

# Building and Using Comparable Corpora

## Workshop Programme

- 09:00–10:00 **Session Opening: (9:00-10:00) Invited talk**  
*Crowdsourcing Translation*  
Chris Callison-Burch
- 10:00–10:30 **Session B: (10:00-12:30) Building corpora**  
*Construction of a French-LSF corpus*  
Michael Filhol and Xavier Tannier
- 10:30–11:00 *Merging Comparable Data Sources for the Discrimination of Similar Languages: The DSL Corpus Collection*  
Marcos Zampieri, Nikola Ljubesic and Jorg Tiedemann
- 10:30–11:00 **Coffee break**
- 11:30–12:00 *Building comparable corpora from social networks*  
Marwa Trabelsi, Malek Hajjem and Chiraz Latiri
- 12:00–12:30 *Twitter as a Comparable Corpus to build Multilingual Affective Lexicons*  
Amel Fraisse and Patrick Paroubek
- 12:30–14:00 **Lunch break**
- 14:00–14:30 **Session MT: (14:00-16:00) Machine Translation**  
*Comparability of Corpora in Human and Machine Translation*  
Haïthem Afli, Loïc Barrault and Holger Schwenk
- 14:30–15:00 *Extended Translation Memories for Multilingual Document Authoring*  
Jean-Luc Meunier and Marc Dymetman
- 15:00–15:30 *Using partly multilingual patents to support research on multilingual IR by building translation memories and MT systems*  
Lingxiao Wang, Christian Boitet and Mathieu Mangeot
- 15:30–16:00 *Comparability of Corpora in Human and Machine Translation*  
Ekaterina Lapshinova-Koltunski and Santanu Pal
- 16:00–16:30 **Coffee break**
- 16:30–17:00 **Session T: (16:30-17:30) Terminology**  
*Identifying Japanese-Chinese Bilingual Synonymous Technical Terms from Patent Families*  
Zi Long, Lijuan Dong, Takehito Utsuro, Tomoharu Mitsuhashi and Mikio Yamamoto
- 17:00–17:30 *Revisiting comparable corpora in connected space*  
Pierre Zweigenbaum
- Session P: (17:30-18:00) Panel on a shared task**

**Editors & Workshop Organising Committee:**

Pierre Zweigenbaum, LIMSI, CNRS, Orsay, France (Chair)  
Serge Sharoff, University of Leeds, UK  
Reinhard Rapp, Universities of Mainz, Germany, and Aix-Marseille, France  
Ahmet Aker, University of Sheffield, UK  
Stephan Vogel, QCRI, Qatar

**Workshop Programme Committee:**

Ahmet Aker (University of Sheffield, UK)  
Srinivas Bangalore (AT&T Labs, US)  
Caroline Barrière (CRIM, Montréal, Canada)  
Chris Biemann (TU Darmstadt, Germany)  
Hervé Déjean (Xerox Research Centre Europe, Grenoble, France)  
Kurt Eberle (Lingenio, Heidelberg, Germany)  
Andreas Eisele (European Commission, Luxembourg)  
Éric Gaussier (Université Joseph Fourier, Grenoble, France)  
Gregory Grefenstette (Exalead, Paris, France)  
Silvia Hansen-Schirra (University of Mainz, Germany)  
Hitoshi Isahara (Toyohashi University of Technology)  
Kyo Kageura (University of Tokyo, Japan)  
Adam Kilgarriff (Lexical Computing Ltd, UK)  
Natalie Kübler (Université Paris Diderot, France)  
Philippe Langlais (Université de Montréal, Canada)  
Emmanuel Morin (Université de Nantes, France)  
Dragos Stefan Munteanu (Language Weaver, Inc., US)  
Lene Offersgaard (University of Copenhagen, Denmark)  
Ted Pedersen (University of Minnesota, Duluth, US)  
Reinhard Rapp (Université Aix-Marseille, France)  
Sujith Ravi (Google, US)  
Serge Sharoff (University of Leeds, UK)  
Michel Simard (National Research Council Canada)  
Richard Sproat (OGI School of Science & Technology, US)  
Tim Van de Cruys (IRIT-CNRS, Toulouse, France)  
Stephan Vogel, QCRI (Qatar)  
Guillaume Wisniewski (Université Paris Sud & LIMSI-CNRS, Orsay, France)  
Pierre Zweigenbaum (LIMSI-CNRS, France)

**Invited Speaker:**

Chris Callison-Burch, University of Pennsylvania, US

# Contents

<b>1</b>	<b>Crowdsourcing Translation</b> Chris Callison-Burch	<b>1</b>
<b>2</b>	<b>Construction of a French-LSF corpus</b> Michael Filhol and Xavier Tannier	<b>2</b>
<b>3</b>	<b>Merging Comparable Data Sources for the Discrimination of Similar Languages: The DSL Corpus Collection</b> Marcos Zampieri, Nikola Ljubesic and Jorg Tiedemann	<b>6</b>
<b>4</b>	<b>Building comparable corpora from social networks</b> Marwa Trabelsi, Malek Hajjem and Chiraz Latiri	<b>11</b>
<b>5</b>	<b>Twitter as a Comparable Corpus to build Multilingual Affective Lexicons</b> Amel Fraisse and Patrick Paroubek	<b>17</b>
<b>6</b>	<b>Multimodal Comparable Corpora for Machine Translation</b> Haithem Afli, Loïc Barrault and Holger Schwenk	<b>22</b>
<b>7</b>	<b>Extended Translation Memories for Multilingual Document Authoring</b> Jean-Luc Meunier and Marc Dymetman	<b>28</b>
<b>8</b>	<b>Using partly multilingual patents to support research on multilingual IR by building translation memories and MT systems</b> Lingxiao Wang, Christian Boitet and Mathieu Mangeot	<b>38</b>
<b>9</b>	<b>Comparability of Corpora in Human and Machine Translation</b> Ekaterina Lapshinova-Koltunski and Santanu Pal	<b>42</b>
<b>10</b>	<b>Identifying Japanese-Chinese Bilingual Synonymous Technical Terms from Patent Families</b> Zi Long, Lijuan Dong, Takehito Utsuro, Tomoharu Mitsuhashi and Mikio Yamamoto	<b>49</b>
<b>11</b>	<b>Revisiting comparable corpora in connected space</b> Pierre Zweigenbaum	<b>55</b>

# Author Index

Afi, Haithem, 22

Barrault, Loïc, 22  
Boitet, Christian, 38

Callison-Burch, Chris, 1

Dong, Lijuan, 49  
Dymetman, Marc, 28

Filhol, Michael, 2  
Fraisse, Amel, 17

Hajjem, Malek, 11

Jean-Meunier Luc, 28

Lapshinova-Koltunski, Ekaterina, 42  
Latiri, Chiraz, 11  
Ljubescic, Nikola, 6  
Long, Zi, 49

Mangeot, Mathieu, 38  
Mitsubishi, Tomoharu, 49

Pal, Santanu, 42  
Paroubek, Patrick, 17

Schwenk, Holger, 22

Tannier, Xavier, 2  
Tiedemann, Jorg, 6  
Trabelsi, Marwa, 11

Utsuro, Takehito, 49

Wang, Lingxiao, 38

Yamamoto, Mikio, 49

Zampieri, Marcos, 6  
Zweigenbaum, Pierre, 55



## **Introduction to BUCC 2014**

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 six previous editions of the workshop which took place in Africa (LREC’08 in Marrakech), America (ACL’11 in Portland), Asia (ACL-IJCNLP’09 in Singapore), Europe (LREC’10 in Malta and ACL’13 in Sofia) and also on the border between Asia and Europe (LREC’12 in Istanbul), the workshop this year is co-located with LREC’14 in the middle of the Atlantic in Reykjavík, Iceland. The main theme for the current edition is “Comparable Corpora and Machine Translation”. This topic reminds of the very origin of research in comparable corpora, which stemmed from the scarcity of parallel resources for Machine Translation (and also for Term Alignment).

We would like to thank all people who in one way or another helped in making this workshop once again a success. Our special thanks go to Chris Callison-Burch for accepting to give the invited presentation, to the members of the program committee who did an excellent job in reviewing the submitted papers under strict time constraints, and to the LREC’14 workshop chairs and organizers. Last but not least we would like to thank our authors and the participants of the workshop.

Pierre Zweigenbaum, Serge Sharoff, Reinhard Rapp, Ahmet Aker, Stephan Vogel

# Crowdsourcing Translation

**Chris Callison-Burch**  
University of Pennsylvania  
ccb@cis.upenn.edu

Modern approaches to machine translation are data-driven. Statistical translation models are trained using parallel text, which consist of sentences in one language paired with their translation into another language. One advantage of statistical translation models is that they are language independent, meaning that they can be applied to any language that we have training data for. Unfortunately, most of the world's languages do not have sufficient amounts of training data to achieve reasonable translation quality.

In this talk, I will detail my experiments using Amazon Mechanical Turk to create crowd-sourced translations for "low resource" languages that we do not have training data for. I will discuss the following topics:

- Quality control: Can non-expert translators produce translations approaching the level of professional translators?
- Cost: How much do crowdsourced translations cost compared to professional translations?
- Impact of quality on training: When training a statistical model, what is the appropriate trade-off between small amounts of high quality data v. larger amounts of lower quality data?
- Languages: Which low resource languages is it possible to translate on Mechanical Turk? What volumes of data can we collect, and how fast?
- Implications: What implications does this have for national defense, disaster response, computational linguistics research, and companies like Google?

## Bio

Chris Callison-Burch is an assistant professor in the Computer and Information Science Department at the University of Pennsylvania. Before joining Penn, he was a research faculty member for 6 years at the Center for Language and Speech Processing at Johns Hopkins University. He was the Chair of the Executive Board of the North American chapter of the Association for Computational Linguistics (NAACL) from 2011-2013, and he has served on the editorial boards of the journals Transactions of the ACL (TACL) and Computational Linguistics. He has more than 80 publications, which have been cited more than 5000 times. He is a Sloan Research Fellow, and he has received faculty research awards from Google, Microsoft and Facebook in addition to funding from DARPA and the NSF.

# Construction of a French–LSF corpus

Michael Filhol, Xavier Tannier

LIMSI-CNRS

B.P. 133, 91403 Orsay cedex, France

michael.filhol@limsi.fr, xavier.tannier@limsi.fr

## Abstract

In this article, we present the first academic comparable corpus involving written French and French Sign Language. After explaining our initial motivation to build a parallel set of such data, especially in the context of our work on Sign Language modelling and our prospect of machine translation into Sign Language, we present the main problems posed when mixing language channels and modalities (oral, written, signed), discussing the translation-vs-interpretation narrative in particular. We describe the process followed to guarantee feature coverage and exploitable results despite a serious cost limitation, the data being collected from professional translations. We conclude with a few uses and prospects of the corpus.

**Keywords:** Sign Language; text–video parallelism; elicitation

## 1. Motivation for a French–LSF corpus

Sign languages are part of the less-resourced languages of the world, which means that very little data is available, and indeed linguistic knowledge all together remains limited. The Sign linguistics field has reached no agreement comparable to the more or less stable theories describing a language like French or English. Significant matters such as where and how—and even whether—to draw a line between the language construction layers, e.g. lexicon and syntax (which though not definitely do more obviously appear in written languages), remain open questions.

As for any such language, one can hardly hope to find sufficient data on a specific language feature without building an elicited corpus beforehand to serve the study. For French Sign Language (LSF), a few accessible corpora are available (LS-COLIN, 2000; Matthes et al., 2010; Balvet et al., 2010), but the community is still strongly confronted to the data limitation. Moreover, in our context of automatic or assisted translation, we felt we required not only Sign Language data, but language data for both French and LSF, in view of comparing linguistic features and structures between the two languages.

The DEGELS corpus (Braffort and Boutora, 2012) would be a little closer to our needs than SL-only data, as it involves two languages, namely spoken French and LSF. It is a comparable audio-visual corpus built for a comparative study of gestures in vocal and signed languages in face-to-face communication. To our knowledge, the only bilingual data available including written French is the feed of written news items selected and reduced from the AFP newswire, published daily on WebSourd’s<sup>1</sup> website together with their equivalent version in LSF (cf. fig. 1). The signed version is translated, signed and recorded by professional French-to-LSF translators.

However, the WebSourd data is intended for short-term online viewing, not for academic research. Besides the data collection problem requiring that we save the few videos



Figure 1: WebSourd’s website with the daily list of news items

daily with no control on the contents, the videos come in a lossy Flash encoding format, which is a problem when analysing finer details such as the direction of the eye gaze. A better geometric and time resolution would be a requirement for any thorough study on such feature. Also, it is important for corpora to enclose relevant meta-data, for example on the informants’ connection with the language to enable regional variation awareness, Sign Language (SL) still not being well documented in that respect. These were enough reasons to motivate us to build a reference corpus joining written French and LSF, for academic research and sharing.

## 2. Problems with text–SL parallelism

The oral (live production) nature and the oral-only status (no written form) of SL together have significant consequences on the way one can address translation.

First, when working with text for both source and target languages, the translator is enabled to produce a first wording of the source meaning, and work from it iteratively. Alternatively interpreting both texts, he can modify the target translation until its distance in meaning and effect to the source is satisfactorily low. This convergence process is

<sup>1</sup>A company providing accessibility services to the deaf public. <http://www.websourd.org>

to us what defines translation. It contrasts with captioning and interpretation, whether live or consecutive, which only allow one shot for output delivery. The “one shot” criterion we define here brings a contrast to the common use of the terms, where *translation* is written and *interpretation* is oral. Of course, a reason is that such distinction is not applicable to SL as no written form exists for the language, but we also think that the two activities are different in nature, and that both are possible in SL: translation if the output can be reworked on (refilmed for example); captioning and interpretation otherwise.

