistical machine translation. In *Proceedings of ACL*, pp. 160-167.

Och, F.J., & Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, *29*(1), 19-51.

Och, F.J., & Ney, H. (2004). The alignment template approach to machine translation. *Computational Linguistics*, *30*(4), 417-449.

Ozdowska, Sylwia and Way, Andy. (2009) Optimal bilingual data for French-English PB-SMT. In *Proceedings of 13th Annual Conference of the European Association for Machine Translation (EAMT 2009)*.

Resnik, P., & Smith, N.A. 2003. The Web as a parallel corpus. *Computational Linguistics*, 29(3):349–380.

Utiyama, M., & Isahara, H. (2007). A Japanese-English patent parallel corpus. In *Proceeding of MT Summit XI*. pp. 475–482.

Simard, M., & Plamondon, P. (1998). Bilingual sentence alignment: Balancing robustness and accuracy. *Machine Translation*, *13*(1), 59-80.

Sun, Y. (2003). Determinants of foreign patents in China. *World Patent Information,* 25, pp. 27-37.

Tsou, B.K., & LU, B. (2010). Automotive patents from Mainland China and Taiwan: A preliminary exploration of terminological differentiation and content convergence. *World Patent Information*. (to appear)

Wu, D., & Fung, P. (2005). Inversion Transduction Grammar Constraints for Mining Parallel Sentences from Quasi-Comparable Corpora. In *Proceedings of IJCNLP2005*.

Zhao, B., & Vogel, S. (2002). Adaptive Parallel Sentences Mining from Web Bilingual News Collection. In *Proceedings of Second IEEE International Conference on Data Mining (ICDM'02)*.

# Automatic Parallel Corpora and Bilingual Terminology extraction from Parallel WebSites

**José João Almeida**[1], **Alberto Simões**[2]

[1] Departamento de Informática, Universidade do Minho, Portugal
[2] Escola Superior de Estudos Industriais e de Gestão, Instituto Politécnico do Porto, Portugal
jj@di.uminho.pt, alberto.simoes@eu.ipp.pt

**Abstract**

In our days, the notion, the importance and the significance of parallel corpora is so big that needs no special introduction. Unfortunately, public available parallel corpora is somewhat limited in range. There are big corpora about politics or legislation, about medicine and other specific areas, but we miss corpora for other different areas. Currently there is a huge investment on using the Web as a corpus.
This article uncovers GWB, a tool that aims automatic construction of parallel corpora from the web. We defend that it is possible to build high quality terminological corpora in an automatic fashion, just by specifying a sensible Internet domain and using an appropriate set of seed keywords. GWB is a web-spider that works in conjunction with a set of other Open-Source tools, defining a pipeline that includes the documents retrieval from the web, alignment at sentence level and its quality analysis, bilingual dictionaries and terminology extraction and construction of off-line dictionaries.

## 1. Introduction

As it is already well known, parallel corpora is relevant for different natural language studies, as translation studies, machine translation and other important tasks.

One of the many uses for parallel corpora is the extraction of bilingual resources, like bilingual dictionaries or bilingual terminology. Unfortunately not all parallel corpora are suitable for terminology extraction. In fact, first developed parallel corpora were mainly devoted to literary translation studies and did not include relevant quantities of terminology. Recently corpora started to include other types of texts, like juridical or law texts. Examples are EuroParl (Koehn, 2005) or JRC-Acquis (Steinberger et al., 2006).

If we focus on languages like the Portuguese, we notice that other than these big corpora there are not much more choices. There are a few technical corpora compiled in the OPUS project (Tiedemann and Nygaard, 2004) like OpenOffice or Apache documentation, or literary corpus like (Frankenberg-Garcia and Santos, 2003).

These corpora include some terminology and are relevant for terminology analysis and extraction. But they have problems. EuroParl is mainly oral, that results in a bad quality alignment. JRC-Acquis include some more interesting terminology and has good alignment quality. But the range of terminological terms found is quite limited. JRC-Acquis includes the basic norms for every country joining the European Union. These norms focuses mainly on social and economic behavior laws. Finally, the technical corpora from OPUS are mainly in the computer science area. There is also a medicine corpus and a subtitles corpus.

Therefore, methods to create automatically closed-domain corpora are relevant, especially if one can construct it fast and easily.

With that in mind we present a tool, GWB(GetWebBitext), to lookup for parallel documents in the web and create parallel corpora, from a closed-domain area of knowledge and rich on terminology. As main design principle, all this process should be completely automatic.

Our system is based on a set of seed keywords (normally, a couple of terminology term examples) and one or more Internet domains where the tool will search for the texts that will comprise the parallel corpus.

While we present the full pipeline of GWB, this article will focus essentially the parallel page candidates detection, their download and analysis. The remaining part of the pipeline uses a set of tools that where chosen for being open-source and freely available, but can be easily swapped by other similar tools.

### 1.1. GWB Design Principles

GWB was develop with the following design principles:

**Control over the text sources:** the user provides the set of Internet domains in which the search process will be performed;

**Full pipeline for terminology extraction:** GWB is not designed just to download the text that comprises the parallel corpus, but includes a complete pipeline of corpora processing that ends with the automatic extraction of bilingual resources;

**Modularity:** it is important to have a full pipeline of tools that work correctly as an unique tool. But is is also important that all the tools of the pipeline can be used as a standalone application[1]. Thus, it should be possible to make GWB perform just part of the pipeline, accordingly with the user needs. Also, some of the GWB modules depend on other specific languages, or use a specific tools. Being modular, GWB lets the user substitute any of the modules by any other tool.

**Reuse:** GWB does not try to reinvent the wheel, but instead, use already available Open-Source tools, like Open-Corpus-Workbench, Easy-Align, NATools, *Yahoo!* API or

[1] Following the Unix tradition: each command should do only one thing but do it well (in our case, we have a lot of space for improvement)



*Proceedings of the 3rd Workshop on Building and Using Comparable Corpora, LREC 2010*, pages 50–55
Malta, 22 May 2010

StarDict[2] (and others).

### 1.2. Other Tools

The idea of using the Web as a Corpus is not new (Bernardini et al., 2006) and there are a couple of well known applications for automatic corpora construction from the Web.

#### 1.2.1. BootCat and WebBootCat

Probably, the most well known application for corpora construction is BootCat (Baroni and Bernardini, 2004), also available as a web application (Baroni et al., 2006).
BootCat was originally designed for automatic building of disposable corpora, using a set of seed terms. Web pages retrieved did not had to contain all the terms specified, but at least some combination.
BootCat is not just a retrieval tool. It includes a rich set of tools used for terminology extraction and statistical manipulation of terms, n-grams and others.
In order to connect all the small available tools in one single task some Unix expertise may be useful. This has some advantages and drawbacks. In one hand it makes the system flexible, in the other hand, it makes the system hard to use for less knowledge users.
GWB deals with a different problem (parallel corpora) but it try to reuse part of the BootCat principles, but adding an extra layer: a "work-flow" level command — a single command that can hide some of the typical tools combination. This tool defines a set of rules that specify how to run a pipeline of tools until the intended results are achieved.

#### 1.2.2. STRAND

Another tool for Parallel Corpora retrieval and construction from the web is STRAND (Resnik and Smith, 2003). STRAND approach is completely different from GWB or BootCat. STRAND does not search for specific terms. It just searches for parallel pages from the Web (or a specific domain). The procedure is simple: after retrieval, each page is checked for one of the following two properties:

- an entry page, with links to different language websites (thus, links with language names);
- check pages that link for the respective translated page.

GWB parallel page detection system is faster as it does not need to parse the HTML files neither to download all the document from the web. Also, GWB detects non-HTML documents that would not be detected using the above mentioned heuristics.

## 2. Architecture

GWB main algorithm might be defined as the following steps:

1. from a set of user-provided keywords $K$, a pair of languages, $L_1$ and $L_2$ and a set of valid Internet domains $D$, retrieve the first $N$ document URLs that contain all the keywords $K$ and is cataloged by the search engine (for instance, *Yahoo!*) as being in language $L_1$.

$$Docs_{L_1} = yahoo(K, site : D, lang : L_1)$$

2. for each URL in $Docs_{L_1}$ try to guess the corresponding URL with the document in language $L_2$ (this process is similar to the described by Mohler and Mihalcea (2008)):

$$Docs_{L_2} = parguess(Docs_{L_1}, L_2)$$

3. retrieve all documents pointed by the obtained URL (if they exist) and convert them to a textual format (PML):

$$Bitexts = retrieve(Docs_{L_1}, Docs_{L_2})$$

4. build a parallel corpora $PC$ aligning at the sentence level the retrieved documents. Note that this is done for each document pair.

$$PC = align(Bitexts)$$

5. filter the parallel corpora discarding translation units or documents with low alignment quality:

$$PC = filter(PC)$$

6. extract probabilistic translation dictionaries from the aligned corpora:

$$PTD = extractPTD(PC)$$

7. extract bilingual terminology using the probabilistic dictionaries and a set of alignment patterns:

$$Terms = terms(PC, PTD, Patterns)$$

8. create a StarDict dictionary for off-line usage based on the bilingual terminology and dictionaries extracted:

$$StarDict = mkSD(PTD, Terms)$$

The GWB modules work in pipeline as shown in figure 1.

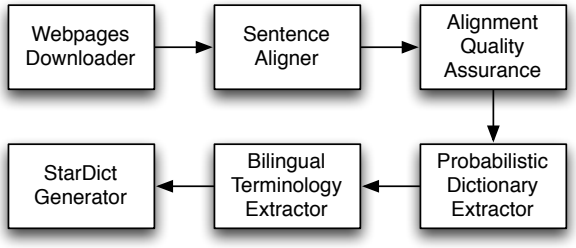



Figure 1: GWB Architecture

The next sections describe each one of these modules in greater detail.

[2] Available from `http://stardict.sourceforge.net/`

### 2.1. Web-pages downloader

The downloader module relies on an Applications Programmer Interface (API) to a web search engine, like *Google* or *Yahoo!*. Given the current limitations imposed by *Google* for its API, GWB uses *Yahoo!*.

#### 2.1.1. Find Documents in language 1

GWB receives a pair of language identifiers, a set of keywords (terminology examples) in the first language and a set of Internet domains to crawl. GWB lets the user to specify a closed set of keywords or to let the application generate more seeds based on morphological information. Therefore, any keyword followed by an asterisk will be lemmatized and its lemma will be used for generation. For the Portuguese and English language we have been using jSpell morphological analyzer and respective dictionaries (Simões and Almeida, 2001).

GWB will search for pages with all those keywords (generated words will be OR'ed automatically) under the specified Internet domains.

By default, GWB searches for PDF and HTML pages. Other document types can be defined, being just a matter of writing a plug-in to convert that document type to plain text.

The number of pages retrieved can be defined (as a command line option or by configuration). By default, GWB retrieves 100 pages.

#### 2.1.2. Calculate URL candidates for language 2

The next step is to find the translation pages. The URL candidates are calculated using a technique described in (Almeida et al., 2002) that relies on systematic web-pages organization (as *good* web-masters usually do): it is natural that a page in Portuguese includes a substring in the URL that specifies that language, like `Portuguese`, `pt` or `port`[3].

Therefore, one can rewrite that substring in the URL for a set of possible equivalences in the target language and check if any of the pages exist. There is a list of common keywords for each language which makes it easy to any user to use the tool without further configuration. In any case, it is possible for the user to add new languages or new language keywords.

Note that some caution should be taken during this substitution, as the domain portion of the URL should not be adapted.

Each document pair successfully retrieved, becomes a bitext candidate, is converted to plain text and sent to the sentence aligner.

### 2.2. Sentence aligner

Each plain text pair needs to be processed before alignment. It is necessary to detect sentence boundaries (segmentation) and detect word boundaries (tokenization). In our experiments we are using `Lingua::PT::PLNbase`, a Perl modules written for segmentation and tokenization of Portuguese. While the module was written with Portuguese in mind it supports some constructs from English, French and Spanish. In any case, it is easy to plug-in any other tool to segment and tokenize the text.

The segmented and tokenized texts are stored in a specific XML-based format, named PML, where just texts, paragraphs and sentences are annotated. Words are separated from each other with a blank.

These PML files are then sentence-aligned using easy-align. This aligner is part of Corpus Workbench (Christ et al., 1999). It uses the usual sentence size information to perform the alignment, but it also supports external bilingual dictionaries to help the synchronization. As easy-align relies on CWB, the PML files are firstly encoded as two separate monolingual corpora and then aligned.

The alignment result is then exported in TMX (Translation Memory Exchange) format files. While we are aware that this is not the most usual format for parallel corpora we find it more usable than TEI (Text-Encoded Initiative) or XCES (XML Corpus Encoding Standard).

Note that at this moment we still have several different TMX files (one for each retrieved document).

### 2.3. Alignment quality assurance

Each TMX file is analyzed in terms of quality. This can make the full document to be rejected, or some specific translation units to be deleted.

For translation unit quality analysis GWB uses the following heuristics[4]:

- sentence length comparison: while the main algorithm for easy-align is based on sentence length, some times the algorithm results include alignments with big sentence length differences. More precisely, the system will discard any translation unit with more than 20 characters for both languages, and with one language length greater than two times the length of the other.
- non-words preservation: numbers are extracted and compared. While the sequence is not required to be the same, they must all be preserved.
- punctuation analysis: while it is natural that punctuation changes (sentences are split, some languages use more commas than other and so on), some specific punctuation should be preserved.
- word translation probabilities: GWB is also able to evaluate translation units quality using bilingual dictionaries (or probabilistic translation dictionaries). As this subject is not the main topic for this article details will not be presented.

Full translation memories will be discarded if:

- more than half of the translation units were discarded by the previous heuristics;
- the majority (80%) of the alignments are not one-to-one sentence alignments.

[3] We know not all web sites follow these convention. Also, there is the possibility of false positives. But the pipeline of tools take the needed care to check languages and alignment possibilities.

[4] All these values can be user configured. This is relevant as different language pairs will have different ratios.

The TMX files that are not discarded are then concatenated together in a single file using the `XML::TMX` Perl module. This TMX file is the final corpus that will be processed by the next modules to produce bilingual resources.

### 2.4. Probabilistic dictionary extractor

The resulting parallel corpus is processed by NATools toolkit (Simões and Almeida, 2007) for the extraction of probabilistic translation dictionaries (Simões and Almeida, 2003). As the NATools extractor handles TMX files directly, this step is nothing more than the NATools corpus creation application and the final treatment of probabilistic translation dictionaries.

### 2.5. Bilingual terminology extractor

The same NATools toolkit includes an application for parallel terminology extraction (Simões and Almeida, 2008). This extraction is guided by a set of translation patterns, where the user can specify what kind of constructions he/she is searching. Therefore, this method can be used to extract terminology but also to extract specific linguistic constructions that are under analysis.

### 2.6. StarDict generator

It is our conviction that results extracted automatically should be made available to the end-user in a legible format. While to extract resources and have them available in textual format is useful when statistics are to be calculated, or the resources are to be integrated in other tools, for translation or linguistic studies it is easier to consult the resources as if they were a dictionary. With this in mind, GWB final module grabs the probabilistic translation dictionaries and the terminology extracted in the previous steps and constructs a StarDict dictionary for off-line viewing and querying.

## 3. Experiments

In this section we will discuss some experiments in order to give a better picture of what we can do with this tools (how we are using it) and show some simple metrics. The presented case-studies are:

- extract a translation memory from a small-size Web-site (a call-for-papers web-site);
- build a narrow domain parallel corpus (about alcoholic beverages) following the complete pipeline.

### 3.1. Small parallel corpus for a simple terminologically rich Web-Page

The first experiment corresponds to the following situation:

- we spotted a well written call for papers, with a good introduction to the area and a large set of *central topics*[5];
- we are in the presence of a small-size Web-site, with suitable translation quality;
- the web-site is available both in Portuguese and Spanish.
- we do not expect to obtain a real-size parallel corpus, but just a very small translation memory file (TMX) with a specific list of topics.