Because of corpus shortage, some projects have made use of interpreter’s data as parallel data for translation research (Forster et al., 2012). The problem with such use is precisely that interpreted recordings are one-shot deliveries, i.e. not reviewed, not corrected. While interpretation services remain crucial and the best solution for accessibility and to enable cross-language discussions, their one-shot property makes them subject to undetected mishearings and source language bias. For example, simultaneous interpretation will at least have to follow the sentence-level chunking of the source, which is not necessarily appropriate in the target language. To avoid this bias, we make the claim that building a French–LSF parallel corpus must allow the to-ing and fro-ing between the text and the signed output, i.e. in our terms requires a translation process.

But a second problem exists when translating to the oral modality: the result eventually needs to be memorised and delivered by heart. Regardless of how prepared the output is, video capture of the translation requires that the signer performs it live, from the beginning of the message to its end. Use of a white board with personal notes behind the camera or allowing segmented production are possible tools to cope with somewhat longer texts and avoid omissions, but this is essentially a problem to which no real solution is known yet.

For now, we choose to translate texts that are short enough to remain within the limits of memorised productions, thus clear of hesitations and not requiring post video edition. In this way, our corpus tends to be a fully parallel corpus. But, as we have just seen and because of unavoidable memorising, perfect parallelism is arguably unreachable. Moreover, the community of researchers interested in corpus parallelism usually include chunk, sentence or lexical alignment, which does not apply well here. In this sense, our corpus is not a fully parallel one. This classification problem already emerged in an earlier paper where Segouat and Brafport (2009) attempted to categorise existing SL corpora. For these reasons, we prefer to situate our corpus somewhere between a comparable and a parallel set.

### 3. Preparing for the corpus

WebSourd textual documents are short summaries of AFP newswire articles. They contain one or two sentences for an average of 39 words. They normally describe the five ‘W’s of the reported event: *what*, *when*, *where*, *who* and, as much as possible, *why*. For example:

- (1) “Abidjan, la capitale économique ivoirienne, était à nouveau paralysée mercredi, pour le troisième jour

consécutif, par des jeunes partisans du président Laurent Gbagbo qui tiennent de nombreux barrages dans la plupart des quartiers, rendant la circulation quasiment impossible.” (*Abidjan, the economic capital of Ivory Coast, was again paralysed on Wednesday for the third consecutive day, by young supporters of president Laurent Gbagbo, barricading most of the town districts and almost blocking the traffic.*)

News items were judged the ideal genre for our purpose, for different reasons:

- the domain is not restricted, the news reporting about events in virtually all contexts;
- the language is standard (no grammatical errors), and meant to be concise (no bloat or repetition) and unambiguous;
- productions involve times, places, protagonists and events, with clear relationships between them, which typically triggers heavy use of signing space, a SL specificity requiring scientific attention;
- our lab had worked with the AFP newswire feed in different projects, so we could benefit from local expertise and systems.

Our goal being to provide a corpus of reference translations, we have used the professional service of native deaf translators whose SL performance is acknowledged by the community. Professional service being costly (and currently about 10 times more by the word into SL than into a written language), it is important to select the source material and control redundancy in a way that limits noise but not linguistic use cases. A point was made to work on real-life text excerpts to avoid any fake language intrusion in the source. Hence, we decided to select a set of 40 articles among the textual news archive from WebSourd, and for cross-informant comparison, have each one signed by 3 different informants (translators). The way we chose the texts is one of the main points of our contribution, and presented in the remainder of this section.

First, we restrained the domains of linguistic features to appear, to give us a chance of building a model of a language subset. Otherwise, we would barely have collected a list of positive examples with too few of each feature to enable generalisation. However, to avoid all texts to look alike and lead our informants to guess too much of what is being analysed because of a too narrow focus, we chose four elements of focus, related to events and temporality. This choice was partly due to the fact that we already had expertise on time expressions and events from prior work in text analysis (Moriceau and Tannier, 2014; Arnulphy et al., 2012), which gave us background on the related theoretical aspects as well. Also, results on the expression of time in SL had been published<sup>2</sup> and showed a relevant space mapping of time anchors on all spatial axes (vertical, sagittal and horizontal left-to-right), dictated by certain semantic criteria.

The four studied features, non mutually exclusive in a single article, are the following:

<sup>2</sup>many referenced by Fusellier-Souza (2005)

- Event precedence: one happening before, just before or after another or a date;

(2) “Un éleveur français de 62 ans, Christophe Beck, enlevé il y a un peu plus d’un an au Venezuela et dont le nom était depuis tombé dans l’oubli, a regagné la France dimanche, cinq jours après sa libération contre rançon par ses ravisseurs.” (*Christophe Beck, a French 62-year-old breeder kidnapped in Venezuela just over a year ago [...], came back to France on Sunday, five days after being released for a ransom.*)

- Durations of events or of periods separating/preceding/following events;

(3) “Un homme de 25 ans qui voulait braquer un bureau de Poste à Limay (Yvelines) a retenu jeudi pendant trois heures cinq personnes en otages avant d’être tué par la police.” (*A 25-year-old man who was trying to hold up a post office [...] took five people hostage during three hours, before the police finally killed him.*)

- Causal relationships between events;

(4) “Au moins 525 personnes ont été tuées en Indonésie par un tsunami causé lundi par un séisme sous-marin, selon un nouveau bilan annoncé mercredi par le gouvernement.” (*At least 525 people have been killed in Indonesia by a tsunami caused Monday by an underwater earthquake, according to [...]*)

- Repeated—or repetition of—events.

(5) “Le lancement de la navette Atlantis est prévu ce mercredi à 16h29 GMT de Cap Canaveral (Floride), après avoir été reporté trois fois depuis le 27 août à cause d’orages puis de la tempête tropicale Ernesto.” (*The launch of space shuttle Atlantis is expected Wednesday [...], after being cancelled three times since August 27 because of [...]*)

All these relations were marked as true only if made explicit. For instance, causal relations are difficult to distinguish from simple precedence of events in the case of expressions such as “suite à” (*following*) or “à la suite de” (*in the wake of*), which often require pragmatic or expert knowledge. Too strong an ambiguity would impair the comparison of the three collected translations, as different productions could either be imputed to different ways of expressing the same relation or to different understandings of the relation by our informants. As our resources are limited, we cannot afford this ambiguity.

To guide our selection, we listed a number of semantic criteria to discriminate items in each category, for example:

- whether or not the date of each event is made explicit in the text;
- whether or not the duration between two related events is made explicit in the text (*three days after...*);

- for a repeated event, whether the number of repetitions is made explicit (*three times*) or not (*again*).

The idea behind these criteria is that we will elicit more different structures in SL if we cover more qualitative semantic distinctions. Though the language still remains little documented in that respect, its iconic power in concision undoubtedly makes its underlying system sensitive to semantic variations more than merely to syntax.

Finally, a set of 40 texts was chosen to balance all the criterion values and create a sample as representative as possible. Table 1 provides more information concerning the distribution of studied phenomena in the corpus.

#### 4. Filming and editing

We then proceeded to the actual corpus capture, arranging studio sessions to collect the translations of the same text material by three different translators, totalling 120 signed clips and an hour of signing. To enable movement analyses on all three axes, we filmed the informants from both facial and side views.

Along the same lines as the choice of real-life texts and to avoid any undesired discomfort which might inflict on their language, we did everything to place the translators in their usual set-up. They discovered the news in the morning, and would be left to work on it until they signed it in their own studio. The only notable difference was the additional side-view camera, which they were aware of, and the requirement to clap their hands, arms extended with horizontal palms, before every translation for video synchronisation purposes.

Once synchronised, the two camera views were rendered side by side in a single video file, as shown in the still picture of figure 2. The resulting corpus is a set of 40 news texts in French and 120 LSF video files totalling 1 hour of elicited signed production, where each text is translated by 3 informants. The whole contents will be made available during the year, together with their respective text equivalents.



Figure 2: Still shot of a corpus video file

#### 5. Uses and prospects

The first usage of a corpus, especially in the case of almost undocumented languages, is to be searched for patterns which may lead to the establishment of rules, together forming a grammar. The video part of this corpus has actually just served to formalise a rule system for past event

Date of the event	
Date is made explicit	Date is not made explicit
30	44

Gap between two events	
Gap is known and precise	Gap is fuzzy or unknown
4	5

Repeated events		
First occurrence of a repeated event	Other occurrences	
	Number is known	Number is unknown
4	5	5

Table 1: Number of occurrences of each criterion in the corpus. The total is higher than 40 because several events can be described and several phenomena can occur in the same document.

chronologies (dates, precedence and durations) in LSF (Filhol et al., 2013).

As for its bilingual property, an immediate prospect of this corpus lies in machine translation research, a domain on which several efforts have been summarised in Morrissey (2008). From the statistical point of view of course, data of this size will not be helpful enough to fully train a translation or a language model. Furthermore, statistical learning will normally need pre-aligned bitexts, whereas the video nature of the translated part (unsegmented and continuous stream of pixels) and the non-sequential syntax (simultaneity) of Sign Language together make this difficult. Thus even big enough such type of corpus may not serve the approach.

However, this corpus can be very useful for text-to-SL machine translation evaluation, whether based on statistical learning or on linguistic rules. Translations not being unique, we must rule out a simple comparison between the corpus data and the system’s output, but such corpus can serve as a validation by positive comparison of similar output. Also, the fact that we have three productions for every text can help elaborate new metrics with a philosophy similar to BLEU, a typical score measure of statistical text-to-text translation systems based on edit distances to a set of human reference translations (Papineni et al., 2002).

Future work is required to address longer texts in bilingual corpora involving a Sign Language, especially when a parallel status is desired. We propose to work the other way around and build a corpus from signed production as input translated into text. This would allow the iterative process of translation to rather apply on the text, and indeed guarantee that no bias from the text is carried into the sign discourse, by design.

## 6. References

- B. Arnulphy, X. Tannier, and A. Vilnat. 2012. Automatically Generated Noun Lexicons for Event Extraction. In *Proceedings of the 13th International Conference on Intelligent Text Processing and Computational Linguistics (CicLing 2012)*, volume 2, pages 219–231, New Delhi, India, March.
- A. Balvet, C. Courtin, D. Boutet, C. Cuxac, I. Fusellier-Souza, B. Garcia, M.-T. LHuillier, and M.-A. Sallandre. 2010. The creagest project: a digitized and annotated corpus for french signlanguage (lsf) and natural gestural languages. In *Proceedings of the International Language Resources and Evaluation Conference (LREC)*, Malta.
- A. Braffort and L. Boutora. 2012. Degels1: A comparable corpus of french sign language and co-speech gestures. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uur Doan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC’12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).
- M. Filhol, M. N. Hadjadj, and B. Testu. 2013. A rule triggering system for automatic text-to-sign-translation. In *Sign Language translation and avatar technology (SLTAT)*, Chicago, IL, USA.
- J. Forster, C. Schmidt, T. Hoyoux, O. Koller, U. Zelle, J. Piater, and H. Ney. 2012. Rwth-phoenix-weather: A large vocabulary sign language recognition and translation corpus. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*, Istanbul, Turkey.
- I. Fusellier-Souza. 2005. L’expression de la temporalité en langue des signes française. *La nouvelle revue AIS*, 31.
- LS-COLIN. 2000. <http://www.irit.fr/lc-colin>. Project website (final report available).
- S. Matthes, T. Hanke, J. Storz, E. Efthimiou, A.-L. Dimou, P. Karioris, A. Braffort, A. Choisier, J. Pelhate, and É. Sáfár. 2010. Elicitation tasks and materials designed for dictasign’s multi-lingual corpus. In *Proceedings of the 4th LREC Workshop on Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, Malta.
- V. Moriceau and X. Tannier. 2014. French Resources for Extraction and Normalization of Temporal Expressions with HeidelTime. In *Proceedings of the 9th International Language Resources and Evaluation (LREC)*, Reykjavik, Iceland, May.
- S. Morrissey. 2008. *Data-driven Machine Translation for Sign Languages PhD Thesis*. Ph.D. thesis, Dublin City University, Dublin, Ireland.
- K. Papineni, S. Roukos, T. Ward, and W. J. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *40th annual meeting of the Association for Computational Linguistics (ACL)*, pages 311–318.
- J. Segouat and A. Braffort. 2009. Toward categorization of sign language corpora. In *Building and Using Comparable Corpora, LREC*, pages 64–67, Singapore.

# Merging Comparable Data Sources for the Discrimination of Similar Languages: The DSL Corpus Collection

Liling Tan<sup>1</sup>, Marcos Zampieri<sup>1</sup>, Nikola Ljubešić<sup>2</sup>, Jörg Tiedemann<sup>3</sup>

Saarland University, Germany<sup>1</sup>

University of Zagreb, Croatia<sup>2</sup>

University of Uppsala, Sweden<sup>3</sup>

liling.tan@uni-saarland.de, marcos.zampieri@uni-saarland.de

nikola.ljubestic@ffzg.hr, jorg.tiedemann@lingfil.uu.se

## Abstract

This paper presents the compilation of the DSL corpus collection created for the DSL (Discriminating Similar Languages) shared task to be held at the VarDial workshop at COLING 2014. The DSL corpus collection were merged from three comparable corpora to provide a suitable dataset for automatic classification to discriminate similar languages and language varieties. Along with the description of the DSL corpus collection we also present results of baseline discrimination experiments reporting performance of up to 87.4% accuracy.