While this is a small text that will not be suitable for terminology extraction, our main purpose here is the extraction of a small translation memory that can be later used together with other to translate or create a bigger parallel corpus.
To create the translation memory we can use GWB as follows:

```
getwebbitext
 -s "ciawi-conf.org"   #  site-sources
 -l pt:es              # language pair
 -until tmx    # stop when TMX is done
 trabalhos
```

The keyword used — **trabalhos** (*call for papers* = *chamada de trabalhos*) — is present in the text and is valid in just one of the languages (Portuguese).
After near 10 seconds of network activity, we obtained 7 bitexts. GWB found 8 documents matching "*trabalhos*", but only 7 parallel documents. Follows an example of a retrieved URL (Portuguese) and the respective rewritten URL for the target language (Spanish).

`www.ciawi-conf.org/pt/cfp.asp`

⇓

`www.ciawi-conf.org/es/cfp.asp`

After bitexts extraction the alignment process takes place, aligning the documents and building a TMX file, with about 305 translation units (1 987 words for the source language and 2 067 for the target language).)
This experiment took less then 20s. In this case we decided not to generate dictionaries or terminology as the corpus size is too small.
In any case, the TMX file is still useful and can be used directly in common computer aided translation (CAT) tools like SDL-Trados, POedit or Omega-T.
The exercise is not complex – all the bitext candidate pairs passed in the quality control (100%) and the TMX had 4 alignment errors (98.6%)

### 3.2. Terminology on alcoholic beverages

In this second experiment we built a parallel corpus for a specific narrow domain (wine, spirit drinks and similar).
To start using GWB the user needs an Internet domain where there is texts on the chosen area. European laws include sections related to that subject, so we used `http://eur-lex.europa.eu/` as the source web-site.
In our experiments we concluded that EurLex web-site is both, one of the biggest multilingual quality sources and a good source for terminologically rich documents. That is why EurLex is the default source for GWB.
In order to select relevant information we used some terms in domain we are searching. In this case they keywords were *cerveja* (beer) and *vodka* (in Portuguese we use the same word as in English):

[5]For example, `http://www.ciawi-conf.org/`

```
getwebbitext
   -s eur-lex.europa.eu
   -l pt:en
   cerveja  vodka
```

By default GWB will get the first 100 documents[6] in the source language. GWB took 3m22s[7] to execute this task, and we obtained 37 MB of bitext candidates (12 bitexts in HTML format and 22 in PDF)[8].

#### 3.2.1. Bitext to TMX

We made two align experiments, one excluding the PDF files and another one including all retrieved files.
Excluding the PDF documents the final TMX had 32 941 translation units (about 9MB). Including PDF documents (several of the PDF documents were rejected given format or alignment problems) the final TMX had 81 844 translation units (about 22MB of text — about 1 300 000 tokens per language).

#### 3.2.2. Probabilistic translation dictionary extraction

Follows some (hand-selected) example entries from the extracted probabilistic translation dictionary. Each entry includes the term, its number of occurrences in the source language corpus and a set of probable translations.

$$cervejas\ (29)\ \begin{cases} beer & \rightarrow & 98\% \\ actual & \rightarrow & 2\% \end{cases}$$

$$cerveja\ (53)\ \begin{cases} beer & \rightarrow & 62\% \\ brewing & \rightarrow & 24\% \\ distilling & \rightarrow & 3\% \\ coloured & \rightarrow & 1\% \end{cases}$$

$$vodka\ (139)\ \begin{cases} vodka & \rightarrow & 94\% \\ flavoured & \rightarrow & 2\% \\ vodkas & \rightarrow & 1\% \end{cases}$$

$$licor\ (73)\ \begin{cases} liqueur & \rightarrow & 95\% \\ licor & \rightarrow & 2\% \\ liqueurs & \rightarrow & 1\% \end{cases}$$

$$rum\ (99)\ \begin{cases} rum & \rightarrow & 96\% \\ produced & \rightarrow & 1\% \\ word & \rightarrow & 1\% \\ solbaerrom & \rightarrow & 1\% \end{cases}$$

$$vinho\ (271)\ \begin{cases} wine & \rightarrow & 81\% \\ vinho & \rightarrow & 7\% \\ aromatised & \rightarrow & 2\% \\ wines & \rightarrow & 2\% \\ wine-based & \rightarrow & 1\% \end{cases}$$

$$vinagres\ (38)\ \{ vinegar \rightarrow 96\%$$

$$malte\ (208)\ \{ malt \rightarrow 95\%$$

$$aguardente\ (226)\ \begin{cases} spirit & \rightarrow & 70\% \\ aguardente & \rightarrow & 14\% \\ spirits & \rightarrow & 13\% \\ diluted & \rightarrow & 1\% \\ distilled & \rightarrow & 1\% \end{cases}$$

$$porto\ (42)\ \begin{cases} porto & \rightarrow & 89\% \\ port & \rightarrow & 6\% \\ reserva & \rightarrow & 3\% \\ doce & \rightarrow & 2\% \end{cases}$$

The full process took near 30 minutes but was completely automatic. In the end we obtained:

- a 81K translation unit TMX file (22MB);
- a pair of probabilistic translation dictionaries;
- a StarDict dictionary (check figure 2 for an example);

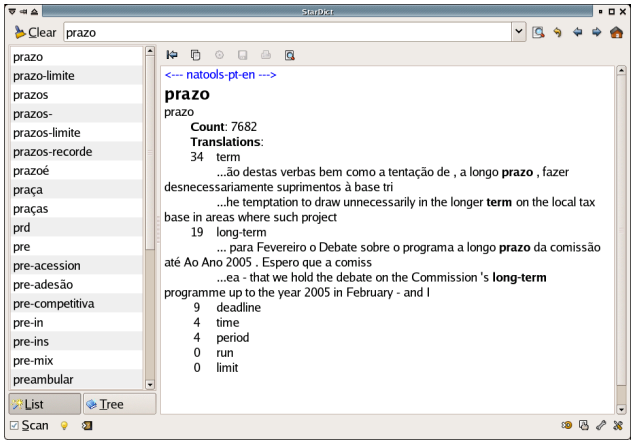



Figure 2: StarDict screenshot for an automatically built dictionary.

### 3.3. Some practical tips

In order to make useful things with GWB, some practical tips may help:

1. whenever possible choose sources that you know.
2. parallel corpora do not need to be built all at once. You can build small translation memories with good translation quality and join them later to create bigger resources.
3. when selecting the seed terms, try to use words that just belong to one of the languages (if possible the most less represented one).
4. use websites where you suspect there are annexed/linked documents like legislative texts;
5. chose a set of domain specific seed terms (eg. “ácido sulfúrico” and “nitrato de prata”);
6. always save the set of seed terms and domain sources, in order to be easy to continue the parallel corpora extraction task in the future.
7. sometimes is difficult to find good site with bitexts about some domains. Techniques like the ones presented in (Resnik and Smith, 2003) prove to be useful in finding sites of bitext sources.

[6] There is a command line option to redefine this value.
[7] real 3m21.913s
[8] It is possible to select the type of documents to be retrieved.

## 4. Conclusions

Current web-life makes it easy to find interesting pages, rich in terminology. Main problem would be how to retrieve those pages and create some kind of lexicon from it. We defend that GWB is a suitable tool to attack this kind of web-sites and construct bilingual resources automatically.

Our main objective is not quantity but quality. That explains why GWB requires a specific domain for the documents to be searched and why it also requires a full set of terms (and not just a subset).

GWB is mainly designed for small knowledge areas. A well defined set of seed terms is the key for the quality of the obtained corpora.

Main future issue for GWB is the creation of a distribution package, and put the tool available in CPAN for easy dissemination and installation.

## 5. Acknowledgments

This work was partially funded by project *Português em paralelo com seis línguas (Português, Español, Russian, Français, Italiano, Deutsch, English)* grant PTDC/CLE-LLI/108948/2008 from *Fundação para a Ciência e a Tecnologia*.

## 6. References

José João Almeida, Alberto Manuel Simões, and José Alves Castro. 2002. Grabbing parallel corpora from the web. *Procesamiento del Lenguaje Natural*, 29:13–20, September.

Marco Baroni and Silvia Bernardini. 2004. BootCaT: Bootstrapping corpora and terms from the web. In *In Proceedings of LREC 2004*, pages 1313–1316.

Marco Baroni, Adam Kilgarriff, Jan Pomikálek, and Pavel Rychlý. 2006. WebBootCaT: instant domain-specific corpora to support human translators. In *Proceedings of EAMT 2006 - 11th Annual Conference of the European Association for Machine Translation*, pages 247–252, Oslo. The Norwegian National LOGON Consortium and The Deparments of Computer Science and Linguistics and Nordic Studies at Oslo University.

Silvia Bernardini, Marco Baroni, and Stefan Evert. 2006. A wacky introduction. In Marco Baroni and Silvia Bernardini, editors, *WaCky! Working Papers on the Web as Corpus*, pages 9–40. Gedit Edizioni, September.

Oliver Christ, Bruno M. Schulze, Anja Hofmann, and Esther König, 1999. *The IMS Corpus Workbench: Corpus Query Processor (CQP): User's Manual*. Institute for Natural Language Processing, University of Stuttgart, March.

Ana Frankenberg-Garcia and Diana Santos. 2003. Introducing COMPARA, the portuguese-english parallel translation corpus. In Silvia Bernardini Federico Zanettin and Dominic Stewart, editors, *Corpora in Translation Education*, pages 71–87. Manchester: St. Jerome Publishing.

Philipp Koehn. 2005. EuroParl: A parallel corpus for statistical machine translation. In *Proceedings of MT-Summit*, pages 79–86.

Michael Mohler and Rada Mihalcea. 2008. Babylon parallel text builder: Gathering parallel texts for low-density languages. In Nicoletta Calzolari et at., editor, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

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

Alberto Manuel Simões and José João Almeida. 2001. `jspell.pm` — um módulo de análise morfológica para uso em processamento de linguagem natural. In *Actas da Associação Portuguesa de Linguística*, pages 485–495.

Alberto M. Simões and J. João Almeida. 2003. NATools – a statistical word aligner workbench. *Procesamiento del Lenguaje Natural*, 31:217–224, September.

Alberto Simões and José João Almeida. 2007. Parallel corpora based translation resources extraction. *Procesamiento del Lenguaje Natural*, 39:265–272, September.

Alberto Simões and José João Almeida. 2008. Bilingual terminology extraction based on translation patterns. *Procesamiento del Lenguaje Natural*, 41:281–288, September.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufiş, and Dániel Varga. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *5th International Conference on Language Resources and Evaluation (LREC'2006)*, Genoa, Italy, 24–26 May.

Jörg Tiedemann and Lars Nygaard. 2004. The opus corpus - parallel & free. In *Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, May 26–28.

# Foreign Language Examination Corpus for L2-Learning Studies

**Piotr Bański, Romuald Gozdawa-Gołębiowski**

Institute of English Studies
University of Warsaw
Nowy Świat 4
00-497 Warszawa, Poland
pkbanski@uw.edu.pl, r.gozdawa@uw.edu.pl

**Abstract**

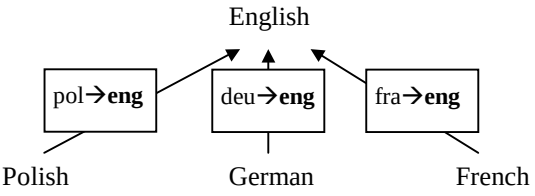
We describe the structure and the features of the Foreign Language Examination Corpus, a University of Warsaw project, launched on the initiative of the University Council for the Certification of Language Proficiency. The FLEC is an innovative comparable learner corpus, which will ultimately contain data from nearly forty languages offered for examination purposes at the University of Warsaw at five different levels of proficiency. The main stress will be on the linguistic behaviour of Poles studying any foreign language, error patterns exhibited, possible transfer/interference from the L2 learners' native language, emerging interlanguage properties as well as language traits stable across a population of test-takers. It will also be possible to assess the performance of the examiners and determine inter-rater stability. The corpus is encoded in TEI XML and uses stand-off architecture.

## 1. Introduction

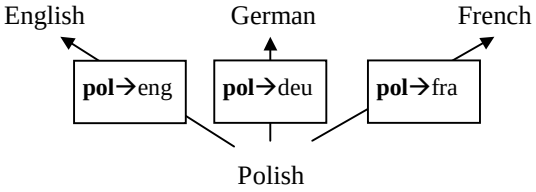
The present paper describes the structure and the features of the Foreign Language Examination Corpus (FLEC), a University of Warsaw project, launched on the initiative of the University Council for the Certification of Language Proficiency. The FLEC will ultimately contain data from nearly forty languages offered for examination purposes at the University of Warsaw at five different levels of proficiency – A2, B1, B2, C1 and C2, according to the Common European Framework of Reference for Languages (CEFR). CEFR guidelines, put together by a team of Council of Europe authors (D. Coste, B. North, J. Sheils, J. Trim), after more than twenty years of research, have been adopted all over Europe for designing language syllabuses and curricula, training examiners, preparing exams and exam formats, writing manuals and teaching materials. At the same time the structure of language courses, examination standards, proficiency levels become comparable for a wide range of languages, thereby offering a stable and transparent system of assessment and evaluation of language abilities, irrespective of the context in which these abilities may have been acquired.

The FLEC will be a multi-lingual comparable learner corpus with applications for the study of interlanguage growth in the process of L2 Learning, as well as for the improvement of inter-rater reliability. Part of the inspiration for the FLEC came from the ICLE (International Corpus of Learner English) project, cf. (Granger, 2008; Granger *et al.*, 2009).[1] However, while the ICLE concentrates on English and measures the interlanguage depending on the student's native language (cf. 1), the FLEC focuses on Polish as the native language and looks at the resulting range of L2 approximations, acknowledging, but not limiting itself to the identification of transfer and interference phenomena in the test-takers' written output (cf. 2).

(1) ICLE: focus on the interlanguage from the point of view of the target (English)




(2) FLEC UW: (primary) focus on the interlanguage from the point of view of the source (Polish)




The data for the corpus will come from the Warsaw University Certification Exams that are held every semester and whose format conforms to the requirements and guidelines formulated in the CEFR. The written part of the exam has the same format for all the languages offered in any examination session and is assessed according to the same set of criteria, which should enable cross-linguistic comparisons.

The corpus is going to have a multi-layer and multi-module architecture, allowing for many research tasks to be carried out on it: among others, it will allow for measuring the influence of the Polish language on the acquisition of target-language structures, thanks to the possibility of comparing the kinds of errors made by

[1] http://cecl.fltr.ucl.ac.be/Cecl-Projects/Icle/icle.htm



*Proceedings of the 3rd Workshop on Building and Using Comparable Corpora, LREC 2010*, pages 56–64
Malta, 22 May 2010

students; it will also allow for measuring and correcting the discrepancies among the raters. This should result in a more objective and effective learning and teaching process.

A learner corpus is often used in measurements concerning the phenomenon of distinct non-native quality of written compositions by L2 students. The results of such measurements make it possible to correct this negative effect by upgrading the teaching materials (textbooks, teacher's books, dictionaries, etc.) and by helping to select better source materials. In the case of the ICLE, the focus is on capturing the common traits that students of English of many different backgrounds display. In the case of the FLEC, this will also be possible (including the comparison against the results of the PICLE project carried out at Adam Mickiewicz University in Poznań[2]), but the main stress will be on the linguistic behaviour of Poles studying any foreign language.

In section 2, we talk about the reasons for the adoption of TEI XML as the encoding format, section 3 looks at the architecture of the corpus, and section 4 reviews the projected uses.

## 2. Corpus encoding format: TEI XML

Unlike the ICLE, the FLEC will be encoded in XML (Bray *et al.*, 1998), and more specifically, in its application known as the Text Encoding Initiative (TEI Consortium, 2010). Using this self-documenting format will ensure the versatility of the corpus, as well as its sustainability against the changing technology, at the same time increasing the chance for interoperability with other resources and tools (cf. Bird and Simons, 2003).