**Keywords:** language identification, language discrimination, comparable corpus, similar languages, language varieties

## 1. Introduction

The interest in building language resources for similar languages, dialects and varieties (*SimDiVa*) has been growing significantly in the past few years. Along with these resources, we have recently seen a substantial growth in studies creating NLP tools to process and analyse *SimDiVa*; for instance, adapting character and word-level models for machine translation between similar languages (Nakov and Tiedemann, 2012), lexicon extraction from comparable corpora for closely related languages (Fišer and Ljubešić, 2011), identification of lexical variation between language varieties (Piersman et al., 2010) and automatically extracting comparable lexical and syntactic differences between language varieties (Anstein, 2013).

Along with recently published studies, the growth of interest in varieties and dialects within the NLP community is evidenced by recent events held at international NLP conferences such as the DIALECTS workshop<sup>1</sup> at the 2011 edition of EMNLP and ‘*Adaptation of Language Resources and Tools for Closely Related Languages and Language Variants*’ held at the latest RANLP2013 in Bulgaria<sup>2</sup>.

In like manner, forthcoming workshops such as LT4CloseLang<sup>3</sup> at EMNLP 2014 and the VarDial workshop at COLING 2014 express the same interest in *SimDiVa*. The VarDial workshop will host the *Discriminating Similar Language* (DSL) shared task which uses the corpus collection that this paper describes.

### 1.1. DSL Shared Task

Within the scope of the DSL shared task and also the VarDial workshop, we do not make a distinction between similar languages, dialects and language varieties and we aim

to discuss them collectively.

From a computational perspective, language processing and tools adaptation for *SimDiVa* is analogous; the task of adapting tools to process similar languages (e.g. Croatian and Serbian) is not unlike adapting tools for dialects/language varieties (e.g. Dutch and Flemish; Brazilian and European Portuguese).

The DSL shared task aims at discriminating similar languages and language varieties. We treated similar languages and varieties as classes and grouped by similarity (see section 2). Similar shared tasks have dealt with language identification or discrimination for a specific language/variety group and generic language identification evaluation. For instance, the DEFT2010 attempted to discriminate the country of origin of French texts (e.g. Belgium, France, Quebec, Switzerland, etc.) (Grouin et al., 2010) and the Multilingual Language Identification (MLI) shared task focusing on general purpose language identification rather than on similar languages or language varieties (Baldwin and Lui, 2010b). The main motivation of the DSL shared task is to provide a non-partisan platform for comparing classification systems using the same dataset.

For the purpose of the shared task we had to collect datasets for training, development and testing. There was no corpus compiled specifically for the purpose of discriminating similar languages or language varieties. However, there were existing corpora that held data for various languages/varieties of interest to the DSL shared task. Short of collecting data to build a new corpus, we collected corpus subsets from various corpora to build the DSL corpus collection.

To ensure that the systems participating in the shared task were actually distinguishing classes (languages or varieties) rather than text types or genres, we opted for comparable journalistic texts as this is the most common text type that has been used for previous studies on similar language discrimination (as evidenced in 1.2.). Beyond the DSL shared task, the DSL corpus collection is a useful resource for fu-

<sup>1</sup><http://www.ofai.at/dialects2011/>

<sup>2</sup><http://c-phil.informatik.uni-hamburg.de/view/Main/RANLPLangVar2013>

<sup>3</sup><http://www.c-phil.uni-hamburg.de/view/Main/LTforCloseLang2014>

ture experiments in language identification/discrimination.

## 1.2. Identifying Similar Languages and Varieties

Distinguishing similar languages is an obstacle in language identification. The DSL shared task aims to fill this gap by providing a dataset for researchers to test their systems in different language groups containing closely related languages or varieties. This aspect of language identification received more attention from the NLP community in the last few years.

One of the first studies to explore this issue is the by Ljubešić et al. (2007). This study proposes a computational model for the identification of Croatian texts in comparison to other closely related South Slavic languages. The study reports 99% recall and precision in three processing stages. One of these processing stages, includes a list of forbidden words, a 'black list', that appear only in Croatian texts. Tiedemann and Ljubešić (2012) improve this method and apply it to Bosnian, Serbian and Croatian texts. The study reports significantly higher performance than the accuracy of general-purpose methods, such as *TextCat* (Cavnan and Trenkle, 1994) and *langid.py* (Lui and Baldwin, 2012). Bosnian, Serbian and Croatian datasets provided are included in the DSL corpus collection as group A.

Another study presents a semi-supervised character-based model to distinguish between Indonesian and Malay (Ranaivo-Malancon, 2006), two closely related languages from the Austronesian family also represented in our dataset. The study uses different features such as the frequency and rank of character trigrams extracted from the most frequent words in each language, lists of exclusive words in each of the classes, and the format of numbers (Malay uses decimal point and Indonesian uses comma). The authors compare the performance obtained by their approach with the one obtained by *TextCat*. From the previously mentioned DEFT 2010 shared task, Mohkov (2010) proposes a classification method based on the MARF framework.

One of the methods proposed to identify language varieties is by Huang and Lee (2008). This study presented a bag-of-words approach to distinguish Chinese texts from the mainland and Taiwan. Authors report results of up to 92% accuracy. Another study is the one presented by Zampieri and Gebre (2012) for Portuguese. In this study, the authors proposed a log-likelihood estimation to identify two varieties of Portuguese (Brazilian and European). Their approach was trained and tested using journalistic texts with accuracy results above 99.5% for character n-grams. The algorithm was later adapted to classify Spanish texts using not only the classical word and character n-grams but also POS and morphology information (Zampieri et al., 2013).

The most recent experiments, to our knowledge, aim to distinguish between Australian, Canadian and British English (Lui and Cook, 2013). This study investigates the performance of classifier across different domains and the results obtained suggest that the characteristics of each variety are consistent across them. Portuguese, Spanish and English are also represented in the DSL dataset with two varieties for each language.

## 2. DSL Corpus Collection

The availability of adequate language resources has been a bottleneck for most language technology applications. Reusing and merging existing resources is not altogether unknown (Pustejovsky et al., 2005; Silvia et al., 2011; Eckle-Kohler and Gurevych, 2012). Since there was no existing resources specifically designed for discriminating similar languages or language varieties, we merged different corpora subsets for the purpose of the DSL shared task. The DSL corpus collection comprises news data from various corpora to emulate the diverse news content across different languages, viz. SETimes Corpus<sup>4</sup> (Ljubešić, 2011; Tyers and Alperen, 2010), HC Corpora (Christensen, 2014) and Leipzig Corpora Collection (Biemann et al., 2007).

### 2.1. Corpora Cleaning

Although the source corpora for the DSL corpora used a standardized Unicode encoding (UTF-8), the web-crawled nature of news texts from Leipzig Corpora Collection and HC Corpora contains various (X)HTML markups (e.g. `&mdash;` and `&rsquo;`) and control-characters (e.g. `U+0091` to `U+009F`), that requires cleaning prior to data usage for the DSL task. The HTMLParser<sup>5</sup> was used to resolve the (X)HTML markups and a python code snippet<sup>6</sup> was used to replace control characters with a null string.

Group	Language/Variety	Code
A	Bosnian	<i>bs</i>
	Croatian	<i>hr</i>
	Serbian	<i>sr</i>
B	Indonesian	<i>id</i>
	Malay	<i>my</i>
C	Czech	<i>cz</i>
	Slovak	<i>sk</i>
D	Brazilian Portuguese	<i>pt-BR</i>
	European Portuguese	<i>pt-PT</i>
E	Argentine Spanish	<i>es-AR</i>
	Castilian Spanish	<i>es-ES</i>
F	British English	<i>en-GB</i>
	American English	<i>en-US</i>

Table 1: Closely Related Language/Language Variety Groups

### 2.2. Size, Format and Representation

For each language/variety, the DSL corpus collection contains 18,000 randomly sampled training sentences, 2,000 development sentences and 1,000 test sentences; each sentence contains at least 20 tokens. We note that our naive notion of "tokens" here refer to orthographic units delimited by white spaces and this is not necessarily scalable to disambiguate language/variety groups that do not overtly mark word boundaries such as Chinese *vs* Cantonese. But for the purpose of the shared task, tokenization at codepoint

<sup>4</sup>published in OPUS (Tiedemann, 2012)

<sup>5</sup>[www.docs.python.org/2/library/htmlparser.html](http://www.docs.python.org/2/library/htmlparser.html)

<sup>6</sup>[www.pastebin.com/1aR11vaR](http://www.pastebin.com/1aR11vaR)



is sufficient because (i) the datasets are of a single encoding and (ii) all languages involved use white spaces in their orthography.

These sentences were randomly selected from the corpora collections for each language/variety, the dataset compiled can be treated as a balanced comparable corpora sample of the news domain. To distinguish between the languages we refer to them by the language code using ISO 639-1 convention<sup>7</sup> and for language varieties, we use a common convention in localization, where the country code is appended to the ISO code, e.g. *en-GB* refers to the British variety of English.

The DSL Corpus Collection are in tab delimited format; the first column presents a sentence in the language/variety, the second column states its group and the last column refers to its language code. Table 1 summarizes the language/variety groups and their respective sources.

### 3. Baseline Discrimination Experiment

Using all 234,000 sentences of the training dataset, we trained the Naive Bayes classification models with character and word ngrams features to discriminate between the datasets. And we report the accuracy of the baseline system on the 13,000 test sentences (1000 from each language/variety).

#### 3.1. Models

We used a lightweight Naive Bayes classification model that was previously described in language identification studies (Baldwin and Lui, 2010a; Zampieri and Gebre, 2012; Tiedemann and Ljubešić, 2012). Naive Bayes is a popular classification model due to its robustness and speed. The language of test document  $D$  is predicted by maximizing the sum of the logarithmic probability of a feature (i.e. word/character ngrams frequency)  $w$  given a language  $l$ :

$$\hat{l}(D) = \underset{l_i \in L}{\operatorname{argmax}} \sum_{j=1}^{|V|} \log P(w_j | l_i) \quad (1)$$

where  $L$  is the set of languages/varieties in each language group,  $N$  is the frequency of the  $j$ th word/character ngram in  $D$  and  $V$  is the set of all word/character ngrams in the training data. We use the `sklearn` implementation of multinomial Naive Bayes in our experiments<sup>8</sup> (Kibriya et al., 2004), which calculates:

$$P(w | l_i) = \frac{\sum_{k=1}^{|\delta|} N_{k,w} + \alpha}{|V| + \sum_{j=1}^{|V|} \sum_{k=1}^{|\delta|} N_{k,w_j}} \quad (2)$$

where  $\delta$  is the set of features from the test document  $D$  and  $\alpha$  is the smoothing factor; setting  $\alpha=1$  results in Laplace smoothing and  $\alpha<1$  for Lidstone smoothing. We used Laplace smoothing for our experiments.

### 3.2. Preliminary Results

The best results obtained in our baseline experiments reported 87.4% accuracy when training on character 5grams features (Table 2) and 87.1% when training on word unigrams features (Table 3).

Character Ngrams	Accuracy
2grams	0.763
3grams	0.837
4grams	0.867
5grams	<b>0.874</b>
6grams	0.873

Table 2: Discrimination Results with Character Ngrams Features

Word Ngrams	Accuracy
unigrams	<b>0.871</b>
bigrams	0.841
trigrams	0.736
uni+bigrams	0.857

Table 3: Discrimination Results with Word Ngrams features

As a sanity check, we selected a subset of the training data (108,000 sentences) and the testing data (6000 sentences) from the first language of each language/variety group (i.e. *bs*, *id*, *cz*, *pt-BR*, *es-AR*, *en-GB*) and ran the same Naive Bayes classification training on the subset and we achieved **99.97%** accuracy (5998 out of 6000 instances) with only character 5grams feature. The contrasting increase in accuracy without the need for discrimination of similar languages reiterates the need for language identification tools to incorporate devices to discriminate similar languages.

	Precision	Recall	F-score
<i>bs</i>	0.908	0.915	0.911
<i>hr</i>	0.957	0.944	0.950
<i>sr</i>	0.947	0.954	0.950
<i>id</i>	0.993	0.994	0.993
<i>my</i>	0.995	0.993	0.994
<i>cz</i>	1.000	1.000	1.000
<i>sk</i>	1.000	1.000	1.000
<i>pt-BR</i>	0.934	0.944	0.939
<i>pt-PT</i>	0.943	0.934	0.938
<i>es-AR</i>	0.927	0.744	0.825
<i>es-ES</i>	0.787	0.941	0.857
<i>en-GB</i>	0.600	0.602	0.601
<i>en-US</i>	0.600	0.598	0.598
<b>Overall</b>	0.889	0.889	0.889

Table 4: Precision, Recall and F-score of best performing system

Table 4 reports the precision, recall and f-score of the the 5-gram classifier for the individual languages/varieties. In the following section, we provide a brief error analysis on the preliminary results from the best performing baseline system.