Just like XML is a *de facto* interchange format for many kinds of data on the Internet, the TEI has become a *de facto* documentation and interchange standard for many kinds of texts used or produced by the Humanities. In the field of language corpora with linguistic annotation, it nowadays competes primarily with the XML variant of the Corpus Encoding Standard (XCES, Ide *et al.*, 2000)[3] and the PAULA (Dipper, 2005), a well-thought *best practice* for the description of multi-modal and multi-level corpora. Additionally, a tightly interrelated set of Language-Resource-description standards is at various stages of preparation by the ISO TC 37 SC 4 committee; they are often referred to as the "LAF-family of standards", where "LAF" stands for the Linguistic Annotation Format (Ide and Romary, 2007) that defines a pivot representation for various layers of annotation. As Przepiórkowski and Bański (2010) show, the current TEI toolkit has all the advantages of the XCES without its shortcomings (among others, restriction to morphosyntactic markup, lack of devices to handle alternatives and discontinuity, incomplete documentation, lack of further development), while at the same time being easily mappable to the developing ISO-LAF standards and the LAF-inspired PAULA. Additionally, the TEI provides a widely recognized way of encoding text metadata, in terms of the so-called TEI headers (which many corpus projects use even if they do not use TEI schemas for text encoding). The point that Przepiórkowski and Bański (2010) make is that it is in essence an economical and pragmatic decision to adopt a toolkit that is able to encompass the current best-practice recommendations within a single set of well-constrained and well-defined schemas, configured in the so-called ODD files that make it possible to both define the constraints for annotations (schemas) and at the same time to document them. Adopting TEI XML will also allow us to process and visualize the corpus data and dependencies by means of multiple free and often open-source tools that have been created by the XML and TEI communities.

The FLEC is the second resource after the Open-Content Text Corpus[4] to closely follow the example of the National Corpus of Polish (NCP, http://nkjp.pl/) of using TEI XML for the purpose of creating large multi-layer text corpora (cf. Przepiórkowski and Bański, 2009). The OCTC explores an open-content strategy of development, adding a comparable component to the NCP-style of architecture, while the FLEC is a practical application in the area of L2 Learning. A project with similar aims, based on similar principles and using similar architecture is FALKO (*Fehlerannotiertes Lernerkorpus des Deutschen*, cf. Lüdeling *et al.*, 2005).

## 3. Structure of the corpus

The corpus prototype comprises four basic layers of annotation: the text layer, the segmentation layer, the grammatical layer, and the error-identification layer.[5] Each of them is stored in a separate file and references the others by a system of pointers (cf. 3 below). In this way, they create a stand-off annotation system, where the annotation documents are stored separately from the text that they refer to (see e.g. Ide and Romary, 2007). Stand-off markup is typically contrasted with inline markup, but it has to be borne in mind that the distinction is rarely binary. A truly radical stand-off approach was advocated by the XCES (*Ide at al.*, 2000) and nowadays the Ameri-

[2] http://ifa.amu.edu.pl/~kprzemek/PICLE_search.php, cf. (Kaszubski, 1999). Another learner corpus project in carried out in Poland was PELCRA, cf. (Lewandowska-Tomaszczyk *et al.*, 2000).

[3] The XCES is an XML-ised version of the SGML-based CES, which was a specialization of an early version of the TEI (TEI P3), for the purpose of encoding language corpora. The TEI (nowadays at P5) has reabsorbed many of the valuable innovations that the (X)CES proposed and offers them as a separate module (mostly in chapter 15 of the TEI Guidelines).

[4] https://sourceforge.net/projects/octc/, see Bański and Wójtowicz (2010).

[5] In the spirit of Goecke *et al.* (2009), we use the term *annotation level* to refer to the concept being annotated (e.g. segmentation or morphosyntax), and the term *annotation layer* to refer to the particular technical realization of that concept. In stand-off markup, each annotation layer tends to be located in a separate XML document.

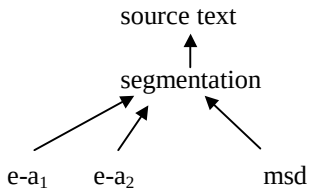
can National Corpus (ANC, cf. Ide and Suderman, 2006) inherits this idea, which in itself is very attractive, because it cleanly separates the object to be annotated (in our case, text) from its description, allowing the pure text to be made read-only and thus immutable and stored in an un-annotated form.[6] This approach duly sanctifies the electronic text as the concrete essence of an author's linguistic creation, possibly copyright-restricted and deservedly secure, and shifts the entire weight of processing towards the annotation documents.

There is no need or basis in the FLEC for this kind of approach. Its sources are not electronic documents – they need to be transcribed into the electronic form by hand.[7] Thus, there is no previous electronic creation to be preserved without markup, and additionally, the transcribed text must enter the corpus with spelling errors corrected, for practical reasons: spelling-error-aware tagging is still at the early stage, and not available for all the 40 languages that the FLEC will eventually contain. This is why we need to take a more conservative approach and make sure that, while the original misspelt forms are preserved, the basic text layer contains text prepared for at least morphological analysis (morphosyntactic analysis may still go astray due to various grammatical errors). This will be further discussed in section 3.1 below. While entering text, the transcribers will also introduce sentence boundaries (thus avoiding another potential trap in automatic tagging, especially of potentially malformed text), gap markers for incomplete text (this in fact could be very difficult on the basis of the electronic version alone), and possibly highlight markers. Thus, the encoders who retype the essays will at the same time introduce some inline markup.

This non-radical stand-off approach is motivated by two more factors: firstly, the tools to merge the pure text with its annotations in the TEI do not exist (the above-mentioned ANC has to use its own tools tailored to the specific annotation format adopted there), so even if the pure text is to be addressed by means of character offsets, it has to be enclosed in at least one XML element, in order to be processable by standard XML tools. Secondly, the TEI imposes minimal structure on each corpus file, and once we conform to this minimal structure, it is a matter of convenience and pragmatism to allow the encoders to use some of the basic TEI tags to flesh out some more of the nuances of the encoded text.

After the layer of source text is created, it is operated on by an indexer and a tokeniser that create another annotation layer that is going to serve as the basis for further analysis. The next layer is the layer of error annotation, created on the basis of rater assessments. Stand-off approaches often use multi-level hierarchies for annotation layers (cf. Dipper, 2005; Przepiórkowski and Bański, 2009) – the FLEC uses a variant of such a hierarchy, as shown in example (3) below, where it can be seen that both the error-assessment layer and the morphosyntactic layer reference the segmentation layer.

(3) Multi-level annotation hierarchy in the FLEC




e-a: error-assessment layer
msd: morphosyntactic description layer

Note that the error-assessment layers are located at the same level as the morphosyntactic layer – this is because both layers address the segmentation layer.

Apart from that, the metadata concerning each text will be stored in a header file, included by each of the annotation layers. The header contains information on the source of the text (date and level of the exam, the style and the topic of the composition, auxiliary materials used, etc.) as well as a partial (anonymous) author profile – stating the estimated level of command of the target language, the degree of the learner's immersion in the target language, possibly their age, etc.

Each essay is ca. 200 to 400 word long. The numbers of students taking the test in English – the language most widely represented in the corpus – for the B2 and C1 level exams, respectively, are as follows: 2007: 1582+303, 2008: 3281+304, 2009: 4971+359. In 2008 and 2009, respectively 8 and 12 students took the C2-level test in English. Overall, in 2009, the Foreign Language Certification Exam was taken by slightly over 6,000 students. These numbers show that it is reasonable to construct the corpus on the basis of English data and then, after all architectural issues are solved, to periodi-

[6] If newline markers – also a type of structural annotation – are ignored, as is commonly the case. Though already at this point, it is worth highlighting one difficulty in handling raw text files: they need to be consistently normalized in terms of the operating system newline markers, which consist of two characters in the Microsoft Windows operating system, and a single character in Unix-like systems or MacOS. The FLEC avoids this by keeping source text in lightly tagged XML files.

[7] We dismiss the option of performing any kind of OCR (optical character recognition) on these exams. There is no way for any currently available system known to us to help us in this task – the text pieces are too short for an intelligent OCR system to learn and often, they are a challenge even for human eyes – we cannot afford multiple sessions scanning and re-scanning single exams and then verifying the ca. 400 words of text. However, the transcribers' job will be made as easy as possible: the source text file templates come already filled with XML down to the level of the <s>entence, and the transcribers will only have to fill in the <s> elements, possibly using a few other tags to mark up highlighted spans, gaps, unclear words or spelling errors (cf. sect. 3.1).

cally add the other languages.[8]
In what follows, we briefly look at each of the annotation layers, providing examples from a corpus prototype that we have built for testing purposes.

### 3.1. Text layer

The text layer comprises the individual essays. The essays from each kind of exam, each level and each language will constitute separate subcorpora, to provide for cross-section and pseudo-longitudinal measurements. Recall that the essays consist of 200 to 400 words depending on the exam level. This layer will have minimal markup: apart from the obligatory structure imposed by the TEI, it will contain paragraph (<p>) and sentence (<s>) markers, in addition to several other XML tags, indicating highlighted (underlined, capitalized) spans (<hi>), quotations (<q>) and indecipherable words (<unclear>), incomplete/unfinished passages (<gap>). An additional mechanism that has to be employed at this level is normalization of spelling mistakes. This is handled by special TEI elements, as illustrated below.

(4) Corrections at the source-text level, as introduced by the transcriber

```
<s>I <choice>
      <sic>whoud</sic>
      <corr cert="high"
       resp="#bansp">would</corr>
   </choice> black-mail him?</s>
```

The <choice> element signals that only one of the segments that it contains belongs to the narrative stream and the user (in fact, the visualising application or query engine) has to choose between the erroneous form within the <sic> element or the corrected form within the <corr>. In this way, the automatic grammatical description tools can look at the corrected forms, while statistical measurements can be carried out on the misspellings.[9] Note that the correction carries attributes identifying the encoder (listed in the header) and provides information on the certainty of the judgement. Note also that our purpose is not editorial but linguistic: we want to record the final output of the test-taker, and therefore additions or deletions of text performed by the student are not marked up in any way – what is encoded is the final text stream of the essays.

### 3.2. Segmentation layer

Example (4) above shows the output of semi-manual encoding that it further processed by tokenising and indexing tools. A partial, simplified output of tokenization and indexing is shown below.

(5) Source-text level after automatic tokenization and indexing[10]

```
<s xml:id="_1.15-s">
   <seg xml:id="_1.15.1-seg">I</seg>
   <choice>
      <sic xml:id="_1.15.2.1-sic">
         <seg
          xml:id="_1.15.2-seg">whoud</seg>
      </sic>
      <corr
        xml:id="_1.15.2.2-corr"
        cert="high" resp="#bansp">
         <seg xml:id="_1.15.2.2.1-seg"
            >would</seg>
      </corr>
   </choice>
   <seg xml:id="_1.15.3-seg">
      <seg  xml:id="_1.15.3.1-seg"
            type="sub-token">black</seg>
      <seg  xml:id="_1.15.3.2-seg"
            type="sub-token">-</seg>
      <seg  xml:id="_1.15.3.3-seg"
            type="sub-token">mail</seg>
   </seg>
   <seg xml:id="_1.15.4-seg">him</seg>
   <seg xml:id="_1.15.5-seg">?</seg>
</s>
```

As can be seen above, each segment (including non-letter characters) is identified and equipped with an xml:id attribute, which will identify it at other layers of annotation. Tokenization in our corpus prototype for English is fairly radical, as can be seen in the example of *black-mail* treated as three segments. Similarly, all contracted *n't* sequences are tokenised separately, after the example of the CLAWS tagger (Garside and Smith, 1997; http://ucrel.lancs.ac.uk/claws/). This demonstrates the trivial truth that tokenization is often (i) language-dependent, and (ii) may be done in several ways (some taggers would tokenise *can't* as a single segment, some would split it into *can* and 't, and some would even lemmatise into *can* and *n't* (or *not*). This forces us to keep the segmentation document separate from the source text level – this way, we may supply a different tokenization document if we happen to use a different tagger in the future. At the same time, example (5) above shows that we do not obey stand-off principles closely at

[8] German is next in line (315 students at B2 in 2009), followed closely by French, Russian, and Spanish (53 at B2 in 2009). It has to be stressed that the uniform FLE system is still in the process of being instituted in all philological departments.

[9] This is also the level where various NLP methods can be applied. We are grateful for a reviewer's suggestion concerning large-scale text-mining, as described in e.g. (Turney, 2001).

[10] Note that the index on the segment *whoud* is the same as it would be on an unregularized segment. This is done to minimize the impact of corrections made in the source text files after the error-identification layers have been created – they will not be affected. The morphosyntactic layer will have to be regenerated automatically, after each modification of the source text.

this level, copying tokens from the source text level rather than using references. This is dictated by practical reasons: the TEI stand-off system is not capable of addressing spans of text because the appropriate tools do not exist (cf. (Bański, 2010), for a summary of the issues involved). Instead of constructing our own merge tools, we prefer to use what Goecke *et al.* (2010) call a "multiply-annotated text" system, at least temporarily.

### 3.3. Grammatical layer

The grammatical layer, as of yet non-existent in the corpus prototype, will contain the basic morphosyntactic description, to facilitate searches for grammatical patterns. It will be created automatically, by taggers that first identify the possibly disparate morphological interpretations of single words and then, on the basis of their syntactic contexts and statistical measurements or grammatical rules, choose the most likely interpretation(s). We envision experimenting with various taggers for individual languages, and having possibly more than one morphosyntactic annotation document addressing a single source text (or rather its segmentation layer). Recall that the source undergoes radical tokenization into indexed segments. Elements of the grammatical layer refer to the elements of the source by their xml:id attributes. This is presented in our test example in (6) below, on two versions of the CLAWS tagset (we suppress the xml:id attributes of segments as well as the xml:base attribute that redirects all corresp attributes to the appropriate file).

(6) a. CLAWS c5

```
<s xml:id="morph_1.1-s">
  <seg ana="PNP"
    corresp="segm.xml#_1.15.1-seg"/>
  <seg ana="VM0"
    corresp="segm.xml#_1.15.2.2.1-seg"/>
  <seg ana="VVI"
    corresp="segm.xml#_1.15.3-seg"/>
  <seg ana="PNP"
    corresp="segm.xml#_1.15.4-seg"/>
  <seg ana="?"
    corresp="segm.xml#_1.15.5-seg"/>
</s>
```

b. CLAWS c7

```
<s xml:id="morph_1.1-s">
  <seg ana="PPIS1"
    corresp="segm.xml#_1.15.1-seg"/>
  <seg ana="VM"
    corresp="#segm.xml_1.15.2.2.1-seg"/>
  <seg ana="VVI"
    corresp="#segm.xml_1.15.3-seg"/>
  <seg ana="PPHO1"
    corresp="segm.xml#_1.15.4-seg"/>
  <seg ana="?"
    corresp="#segm.xml_1.15.5-seg"/>
</s>
```

### 3.4. Error-identification layer

The fourth layer of annotation, the error-identification layer, holds pointers to individual words or spans thereof and the description of the grammatical errors that they illustrate. These will be created in the process of grading the tests (the raters will be anonymised but will also have identifiers, needed for the purpose of inter-rater comparisons). Note that a test is typically graded by at least two persons. Thus, there will usually be two instantiations of the error-identification layer per text.

The documents in this layer consist of several <spanGrp> elements, one for each value of several error categories, and a general  element addressing issues relevant to the entire essay, such as cohesion. Below are fragments of selected annotations.

(7) Fragments of the error-identification layer

```
<spanGrp resp="#bansp"
                type="gram" n="art">
  <span from="#segm.xml_1.1.1-seg"
        to="segm.xml#_1.1.1-seg"
        cert="high"
        rend="add">the $1</span>
  <span from="segm.xml#_1.5.7-seg"
        to="segm.xml#_1.5.7-seg"
        cert="high" rend="del"/>
</spanGrp>
<spanGrp resp="#bansp"
                type="gram" n="w/o">
  <span from="segm.xml#_1.15.1-seg"
        to="segm.xml#_1.15.2-seg"
        cert="high"
        rend="change">$2 $1</span>
</spanGrp>
```

The desired change is expressed by means of the @rend attribute (for additions, deletions, replacements or changes) and backreferences to the segments within the span. The first example illustrates the tagging for article omission (*role → the role*) and overuse (*the pupils → pupils*) and the second the lack of inversion in example (4). Note that keeping different error categories separately allows us to multiply annotate a single segment or span. A fairly flat structure for error categories is assumed at this level, so that we can remain largely agnostic with respect to the various error taxonomies (cf. e.g. (Díaz-Negrillo and Fernández-Domínguez, 2006) for an overview).

## 4. Projected uses of the corpus

The FLEC has the potential to become the nexus for several types of L2-related grammatical systemic analysis: it could be helpful in making straightforward L1:L2 comparisons, in the spirit of Contrastive Analysis, but without its behavioural underpinnings. It will examine a set of interlanguage properties against the corresponding L2 features, thereby doing Error Analysis. By projecting the native Polish patterns against the emerging interlanguage regularities, we get involved with Transfer Analy-

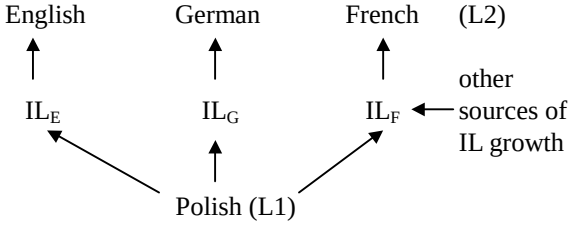
sis. Finally, comparing individual interlanguages helps identify the national traits of English (French, Swahili, etc) as it is used by native speakers of Polish, shedding light on L1:L2 transfer, commonly adopted learning and communication strategies, etc. Comparing the error identification results for groups of students sharing the same target language but being at different stages of language training will allow for measuring the rate of success of the teaching process.

Below, we first look more closely at the pedagogical applications of the FLEC, and then move on to show how the architecture of the corpus facilitates its projected uses.

### 4.1. Interlanguage studies

Unlike the ICLE model, the FLEC focuses on Polish as the native language and looks at the resulting range of L2 approximations, acknowledging, but not limiting itself to the identification of transfer and interference phenomena in the test-takers' written output (cf. 2). Thus an interlanguage text will be checked for possible L1 (Polish) influence, overgeneralizations and undergeneralizations of the target (L2) rules, creative formations, which do not conform to either L1 or L2 norms but result from the process of hypothesis formation/hypothesis testing. This may be juxtaposed against the amount of interlanguage which is grammatical and acceptable by L2 standards and the amount which is grammatical but lacks the native flavour, for example because it violates the subtle (or not-so-subtle) collocational preferences of individual lexical items.

(8) Interlanguage as an approximative system




The ratio of rule-governed patterns to the prefabricated chunks, found in any text will be a powerful indicator of the preferred production strategy: holistic or analytic. Possibly, evidence may be accumulated in favour of a third, middle-of-the road mode of sentence processing – the contentive mode, with its focus on content words, the main carriers of meaning, to the (partial) exclusion of analytic or formulaic considerations.

An analysis of interlanguage samples should lead to identifying other sources of interlanguage growth, above all the role of form-focussed instruction (cf. (Gozdawa-Gołębiowski, 2003) for some discussion of overt pedagogical intervention and a number of other mechanisms that drive forward interlanguage growth, while remaining insignificant in the process of first language acquisition). The majority of our test-takers are BA candidates, who need to complete a two-year language course to earn their bachelor's degree. The language course syllabi need to be approved by the University Council for the Certification of Language Proficiency, and depending on the emerging patterns of weak and strong points in the use of the L2 system by our students under exam conditions, course contents may change accordingly.

The FLEC's findings, in the sense of statistically significant morphosyntactic or lexical regularities as well as patterns of translational (non-)equivalence can serve as a powerful tool in modifying, updating and constructing language teaching syllabi, providing an accurate description of the strength and weaknesses of language use by Polish L2 learners. Precise questions can be asked and answered about the mechanisms of discourse management, text organisation, the use of linking devices, paragraph development, register appropriacy. This is the sort of data that contrastive linguists have always sought with a view to constructing reliable pedagogical grammars, and teacher trainers need to prepare teaching materials.

### 4.2. Corpus architecture vis-à-vis its uses

There are numerous advantages of the kind of corpus architecture described above. Firstly, it makes it possible to keep the original text virtually unmodified (note that spelling correction described above leaves the original forms as the content of <sic/> elements), so that it is open to new uses in the future. At the same time, it is open to critical evaluation of the various corrections and interpretations proposed in the higher levels of annotation. This is also advantageous for practical purposes, when the process of corpus-building is considered: source-text encoding will be performed by one group of annotators, with peaks after each exam session. After this is over, the segmentation layer will be created automatically and at that point, error encoding will be able to begin, the derivation of the grammatical layer being a procedure orthogonal to the encoding, the new corpus material will be usable for statistical purposes (spelling error rate, simple word-based concordancing).

Secondly, it will be possible to visualize the corrections introduced by the raters independently of each other or together – note that the raters may identify different spans of the source text and that their evaluations will often differ – by keeping the data from each rater in separate files, it will be possible to maintain the logical separation between the object that is described and the description itself. It also has to be borne in mind that these levels of annotation will be created later than the level of the source text – thanks to the fact that error-identification layers are separate, the integrity of the source data is ensured.

Thirdly, the rater decisions will be open to measurements against each other, and their intersection (the spans and decisions on which they agree) will create a common description of errors performed by the given group of students. This will allow for comparisons among different groups of students: those sharing the same target language (in a way recreating the goal of the ICLE but extending it to various target languages), as well as those with different target languages.

Next, it will be possible to examine the rater decisions

and see what the points of disagreement are. This may lead to creating guidelines for raters, thus ensuring a more objective grading process.

Finally, a corpus structured in the way described here is open-ended not only at the level of new texts and new target languages, but also new layers of annotation. It may be possible, in a future project, to add other kinds of description: e.g. syntactic or semantic, to e.g. measure the usage of the various senses of polysemous words and look for alignments with the ranges of meanings that their Polish equivalents display.

## 5. For and against a FLEC-based approach to the study of interlanguage

It should be clear by now that the proposed corpus has more potential than merely to support the three tradition-sanctioned approaches to learner language: contrastive analysis, transfer analysis and error analysis (cf. James 1997 for a thorough discussion of how the three approaches evolved and competed with each other). It elevates interlanguage to the status of an independent linguistic system, one which merits attention on its own, as rule-governed behaviour. This – inevitably – raises the question of the stability of the data. Any language token which appears in the data might be an accidental construct or a slip of the pen, due to inattention, stress, tiredness, perhaps even eligible for self-correction. A follow-up measurement of any learner's interlanguage would, of course, shed more light on the issue and help the researcher to sift the incidental "noise" in the data from the systematic properties of the emerging system. Since no post-tests can easily be arranged, the stable interlanguage properties can be identified with some degree of certainty by looking for repetitive patterns in a larger population of test takers. Needless to say, we are aware of a certain difficulty here: an interlanguage is a highly individual system (what James 1997 calls an idiosyncratic dialect) and its growth and evolution are determined by examining language samples coming from the same learner in a carefully designed longitudinal study. The idiosyncratic aspect of the evolution of any one interlanguage is unavailable for FLEC-based inspection. Instead, the cross-sectional patterns which the FLEC permits us to capture, will characterise interlanguage as a property of a wider group of users. After all, an interlanguage is not just a bundle of idiosyncratic patterns. As with any natural communication system it must belong to and be shared by a group. Let us take this to be a pedagogically oriented definition of interlanguage, with a Saussurean twist. It is to be hoped that this is the right way to heed Douglas' (2001:453) warning that "we provide empirical evidence for the consistency of the performances we observe, and that we further provide empirical and logical evidence that the interpretations we make of those performances are justified."

This brings us to another potential weakness of corpus-based interlanguage studies: we are prepared to draw conclusions about the learners' competence based on their performance. To the extent that the competence/performance distinction is real, we do not see any alternative to performance-based analysis of interlanguage competence. As Ellis and Barkuizen (2005: 21) aptly put it, "competence" can only be examined by investigating some kind of performance and (…) the key methodological issue is what kind of performance provides the most valid and reliable information about competence." It is a sound methodological assumption that samples of written production, divided according to the level of linguistic proficiency and reflecting topics that are part and parcel of our test takers' everyday experience, do qualify as the right kind of evidence to draw conclusions about interlanguage competence in a population.

## 6. Conclusion

Tono (2003) remarks on the poor design in some learner corpora, resulting in their data not being able to be fully exploited due to insufficient metadata, un-sustainable and un-interchangeable formats or missing kinds of annotation. The FLEC is designed to avoid such shortcomings: the metadata will be kept in a well-constrained TEI header, and the corpus itself will have all the flexibility of XML applications with the added markup semantics of the TEI. In addition, the corpus is open-ended in terms of annotation layers: syntactic, semantic, stylistic, etc. annotation layers can be added to the corpus at any time.

The FLEC is still within the 80% part of its creation: planning. Compared to planning, the building phase of the initial version of the corpus is expected to be reasonably short, given the expertise and tools coming from the NCP and the OCTC projects. During the coming certification exams, students will receive additional forms to fill out with their profiles for the text metadata, and the raters – guidelines concerning the visual aspects of the grading process (clear indication of text spans, unified taxonomy of error identifiers, etc.). After the English part of the corpus is tested and tuned, we will proceed to encode the other languages, to test the tools for cross-language (in effect, cross-subcorpus) query and visualization.

Measuring the influence of the native linguistic system upon the target language system requires a firmly defined starting point – in order to measure the degree of divergence or convergence, one has to know what to measure against. A solid basis for establishing the patterns of behaviour of native speakers of Polish will be provided by the National Corpus of Polish – a $10^9$-word resource that will be completed this year. Thanks to its stand-off architecture, the FLEC will be able to re-use the query and manual annotation tools produced by the NCP, such as Poliquarp (Janus and Przepiórkowski, 2009) or Anotatornia (Przepiórkowski and Murzynowski, forthcoming). We also intend to put the claims of Bański and Przepiórkowski (2010) to test, by transducing TEI to the PAULA format for visualisation and search purposes (cf. Chiarcos *et al.*, 2008), which is the format that the German FALKO corpus uses.

Note that the comparable nature of the FLEC will be

realised on more than one level and in more than one direction: apart from horizontal comparisons among students within the same exam component, vertical comparisons among the takers of exams of different levels of difficulty (e.g. A2 vs. B2). These comparisons concern both the textual output produced by students and the error-identification information produced by the raters.

## 7. Acknowledgements

We are grateful to the four anonymous BUCC reviewers for their constructive remarks. We were unable to answer one point of criticism, concerning the unfinished status of the corpus. We are hopeful, however, that even when illustrated by a small prototype, the paper is still useful as a report on how TEI markup and stand-off architecture in general can be employed for the purpose of creating a resource such as the FLEC.

## 8. References

Bański, P. (2010). Why TEI stand-off annotation doesn't quite work. Manuscript, University of Warsaw.

Bański, P.; Wójtowicz, B. (2010). The Open-Content Text Corpus project. In proceedings of the LREC workshop on "Language Resources: From Storyboard to Sustainability and LR Lifecycle Management" (LRSLM2010).

Bird, S., Simons, G. (2003). Seven dimensions of portability for language documentation and description. *Language*, 79, pp. 557--582.

Bray, T., Paoli, J., Sperberg-McQueen, C.M. (eds.) (1998). Extensible Markup Language (XML) Version 1.0. W3C Recommendation. Available from http://www.w3.org/TR/REC-xml/

Chiarcos, Ch.; Dipper, S.; Götze, M.; Leser, U.; Lüdeling, A.; Ritz, J.; Stede, M. (2008). A Flexible Framework for Integrating Annotations from Different Tools and Tagsets. In *Traitement Automatique des Langues* 49(2), pp. 271--293.

Díaz-Negrillo, A.; Fernández-Domínguez, J. (2006). Error Tagging Systems for Learner Corpora. *Revista Española de Lingüística Aplicada* (RESLA), 19, pp. 83--102.

Dipper, S. (2005). XML-based stand-off representation and exploitation of multi-level linguistic annotation. In *Proceedings of Berliner XML Tage 2005* (BXML 2005). pages 39--50, Berlin.

Douglas, D (2001). Performance consistency in second language acquisition and language testing research: a conceptual gap. *Second Language Research* 17: 442--456.

Ellis, R and Barkhuizen, G. (2005). Analysing learner language. Oxford: Oxford University Press.

Garside, R.; Smith, N. (1997). A Hybrid Grammatical Tagger: CLAWS4. In Garside, R.; Leech, G.; McEnery, A. (eds.), Corpus Annotation: Linguistic Information from Computer Text Corpora. London: Longman, pp. 102--121.

Goecke, D.; Metzing, D.; Lüngen, H.; Stührenberg, M.; Witt, A. (2010). Different views on markup. distinguishing levels and layers. In *Linguistic modeling of information and markup languages. Contributions to language technology.* Springer Netherlands, pp. 1--21.

Gozdawa-Gołębiowski, R. (2003) *Interlanguage formation. A study of the triggering mechanisms*. Warszawa: Instytut Anglistyki UW.

Granger, S. (2008) Learner Corpora. In Lüdeling A.; Kytö, M. (eds) *Corpus linguistics: an international handbook*, vol. 1. Mouton de Gruyter. pp. 259--275.

Granger, S.; Dagneaux, E.; Meunier, F.; Paquot, M. (eds.) (2009). *International Corpus of Learner English V2*. Louvain-la-Neuve: Presses universitaires de Louvain. Available from http://www.i6doc.com/.

Ide, N.; Bonhomme, P.; Romary, L. (2000). XCES: An XML-based Standard for Linguistic Corpora. Proceedings of the Second Language Resources and Evaluation Conference (LREC), Athens, Greece, pages 825--830.

Ide, N., Suderman, K. (2006). Integrating Linguistic Resources: The American National Corpus Model. *Proceedings of the Fifth Language Resources and Evaluation Conference (LREC)*, Genoa, Italy.

Ide, N., Romary, L. (2007). Towards International Standards for Language Resources. In Dybkjaer, L., Hemsen, H., Minker, W. (eds.), Evaluation of Text and Speech Systems, Springer, pages 263--284.

James, C (1997). Errors in language learning and user. Exploring error analysis. London: Longman.

Janus, D.; Przepiórkowski, A. (2007). Poliqarp: An open source corpus indexer and search engine with syntactic extensions. In *Proceedings of the ACL 2007 Demo and Poster Sessions*, Prague, pages 85--88.

Kaszubski, P. (1999). English Learner Corpora And The Polish Learner. Humanising Language Teaching, Year 1, Issue 8, December 1999; available from http://www.hltmag.co.uk/dec99/idea.htm.

Lewandowska-Tomaszczyk, B.; McEnery, A.; Leńko-Szymańska, A. (2000). Lexical problem areas in the PELCRA Learner Corpus of English. In Lewandowska-Tomaszczyk, B.; Melia, P.J. (eds) *PALC'99. Practical Apllications in Language Corpora*. Hamburg, Peter Lang, pp. 303--312.

Lüdeling, A.; Walter, M.; Kroymann, E.; Adolphs, P. (2005). Multi-level error annotation in learner corpora. *Proceedings of the Corpus Linguistics 2005 Conference*, Birmingham.

Przepiórkowski, A., Bański, P. (forthcoming). XML Text Interchange Format in the National Corpus of Polish. In S. Goźdź-Roszkowski (ed.) *The proceedings of Practical Applications in Language and Computers PALC 2009*, Frankfurt am Main: Peter Lang.

Przepiórkowski, A.; Bański, P. (2010). TEI P5 as a text encoding standard for multilevel corpus annotation. In Fang, A.C., Ide, N. and J. Webster (eds). *Language Resources and Global Interoperability. The Second International Conference on Global Interoperability for Language Resources (ICGL2010).* Hong Kong: City University of Hong Kong, pages 133--142.

Przepiórkowski, A.; Murzynowski, G. (forthcoming).

Manual annotation of the National Corpus of Polish with Anotatornia. In Goźdź-Roszkowski, S. *The proceedings of Practical Applications in Language and Computers (PALC-2009)*. Frankfurt: Peter Lang.

TEI Consortium (eds) (2010). TEI P5: Guidelines for Electronic Text Encoding and Interchange. Version 1.6.0. Last updated on February 12th 2010. TEI Consortium. http://www.tei-c.org/Guidelines/P5/

Tono, Y. (2003). Learner corpora: design, development and applications. In *Proceedings of the Corpus Linguistics 2003 Conference*. Lancaster, UK, pp. 800--809.