<sup>7</sup>[http://www.loc.gov/standards/iso639-2/php/English\\_list.php](http://www.loc.gov/standards/iso639-2/php/English_list.php)

<sup>8</sup>[www.scikit-learn.org](http://www.scikit-learn.org)

	<i>bs</i>	<i>hr</i>	<i>sr</i>	<i>id</i>	<i>my</i>	<i>cz</i>	<i>sk</i>	<i>pt-BR</i>	<i>pt-PT</i>	<i>es-AR</i>	<i>es-ES</i>	<i>en-GB</i>	<i>en-US</i>
<i>bs</i>	<b>915</b>	35	50										
<i>hr</i>	53	<b>944</b>	3										
<i>sr</i>	39	7	<b>954</b>										
<i>id</i>				<b>994</b>	5							1	
<i>my</i>				7	<b>993</b>								
<i>cz</i>						<b>1000</b>	-						
<i>sk</i>						-	<b>1000</b>						
<i>pt-BR</i>								<b>944</b>	56				
<i>pt-PT</i>								66	<b>934</b>				
<i>es-AR</i>										<b>744</b>	255		1
<i>es-ES</i>										59	<b>941</b>		
<i>en-GB</i>												<b>602</b>	398
<i>en-US</i>												402	<b>598</b>

Table 5: Confusion Matrix for Character 5grams Naive Bayes Discrimination Classifier on Language varieties

### 3.3. Error Analysis

Table 5 presents the confusion matrix of the error analysis for the character 5gram classifier performance. The table is to be understood as such, when classifying 1000 Bosnian (*bs*) test sentences, the classifier correctly tagged 915 instances (i.e. *true positives*), wrongly tagged 35 and 50 Bosnian sentences as Croatian (*hr*) and Serbian respectively (i.e. *false negatives*) and wrongly tagged 53 Croatian sentences and 39 Serbian as Bosnian (i.e. *false positives*). We provide a brief error analysis to emphasize the need for language discrimination among similar languages and language varieties. From the confusion matrix, the Naive Bayes classifier overfits in discriminating languages from group A and cast Bosnian features on the other two similar languages, thus resulting in high false negatives and false positives.

For group B and C, Table 4 and 5 suggest that languages that are thought to be similar are not so similar after all; Czech and Slovak (group C) though sharing the same alphabet and Slavic roots can be easily classified using the baseline system. Also, for Indonesian and Malaysian (group B), the common orthography and Austronesian origin did not hinder the performance of the baseline system<sup>9</sup>.

Looking at the Portuguese varieties (group D), the baseline classifier performed reasonably well but it still falls behind the state-of-art accuracy (>95%) as reported in classical language identification literature (Cavnar and Trenkle, 1994; Baldwin and Lui, 2010a; Lui and Baldwin, 2012).

From Group E, the Castilian Spanish features overfits and when the classifier tagged Argentine Spanish instances, ~25% of the time, it wrongly tagged them as Castilian Spanish. Group F consisting of British (*en-GB*) and American (*en-US*) English also suffers from classification performance; ~40% of the time the classifier makes mistakes and tags an American test sentence as British and vice versa.

Prior to the DSL shared task, we might consider adding more similar languages to group B and C so as to increase the complexity of DSL task or replace the groups with other groups of similar languages (e.g. Danish and Norwegian

(Bokmål) or Dutch and Flemish).

## 4. Conclusion

In this paper, we described the compilation of the DSL corpus collection for the DSL shared task. This was done through merging subsets of existing comparable corpora. Using the DSL corpus collection, we run a simple Naive Bayes discrimination system at the character and word levels to serve as baseline for the shared task. This method achieved an overall accuracy of 87.4% on the whole dataset. The task of distinguishing similar languages and varieties is by no means trivial and with this preliminary baseline results, we would like to encourage the participation of researchers and developers in the DSL shared task. The DSL corpus collection and shared task are aimed at improving the state-of-art language identification systems by tackling a known bottleneck of this task: discriminating similar languages and varieties.

The compilation of the DSL collection fills an important gap as no equivalent resource focusing on similar languages and varieties was available prior to the compilation of this collection. The resource and baseline system presented in this paper can be used beyond the context of the shared task to improve/evaluate language identification systems as well as for related NLP tasks.

## Acknowledgements

The authors would like to thank the original data source providers for the free and open access to their datasets and also the anonymous reviewers who provided important feedback to increase the quality of this paper.

<sup>9</sup>Note that one Indonesian test sentence was wrongly identified as British English *en-GB* and one Argentine Spanish test sentence was wrongly identified as American English *en-US*

## 5. References

- Stefanie Anstein. 2013. *Computational approaches to the comparison of regional variety corpora : prototyping a semi-automatic system for German*. Ph.D. thesis, University of Stuttgart.
- Timothy Baldwin and Marco Lui. 2010a. Language identification: The long and the short of the matter. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 229–237. Association for Computational Linguistics.
- Timothy Baldwin and Marco Lui. 2010b. Multilingual language identification: Altw 2010 shared task data. In *Proceedings of Australasian Language Technology Association Workshop*, pages 4–7.
- Chris Biemann, Gerhard Heyer, Uwe Quasthoff, and Matthias Richter. 2007. The leipzig corpora collection-monolingual corpora of standard size. *Proceedings of Corpus Linguistic*.
- William Cavnar and John Trenkle. 1994. N-gram-based text categorization. *3rd Symposium on Document Analysis and Information Retrieval (SDAIR-94)*.
- Hans Christensen. 2014. Hc corpora. <http://www.corpora.heliohost.org/>.
- Judith Eckle-Kohler and Iryna Gurevych. 2012. Subcatlmf: Fleshing out a standardized format for subcategorization frame interoperability. In *EACL*, pages 550–560.
- Darja Fišer and Nikola Ljubešić. 2011. Bilingual lexicon extraction from comparable corpora for closely related languages. *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 125–131.
- Cyril Grouin, Dominic Forest, Lyne Da Sylva, Patrick Paroubek, and Pierre Zweigenbaum. 2010. Présentation et résultats du défi fouille de texte deft2010 où et quand un article de presse a-t-il été écrit? *Actes du sixième Défi Fouille de Textes*.
- Chu-ren Huang and Lung-hao Lee. 2008. Contrastive approach towards text source classification based on topbag-of-word similarity. In *Proceedings of PACLIC 2008*, pages 404–410.
- Ashraf M. Kibriya, Eibe Frank, Bernhard Pfahringer, and Geoffrey Holmes. 2004. Multinomial naive Bayes for text categorization revisited. In *Proc 17th Australian Joint Conference on Artificial Intelligence*, Cairns, Australia, pages 488–499. Springer.
- Nikola Ljubešić, Nives Mikelic, and Damir Boras. 2007. Language identification: How to distinguish similar languages? In *Proceedings of the 29th International Conference on Information Technology Interfaces*.
- Nikola Ljubešić. 2011. Setimes corpus. <http://nlp.ffzg.hr/resources/corpora/setimes/>.
- Marco Lui and Tymothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proceedings of the 50th Meeting of the ACL*.
- Marco Lui and Paul Cook. 2013. Classifying english documents by national dialect. In *Proceedings of Australasian Language Technology Workshop*, pages 5–15.
- Sergei Mokhov. 2010. A marf approach to deft2010. In *Proceedings of TALN2010*, Montreal, Canada.
- Preslav Nakov and Jörg Tiedemann. 2012. Combining word-level and character-level models for machine translation between closely-related languages. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 301–305. Association for Computational Linguistics.
- Yves Piersman, Dirk Geeraerts, and Dirk Spelman. 2010. The automatic identification of lexical variation between language varieties. *Natural Language Engineering*, 16:469–491.
- James Pustejovsky, Adam Meyers, Martha Palmer, and Massimo Poesio. 2005. Merging probank, nombank, timebank, penn discourse treebank and coreference. In *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*, pages 5–12. Association for Computational Linguistics.
- Bali Ranaivo-Malancon. 2006. Automatic identification of close languages - case study: Malay and indonesian. *ECTI Transactions on Computer and Information Technology*, 2:126–134.
- Necsulescu Silvia, Núria Bel, Muntsa Padró, Montserrat Marimón, and Eva Revilla. 2011. Towards the automatic merging of language resources.
- Jörg Tiedemann and Nikola Ljubešić. 2012. Efficient discrimination between closely related languages. In *Proceedings of COLING 2012*, pages 2619–2634, Mumbai, India.
- Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).
- Francis M. Tyers and Murat Alperen. 2010. Setimes: a parallel corpus of balkan languages. In *Proceedings of the multiLR workshop at LREC*.
- Marcos Zampieri and Binyam Gebrekidan Gebre. 2012. Automatic identification of language varieties: The case of Portuguese. In *Proceedings of KONVENS2012*, pages 233–237, Vienna, Austria.
- Marcos Zampieri, Binyam Gebrekidan Gebre, and Sascha Diwersy. 2013. N-gram language models and POS distribution for the identification of Spanish varieties. In *Proceedings of TALN2013*, Sable d’Olonne, France.

# Building comparable corpora from social networks

Maroua Trabelsi\*, Malek Hajjem†, Chiraz Latiri\*

LIPAH\*, LISI†

FST Campus Universitaire El Manar Tunis; Tunisia, INSAT Centre Urbain Nord Tunis; Tunisia  
trabelsimarou@live.com, malek.hajjem@gmail.com, chiraz.latiri@gnet.tn

## Abstract

Working with comparable corpora becomes an interesting alternative to rare parallel corpora in different natural language tasks. Therefore many researchers have accentuated the need of large quantities of such corpora and the need to work on their construction. In this paper, we highlight the interest and usefulness of textual data mining in social networks. We propose the extraction of tweets from the microblog Twitter in order to construct a comparable corpus. This work aims to develop a new method for the construction of comparable corpus from twitter that could be used in forthcoming work to construct a bilingual dictionary, using text mining approach.

**Keywords:** Social networks, Text mining, Comparable corpora, Comparability metrics

## 1. Introduction

Social networks are dynamic structures formed by individuals or organizations. They have been developed and diversified on the web allowing large audiences to express their thoughts and reactions throughout multiple platforms such as blogs, micro-blogs, facebook and wikis in various languages. Recently, social networks have even played an instrumental role in popular revolution, social movement and participated to entire governmental policy changes (Eltantawy and Wiest, 2011). As a result, a large multilingual collection of posts became publicly available. This has made text mining in social networks the subject of many recent researches. In this work, we conduct an exploratory study of the construction of a multilingual resource from these new modes of communication. In fact, multilingual corpora are useful in different areas such as multilingual text mining, bilingual lexicons extraction, cross-lingual information retrieval and machine translation. Multilingual corpora are either **parallel** corpora: corpus that contains source text and their translations (McEnery and Xiao, 2007), or **comparable** corpora : collections of documents in the same or in different languages made up of similar texts.

Although parallel corpora are very effective and used, they have several disadvantages: firstly, their language coverage remains insufficient. Besides, parallel texts freely available are few. They are expensive to produce as they need human translation. Then, comparable corpora are the best alternative, because they are less expensive and more productive. It is clearly easier to find document collections with similar topics in multiple languages than to find parallel corpora (Talvensaar et al., 2007). However, it remains to note that researchers like (Morin et al., 2006) and (Li et al., 2011) are more interested by the exploitation of comparable corpora than creating new methods for their automatic construction.

Our work consists in analyzing and exploiting the huge data text from Twitter in order to build a comparable corpus. Our goal is proving the feasibility of the new method for the construction of comparable corpus using tweets. We focus, in this work, on Arabic and French language seeing that

there are few Arabic/other languages pair comparable corpora. For that, we decided to collect French/Arabic tweets about **Arab Spring** posted from May 2013 to September 2013 and to calculate a comparability measure (CM) between collected posts.

In fact, a comparability metric is a key issue in field of building comparable corpora. Its function is to estimate the quality of corpus built on similar topics and different languages. Recent works refer to three ways to calculate comparability measures:

- **Statistical measures:** they are based on the quantity of the common vocabulary. It includes (Li and Gaussier, 2010) who used a translation table and (Su and Babych, 2012) who used a bilingual dictionary, given a comparable corpus  $P$  consisting of a source part  $P_s$ , and a target part  $P_t$ , the degree of comparability of  $P$  is defined as the expectation of finding the translation of any given source/target words in the target/source corpus vocabulary. Regarding (Yapomo et al., 2012), their work described a CLIR- based method to calculate similarity between texts. We cite also (Saad et al., 2013) who have proposed two different comparability measures based on binary and cosine similarity measures. Their work is closer to (Li and Gaussier, 2010). Unlike (Li and Gaussier, 2010), their work was based on the bilingual dictionary Open Multilingual WordNet (OMWN) word alignment.
- **Semantic measures:** they are based on the exploitation of semantic resources to calculate word similarity and still basically used for a monolingual collection (Corley and Mihalcea, 2005). This measure can be adapted to multilingual environment by using resources like global wordNet<sup>1</sup>.
- **Hybrid measures:** they are based on the use of both information from corpora and a semantic resource such as the work of (Mohammad et al., 2007) who presented the idea of estimating semantic distance in one language using a knowledge source in another.

---

<sup>1</sup><http://www.globalwordnet.org>

Concerning our work, we discuss in Section 4, the result of two different statistical comparability measures applied to our collected corpus from twitter, which are based on binary and cosine similarity measures. Our work is close to (Saad et al., 2013) who proposed the same comparability measure for Wikipedia corpus. Moreover, (Saad et al., 2013) used a bilingual dictionary, we propose to use machine translation (MT). In fact, MT seems to be more appropriate with the noisy nature of data processed (twitter data). The rest of the paper is organized as follows. We first present some related work in the next section. Section 3 introduces our proposed approach. In section 4, we discuss different evaluations used in this work. Finally, conclusions and some prospects are stated.