Turney, P. (2001). Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL. In *Proceedings of the 12th European Conference on Machine Learning*, pp. 491--502. (Available from http://cogprints.org/1796/)

# Lexical analysis of pre and post revolution discourse in Portugal

**Michel Généreux, Amália Mendes, L. Alice Santos Pereira, M. Fernanda Bacelar do Nascimento**

Centro de Linguística da Universidade de Lisboa
Av. Prof. Gama Pinto, 2
1649-003 Lisboa - Portugal

**Abstract**

This paper presents a lexical comparison of pre (1954-74) and post (1974-94) revolution parliamentary discourse in four comparable sub-corpora extracted from the Reference Corpus of Contemporary Portuguese (CRPC). After introducing the CRPC, including annotation and meta-data, we focus on a subset of the corpus dealing with parliamentary discourses, more particularly a time frame of forty years divided into four comparable sub-corpora, each covering a ten-year period, two pre revolution and two post revolution. We extract lexical density information as well as salient terms pertaining to each period to make a comparative evaluation of the periods. Our results show how a linguistic analysis essentially based on the use of simple n-gram statistics can produce key insights into the use, change and evolution of the Portuguese language around a critical time period in its history.

## 1. Introduction

This paper presents the Reference Corpus of Contemporary Portuguese and how we use it to compare the lexicons of pre (1954-74) and post (1974-94) revolution parliamentary discourse in Portugal. The question we are addressing is to what extent a political change (Portuguese revolution in 1974) is reflected in a change in vocabulary usage in speeches of the national assembly. The period covered by our corpora, 1954-1994, was chosen in accordance to the political situation in Portugal over this period of time. The date of the revolution, April 25th 1974, marks a deep change in the political regime, when a dictatorship who lasted almost 50 years was replaced by a democratic state. The first period, from 1954 to 1963, follows the second World War and brought some innovation, but is mainly marked by the McCarthyism, appreciated by the dictator Salazar, and also by the beginning of the war for liberation in the African territories occupied by Portugal. The following period, from 1964 to 1974, was dominated by the colonial wars, especially in Angola, Guinea-Bissau and Mozambique, and had a very negative imprint on the Portuguese population, leading to an increase in revolutionary activities, and an increasingly violent repression by the political police of the regime. During these two periods, there were no free elections and the political regime was based on the autocratic power of Salazar. The Parliament, called at the time *Assembleia Nacional*, could only discuss the legislation proposals of the Government, and political parties, like the Communist and Socialist Parties, could not openly exercise their activities. After the revolution, in 1974, there was a strong rupture with the ideology of the dictatorship, the colonial wars ended abruptly and the African colonies gained their independence. The political parties were legalized and the first free election took place exactly one year after the revolution. In the period between 1984 and 1994, the Portuguese State entered into a stable democracy and joined the European Community.

### 1.1. The CRPC corpus

The CRPC is the result of numerous efforts at the Centro de Linguística da Universidade de Lisboa (CLUL)[1] to produce an electronically based linguistic corpus containing, after cleaning, 301 million tokens, taken by sampling from several types of written texts (literature, newspapers, science, economics, law, parliamentary debates, technical and didactic documents, etc.) as well as spoken texts, both formal and informal (2,5M tokens). These samplings pertain to national and regional varieties of Portuguese European Portuguese and also Portuguese spoken in Brazil, in the countries where Portuguese is the official language (Angola, Cape Verde, Guinea-Bissau, Mozambique, São Tomé and Principe, East-Timor), and in Macao and Goa. From a chronological point of view, our corpus contains texts from the second half of the XIX century up until 2008, albeit mostly after 1970 (Bacelar do Nascimento et al., 2000; Bacelar do Nascimento, 2000). Therefore, the CRPC is very-well suited for comparative studies.

The compilation of the CRPC started in 1988 and its main goals are to keep an up-to-date and balanced version of the corpus, disseminate information related to it and make it available on-line so that the resource is friendly and easily accessible. The CRPC is a resource and knowledge database made of authentic linguistic documents, organized in an electronic format accessible to researchers, teachers, and translators and to all, national and foreign, working on the Portuguese language to whom there is a need for reliable linguistic data. These specific linguistic resources constitute an essential prerequisite for a large number of research projects and several types of development and applications.

Two examples of contrastive studies based on comparable corpora partially extracted from CRPC are the results obtained under the scope of the projects VARPORT-Contrastive Analysis of Portuguese Varieties[2], and African Varieties of Portuguese[3], both on the analysis of geograph-

[1] http://www.clul.ul.pt/english/sectores/linguistica_de_corpus/projecto_crpc.php

[2] VARPORT is a joint project of CLUL and UFRJ-Rio de Janeiro, Brazil http://www.letras.ufrj.br/varport

[3] http://www.clul.ul.pt/english/sectores/



*Proceedings of the 3rd Workshop on Building and Using Comparable Corpora, LREC 2010*, pages 65–71
Malta, 22 May 2010

ical varieties of Portuguese and, in the case of VARPORT, combining a diachronic approach (M. F. Bacelar do Nascimento, 2008; S. F. Brandão and Mota, 2003).
Once the corpus collected, our methodology for segmenting and processing the corpus follows widely accepted principles and its recent update is largely inspired by (Wynne, 2005). The written corpus, which is the focus of this paper, contains 368k files, a large number of them extracted from potentially noisy web sources (html, asp, sgml, php).

### 1.2. CRPC Meta-data

The richness of the meta-data included in the CRPC allows us to select a subset of documents suitable for our comparative study. Here we describe the main meta-data in some detail to give insights in the variety of information available and how the CRPC can be tailored to our needs. Each document in the CRPC is first classified according to a broad categorization distinguishing written from spoken materials, to which a specific set of meta-data applies. Written texts are classified in terms of analytic meta-data regarding source, text type (book, review, newspaper, parliamentary discourses, etc.) and topic. For each major type a particular combination of text-descriptive features is assigned: for example, the set of descriptive meta-data for newspapers includes information on the sections, while for didactic books it covers the course name and the curricular year. Other general descriptive meta-data address a set of bibliographic information like title, editor, country of edition, date of edition and the author's name. Since the corpus covers different time periods and national varieties of Portuguese, a set of descriptive meta-data give detailed information on the year and country of birth of the author, as well as on his first language and on the country whose variety he represents (for example, some authors born in Portugal and whose first acquired variety might be European Portuguese are in fact living in Mozambique and their works are to be classified as pertaining to the Mozambique variety in the corpus). Other descriptive meta-data focus on the file properties: its name, size in tokens, location in the corpus directories. And finally editorial meta-data describe the status of the file in terms of its correction and normalization (e.g, there are two levels of correction for texts that are digitalized: corrected and revised). The meta-data are stored in an Excel database and have recently been revised regarding the main fields. The meta-data will soon migrate to a MySQL database.

### 1.3. Cleaning the corpus

The CRPC has been cleaned using the publicly available software *NCLEANER* (Evert, 2008). In the first step, *NCLEANER* removes HTML tags and produces segments, essentially paragraphs. To remove segments which contains unwanted texts (boilerplate, announcements, spam, etc), the second step requires a language model that can be produced by training the system on a relatively small number of annotated documents (with *good* and *bad* segments). We have trained *NCLEANER* on 200 documents selected randomly in the CRPC. The segments produced by the first *NCLEANER* step were annotated as being either *good* or *bad*. This resulted in 4986 *good* segments and 1474 *bad* segments, that we used to train *NCLEANER* with no text normalization (-m 0) to preserve accented characters. The language model created was evaluated using ten-fold cross validation on all 6460 annotated segments, obtaining a F-score of 90%. This language model was used to clean the entire CRPC, which shrank from 433 to 301 millions token. This cleaned corpus was used for our diachronic study. The cleaned corpus has been POS annotated with Treetagger(Schmid, 1994).

## 2. Experiments: diachronic variation of Portuguese around the revolution

In this section we present the experiment aiming at discovering how political and social turmoil initiated by the revolution in 1974 in Portugal changed the discourse in parliamentary sessions of the Portuguese national assembly. The general idea is to compare sub-corpora representing parliamentary discourses in four consecutive decades around the revolution in Portugal with one reference corpora (RC). In what follows we first describe the sub-corpora of the CRPC used, then the approach adopted and finally the results.

### 2.1. The sub-corpora

The CRPC corpus includes parliamentary speeches from the 19th and the 20th centuries. To examine changes that occurred in the parliamentary sessions at the time of the revolution, we have limited ourselves to a period of 40 consecutive years, spamming from 1954 to 1994. In order to make a pre/post revolution comparison, the 40 years were divided into 4 ten year periods with an approximately equal number of tokens: 1954-63 (PER1), 1964-74 (PER2), 1974-84 (PER3) and 1985-94 (PER4). The 1974 split was made on April 25th, when the dictatorship ended. Each of the four 10-year periods was made of a random selection of parliamentary speeches from the CRPC pertaining to that period. A reference corpus (RC), built from a random selection of files pertaining to the written CRPC that originates from Portugal, serves as a basis for the comparisons. It must be said that because the CRPC itself has more documents after than before the revolution, the RC also shares this characteristic, which means that even though it does not affect the interpretation of relative values, absolute values should be interpreted with caution. Table 1 gives more detailed information about the corpora, where there appears to be no significant distribution discrepancy between the reference corpus (RC) and the four sub-corpora.

### 2.2. The approach

Table 1 already provides some useful information to compare the periods. However, providing a more exhaustive comparison requires an analysis of the statistics of words and multi-word expressions (MWs in short), as in (Belica, 1996). Statistics about words are rather straightforward to collect, while MWs and statistics about them are more challenging to acquire. We investigated two methods for MW extraction. We did not lemmatize the texts, so we performed extraction directly on lexis. Our first approach is based on the one presented in (Baroni and Bernardini,

`linguistica_de_corpus/projecto_variedades_africanas.php`

| Info | RC | PER1 | PER2 | PER3 | PER4 |
|---|---|---|---|---|---|
| Nb doc. | 10k | 6k | 6k | 7k | 8k |
| Types | 116k | 70k | 73k | 61k | 58k |
| Tokens | 5768k | 3643k | 3698k | 3589k | 3552k |
| V | 16% | 14% | 14% | 17% | 17% |
| ADV | 5% | 5% | 5% | 5% | 5% |
| NOM | 32% | 31% | 30% | 33% | 34% |
| ADJ | 8% | 10% | 11% | 7% | 7% |

Table 1: Lexical density

2004). The advantage of this approach is that it can generate automatically a list of MWs with little supervision. This approach first uses the log odds ratio measure to compare token and document frequencies between a target corpus (here one of the four periods) and a reference corpus (here RC) producing a list of candidate unigrams. Then a list of connectors is collected from the reference corpus by looking for words and bigrams that frequently occur between the candidate unigrams (e.g. *de*, *a*, *para o*) which are subtracted from the list of stop-words for Portuguese. Starting with bigrams, a procedure is applied recursively to find MWs that must satisfy a number of linguistic (they contain at least one candidate unigram but no stop words and no connectors at the edges or adjacent to each other) and statistical (they satisfy a threshold of frequency and cannot be part of a longer term with frequency close to their own) constraints. Applying this method to each of our four periods provided us with a list of MWs of variable quality[4], although a vast majority of them has a positive log-odds ratio (salience). As this automated list did not provide enough information to form a solid basis for a comparative evaluation of the four periods, we turned our efforts towards the extraction of all n-grams ($n < 6$) from the texts of the periods, the only constraint being that no stop-words should appear at the edge of the n-grams. We also use the log odds ratio as a statistical measure of salience or prominence of a MW, so the salience for each expression was then computed and sorted from the highest to the lowest. Those lists were then inspected by humans top-down, so that potentially more informative expressions were examined first. The BootCat (Baroni and Bernardini, 2004) tool was used and adapted for our needs in both approaches. Other approaches generating lists of keywords are possible, for example (Smith, 1996).

### 2.3. Results and Analysis

#### 2.3.1. N-grams sorted by salience

The results of the n-grams lists sorted according to salience provided most information for identifying significant word forms or MWs in the different periods. For example, the adjective *ultramarina* 'overseas' which qualified territories ruled by Portugal but which were located outside its European frontiers, as well as services or institutions in those territories, show relatively high salience in corpora 1 and 2 (values 3.9 and 3.8, respectively). The same is true for MWs related to Portugal's colonies, which gained their independence after the revolution of 1974, like *territórios ultramarinos* 'overseas territories' (3.6 and 3.3) and, indirectly, *missão civilizadora* 'civilizing mission' (5.8 and 3.3) and for MWs related to the concept of 'corporatism', defended by the regime before 1974, like *nosso corporativismo* 'our corporativism' (5.9 and 4.3), *enquadramento corporativo* 'corporatist frame' (5.5 and 2.5). Other MWs are, on the contrary, highly salient in corpus 3, from 1974 to 1984, like *democraticamente eleitos* 'democratically elected' (5.2), related to the new democratic state, and *prédios nacionalizados* 'nationalized buildings' (2.8), evoking the high number of nationalization of industries and holdings after the revolution.

#### 2.3.2. Diachronic contrast in collocational profile

After identifying a unit as significant based on salience, in many cases a more detailed analysis showed a different collocational profile (Sinclair, 1991) of the lemma for each period. This collocational profile can be analysed in certain cases as related to semantic prosody, i.e., the notion that words associate with collocates that belong to a specific semantic set and that particular collocations receive specific attitudinal semantics. We will discuss three cases, *comunista* 'communist', *democracia* 'democracy' and its derived forms, and the adverb *publicamente* 'publicly'. The collocates of *comunista* revealed quite opposite perspectives regarding this ideology. In corpora 1 and 2 we find *bloco comunista* 'communist bloc', *China comunista* 'communist China' and *propaganda communist* 'communist propaganda', while in corpora 3 and 4 we encounter *Juventude comunista* 'communist Youth', *Partido Comunista* 'Communist Party', *comunistas portugueses* 'Portuguese communists', *nós comunistas* 'we communists', among many others. The collocational profile in the first two periods reflects an ideology which is alien to the Portuguese state at that point in time and contrasts deeply with the high level of involvement of the later collocates. A similar contrast is found with the noun *democracia* and the adjective *democrático* 'democratic'. The two occur in exactly two MWs in corpora 1 and 2: *chamadas democracias* 'so called democracies' (salience 3.3), *totalitarismo democrático* 'democratic totalitarianism' (salience 1.01), but are highly frequent in MWs in corpora 3 and 4, and, as the following selection shows, with a quite different semantic prosody: *regras democráticas* 'democratic rules', *ética democrática* 'democratic ethic', *governo democrático* 'democratic government', *sociedades democráticas* 'democratic societies', *escola democrática* 'democratic school', *jovem democracia* 'young democracy', *democracia avançada* 'advanced democracy', *democraticamente legitimados* 'democratically legitimized'. Finally, in corpora 3 and 4, the adverb *publicamente* 'publicly' occurs in the highly salient MW *aqui publicamente* 'here publicly' (which is part of a larger expression start-

[4]One problem is the overly significant number of proper names. In fact, (Baroni and Bernardini, 2004) reports a precision of 73% with a recall of 68% for English, and a precision of 32% with a recall of 5% for Italian. However, the quality of the texts on which extraction was performed was somewhat lesser than that of the CRPC, not the least because the English and Italian texts were all harvested from the web.

ing with a declarative verb: *declaro/afirmo aqui publicamente* 'I declare/affirm here publicly') and is very productive in collocations of the last two periods: *denunciar publicamente* 'to denounce publicly', *anunciar publicamente* 'to announce publicly', *assumidas publicamente* 'assumed publicly', etc. But there are no well formed MW with this adverb in the first two corpora. The diachronic contrast that we observed in the collocations behaviour answers the need for more diachronic studies of semantic prosody: "A diachronic approach, on the other hand, could try to establish how the meaning of the unit changes over the years or centuries, or it could investigate how words bestow meanings upon each other over time within that unit" (Stewart, 2009) and is an interesting subject to further explore in the analysis of the 4 comparable corpora.