## 2. Related work

### 2.1. Data sources of comparable corpora

Comparable corpora can be obtained easily from multilingual textual contents. Initially comparable corpora were made from newspapers, in this case the corpora does not target a particular area and cover different topics (Fung and McKeown, 1997), (Rapp, 1999). Scientific articles are considered as an interesting source for comparable corpora, because they cover many languages and topics. For example (Déjean and Gaussier, 2002) built a comparable corpus composed of medical records. For their part, (Chiao, 2004) used specialized websites in the medical field (CISMeF<sup>2</sup> for French corpora and CliniWeb<sup>3</sup> for the English corpora) rather than using general search engines.

Comparable corpora can also be acquired from the web, which is considered as large source of data. Among the studies that have used the web, we cite (Issac et al., 2001) which built a corpus based on syntactic and semantic criteria from the web. (Goeriot, 2009) has built comparable corpus for language pairs with great linguistic distance (Japanese/French) based on an automatic classification system.

Other approaches like (Laroche and Langlais, 2010), (Rebout, 2012), (Sellami et al., 2013) and (Saad et al., 2013) work on the online encyclopedia, wikipedia to extract comparable articles. Recently, work such as (Gotti et al., 2013), invested in automatic translation of tweets, they exploit the great potential of tweets published by canadian government agencies and organizations to construct a bilingual tweet feeds used to create a tuning and training material for Statistical Machine Translation. (Jehl et al., 2012) also focused on automated translation of microblogging messages, they provide a bilingual sentence pair data from twitter in English and Arabic about Arab spring for training SMT system.

### 2.2. Construction methods of comparable corpora

Construction methods allow the acquisition and structuring of multilingual data. They depend on the selected data sources :

- **Thematic crawling** or focused crawling is a method adopted for automatic construction of comparable cor-

pora from the web. It consists in using links between pages to collect documents. This method was used by (Talvensaari et al., 2008) to extract English-Spanish-German comparable corpora mined from the web and concentrate on a specific domain. Thematic crawling has as objective to minimize the number of pages which are not related with the area studied. We note that even if the web is a large volume of data, the automatic acquisition of comparable corpora is still a challenging task.

- **Cross-language information retrieval** is a method which is an independent method from the web. It consists in using the translated keywords of a source collection as a query to the target collection. It was operated by (Talvensaari et al., 2007) who have proposed a new approach using CLIR to extract Swedish-English comparable corpus. In this approach, the keywords were extracted using the RATF<sup>4</sup> measure. Their translations are executed as query on the target collection by the Indri<sup>5</sup> information retrieval system. This method may extract pertinent documents from the target collection but it has a disambiguation issue in the choice of the best translation of keywords.
- **Clustering** is defined as the distribution of a set of texts in groups according their similarity and without a priori knowledge. It has been used by (Li et al., 2011) to obtain bilingual clusters from a part of an initial corpus. This part includes texts above a minimum threshold of similarity that will be used to form a comparable corpora. The same procedure is reproduced on the rest of the corpus. This method of construction is simple and organized but it can be slow.

## 3. Proposed approach to construct comparable corpora from Twitter

### 3.1. Textual data collection

In this section, we present our textual data collection extracted from the popular social network Twitter. Twitter is an online social networking and microblogging service that allows users to send and read Twitter messages (tweets), limited to 140 characters. An important role was played by Twitter in the socio-political events, such as the Arabic spring, the theme of our corpus. In fact, since the Arabic revolutions, this media presents itself as a vehicle for the voice of politicians, artists, and especially young people.

This choice of source data was made because of the massive volume of data posted on twitter and available through the Twitter API which allows queries against specific topics. Also, Twitter data can respect criteria of comparability like theme, date proximity and document length.

Tweets about the Arab spring were retrieved using Twitter's Search Api<sup>6</sup> feature which is offered by Twitter to give developers access to tweet data servers. The search API is focused on relevance and not completeness. It usually serves only tweets from the past week.

<sup>2</sup>[www.chu-rouen.fr/cismef](http://www.chu-rouen.fr/cismef)

<sup>3</sup>[www.ohsu.edu/clinweb](http://www.ohsu.edu/clinweb)

<sup>4</sup>Relative Average Term Frequency

<sup>5</sup><http://www.lemurproject.org/indri.php>

<sup>6</sup><http://dev.twitter.com>

This API can filter tweets based on queries. For example, to retrieve tweets that report on the movement Occupy Wall Street, you have just to use the keywords that describe this movement and specify the language/period of this movement. After collecting the data, we specify in the following subsection the various forms of data preprocessing performed on the collected corpus.

### 3.2. Preprocessing of collected corpora

After collecting the data we have employed a number of preprocessing techniques. This phase is a succession of three steps, the result of each step will be used by the next. The three steps are performed on each of the Arabic and French corpus separately.

First, we have eliminated special characters and numbers of each collection to just obtain the textual content of tweets (for example remove the names of users, the punctuation, smileys, etc). Second we have eliminated redundancy by deleting retweets. Retweets is a copy of someone else's tweet broadcasted by a second user to their followers, they do not generally add any new information (McMinn et al., 2013).

The last step is the morpho-syntactic labeling of the tweet corpora. This task associates to each word of the collected corpora a label which recapitulates its morpho-syntactic proprieties in the text. Morpho-syntactic labeling has been made in this step using TreeTagger (Schmid, 1994) for French tweets and MADA (Habash and Rambow, 2005) for Arabic .

### 3.3. Normalisation of tweet corpora

The variety of linguistic phenomena existing in the textual data and the lack of conventions and spelling standards in social networks require a phase of standardization. In fact, building comparable corpora from this media raises a number of challenges.

Indeed, the recovered data could not be used directly. The writing style, used in social networks and microblogs, is sometimes incomprehensible. The users frequently make spelling and grammar mistakes and create short texts that are difficult to analyze. Our normalisation process is focused, in this work, on the French collection. It was based on a spellchecking approach for normalising short text as works that have been conducted on normalising social media in French language were scarce except some attempts like (Fairon et al., 2006), (Yvon, 2008) and (Beaufort et al., 2010). Our implementation involves the following steps:

- First, we have used a short text messages (SMS) dictionary<sup>7</sup> which covers global spelling mistakes used with SMS and their standard lexical forms. In other words, it provide translations from SMS expressions to plain language expressions. This dictionary was used to identify candidate token (OOV) for lexical normalisation. We note that the coverage of SMS dictionary used, was incapable to identify all OOV words in tweet corpora. For the purpose of this work we have employed a personalized dictionary manually built from a training corpus collected through topsy

<sup>8</sup> a tweet search engine. This, personalized dictionary check the OOV words of tweet relative to our theme corpora (for example: manif→ manifestation, mvt→mouvement, jan→janvier).

- Second, our two dictionaries were automatically applied to the corpus, then ill-formed words were transformed to their standard format.

### 3.4. Description of the built corpus

The constructed textual resource is an Arabic/French bilingual corpus consisting of a total of 52000 tweets which were published on Twitter's public message board during May 2013 to September 2013. We collected tweets that contained the keywords respectively in Table 1 and Table 2. The tweets are then subjected to Pre-processing and standardization resulted in a total of **20025** tweets in Arabic and **20023** tweets in French .

Keywords	Translation	number of tweets
Printemps arabe	Arab spring	4003
Révolution arabe	Arabic revolution	110
Syrie	Syria	9110
Egypte	Egypt	4600
Révolution tunisienne	Tunisian revolution	2200

Table 1: Number of French tweets by keywords (after Pre-processing)

Keywords	Transliteration	Number of tweets
الربيع العربي	Alrrabyç Alçrbyy	14285
الثورة العربية	Alθwrĥ Alçrbyyĥ	20
سوريا	swryA	2100
مصر	mSr	2300
الثورة التونسية	Alθwrĥ Altwnsyĥ	1320

Note: The transliteration consists on writing Arabic with latin characters to help non Arabic speakers to read Arabic. In this paper, Arabic orthographic transliteration is presented in the HSB scheme (Habash et al., 2007): (in alphabetical order)

ي و ه ن م ل د ق ف غ ع ظ ط ض ص ش س ز ر ذ د ح ج ث ت ب ا  
A b t θ j H x d d r z s š S D T D ç γ f q k l m n h w y  
and the additional letters: ء، آ، إ، أ، ل، آ، و، ؤ، ى، ة، هـ، حـ، يـ.

Table 2: Number of Arabic tweets by keywords (after Pre-processing)

## 4. Evaluation of the comparability

As we stated, comparability is the key concept in the process of building comparable corpora. However, there has been no widely accepted definition of comparability (Liu and Zhang, 2013). Even if, tweets that talk about the same event in the same period but in different languages were extracted, thus respecting comparability's criteria, we need to evaluate similarity between the Arabic and French data collected from Twitter. During this step, methods based on word frequency have been processed to measure corpus homogeneity between French and Arabic collections. In fact,

<sup>7</sup><http://www.langagesms.com/dictionnaire.html>

<sup>8</sup><http://topsy.com/tweets>

comparability is defined according to an application. As we aim to use our corpora in extracting bilingual lexicons, these methods are the best alternative because they are generally focused on the amount of common vocabulary in the document. So, the comparability measures used in our approach are statistical measures based on CLIR. Two information retrieval models were considered: binary and vector space model (cosine similarity).

#### 4.1. Binary measure of comparability

In binary measure, the source and target (Arabic and French) collections are represented as a bag of words. In this case, the degree of comparability reflects the absence or presence of keywords (or index) translation from the source vocabulary (respectively target) in the target vocabulary (respectively source).

To extract the index of the two collections (Arabic and French), we have used the Lemur<sup>9</sup> information retrieval system. The resulting indexes are translated with an on-line MT system<sup>10</sup>. Finally, we have verified the absence/presence of index terms in each collection, in other words, we have calculated the degree of comparability of our corpus in a binary way as follows.

Given a corpus P with a source language  $L_s$  and a target language  $L_c$ , the binary function  $trans(W_s, d_t)$  returns 1 if the translation of a Word from the source vocabulary  $W_s$  was found in the target vocabulary  $d_t$  and 0 in the other case. Thus, bin-DC for the source and target documents is calculated as follows:

$$binDC(d_s, d_t) = \frac{\sum_{w_s \in d_s} trans(w_s, d_t)}{|d_s|}$$

We note that,  $binDC(d_s, d_t)$  and  $binDC(d_t, d_s)$  are not symmetrical (Saad et al., 2013), our work was based on the following formula for measuring the total comparability of our comparable corpus :

$$\frac{binDC(d_s, d_t) + binDC(d_t, d_s)}{2}$$

For this measure based on Boolean information retrieval model(bin-DC) the comparability degree is between [0-1]: 1 strongly parallel, 0 neither parallel nor comparable.

#### 4.2. Vector measure of comparability

In the vector information retrieval model, a document is represented as a vector in the vector space. Each vector's document is compound indexing terms. The coordinates of a vector represents the weight of each term. The similarity measure is usually the cosine of the angle that separates the two vectors (Boubekeur-Amirouche, 2008). To represent documents in the vector space model (VSM), we have built the source and target vectors with the following method: we extracted indexes with lemur. The resulting index (in source language) was translated with MT and ran against the target collection with the Lemur retrieval system based its cosine similarity as retrieval model which uses the idf

weighting model to convert documents to vectors. For this second measure which is based on vector model (cosine-DC) the similarity measure logically should therefore be between [-1, 1]:-1 totally opposed, 1 exactly the same and 0 independent. As vectors in our case represent the weight of words in tweets. Since weights of words are always positive values, then the cosine measures ranges also from 0 to 1.

#### 4.3. Results and discussion

To illustrate the evolution of the degree of comparability depending on the amount of data retrieved from *Twitter*, we have created from the French and Arabic corpus several sets containing variable data rates between the first 10% of the corpus and the entire corpus. Then we have calculated the comparability between these datasets through the Boolean model of information retrieval in both Arabic  $\rightarrow$  French and French  $\rightarrow$  Arabic .

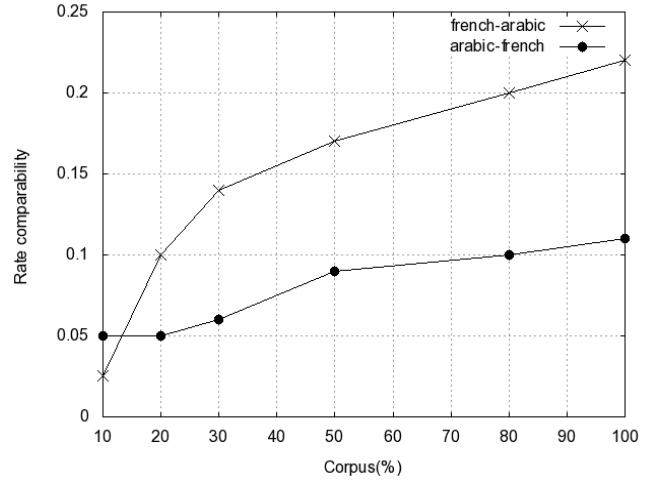


Figure 1: Evolution of comparability with the amount of data

The curve of figure 1 shows that the degree of comparability is proportional to the amount of data in the source and target corpus. As this corpus is exploited automatically, increasing the size of data ensures a lexical coverage. The more vocabulary is used, the more comparability improves.