### 2.3.3. N-grams sorted by differential

The n-grams lists with statistical measures for salience, produced for the period before and after the revolution, give us interesting results, as we can see from the examples above. However, our search for significant MWs in the n-grams lists has shown that their salience values are not necessarily very high in any of the four periods: for example, *causa nacionalista* 'nationalist cause' (1.01 and 1.12), *nacionalismo* 'nationalism' (-0.1 and -0.7), *colónia* 'colony' (0.0 and -1.7). The analysis of the full data shows that it is important to look not only to the salience in each period, but mainly to the contrast between saliences in periods 1-2 and in periods 3-4. Under this contrastive approach, the word form *colónia*, with low salience for periods 1 and 2, becomes much more prominent because its salience decreases significantly in the last period, when Portuguese colonies had gained their independence, and the same accentuated decrease is true for the MWs *colónia portuguesa* 'Portuguese colony' and *territórios ultramarinos* 'overseas territories' (see Figure 1). Other significant word forms in periods 1 and 2 show moderate salience, but a strongly contrastive behaviour over the four periods, like the noun *corporação* 'corporation' and the feminine adjective *católicas* 'catholics' (see Figure 2).

Instead of looking for significant MWs in one or more periods, a different approach is to produce lists of word forms which do not occur in corpora 1 and 2 and do occur in 3 and 4. This immediately singles out new word forms appearing after the revolution, like *Parlamentar* 'Parliamentary', *Constitucionais* 'constitutionals', *Liberdades* 'liberties', *Esquerda* 'Left', as well as terms designing new political parties (*PSD*, *CDS*, *PCP*, *Socialista*), and terms for new concepts like *computador* 'computer' and *euros* 'euros'. The results obtained showed that the first two periods shared a common lexicon, just as the last two periods, and that the best approach was to contrast the salience in the first two corpora with the last two. This led us to produce new statistics sorted according to the difference between saliences 1-2 and 3-4, which we call differential (so differential = S1+S2-S3-S4). The top of the list highlights MWs with a strong contrast between a high salience in 1-2 and a low one in 3-4, and the bottom of the list shows exactly the opposite. An intuitively significant word form in corpora 1-2, like *indígenas* 'indigenous', is difficult to single out based on its individual salience in each period, yet it receives a high differential of 15.5. We present in Table 2 a sample of the top of the list for unigrams and bigrams and in Table 3 a sample of the bottom of the list for unigrams and bigrams: the first column is the differential, followed by columns for the salience in periods 1, 2, 3 and 4. At the top of the unigrams list are word forms like *metrópole* 'metropolis', *Corporativa* 'Corporative', *ultramar* 'Portuguese colonies', *províncias* 'provinces', very significant terms under the dictatorship, while at the bottom are word forms like *Democracia* 'Democracy', *comunista* 'communist', *parlamentar* 'parliamentary', *quórum* 'quorum', highly representative of the politics after the Revolution. This is certainly the most promising approach for the analysis of the four corpora.

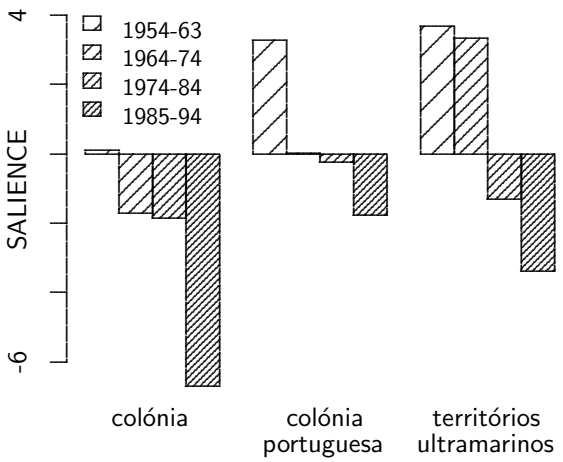



Figure 1: Diachronic behaviour of *colónia*, *colónia portuguesa* and *territórios ultramarinos*

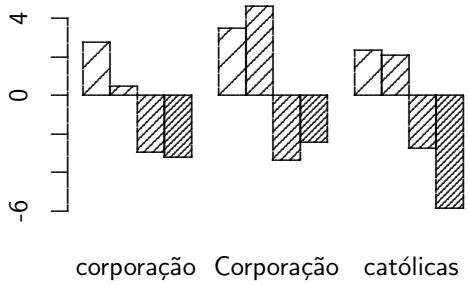



Figure 2: Diachronic behaviour of *corporação* and *católicas*

| Diff. | S1 | S2 | S3 | S4 |
|---|---|---|---|---|
| **unigrams** | | | | |
| *metrópole* 'metropolis' | | | | |
| 24.21 | 4 | 4 | -8 | -9 |
| *Corporativa* 'Corporative' | | | | |
| 24.08 | 4 | 3 | -7 | -10 |
| *ultramar* 'Portuguese colonies' | | | | |
| 22.41 | 4 | 4 | -7 | -7 |
| *províncias* 'provinces' | | | | |
| 22.36 | 4 | 4 | -7 | -9 |
| *Colonização* 'Colonization' | | | | |
| 16.13 | 4 | 2 | -4 | -6 |
| *ultramarinos* 'overseas' | | | | |
| 15.41 | 4 | 4 | -3 | -5 |
| **bigrams** | | | | |
| *Câmara Corporativa* 'Corporate Board' | | | | |
| 23.93 | 4 | 3 | -7 | -10 |
| *Educação Nacional* 'National Education' | | | | |
| 21.66 | 4 | 4 | -6 | -8 |
| *ordem administrativa* 'administrative order' | | | | |
| 18.73 | 5 | 2 | -6 | -5 |
| *Previdência Social* 'Social Security' | | | | |
| 18.44 | 4 | 3 | -6 | -6 |
| *espaço português* 'Portuguese space' | | | | |
| 16.85 | 1 | 5 | -5 | -5 |
| *Fomento Nacional* 'National Development' | | | | |
| 16.81 | 4 | 2 | -6 | -4 |

Table 2: Differential: salience high in 1-2 and low in 3-4

| Diff. | S1 | S2 | S3 | S4 |
|---|---|---|---|---|
| **unigrams** | | | | |
| *Democracia* 'Democracy' | | | | |
| -17.63 | -6 | -8 | 3 | 1 |
| *deputado* 'deputy' | | | | |
| -18.43 | -8 | -8 | 1 | 1 |
| *CEE* 'EEC' | | | | |
| -19.26 | -10 | -9 | 0 | 1 |
| *Comunista* 'Communist' | | | | |
| -21.24 | -9 | -11 | 2 | 0 |
| *abstenções* 'anstentions' | | | | |
| -22.56 | -8 | -11 | 2 | 2 |
| *Parlamentar* 'Parliamentary' | | | | |
| -23.05 | -13 | -8 | 2 | 1 |
| *quórum* 'quorum' | | | | |
| -24.78 | -10 | -9 | 2 | 3 |
| **bigrams** | | | | |
| *sociedade democrática* 'demoocratic society' | | | | |
| -11.34 | -5 | -5 | 2 | 0 |
| *pré escolar* 'preschool' | | | | |
| -12.09 | -5 | -7 | 0 | 1 |
| *salário mínimo* 'minimum wage' | | | | |
| -12.31 | -7 | -5 | 1 | 0 |
| *partidos políticos* 'political parties' | | | | |
| -14.35 | -7 | -5 | 2 | 0 |
| *vamos votar* 'we will vote' | | | | |
| -20.61 | -10 | -7 | 1 | 2 |
| *Partido Comunista* 'Communist Party' | | | | |
| -21.48 | -9 | -11 | 2 | 0 |

Table 3: Differential: salience low in 1-2 and high in 3-4

#### 2.3.4. Other diachronic contrastive patterns

The analysis of the results has mostly showed a strong contrast in lexicon between periods 1-2 and periods 3-4, delimited by the revolution of April 74. Further analysis also reveals other lexical patterns distinguishing period 3 from period 4. The salience of many MWs first decreases from 1 to 2, then increase in 3 and decrease again in 4, as exemplified by *reforma agrária* 'agrarian reform', *Democracia* and *Comunista* in Figure 3. These cases point to an abrupt change in the Parliamentary lexicon related to an equally abrupt political event in 1974, with new parties, ideologies and their application in society, but also to a progressive decrease of radicalism in period 4, when the Portuguese society settled in a stable democracy. In the case of *guerra colonial* 'colonial war', a slightly different pattern appears, with a small increase of salience in 2, followed by a strong increase in 3 and a final decrease in period 4 (see Figure 4). The last observation can be explained by the ongoing wars and the gain of independence. In some very specific cases, there is no contrast between 1-2 and 3-4, but rather between the first three periods and period 4, starting in 1985. An example is the masculine and feminine form of the adjective *Europeu, Europeia* 'european' and the currency *euro* (see Figure 5). This pattern is not related to the pre and post revolution, but instead to the integration of Portugal in the European Community in 1985.

The cases discussed above were either salient in one or two corpora or prominent in terms of differential and could be easily recognized as pertinent because they refer to political realities well known by the Portuguese population. This allowed us to evaluate different approaches towards the identification of lexical units undergoing change. But among the more prominent lexical units we also found words or expressions which seemed to us less obviously representative of any of the periods under study. For example, the high

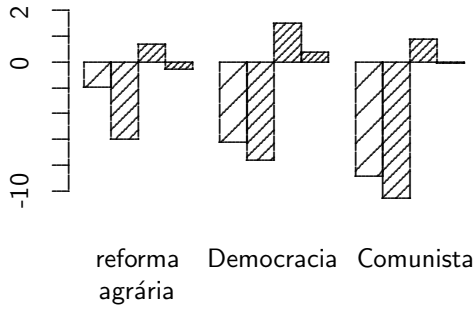



Figure 3: Diachronic behaviour of *reforma agrária*, *Democracia* and *Comunista*

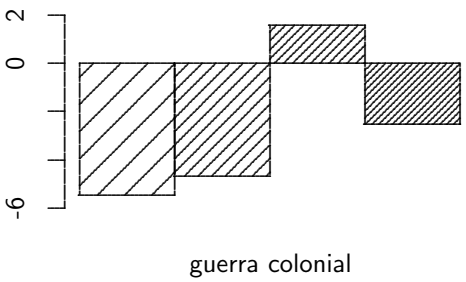



Figure 4: Diachronic behaviour of *guerra colonial*

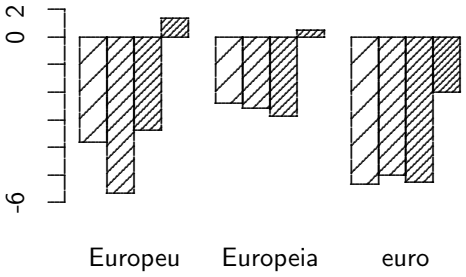



Figure 5: Diachronic behaviour of *Europeu*, *Europeia* and *euro*

number of terms related to agriculture, heavy industries and sports in corpora 1 and 2 requires a collaborative study with experts in recent Portuguese history, sociology and political science. The specialized nature of our corpora requires a terminological approach: these four corpora are not representative of how people talked in the streets or wrote in the newspapers before and after the revolution, but instead are representative of how members of the Parliament expressed themselves in these periods. Statistical information extracted from the corpora is not always in accordance to our intuition as to how a specific word should behave: for example, in the case of *colónia* and *colónia portuguesa* the fact that these MWs, whose usage was disapproved before the revolution (the official term being *província ultramarina*), have the same salience in periods 2 and 3 is unexpected and needs further investigation.

## 3. Conclusion and Future work

We have presented the Portuguese corpus CRPC and the challenges we met in preparing and organising the corpus. After some work on cleaning the CRPC, we explored the diachronic variation of Portuguese during the revolution through careful inspection of an exhaustive list of n-grams. Our main findings are that the most effective method to identify salient lexical units is to compare the four corpora, either by contrasting the lexicon which only occurs in the pre or the post revolution periods, or, and this gives even more interesting results, by using the differential values in corpora 1-2 and 3-4. A follow-up is to contrast the collocates of MWs which are significant in one of the periods. This methodology pointed to lexical units undergoing strong diachronic variation during the periods under study. Several patterns of change were identified: in many cases, the contrast is between the first two corpora and the last two, but the presence of other significant lexical units have shown that there are significant differences in lexical behaviour between the two corpora pre-revolution and even more significant ones between the two periods following the revolution. Future work should look into the optimization of the identification of significant word forms in each period and consider an interdisciplinary approach with social sciences (politics, history, sociology) and law to fully explore the lexicon. Contrasts between the two pre-revolution periods should be fully explored to identify shifts in the dictatorship's ideological and social politics in a time where resistance gained influence. Likewise, a more in depth analysis is required to evaluate lexical changes between period 3, when the revolutionary process gains its full expression, and period 4 when Portugal gradually settles in an established democracy. The methodology produced important and interesting results over the comparable sub-corpora and proved very productive. We plan to apply it to other periods covered by the CRPC.

## 4. References

M. F. Bacelar do Nascimento, L. Pereira, and J. Saramago. 2000. Portuguese Corpora at CLUL. In *Second International Conference on Language Resources and Evaluation (LREC 2000)*, volume II, pages 1603–1607, Athens.

M. F. Bacelar do Nascimento, 2000. *Corpus, Méthodologie et Applications Linguistiques*, chapter Corpus de Référence du Portugais Contemporain, pages 25–30. H. Champion et Presses Universitaires de Perpignan, Paris. Editor: M. Bilger.

M. Baroni and S. Bernardini. 2004. Bootcat: Bootstrapping corpora and terms from the web. In *Proceedings of Language Resources and Evaluation (LREC) 2004*, pages 1313–1316.

C. Belica. 1996. Analysis of temporal changes in corpora. *International Journal of Corpus Linguistics*, 1(1):61–73.

S. Evert. 2008. A lightweight and efficient tool for cleaning web pages. In *6th International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco.

A. Mendes L. A. Pereira M. F. Bacelar do Nascimento, A. Estrela. 2008. On the use of comparable corpora of african varieties of portuguese for linguistic description and teaching/learning applications. In Pierre et al. Zweigenbaum, editor, *Workshop on Building and Using Comparable Corpora, VI Language Resources and Evaluation Conference - LREC2008*, pages 39–46, Marrakech, Morroco, May.

S. F. Brandão and M. A. Mota. 2003. Análise contrastiva de variedades do português: Primeiros estudos. Fólio, Rio de Janeiro.

H. Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Int. Conference on New Methods in Language Processing*, Manchester, UK.

J. Sinclair. 1991. *Corpus Concordance Collocation*. Oxford University Press.

M. Smith. 1996. *WordSmith Tools*. Oxford: Oxford University Press.

D. Stewart. 2009. *Semantic Prosody*. London: Routledge.

M. Wynne. 2005. *Developing Linguistic Corpora: a Guide to Good Practice.* Oxford: Oxbow Books. Available online from http://ahds.ac.uk/linguistic-corpora/ [Accessed 2009-10-26].

# From language to culture and beyond: building and exploring comparable web corpora

**Maristella Gatto**
University of Bari
Via Garruba 6/b – 70122 Bari
E-mail: m.gatto@lingue.uniba.it

**Abstract**

The present paper reports on research and teaching experience based on the creation and exploration of comparable corpora through some of the most interesting tools devised in recent years to make the semi-automated compilation and exploration of web corpora an easier task: Corpus Architect and the Sketch Engine. Section 1 illustrates the compilation of two comparable corpora of medical texts on a specific topic accomplished in the context of classroom activity with a group post-graduate trainee translators and briefly discusses the data retrieved with a specific focus on phraseology. Section 2 introduces the Sketch Engine as a web-based corpus query tool through which a number of recently compiled large general purpose web corpora in several languages can be accessed and explored, and reports on discovery learning classroom activities, carried out with undergraduates, using word 'sketches' within and across languages. Section 3 discusses perspectives for further research arguing that the resources and tools described in the paper might perform very well not only for the rapid extraction of information concerning language use, but also as a source providing fresh insights into discourse and society.

## 1. Building comparable corpora for LSP translation

Despite a widespread conviction that comparable corpora are an invaluable resource for the teaching of LSP translation, in too many cases the task of designing and compiling ad hoc corpora for a specific translation task, especially in the context of translation training, is deemed too time-consuming, and the results are more often than not disappointing. In this context the web has come centre-stage in recent years not only as a corpus in its own right, albeit controversial, to be explored as a corpus 'surrogate' via search engines or meta-search engines (Baroni and Bernardini 2006), but also as the most commonly used resource for the creation of "quick-and-dirty" monolingual and comparable corpora.

With reference to the latter trend, this first section of the paper describes the semi-automated compilation of two comparable corpora of medical texts accomplished in the context of classroom activity with postgraduate trainee translators, and evaluates the information retrieved with specific reference to phraseology.[1]

The tool used to create the two comparable corpora is Corpus Architect. A recent release made available through the Sketch Engine website, Corpus Architect is a system incorporating BootCaT, a tool that, alluding to the well-known metaphor of "bootstrapping", is capable of creating virtually ex-nihilo corpora and term lists from the web in a very short time. In its underlying "philosophy", the tool can be seen, as its predecessor, as the natural development of the widespread practice of building Do-It-Yourself, disposable corpora (Zanettin 2002; Varantola 2003), i.e. corpora created ad hoc from the web for a specific purpose, such as assisting a language professional in some translation task or in the compilation of a terminological database.

It could well be said that the system has a significant bias towards 'customization', in the sense that it is primarily conceived of as a tool helping language professionals build the corpus they need, whenever they need, and as quickly as possible, and as such it is a particularly useful tool to put in the hands of trainee translators. It is perhaps also important to suggest that Corpus Architect has all those characteristics of flexibility, multilinguality, distributed architecture and connection with web-search which have been deemed as characteristic of corpus resources in the 21st century (Wynne 2002). It is therefore more likely than more traditional corpus linguistics resources to accompany the students after University in their real-life profession.

The only thing Corpus Architect needs to start is a number of key words which the linguist considers relevant to the specialized domain for which a corpus is going to be built. The words chosen to start the process are called "seeds" (Baroni – Bernardini 2004) and are transformed by the system into a set of automated queries submitted to an ordinary search engine. The search engine then retrieves and downloads relevant pages, post-processes them, and finally produces a corpus from which a new word list is extracted containing new terms to be used as seeds to build a larger corpus thorugh a cyclical process.

In the present case study the compilation of two comparable corpora (in English and Italian) on "oral squamous cell cancer" started from the four terms "oral", "squamous", "cell", and "cancer", which were used as seeds assuming that each term could to some extent be considered as a key word for this specific domain. The seed terms were used by the system to form the first nucleus of the corpus, from which a wordlist was created and a list of keyword terms was extracted, by comparing the frequency of occurrence of each word in the list with its frequency of occurrence in a reference corpus. The reference corpora to be used for this purpose can be chosen from a number of possibilities, depending on the language used. The keywords extracted can then be turned by the system into new seeds to build a larger corpus via more automated queries.

[1] The activity was carried out with trainee translators within a I Level Master in LSP Translation (Transl.*A.T.E.*) at the University of Bari (Italy), in the A.Y. 2009-2010. It replicates a similar task carried out with WebBootCaT, extensively discussed in Gatto 2009.



*Proceedings of the 3rd Workshop on Building and Using Comparable Corpora, LREC 2010*, pages 72–78
Malta, 22 May 2010

A key feature of Corpus Architect is that, although mainly automated, the process of corpus creation and term list extraction is clearly divided into different phases, allowing the user to interact with the system throughout the process. It is also possible to pre-view web pages that are going to be included in the corpus, and so exclude undesired pages before they are further processed. The latter is a particularly important option in the context of classroom activity because it does not only contribute to enhancing the relevance/reliability of the pages which finally make up the corpus, but also - and more crucially - involves the students in a decision process to some extent comparable to the one underlying the creation and compilation of corpora with more traditional methods. Especially at an early stage in the automated process students are warmly invited to evaluate the texts that are going to be included in the corpus, by considering information inferred from website addresses or even by actually visiting the webpage. In the case of our ORAL CANCER corpus, for instance, many of the pages selected in the first run came from .org or .gov sites, or from portals dedicated to specialized journals such as PubMed (www.ncbi.nlm.nih.gov), with some .com sites leading to web pages devoted to health information. These pages were considered as fairly reliable/relevant, while other pages required further inspection. In the case of dubious or suspect pages, checking for relevance/reliability was anyway very easy because the original page is only one mouse-click away, so the students could have a quick look at it before deciding whether it should be included or excluded from the corpus.

It is not the purpose of the present study to go into further technical detail concerning the pre/post-processing work "going on behind the scenes" of Corpus Architect. It is perhaps useful, though, to consider the decisive importance of the post-processing performed by the system. By filtering out duplicates and near duplicates and by excluding pages which, on the basis of size alone, can be assumed to contain little genuine text (Fletcher 2004), the system does more and better than a student manually can do, while still allowing significant interaction with the machine.

The result of this process is a clean enough text collection which comes to the user in a few minutes. In this case, a corpus of 506,943 tokens was built following an iterative process, in four phases, taking less than 15 minutes in all. This was considered a large enough corpus to allow a rewarding exploration of phraseology in the specific domain under analysis.

On the basis of the same criteria as those followed for the compilation of the English ORAL CANCER corpus, a second corpus composed of comparable Italian texts was created. The Italian CANCRO ORALE corpus is a 434.693 token corpus obtained using the words "cancro", "orale", "cellule" and "squamose" as seed terms. It was compiled in four phases, going through the same steps described for the creation of the English ORAL CANCER corpus. In the process, many similarities between the two corpora emerged, especially concerning key terms automatically extracted by the system. By way of example, Table 1. reports the first content words in each list of keywords (the numbers to the right refer to the number of occurrences):

| | |
|---|---|
| **paziente (1389)** | **cancer (4501)** |
| **tumore (1325)** | **patient (2088)** |
| **trattamento (970)** | **cell (2034)** |
| **cellula (1261)** | **treatment (1539)** |
| farmaco (871) | use (2014) |
| **terapia (631)** | **disease (1016)** |
| **cancro (877)** | oral (1580) |
| **rischio (678)** | surgery (1084) |
| **malattia (557)** | **tissue (794)** |
| **medico (584)** | **tumor (811)** |
| **dolore (1053)** | node (1083) |
| **tessuto (568)** | neck (1079) |
| tipo (521) | lymph (908) |
| **clinico (505)** | **risk (751)** |
| effetto (582) | **study (827)** |
| **studio (624)** | **therapy (797)** |
| venire (881) | **cause (632)** |
| caso (675) | result (611) |
| fattore (345) | blood (799) |
| causa (374) | **medical (595)** |
| carcinoma (488) | **pain (1030)** |
| diagnosi (336) | biopsy (897) |
| aumentare (309) | **clinical (552)** |
| ridurre (267) | body (776) |

Table 1. Sample from the list of keywords extracted from the CANCRO ORALE and ORAL CANCER corpora

As anybody with a knowledge of the two languages could appreciate, over half of the words in one list have their equivalent in the other list (e.g.: paziente/patient; cancer/cancro; treatment/trattamento; cellula/cell and so on). Closer inspection of the complete lists in both languages confirms that they are fairly consistent with each other, even though equivalent words occupy different positions in the two lists, depending on their relative frequency and on grammar differences between the two languages. A comparison between the two lists seemed to suggest, therefore, that the two corpora could well be used as the basis for an exploration of phraseology in the two languages, possibly even leading to the creation of a specific glossary and/or phraseological dictionary. Here is, by way of example, a small sample of information retrieved from the English and Italian corpora for "patient" and its equivalent "paziente". By exploring concordance lines for "patient" in the English ORAL CANCER corpus the students soon noticed an unfamiliar recurring pattern where patients was followed by the past participle of the verb "diagnose", especially in the patterns "patients + (BE) diagnosed with N", where N was almost invariably a disease:

| | | |
|---|---|---|
| study which tested 253 | **patients diagnosed with** | head and neck |
| try, to identify 49,459 | **patients diagnosed with** | metastatic colon |
| About 15% of | **patients diagnosed with** | either oral or |

Turning to the Italian noun "paziente", in the case of

co-occurrence with the verb "diagnosticare" (to diagnose), it was immediately evident that Italian has a different pattern. As the following concordance lines clearly show, in Italian it is the disease (tumori, casi) that is diagnosed (diagnosticato), the basic pattern being "essere/venire diagnosticato", preceded by an indirect object referring to people or preceded/followed by the subject (the disease):

| | | |
|---|---|---|
| A 550 di loro è stato | **diagnosticato** | un carcinoma mamm |
| anni quando le viene | **diagnosticato** | un cancro al polmone |
| a sintomi. Spesso viene | **diagnosticato** | in seguito al riscontro |
| nuovi casi sono stati | **diagnosticati** | nel 1996. L'età media |

While apparently trivial, this example shows how Corpus Architect can make both the compilation and the exploration of two comparable corpora feasible in the limited time-span of classroom activities. The corpora performed extremely well, both as an immediate source of information contributing to a clearer understanding of contrasting lexico-grammar patterns in the specific domain under analysis, and as a source of data to be exploited through long-term activities, such as the compilation of a bilingual glossary or phraseological dictionary.

## 2. Using large comparable corpora in the foreign language classroom

This second case study reports on classroom activities based on the exploration of large web-based comparable corpora with undergraduates[2]. Here again, it is perhaps important to recall that in this context the use of comparable corpora, while in principle advisable (Aston et al. 2004, Zanettin 2009), has been hindered also by a paucity or scarce accessibility of resources. This situation has changed quite recently with the creation of large general purpose corpora created via web crawling, including a nearly 2 billion word corpus of Italian (itWaC) and a 1.5 billion word corpus of English (ukWaC) which were used for the present case study. It is probably worth spending a few words on such corpora, and on the corpus query tool used to explore them, before plunging into the case study.

ItWaC and ukWaC are two very large general-purpose corpora which were compiled between 2005 and 2007 as part of the WaCky Project (Baroni et al. 2009). The ultimate aim when building itWaC, ukWaC and a number of similar general purpose web corpora[3] was to provide a resource that could profit from the immense potential of the web without renouncing high methodological standards for corpus research. It is of course beyond the purpose of the present paper to describe in detail the steps involved in the construction of these corpora, but it is important to remark that the corpora were built by means of automated queries through a process of "bootstrapping" similar, albeit on a larger scale, to the one described in the previous case study. The two corpora are now available through the Wacky Project website as well as through a web-based user interface, including a powerful corpus query tool: the Sketch Engine.

The Sketch Engine is precisely the tool used for the classroom activities proposed in this second case study. It is a corpus query tool specifically designed for offering the linguist "word sketches", i.e. "one-page automatic, corpus-based summaries of a word's grammatical and collocational behaviour" (Kilgarriff 2004 et al.). More specifically, a "word sketch" reports a list of collocates for each grammatical pattern so that, for each collocate, the user can see the corpus contexts in which the node word and its collocate co-occur (Kilgarriff et al. 2004). By way of example, here is a table reporting a small sample of data from the Word Sketch for the lemma "scenery" (see Table 2. below):

| **object_of** | **3846** | **adj_subject_of** | **1241** | **a_modifier** | **11079** |
|---|---|---|---|---|---|
| enjoy | 653 | breathtaking | 72 | spectacular | 1299 |
| admire | 179 | spectacular | 113 | breathtaking | 667 |
| surround | 118 | stunning | 85 | stunning | 1031 |
| boast | 50 | beautiful | 78 | beautiful | 1341 |
| appreciate | 55 | magnificent | 40 | magnificent | 417 |
| savour | 16 | destructible | 9 | coastal | 395 |
| chew | 23 | imaginable | 16 | dramatic | 435 |
| view | 65 | lovely | 37 | rugged | 99 |
| explore | 68 | superb | 32 | wonderful | 308 |
| paint | 34 | amazing | 33 | picturesque | 105 |

Table 2. Word Sketch for "scenery" (sample)

This word sketch was used as the basis of classroom activities with undergraduate students who could quite easily see the words that typically combine with "scenery" in a particular grammatical relation:

- the "object_of" list reports verbs that frequently accompany "scenery", such as "enjoy" or "admire" (e.g. "enjoy the spectacular scenery");
- the "subject_of" column reports in the infinitive verbs that frequently occur in clauses in which scenery is the subject, including passive constructions where scenery is the agent (e.g. "surrounded by spectacular scenery") or participial forms (e.g. "the stunning scenery surrounding the hotel");
- the "adj_subject_of" and "a_modifier" columns report adjectives that frequently accompany scenery in predicative position and in attributive position respectively (e.g.: "the scenery is breathtaking" or "spectacular scenery").

The students found this 'sketch' immediately useful and thought-provoking, indicative as it was – at a glance – of several frequently occurring phraseological patterns revolving around the word "scenery".

Using the Sketch Engine in the classroom proved particular useful also to raise awareness of specific patterns in the students' mother tongue. By way of example we report the results of classroom activities based on the lemma "paesaggio" (meaning both "landscape" and "scenery") from the itWaC corpus of Italian. In this case, the 'sketch' for "paesaggio" called the

[2] The activities reported where carried out with 3rd year students at the Faculty of Foreign Languages and Literatures (University of Bari, Italy), in the A.Y. 2007-2008. For a more detailed analysis of the data see also Gatto 2009.

[3] See the Sketch Engine website for a complete updated list (www.sketchengine.co.uk)

students' attention not only on such predictable phraseological patterns as "paesaggio agrario", "paesaggio incantevole", "paesaggio circostante", "paesaggio urbano", but also on patterns they were less aware of such as "paesaggio da favola" or "paesaggio da cartolina", or even "paesaggi di + N". The students were indeed particularly interested in this latter pattern and their exploration started from "paesaggi di + bellezza":

| | | |
|---|---|---|
| tutto l' anno, e vanta un | **paesaggio** | di straordinaria bellezza. Il |
| corso tra arte e fede in un | **paesaggio** | di grande bellezza. Da |
| li. Però si attraverseranno | **paesaggi** | di rara bellezza naturalistica |
| circa 9 ore ma vedrai dei | **paesaggi** | di una bellezza unica) |
| gione offre una varietà di | **paesaggi** | di una bellezza eccezionale, |
| evoli cime del " Lagorai " e | **paesaggi** | di incomparabile bellezza. La |
| ghi pedemontani sconfinati, | **paesaggi** | di bellezza superlativa, sono i |
| tempo. Il paese gode di un | **paesaggio** | di rara bellezza e suggestione, |
| esto percorso si snoda in un | **paesaggio** | di rara bellezza. Si parte dalle |

By observing the 107 concordances for this collocation they were faced with immediate evidence of a tendency in Italian to resort to the pattern "paesaggi di (una) bellezza + adj" or "paesaggi di (una) + adj + bellezza": all students acknowledged that, although familiar with that pattern, they had not been fully aware of it before the activity, and most of them would have considered the pattern "paesaggi di bellezza + Adj" a probably better equivalent for the English collocation "beautiful scenery", than the plainer translation equivalents "paesaggi belli" "o bei paesaggi".

Besides providing useful information at the level of lexico-grammar and phraseology, the tools performed equally well in teaching contexts whose main focus was not only the exploration of language use but also an insight into culture. This was the case of a brief analysis of sketches for such complex words as "natura" and "nature" obtained from the itWaC and ukWaC corpora carried out in the context of a teaching module aimed at exploring some key words as the starting point for a deeper exploration of the cultural background of tourism discourse.

The extremely high number of occurrences for both "natura" (333722) in itWaC and "nature" (273784) in ukWaC could have hardly been explored without a tool contributing to the extraction of meaningful information. According to data reported by the Sketch Engine, in the itWaC corpus "natura" shows a clear tendency (126605 occurrences) to occur in the pattern Adjective + N, the first modifier in order of statistical significance being "incontaminata", followed by a number of other adjectives connecting "natura" to the legal and economic domain (e.g. "privatistico", "pubblicistico", "giuridico", "patrimoniale", "tributario", "economico, "finanziario", etc.) or to the philosophical domain (e.g. "umano", "divino", "naturata", "naturans",…). Other words taking on again the meanings connected with "incontaminata", and therefore pointing to a more 'concrete' reference to landscape, are "selvaggio" and "lussureggiante".