Measures	bin-DC	cosine-DC
Degree of comparability	0.17	0.22

Table 3: measurement of comparability results

Table 3 summarizes the results of two measures of comparability : bin-DC in the boolean method and cosineDC in the vector method. The results show that the measure of comparability cosine-DC is better than bin-DC. This result was expected since the measure based on vector model includes weighting of terms unlike the Boolean model that uses a binary weighting.

Our experimental results of comparability measures are

<sup>9</sup><http://www.lemurproject.org/>

<sup>10</sup><http://www.bing.com/translator>

promising and show that our corpora has a comparability feature especially if we compare our results with (Saad et al., 2013) who had used articles from Wikipedia which is considered as user content, less noisy than our textual data, and found close results (0.11 for binary measure and 0.18 for vector measure).

## 5. Conclusion and Future Work

Despite the popularity of twitter, we note that few researches have been conducted on the construction of corpora based on tweets. This is due to a number of issues associated with the construction of Twitter corpora, including restrictions on the distribution of the tweets, which prevents us to make our corpus available. In this work, we created an Arabic-French comparable corpus, which is, to the best of our knowledge, the first comparable corpus collected from Twitter. We created the corpus of tweets extracted through the Twitter API based on their topic similarities and close publication dates. Experimental results showed that our calculated comparability measures capture a similarity degree for our comparable corpus. In the future we will improve the normalisation step and we will try to treat a larger tweet corpus. We aim also to improve the comparability evaluation. In closing, building comparable corpus from twitter isn't an end in itself; our goal is to exploit this corpus for bilingual extraction in future works.

## 6. References

- Richard Beaufort, Sophie Roekhaut, Louise-Amélie Cougnon, and Cédric Fairon. 2010. Une approche hybride traduction/correction pour la normalisation des SMS. In *Actes de la 17e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2010)*, Montréal, Canada.
- Fatiha Boubekeur-Amirouche. 2008. *Contribution à la définition de modèles de recherche d'information flexibles basés sur les CP-Nets*. Ph.D. thesis, Université de Toulouse, Université Toulouse III-Paul Sabatier.
- Yun-Chuang Chiao. 2004. *Extraction lexicale bilingue à partir de textes médicaux comparables: application à la recherche d'information translangue*. Ph.D. thesis, Université Pierre et Marie Curie-Paris VI.
- Courtney Corley and Rada Mihalcea. 2005. Measuring the semantic similarity of texts. In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment, EMSEE '05*, pages 13–18, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Hervé Déjean and Eric Gaussier. 2002. Une nouvelle approche à l'extraction de lexiques bilingues à partir de corpus comparables. *Lexicometrica, Alignement lexical dans les corpus multilingues*, pages 1–22.
- Nahed Eltantawy and Julie Wiest. 2011. The arab spring! social media in the egyptian revolution: Reconsidering resource mobilization theory. *International Journal of Communication*, 5(0).
- Cédric Fairon, Jean René, and Sébastien Paumier Klein. 2006. Le Corpus SMS pour la science. Base de données de 30.000 SMS et logiciels de consultation. Presses universitaires de Louvain, Louvain-la-Neuve. Cahiers du Cental, 3.2.
- Pascale Fung and Kathleen McKeown. 1997. Finding terminology translations from non-parallel corpora. In *Proceedings of the 5th Annual Workshop on Very Large Corpora*, pages 192–202.
- Lorraine Goeuriot. 2009. *Découverte et caractérisation des corpus comparables spécialisés*. Ph.D. thesis, Université de Nantes.
- Fabrizio Gotti, Philippe Langlais, and Atefeh Farzindar. 2013. Translating government agencies' tweet feeds: Specificities, problems and (a few) solutions. In *Proceedings of the Workshop on Language Analysis in Social Media*, Atlanta, Georgia, June. Association for Computational Linguistics, Association for Computational Linguistics.
- Nizar Habash and Owen Rambow. 2005. Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 573–580. Association for Computational Linguistics.
- Nizar Habash, Abdelhadi Souidi, and Tim Buckwalter. 2007. On Arabic Transliteration. In A. van den Bosch and A. Souidi, editors, *Arabic Computational Morphology: Knowledge-based and Empirical Methods*. Springer.
- Fabrice Issac, Thierry Hamon, Christophe Fouqueré, Lorne Bouchard, and Louisette Emirkanian. 2001. Extraction informatique de données sur le web. *Revue Distances*, 5(2):195–210.
- Laura Jehl, Felix Hieber, and Stefan Riezler. 2012. Twitter translation using translation-based cross-lingual retrieval. In *Proceedings of the Seventh Workshop on Statistical Machine Translation, WMT '12*, pages 410–421, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Audrey Laroche and Philippe Langlais. 2010. Revisiting context-based projection methods for term-translation spotting in comparable corpora. In Chu-Ren Huang and Dan Jurafsky, editors, *COLING*, pages 617–625. Tsinghua University Press.
- Bo Li and Eric Gaussier. 2010. Improving corpus comparability for bilingual lexicon extraction from comparable corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 644–652. Association for Computational Linguistics.
- Bo Li, Éric Gaussier, Emmanuel Morin, Amir Hazem, et al. 2011. Degré de comparabilité, extraction lexicale bilingue et recherche d'information interlingue. In *18e Conférence sur le Traitement Automatique des Langues Naturelles*, volume 1, pages 211–222.
- Sa Liu and Chengzhi Zhang. 2013. Termhood-based comparability metrics of comparable corpus in special domain. In *Proceedings of the 13th Chinese Conference on Chinese Lexical Semantics, CLSW'12*, pages 134–144, Berlin, Heidelberg. Springer-Verlag.
- A. M. McEnery and R. Z. Xiao, 2007. *Parallel and comparable corpora: What are they up to?* Translating Europe.



- Multilingual Matters. The PDF offprint will be provided when available.
- Andrew J. McMinn, Yashar Moshfeghi, and Joemon M. Jose. 2013. Building a large-scale corpus for evaluating event detection on twitter. In *Proceedings of the 22Nd ACM International Conference on Conference on Information; Knowledge Management, CIKM '13*, pages 409–418, New York, NY, USA. ACM.
- Saif Mohammad, Iryna Gurevych, Graeme Hirst, and Torsten Zesch. 2007. Cross-lingual distributional profiles of concepts for measuring semantic distance.
- Emmanuel Morin, Béatrice Daille, et al. 2006. Comparabilité de corpus et fouille terminologique multilingue. *Traitement Automatique des Langues*, 47(1):113–136.
- Reinhard Rapp. 1999. Automatic identification of word translations from unrelated english and german corpora. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 519–526. Association for Computational Linguistics.
- Lise Rebout. 2012. *L'extraction de phrases en relation de traduction dans Wikipédia*. Mémoire présenté à la Faculté des arts et des sciences. Université de Montréal en vue de l'obtention du grade de Maitre de sciences (M.Sc.) en informatique.
- Motaz Saad, David Langlois, and Kamel Smaïli. 2013. Extracting comparable articles from wikipedia and measuring their comparabilities. In *V International Conference on Corpus Linguistics. University of Alicante, Spain*.
- Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of international conference on new methods in language processing*, volume 12, pages 44–49. Manchester, UK.
- Rahma Sellami, Fatiha Sadat, and Lamia Hadrich Belguith. 2013. Traduction automatique statistique à partir de corpus comparables : application aux couples de langues arabe-français. In *CORIA*, pages 431–440.
- Fangzhong Su and Bogdan Babych. 2012. Measuring comparability of documents in non-parallel corpora for efficient extraction of (semi-)parallel translation equivalents. In *Proceedings of the Joint Workshop on Exploiting Synergies between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation (HyTra)*, pages 10–19, Avignon, France, April. Association for Computational Linguistics.
- Tuomas Talvensaari, Jorma Laurikkala, Kalervo Järvelin, Martti Juhola, and Heikki Keskustalo. 2007. Creating and exploiting a comparable corpus in cross-language information retrieval. *ACM Transactions on Information Systems (TOIS)*, 25(1):4.
- Tuomas Talvensaari, Ari Pirkola, Kalervo Järvelin, Martti Juhola, and Jorma Laurikkala. 2008. Focused web crawling in the acquisition of comparable corpora. *Information Retrieval*, 11(5):427–445.
- Manuela Yapomo, Gloria Corpas, and Ruslan Mitkov. 2012. Clir-and ontology-based approach for bilingual extraction of comparable documents. In *The 5th Workshop on Building and Using Comparable Corpora*, page 121.
- François Yvon. 2008. Réorthographe des sms. LIMSI.

# Twitter as a Comparable Corpus to build Multilingual Affective Lexicons

Amel Fraisse, Patrick Paroubek

LIMSI-CNRS  
Bât 508, Université Paris-Sud XI  
fraisse@limsi.fr, pap@limsi.fr

## Résumé

The main issue of any lexicon-based sentiment analysis system is the lack of affective lexicons. Such lexicons contain lists of words annotated with their affective classes. There exist some number of such resources but only for few languages and often for a small number of affective classes, generally restricted to two classes (*positive* and *negative*). In this paper we propose to use Twitter as a comparable corpus to generate a fine-grained and multilingual affective lexicons. Our approach is based in the co-occurrence between English and target affective words in the same emotional corpus. And it can be applied to any number of target languages. In this paper we describe the building of affective lexicons for seven languages (en, fr, de, it, es, pt, ru).

**Keywords:** Affective Lexicon, Comparable Corpus, Sentiment Analysis

## 1. Introduction

Research in Sentiment Analysis and Opinion Mining, has flourished in the past years. The growing interest in processing emotions and opinions expressed in written text is motivated by the birth and rapid expansion of the Social Web that made it possible for people all over the world to share, comment or consult content on any given topic. In this context, opinions, sentiments and emotions expressed in Social Media texts have been shown to have a high influence on the social and worldwide economic behavior. In spite of the growing body of research in the area in the past years, dealing with affective phenomena in text has proven to be a complex and interdisciplinary problem that remains far from being solved.

As any emergent field, its challenges include the need to develop linguistic resources to perform computational tasks. In our case, we are interested in the sentiment classification task which is performed either with statistical approaches or with lexicon-based approaches. In the two cases, the lack and the scarcity of affective lexicons present a real issue for sentiment analysis system. Multilingual affective lexicons are central components for cross-lingual sentiment analysis systems. Their manual construction is a hard, long and costly process. While often it is impossible to consider for most under-resourced languages because of the scarcity or even lack of experts. Existing affective lexicons are always monolingual and often developed for English. Furthermore, many of these lexicons are very simple, i.e. they consist of a list of words divided into only two classes : *positive* and *negative*. To our knowledge, there is no fine grained affective and multilingual lexicons.

Most previous work addressing the problem of bilingual lexicon extraction are based on parallel corpora. However, despite serious efforts in the compilation of corpora (Armstrong and Thompson, 1995), (Church and Mercer, 1993), to our knowledge, there is no available affective parallel corpus for the field of sentiment analysis.

On the other hand, with the rapidly growing volume of resources on the Web, the acquisition of non-parallel texts is usually much easier. Thus, as mentioned by (Rapp, 1995)

and (Rapp, 1999) it would be desirable to have an approach that can extract lexicons from comparable or even unrelated texts. In this paper, we propose to use Twitter as a comparable corpus to extract multilingual affective corpus. Our approach is motivated by the fact that, nowadays, social media user's and in particular twitter users' express and share their sentiments, opinions and emotions on a variety of topics and discuss current issues over the world. In fact, many people can talk about the same event and describe their emotional state triggered by this event in different languages. Hence, Twitter could be considered as a comparable corpus as we could group tweets (messages written by users) by emotion/opinion/sentiment expressed in different languages. We have tested our approach to build seven affective lexicons for English, French, German, Spanish, Italian, Portuguese and Russian.

## 2. Related Work

There are two ways to cover the lack of sentiment analysis resources. The first way is to create manually a lexicon in a source language as (Bradley and Lang, 1999) who developed the Affective Norms of English Words (ANEW) which is a set of normative emotional ratings for 1034 English words. And then localize the source lexicon into target languages.

(Redondo et al., 2007) have adapted the ANEW into Spanish, (Vo et al., 2009) localized it into German. This approach requires human translators to ensure the quality of the localized resource and therefore is cost expensive and not scalable.

(Strapparava and Valitutti, 2004) developed the WordNet Affect which is a manually created extension of the WordNet, including a subset of synsets suitable to represent affective concepts correlated with affective words. The second approach is automatic construction of a lexicon. The most common method is bootstrapping. This method starts with seed words with a known polarity (e.g. good, happy, wonderful for a positive class, bad, sad, terrible for a negative class). Next, the seed words are used to find related words and assign them the same class or estimate their po-

larity.

(Qadir and Riloff, 2013) present a bootstrapping algorithm to automatically learn English twitter hashtags that convey emotion. (Mohammad, 2012) use the pointwise mutual information to measure the association between a word and a given emotion. So he builds a word emotion association lexicons which are lists of words and associated emotions. For example, the word *victory* may be associated with the emotions of *joy* and *relief*.

(Pak and Paroubek, 2010) proposed to use Twitter to collect a dataset of emotional texts in French. Using the collected dataset, they estimated the affective norms of words present in the corpus and built a polarity classifier. Both for manual and automatic approaches, existing affective lexicons are always monolingual.