A less dominant, yet significant, set of collocates preceding the noun "natura" is found in the Verb + N pattern, featuring verbs that cluster around the concept of respect or protection and suggest such phrases as "rispettare… preservare… salvaguardare… la natura". Also worth exploring in the same pattern is a tendency of the word "natura" to co-occur with verbs of telling/understanding in such patterns as "chiarire, rivelare, svelare, capire, conoscere, specificare, comprendere, precisare, scoprire ... la natura". Indeed, contrary to the students' expectations, most verbs preceding the noun "natura" seemed to be pointing to its abstract meaning as a synonymous of "reality" or "characteristic" (Table 3):

| **AofN** | **126605** | **pp_dell'** | **6162** | **preN_V** | **44313** |
|---|---|---|---|---|---|
| incontaminato | 1027 | uomo | 989 | mutare | 414 |
| privatistico | 888 | appalto | 218 | chiarire | 397 |
| pubblicistico | 439 | attività | 677 | amare | 564 |
| giuridico | 3527 | incarico | 149 | avere | 7645 |
| rivisto | 708 | anima | 139 | cambiare | 971 |
| rerum | 241 | attivita' | 132 | humare | 95 |
| umano | 6059 | handicap | 75 | imitare | 133 |
| vario | 3792 | atto | 269 | rivelare | 428 |
| morto | 810 | infermità | 30 | svelare | 176 |
| ordinatorio | 126 | embrione | 54 | capire | 449 |
| divino | 1136 | i.i. | 5 | conoscere | 740 |
| selvaggio | 687 | intervento | 224 | rispettare | 415 |
| provvedimentale | 124 | oggetto | 128 | specificare | 232 |
| regolamentare | 1028 | affare | 73 | comprendere | 721 |
| intrinseco | 476 | invalidità | 28 | riconoscere | 587 |
| rigoglioso | 192 | opera | 142 | sicardare | 12 |
| patrimoniale | 720 | amianto | 19 | alterare | 137 |
| naturans | 35 | assicurazione | 41 | precisare | 215 |
| naturato | 33 | animo | 33 | scoprire | 380 |

Table 3. Word Sketch for "paesaggio" (sample)

As to "nature" in ukWaC, the sketch reported by the Sketch Engine shows that the word tends to occur as object of verbs of thought such as "understand, reflect, explore, examine, reveal, investigate", which seem to point to a level of high abstraction for the meaning of this word, with a behaviour partly comparable to what was observed for the word "natura". The pattern Adjective + N is characterized by the presence of "human", "true", "divine", pointing to the spiritual/philosophical meaning of the word "nature", whereas no instance is reported of adjectives similar to the ones co-occuring with "natura" in Italian, such as "incontaminato" (unspoilt) or "selvaggio" (wild). The only collocates of nature which seem to point to a meaning of the word connected with the idea of landscape/countryside are those in which nature premodifies such words as "reserve, protection, trail, park, tourism", resulting in such patterns as "nature reserve" (apparently the most frequent collocation) or "nature tourism". On the basis of the information gathered, the students came to the conclusion that the word "nature" does not necessarily cover the same semantic areas of its Italian *prima facie* dictionary equivalent, at least as far as its concrete meaning related to the idea of landscape is concerned. This seemed to point to a gap between the behaviour, and hence the meanings, of "natura" and "nature", generally considered as equivalents in Italian

and English. Such differences, which are to some extent to be considered as genre- and domain-specific, have been partly explored by Manca (2002) with reference to tourism discourse. It is this supposed gap, for instance, that accounts for lack of correspondence between typical phraseology in the language of tourism in Italian and in English, as is the case with such phrases as "circondati dalla natura" or "la tranquillità della natura" in which "natura" cannot be translated with "nature" but might be more appropriately translated with an hyponym (e.g. countryside). This gap, which apparently lays bare interesting differences at the level of language use and context of culture, might deserve further exploration for which the huge amount of data made available by such corpora as ukWaC and itWaC, with the help of information provided by the Sketch Engine, might prove extremely appropriate.

## 3. From language to "culture" and beyond: perspectives for further research

The brief overview of the opportunities offered by such recently developed tools as Corpus Architect and the Sketch Engine along with the large web corpora distributed by the Wacky project group can only suggest the scope and variety of classroom and research activity which can be carried out by building and exploring comparable web corpora through tools capable of speeding up the process of corpus compilation and of summarizing data in a way that is meaningful from the linguist's point of view. By reducing the time spent in the compilation and extraction of information, more time is left for the interpretation of the results, even in the limited span of classroom activities.

The encouraging results obtained in traditional language and translation teaching activities as those reported in the first two sections of this paper lead therefore to the hypothesis that the same resources and tools could be profitably used not only for the rapid extraction of information concerning language use, but also in research and teaching contexts beyond the foreign language and translation classroom.

The background for this research can be found in a growing body of work in which the corpus linguistics approach has been profitably used to attain a deeper understanding of discourse and society, starting from Stubbs's (2001) study of such words as "heritage", "racial", "tribal", to more recent research by Mahlberg (2007) and Pearce (2008), to mention a few contributions in this field. Also important it is to mention the increasing interest in studies which advocate an integration between Discourse Analysis and Corpus Linguistics (e.g. Baker 2006). However, the main focus here is not simply to support Corpus Linguistics as a research domain that is increasingly meeting the challenge of tackling with issues relating to discourse and society, but rather to explore the potential, in this respect, of specific resources and tools.

It is with this in mind that a preliminary study of sketches for the word "culture" from ukWaC was attempted in the context of research and teaching activities in the field of cultural studies (Gatto, forthcoming). The choice for the word "culture" as a case study was indeed provoked by a famous statement by the 'father' of Cultural Studies, Raymond Williams, who argued that culture is "one of the two or three most complicated words in the English language" (Williams 1976).

Culture, Williams reminds us, in all its early uses was the noun of a process: the tending of something, basically crops or animals. This provided a basis for the important next stage when the tending of natural growth was metaphorically extended to a process of human development so that the word culture came to be taken in absolute terms as signifying a process of refinement.

We are not concerned here with the results of research by Williams and his followers, who consistently engaged in discussing the evolution of "culture", and other key words, contributing to the rise of what is now known as the domain of cultural studies. What seems however to be unexpectedly relevant to the present research is the methodology which Williams envisaged. In his introduction to *Culture and Society* Williams argues that his enquiry into the development of this word should be carried out by examining "not a series of abstracted problems, but a series of statements by individuals" (Williams 1972). Furthermore, as recently argued by the authors of *New Keywords*, Williams explored "not only the meanings of words, but also the ways people group or "bond" them together, making implicit or explicit connections that help to initiate new ways of seeing their world [...] so that readers might follow and reflect on the interactions, discontinuities and uncertainties of association that shaped what Williams (1976: 13) called 'particular formations of meaning'" (Bennet et al. 2005). It is this exploration of the way people "bond" or "group" words together which corpus linguistics advocates; and it is this exploration that such tools and resources as those described in the present paper make feasible even for such "overused" and complex words as "culture".

As a matter of fact in ukWaC "culture" occurs 161537 times, definitely a number of occurrences which could have never been explored manually. According to data obtained from the Sketch Engine, in this corpus culture shows a tendency to occur as object of such verbs as "foster", "promote", "experience", "create", "change". It is not unlikely that the analysis of concordance lines for such verbs would provide evidence of discourses and discourse practices about culture in contemporary society: which culture is being, or is simply said to be fostered? promoted? experienced? And who are the actors involved in the process?

A look at concordances for "foster" suggests for instance that the pattern in which culture is an object of the verb "foster" are often found in discourse about institutions or workplaces (schools, departments, the NHS, etc.), with a frequent collocation with "enterprise" and related words. Particularly interesting in this pattern seems to be repeated co-occurrence with the pronoun "we" in such phrases as "we foster a culture of…", which suggests that fostering a certain culture, especially in a workplace, a company, an institution, is something which is not only implicitly done but also explicitly stated, as in the following concordance lines:

| | | |
|---|---|---|
| play. We foster a | **culture** | of successful performance |
| We try to foster a | **culture** | of mutual trust |
| We have fostered ties. | **culture** | which promotes a "can |
| We 're fostering a | **culture** | that prioritises |
| staff. We try to foster a | **culture** | of collaboration |

In this case, taking advantage of the possibility (allowed by the system) of moving from the concordance link to the actual web page from which the concordance was taken proved an invaluable opportunity to shift from text to real-life discourse. There are obvious limitations to this useful shift form concordance to web text, due to the well-known dynamism of the web, which often results in broken links. In most cases, however, at least in this preliminary survey, the links visited were still active, so that the overall communicative environment to which the concordance lines actually belonged could be explored and trigger further considerations. By moving from concordance line to website in the specific case of the co-occurrence of "culture" with the pattern "we foster", our attention was called to the widespread practice of including a "culture and values" or "our culture" link in the home page of many corporation websites. Thus, rather than as mere 'fragments' of evidence of language in use, concordance lines were used as gateways to explore what could be termed with Foucault and critical discourse analysts as the level discursive and social practice (Foucault 1972, Fairclough 1992).

Also deserving particular attention seems to be the huge number of modifiers accompanying the noun "culture", both adjectives and noun modifiers. These seem to provide the most evident proof that the unqualified use of the term "culture" as synonymous with refinement is apparently giving way to countless cultures which need further specification. Especially in the list of noun modifiers, significantly opened by "youth" and "pop" culture, one is struck by the presence of such collocations as "celebrity culture", "hip-hop culture", "gang culture" and even "DIY culture". Although certainly related to the nature of a corpus exclusively made of web texts, this datum nonetheless supports the idea that there is an increasing tendency of culture to be fragmented into myriads of subcultures.

By contrast, the core meaning of "culture" as process/product of refinement and education, both at individual and at national level, is apparently reflected in the pattern "culture and N" where there is a clear tendency of "culture" to co-occur with such words as "history", "language", "art" or "tradition", in such typical patterns as "language and culture", "culture and history", "culture and society".

While not yet completed, the preliminary survey of the data seem thus to confirm the appropriateness of the tools and resources under analysis as a way to provide food for thought in different teaching contexts and for different research aims. And it goes without saying that the possibility, for Italian students, of exploring similar patterns from among the 432397 occurrences for "cultura" in the itWaC corpus, would result in a deeper appreciation of contemporary discourse about culture in a cross-cultural perspective. By comparing, for instance, data from the Adjective + N column form the two corpora, one notices interesting similarities, such as the dominance in both data sets of such adjectives as "popular, western, contemporary, diverse, different, dominant, Jewish" and the equivalent "occidentale, popolare, diversa, contemporanea, dominante, ebraica, differente", all falling, although in different order, within the first 15 positions in the list:

| **a_modifier** | **63673** | | **AofN** | **150204** |
|---|---|---|---|---|
| **popular** | 3327 | | **occidentale** | 3229 |
| **Western** | 881 | | **popolare** | 2688 |
| **contemporary** | 1265 | | **diverso** | 8343 |
| organisational | 755 | | umanistico | 1171 |
| visual | 905 | | musicale | 2099 |
| American | 1201 | | scientifico | 3639 |
| **diverse** | 669 | | **contemporaneo** | 1454 |
| **western** | 499 | | italiano | 7591 |
| indigenous | 371 | | normativo | 1747 |
| **different** | 3580 | | **dominante** | 1023 |
| Japanese | 462 | | **ebraico** | 1034 |
| **dominant** | 366 | | greco | 978 |
| Chinese | 499 | | materiale | 1207 |
| ancient | 553 | | **differente** | 1104 |
| **Jewish** | 450 | | politico | 4458 |

Table 4 – Adj + N pattern from word sketches for "culture" (ukWaC) and "cultura" (itWaC)

It is not the similarities as such, but rather the consistency of the data which emerge from the two lists, that supports the hypothesis of their suitability for a cross-cultural exploration of contemporary notions of culture. Nonetheless it is necessary to be extremely cautious before drawing conclusions, if any, from investigations like these: a rewarding and sound exploration of such data could only be, perhaps, the result of teamwork in the context of a multidisciplinary approach. What is certain therefore, and this is what this paper advocates, is that the potential of these tools and resources deserves further exploration not only by linguists but also by scholars in the humanities in general.

## Conclusion

By way of, obviously provisional, conclusion it could be only argued that this overview of the opportunities offered by tools and resources like Corpus Architect, the Sketch Engine and large web corpora like ukWaC and itWaC definitely open up new horizons. The tools are flexible and user-friendly, the amount of data provided is large enough to promise a rewarding exploration, while the time devoted to the collection of data is reduced to the minimum. Exploration itself is made easier, especially for large corpora, by tools capable of summarizing data in a way that is meaningful for the linguist, and perhaps not only for the linguist.

As to the limitations, the most important seem to be those related to the nature of corpora whose content is not a priori known but needs to be evaluated ex-post. This suggests that in the analysis of the data it may be useful, from time to time, to go back from the concordance line to the text from which the concordance line was taken. And this is where the specific contribution of the tools and resources under analysis in the present paper appears all the more fundamental. Unlike most traditional corpora, these web corpora – whether large or small - provide the linguist with a dynamic collection of living texts, which

makes the identification of the discourse in which each text was produced a feasible and approachable goal, thus giving corpus exploration a real chance to move from language to culture and beyond.

## References


Aston G. et al. (eds.) (2004). *Corpora and Language Learners.* Amsterdam: Benjamin, 271-300.

Baker P. (2006). *Using Corpora in Discourse Analysis.* London: Continuum

Baroni M. et al. (2009), "Wacky Wide Web. A Collection of Very Large Linguistically Processed Web-Crawled Corpora". *Languages, Resources and Evaluation*, 43(3), 209-226.

Baroni, M., Bernardini S. (2006). *Wacky! Working Papers on the Web as Corpus*. Bologna: Gedit

Baroni, M., Bernardini, S. (2004). "BootCaT: Bootstrapping corpora and terms from the web". In *Proceedings of LREC 2004*. Lisbon: ELDA, 1313-1316

Bennet et al. (eds) (2005). *New Keywords. A Revised Vocabulary in Culture and Society*. London: Blackwell Publishing

Fletcher, W. (2004). "Facilitating the Compilation and Dissemination of Ad-Hoc Web Corpora". In G. Aston et al. (eds.). *Corpora and Language Learners*. Amsterdam: Benjamin, 271-300.

Foucault, M. (1972). *The Archeology of Knowledge*. London: Routledge

Gatto M. (2009) *From body to web. An introduction to the web as corpus*, Roma-Bari: Laterza UniversityPressOn-line

Gatto M., (forthcoming). "Sketches of CULTURE from the Web: A Preliminary Study". *Proceedings of the 24th AIA Conference*, Roma, 1-3 October 2009

Kilgarriff A. et al. (2004). "The Sketch Engine", in *Proceedings Euralex*, Lorient, France, 105-116.

Mahlberg M. (2007). "Lexical items in discourse: identifying local textual functions of *sustainable development*". In Hoey et al., *Text, Discourse, Corpora*. London: Continuum, 191-218

Manca, E. (2004). *Translation by Collocation: The Language of Tourism in English and Italian*. Birmingham: Tuscan Word Centre

Pearce M. (2008) "Investigating the Collocational Behaviour of MAN and WOMAN in the BNC using Sketch Engine". *Corpora*, 3,1, 1-29

Stubbs M. (2001). *Words and Phrases*. London: Blackwell Publishing

Varantola, K. (2003). "Translators and disposable corpora". In *Corpora in translator education*. Manchester: St Jerome, 55-70

Williams R. (1972). *Culture and Society*. Oxford: Oxford University Press

Williams R. (1976). *Keywords. A Vocabulary of Culture and Society*. Oxford: Oxford University Press

Wynne, M. (2002). *The Language Resource Archive of the 21th century*, Oxford Text Archive,

Zanettin, F. (2009). "Corpus-based translation activities for Language Learners". *The Interpreter and Translator Trainer*, 3 (2), 209-224

Zanettin, F. (2002). "DIY Corpora: the WWW and the Translator". In Maia B. et al. (eds), *Training the Language Services Provider for the New Millennium*. Porto, 239-248.

Zanettin, F., Bernardini, S. and Stewart, D. (eds) (2003*). Corpora in Translator Education*. Manchester: St. Jerome