### 3. Word-Opinion/Sentiment/Emotion association lexicon

In a previous work (Frasse and Paroubek, 2013), we have presented 20 semantic categories including all types of emotions, sentiments and opinions. Each semantic class correspond to one type of emotion/sentiment/opinion and is referred to by means of a multi-word label that re-groups various subjective words generally associated to one of the various sentiments contained in the considered class (Table 1). For example the *Anger* label includes the *impatience, annoyance, irritation, nervousness, anger, exasperation* semantic categories. For each of the 20 Opinion/Sentiment/Emotion presented in the Table 1, our aim is to build the associated lexicon for each of the seven languages addressed in this paper.

#	Label	Dim.	uComp Semantic Category
1	NEGATIVE SURPRISE	e-	negative surprise / negative amazement
2	DISCOMFORT	e-	discomfort / disturbance / embarrassment / guilt
3	FEAR	e-	shyness / worry / apprehension / alarm fear / terror
4	BOREDOM	e-	boredom
5	DISPLEASURE	e-	displeasure / deception / abuse
6	SADNESS	e-	sadness / resignation / despair / sorrow / hopelessness
7	ANGER	e-	impatience / annoyance / irritation / nervousness / anger / exasperation
8	CONTEMPT	e-	reluctance / contempts / disdain / blame / disgust / hate
9	DISSATISFACTION	s-	disappointment / dissatisfaction / discontent / shame
10	DEVALORIZATION	o-	disinterest / devalorization / depreciation
11	DISAGREEMENT	o-	disapproval / disagreement
12	VALORIZATION	o+	interest / valorization / appreciation
13	AGREEMENT	o+	understanding / approval / agreement
14	SATISFACTION	s+	satisfaction / contentment / pride
15	POSITIVE SURPRISE	e+	positive surprise / positive amazement
16	APPEASEMENT	e+	relief / appeasement / peacefulness forgiveness / thankfulness
17	PLEASURE	e+	pleasure / entertainment / enjoyment / joy / happiness / euphoria / play
18	LOVE	e+	love / affection / care / tenderness / fondness / kindness / attachment / devotion / passion / envy / desire
19	INFORMATION	i	information / announcement / news / demand / query / question
20	INSTRUCTION	i	recommendation / suggestion / instruction / order / command

TABLE 1 – uComp semantic categories of opinion/sentiment/emotion, e=emotion, s=sentiment, o=opinion, i=information, +=positive valence, -=negative valence.

For each label of the Table 1 and for each language, we wish to extract the associated lexicon. Table 2 illustrate an example of comparable tweets in four languages ; the four tweeter talked about the same topic *violence in ukraine* expressing the same emotion *Sadness* in different languages. So, based on such data our approach aims to extract, across different languages, and for each affective label the





	#Ukraine #death toll rises as clashes continue #sad #grief.
	#Ukraine 60 morts aujourd'hui!!! c'est vraiment #triste #chagrin
	Stop the #violence in #Ukraine 60 #tod heute #traurig
	Impresionantes imagenes de #Kiev que pasarian por fotogramas de una pelicula de guerra!! muchos #muertos!! estoy #triste

TABLE 2 – Example of comparable tweets

Affect. Label	Associated Words			
	English	French	German	Spanish
SADNESS	Sad Death Grief	Triste Mort Chagrin	Traurig Tod	Triste Muertos

TABLE 3 – Multilingual affective lexicon associated to the tweets described in the Table 2

associated lexicon (Table 3).

### 4. Our approach for multilingual affective lexicon construction

Hashtags are a distinctive characteristic of tweets (Jackiewicz and Vidak, 2014). They are a community created convention for providing meta-information about a tweet. Hashtags are made by adding the hash symbol # as a prefix to a word. Thus, a hashtag is simply a way for people to search for tweets that have a common topic. In general, the tweeter (one who tweets) use emotion-word hashtag, to notify others of the emotions associated with the message he or she is tweeting. Consider the tweet below :

*Oh okay all the people I fancy are taken ...that's cool watch them be happy as I sit in a corner and cry #sad*

The tweeter has used the emotion word hashtag #sad, to convey that he or she is sad. And as English is considered the reference language on the Web, the tweeter use generally the emotion word hashtag in their native languages and give the corresponding English one as shown in the following French tweet :

*Je suis vraiment #triste aujourd'hui #sad.*

So, our approach is based on the co-occurrence between the English and the target emotion word hashtags in the tweets. To achieve this, we proceed in two steps ; firstly we construct emotional corpora in the following seven languages : English, French, Spanish, German, Italian, Portuguese and Russian. Secondly, We extract affective lexicon

Anger	Fear	Love
#anger	#fear	#love
#rage	#terror	#affection
#irritation	#shyness	#care
#nervousness	#worry	#tenderness
#impatience	#apprehension	#fondness
#annoyance	#terrified	#kindness
#angry	#alarm	#attachment
#edgy	#scare	#devotion
#exasperated	#scared	#passion
#irritated		#envy
#annoyed		#desire

TABLE 4 – Seed Affective word for Anger, Fear and Pleasure affective classes

Affect.Cl.	En	Fr	De	It	Es	Ru	Pt
DISCOMFORT	551	232	65	33	157	10	63
FEAR	1677	156	123	15	488	35	124
DISPLEASURE	1617	645	13	7	74	6	15
SADNESS	283	211	204	209	459	110	272
ANGER	1690	73	9	16	198	102	43
CONTEMPT	506	606	53	43	310	69	68
PLEASURE	2414	1952	1639	1099	2082	664	1198
LOVE	2452	434	595	632	2251	1369	933

TABLE 5 – Number of document per affective class and per language.

from the collected corpora based on the co-occurrence between English and target emotional hashtags in the same affective class.

#### 4.1. Corpora collection

Data collection from the Web usually involves crawling and parsing of HTML pages which is a solvable but at the same time a consuming task. In our case, collecting data from Twitter is much easier since it provides an easy and well-documented API<sup>1</sup> to access its content. In this work, we selected from the Table 1 the 8 prominent affective classes that are frequent in tweets : *Negative surprise, Anger, Sadness, Fear, Displeasure, Boredom, Positive surprise, Pleasure and Love*. For each selected class we have defined a list of English seed emotional words that are commonly used by English speakers to express their affective state on Twitter.

Table 4 presents an extract of English seed emotional words that are used for the three affective classes *Anger, Fear* and *Love*. Then, we supplied the Twitter Search API with the English emotional hashtags queries and collected tweets written in their native languages and containing at least one hashtag of the English list. In fact, we noticed that when a user writes an affective tweet, he or she uses an emotional word hashtag in his or her native language and he or she, also, gives the corresponding English word.

The characteristics of the gathered corpus are presented in the Table 5.

#### 4.2. Lexicon construction

In the preprocessing of the collected corpus, we discarded tweets with the prefix *Rt, RT, and rt*, which indicate that the tweet that follow are re-tweets (re-postings of tweets sent earlier by somebody else).

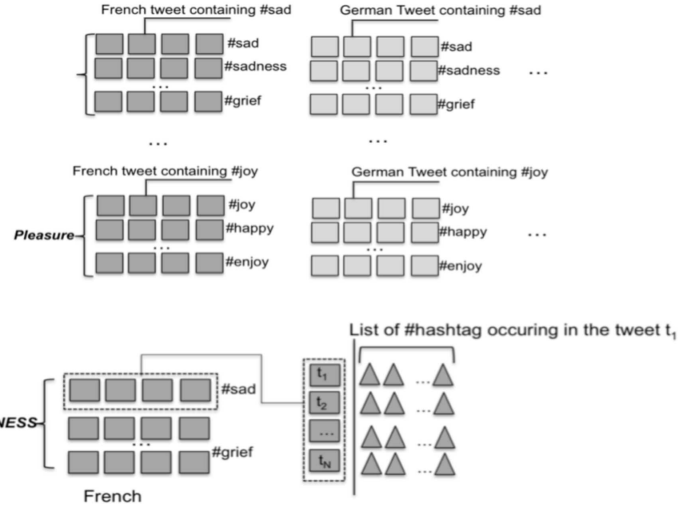


FIGURE 1 – Extraction of Hashtags from the French corpus

Second, we grouped the gathered tweets by language and by emotion (Figure 1). Then, for each emotion  $e$  i.e. *SADNESS, PLEASURE, LOVE*, etc., we extract all co-occurrent hashtags and compute their correlation to  $e$ . In order to compute how much an hashtag  $h$  is correlated to an emotion  $e$ , we compute the Strength of Association (SoA) between an hashtag  $h$  and an emotion  $e$  (Equation 1). We discarded short (less than 2 characters) and numerical hashtags.

$$\text{SoA}(h, e) = \log\left(\frac{\text{freq}(h, e)}{\text{freq}(h) \cdot \text{freq}(e)}\right) \quad (1)$$

Where the  $\text{freq}(h, e)$  is the number of times  $h$  occurs in tweets belonging to the emotion  $e$ . And  $\text{freq}(h)$ ,  $\text{freq}(e)$  are the frequencies of  $h$  and  $e$  in the corpus.

If an hashtag appear in more than one emotion class, we associate it to the most correlated class. The size of the constructed lexicons is about 17.000 entries for the seven languages.

## 5. Conclusion

In this research we have presented a novel approach based on Twitter as a comparable corpus to extract automatically affective lexicons in seven languages (English, French, German, Italian, Spanish, Portuguese and Russian). Our approach was motivated by the fact, that non english speaker's, usually, use bilingual terms in their messages. So, we are based in the co-occurrence between the English and the target affective terms to generate multilingual affective lexicons. The presented approach is generic as it could be applied for any language. Since the number of returned tweets is limited by the Twitter Search API, in a future work, we plan to use the Twitter Streaming API<sup>2</sup>, in order to collect a larger corpus and then obtain larger lexicons. Obtained lexicons, contains not only purely emotio-

1. Twitter API : <https://dev.twitter.com/docs>

2. <https://dev.twitter.com/docs/streaming-apis>

Affective Class	French	German
Anger	en colère fâcher rage irriter rougir nervosité massacre énervé exciter furax	wütend angepisst Wut zerstören Unterbrechung Tollwut massaker Erregung schütteln verärgert
Fear	peur terreur violence trombler mort terrifié appréhender inquiétude timidité anxiété	angst terror befürchten gestrandet Tod erschrocken achtgeben sorge eingeschüchtert ängstlich
Love	amour Valentin coeur mariage manquer aimer adorer envie gentillesse affection	Liebe verheiratet verpassen schön verpassen lieben leidenschaft Neid freundlichkeit zuneigung
Pleasure	heureux content génial bonheur plaisir jouer vacances podium agréable amusant	Vergnügen glücklich spielend Musik schön underschön Ferien erstaunlich reizend lustig

TABLE 6 – The Top-10 entries of the French and German affective lexicons for the *Anger*, *Fear*, *Love* and *Pleasure* emotion classes.

nal words but also some common-sense words that are associated to an affective class ; such as the german word *Tod* which is associated to the *Fear* affective class or the french term *coeur* which is associated to the *Love* class. So, for each language, we plan to divide the obtained lexicon into two sub-categories : purely emotional words and common-sense words.

## 6. References

- S. Armstrong and H. Thompson. 1995. A presentation of mlcc : Multilingual corpora for cooperation. *Linguistic Database Workshop*.
- M. M. Bradley and P. J. Lang. 1999. Affective norms for english words (anew). University of Florida. Gainesville, FL. The NIMH Center for the Study of Emotion and Attention.
- K. W. Church and P. Hanks. 1990. Word association norms, mutual information, and lexicography. volume 16 (1), pages 22–29.
- K. W. Church and R. L. Mercer. 1993. Introduction to the special issue on computational linguistics using large

- corpora. *Computational Linguistics*, 19(1) :1–24.
- A. Fraisse and P. Paroubek. 2013. Toward a unifying model for opinion, sentiment and emotion information extraction. In *In proceedings of the The 9th International Conference on Language Resources and Evaluation*, Reykjavik, Iceland.
- A. Jackiewicz and M. Vidak. 2014. Etude sur les mots-dièse. In *Congrès Mondial de la Linguistique Française.*, Berlin.
- S. M. Mohammad. 2012. Emotional tweets. In *In Proceedings of First Joint Conference on Lexical and Computational Semantics (\*SEM)*.
- A. Pak and P. Paroubek. 2010. Construction d’un lexique affectif pour le français à partir de twitter. In *In Proceedings of TALN (Traitement Automatique des Langues Naturelles) 2010*, Montréal, Canada.
- A. Qadir and E. Riloff. 2013. Bootstrapped learning of emotion hashtags #hashtags4you. In *In the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, Atlanta.
- R. Rapp. 1995. Identifying word translations in non-parallel texts. In *In Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, pages 320–322, Boston. Association for Computational Linguistics.
- R. Rapp. 1999. Automatic identification of word translations from unrelated english and german corpora. In *In Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 519–526, College Park, Maryland, USA. Association for Computational Linguistics.
- J. Redondo, I. Fraga, I. Padron, and M. Comesana. 2007. The spanish adaptation of anew (affective norms for english words). volume 39(3).
- C. Strapparava and A. Valitutti. 2004. Wordnet-affect : an affective extension of wordnet. In *In Proceedings of the 4th International Conference on Language Resources and Evaluation*, Lisbon, Portugal.
- M. L.-H. Vo, M. Conrad, L. Kuchinke, K. Urton, M.J. Hofmann, and A. M. Jacobs. 2009. The berlin affective word list reloaded (bawl-r). volume 41(2).

# Multimodal Comparable Corpora for Machine Translation

**Haithem Afli, Loïc Barrault and Holger Schwenk**

Laboratoire Informatique de l'Université du Maine (LIUM)  
University of Le Mans, France

firstname.lastname@lium.univ-lemans.fr

## Abstract

The construction of a statistical machine translation (SMT) system requires parallel corpus for training the translation model and monolingual data to build the target language model. A parallel corpus, also called bitext, consists in bilingual/multilingual texts. Unfortunately, parallel texts are a sparse resource for many language pairs. One way to overcome this lack of data is to exploit comparable corpora which are much more easily available. In this paper, we present the corpus developed for automatic parallel data extraction from multimodal comparable corpora, from *Euronews* and *TED* web sites. We describe the content of each corpus and how we extracted the parallel data with our new extraction system. We present the methods considered for using multimodal corpora and discuss the results on bitext extraction.

**Keywords:** Multimodal Comparable Corpora, Machine Translation, Parallel Data Extraction.

## 1. Introduction

In recent decades statistical approaches have significantly advanced the development of machine translation (MT). However, the applicability of these methods directly depends on the availability of very large quantities of parallel data. Recent works have demonstrated that a comparable corpus can compensate for the shortage of parallel corpora. However, for some languages, text comparable corpora may not cover all topics in some specific domains. One of the main challenges of our research is to build data and techniques to these under-resourced domains. What we need is to explore other sources like audio to generate parallel texts for each domain.

These kind of data are widely available on the Web for many languages.

In this paper, we present an extraction method used on multimodal comparable corpus. This corpus is then used to adapt and improve machine translation systems that suffer from resource deficiency. We, also, present Euronews-LIUM corpus which has been created within the context of our work on French DEPART<sup>1</sup> project. One of its main objective is the exploitation of multimodal and multilingual data for machine translation.

The methods for improving translation quality proposed in this work rely upon multimodal comparable corpora, that is, multiple corpora in different modalities that cover the same general topics and events. We compare it with (Afli et al., 2013) method built for the same kind of data.

Our main experimental framework is designed to address two situations. The first one is when we translate data from a new domain, different from the training data. In such a condition, the translation quality is generally rather poor. The second one is when we seek to improve the quality of an SMT system already trained on the same kind of data (same domain and/or style). Data is extracted from the available news (video and text modalities) on the *Euronews* website<sup>2</sup>. We also used TED-LIUM (Rousseau et al., 2012) corpus to build our TED multimodal comparable corpus

and test our extraction methods.

This paper is organized as follows: the first two sections present the new corpora. Section 3. contains the general extraction system architecture and some results are presented in Section 4.

## 2. Multimodal comparable corpora

### 2.1. Euronews



Figure 1: Example of multimodal comparable corpora from the *Euronews* website.

Figure 1 shows an example of multimodal comparable data coming from the *Euronews* website. An audio source of a political news and its text version, both in English, are available along with the equivalent news in French (audio and text modalities). The audio content in the videos are not exactly the same for each language, but are dealing with the same subject. Then, audio in one language and the text content in the other language can be considered as comparable data. This corpus can be used to extract parallel data, at the sentence and the sub-sentential level.

<sup>1</sup><http://www.projet-depart.org/>

<sup>2</sup><http://www.euronews.com>



Euronews website clusters news into several categories or sub-domains (e.g. Sport, Politics, etc.). These categories are preserved in the raw version of the provided corpus (but not in extracted versions). Table 1 show the statistics of our English/French *Euronews-LIUM* corpus created from French<sup>3</sup> and English news data from 2010-2012 period. This corpus<sup>4</sup> is composed of a comparable corpus, made of transcriptions (performed with our ASR system, see Section 3.2.) and article content (text found on the webpage). The extracted data performed with the system described in Section 3. are also provided.

## 2.2. TED

TED-LIUM corpus has been created within the context of the IWSLT'11 evaluation campaign. It has been built from some video talks crawled on the TED (Technology, Entertainment, Design) website<sup>5</sup>. The corpus is made of 773 talks representing 118 hours of speech. We used the English audio part of this corpus and the French text part of the *WIT3* parallel corpus<sup>6</sup>, to create the TED multimodal comparable corpus, further called TED-LIUM. Figure 2

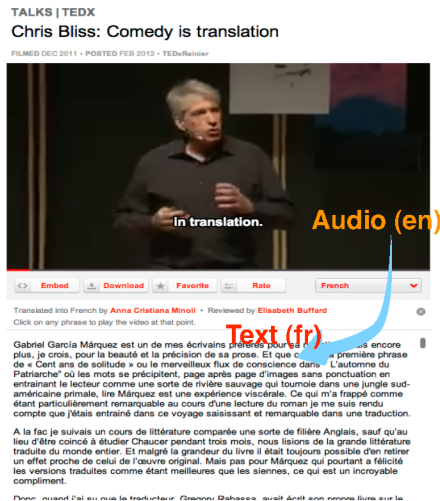


Figure 2: Example of multimodal comparable data from the TED website.

shows an example of such multimodal comparable data.

## 3. Parallel data extraction

### 3.1. System Architecture

The basic system architecture is described in Figure 3. We begin by extracting comparable sentences with the same method of (Afli et al., 2012) called *SentExtract*. We can distinguish three steps in this system: automatic speech recognition (ASR), statistical machine translation (SMT) and information retrieval (IR). The ASR system accepts audio data in language L1 as input and generates an automatic transcription. This transcription is then translated by a baseline SMT system into language L2. Then, we use these translations as queries for an IR system

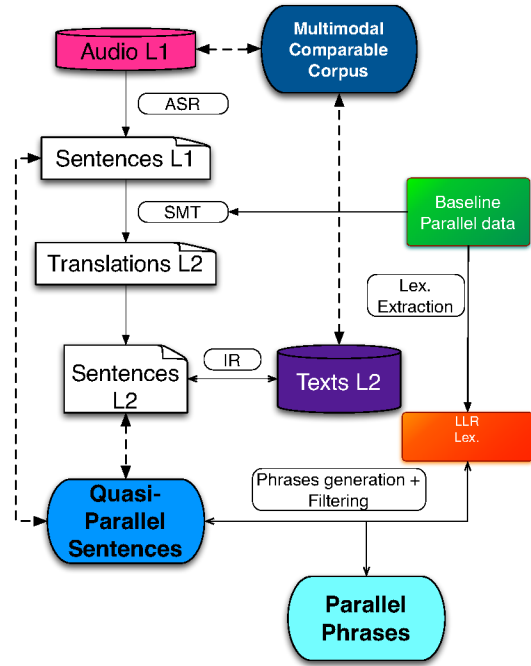


Figure 3: Principle of the parallel data extraction system from multimodal comparable corpora.

to retrieve most similar sentences in the indexed text part of our multimodal comparable corpus. The transcribed text in language L1 and the IR result in language L2 form the comparable sentences.

However, the extracted sentences are of different level of quality, and a filtering step is required in order not to degrade the performance of the baseline system. Previous work make use of different kinds of techniques to filter the extracted sentences, like TER (Snover et al., 2006) between IR query and the returned sentence (Abdul-Rauf and Schwenk, 2011) or bilingual lexicon log-likelihood ratio (Munteanu and Marcu, 2006). One of the drawbacks of filtering is that it can remove a large number of sentences, which often results in a lower impact on the baseline system. Moreover, the withdrawn sentences often

The location of Mohamed Mursi's trial at the police academy on the outskirts of Cairo, was meant to deter his supporters from turning out in large numbers.

But a sizeable number showed up despite a heavy security presence.

One of Mursi's court appointed lawyers said his client was illegally removed from office. The difference between the trial of Dr. Mohamed Mursi and the trial of (Hosni) Mubarak is that Mubarak had stepped down from power however, Mohamed Mursi is still the legitimate leader, legally and constitutionally he is still the president. This is the situation according to the rule of law and according to the constitution.

En Egypte la colère de la rue ne s'est pas fait attendre. Des centaines de manifestants pro-Mursi s'étaient rassemblés devant l'école de police. Le président déchu a refusé la présence d'un avocat, mais celui là s'est porté volontaire. La différence entre le procès de Mohamed Mursi, et de l'ancien président Mubarak, c'est que Mubarak a abandonné le pouvoir alors que Mursi a respecté la légitimité constitutionnelle, explique l'un des avocats. Il est le président de l'Egypte, légalement.

Figure 4: Example of comparable sentences which contain parallels phrases from *Euronews* website.

contain some useful parallel fragments which is interesting to extract.

<sup>3</sup><http://fr.euronews.com/>

<sup>4</sup>available soon on our website

<sup>5</sup><http://www.ted.com>

<sup>6</sup><https://wit3.fbk.eu/>



Sub-Domain	Audio En		Text	
	# words	# sentences	# words Fr	# words En
Business	289909	7898	425001	613684
Sport	81768	2369	112736	102923
Culture	388548	16773	262745	274323
Europe	398675	12531	302665	287178
Life Style	28813	1111	18379	19480
Politics	806607	26002	4932055	4666655
Science	231034	9346	147195	141652
Total	2225354	76030	6213995	6127565

Table 1: Size of the transcribed English audio corpus and English-French texts.

As an example, consider Figure 4, which presents two paragraphs extracted from the news articles presented in Figure 1. Although the articles report on the same event and express overlapping content, the texts cannot be considered as strictly parallel. They contain no fully parallel sentences pairs, but as shown by the boxes in the figure, some parallel phrases do exist in the sub-sentential level.

We developed a parallel phrase extraction system which operates in two steps. First, parallel phrase pair candidates are detected using the IBM1 model (Brown et al., 1993). Then the candidates are filtered with probabilistic translation lexicon (learned on the baseline SMT system training data) to produce parallel phrases using log-likelihood ratio (LLR) method (see (Munteanu and Marcu, 2006) for details). Our technique is similar to that of (Afli et al., 2013) called *PhrExtract*, but we bypass the need of the TER filtering by using a LLR lexicon. We call this new extended system *PhrExtract\_LLRL*.

### 3.2. Baseline systems

The ASR system used in our experiments is an in-house five-pass system based on the open-source CMU Sphinx system (version 3 and 4), similar to the LIUM’08 French ASR system described in (Deléglise et al., 2009). The acoustic models were trained in the same manner, except that we added a multi-layer perceptron (MLP) using the Bottle-Neck feature extraction as described in (Grézl and Fousek, 2008).

To train the language models (LM), we used the SRILM toolkit (Stolcke, 2002). We trained a 4-gram LM on all our monolingual corpus.

The SMT system is a phrase-based system based on the Moses SMT toolkit (Koehn et al., 2007). The standard fourteen feature functions are used, namely phrase and lexical translation probabilities in both directions, seven features for the lexicalized distortion model, a word and a phrase penalty and a target language model. It is constructed as follows. First, word alignments in both directions are calculated with the multi-threaded version of the GIZA++ tool (Gao and Vogel, 2008). Phrases and lexical reorderings are extracted using the default settings of the Moses toolkit. The parameters of our system were tuned on a development corpus using the MERT tool (Och,

Corpus	# words En	# words Fr
nc7	3.1M	3.7M
eparl7	51.2M	56.4M
devEuronews	74k	84k
tstEuronews	61k	70k
devTED	36k	38k
tstTED	8.7k	9.1k

Table 2: MT training and development data.

2003). To train, optimize and test our baseline MT system, we used the data presented in Table 2.

For each comparable corpus (Euronews-LIUM and TED-LIUM), we chose the most appropriate development and test corpus. *devEuronews* and *tstEuronews* are the news corpora used in the, respectively, WMT’10 and WMT’11 evaluation campaigns. *devTED* and *tstTED* are the official dev and test corpora from the IWSLT’11 international evaluation campaign.

We use the Lemur IR toolkit (Ogilvie and Callan, 2001) for the sentence extraction procedure with default settings. We first index all the French text considering each sentence as a document. This allows to use the translated sentences as queries to the IR toolkit. The IR system make use of the bag of word representation of each sentence and returns the most similar to the query. This sentence is then paired with the English query sentence. By these means we can retrieve the best matching sentences from the French side of the comparable corpus.

## 4. Results

For the sake of comparison, we ran several experiments with two methods. The first one, is *PhrExtract\_LLRL* (presented in section 3., and the second one corresponds to the method applied by (Afli et al., 2013) (called *PhrExtract* as in their paper). Experiments were conducted on English to French TED and Euronews tasks.

*PhrExtract* uses TER for filtering the result returned by IR, keeping only the phrases which have a TER score below a certain threshold determined empirically. Thus, we filter the selected sentences in each condition with different TER thresholds ranging from 0 to 100 by steps of 10.

The various SMT systems are evaluated using the BLEU score (Papineni et al., 2002).

Methods	# words (en)	# words (fr)
PhrExtract (TER 60)	16.61M	13.82M
PhrExtract_LLRL	1.68M	2.27M

Table 3: Number of words and sentences extracted from TED-LIUMcorpus with *PhrExtract* and *PhrExtract\_LLRL* methods.

Methods	# words (en)	# words (fr)
PhrExtract (TER 50)	2.39M	1.95M
PhrExtract_LLRL	636.8k	224.1k

Table 4: Number of words and sentences extracted from Euronews-LIUMcorpus with *PhrExtract* and *PhrExtract\_LLRL* methods.

Tables 3 and 4 show the statistics of the bitexts extracted from Euronews-LIUMand TED-LIUM. One can note that the sizes of the two sides of the bilingual text extracted from Euronews-LIUMare very different (English side is almost three times larger than French size). This behaviour is not observed on the TED data, and we do not yet explain this fact which requires a more fine grain analysis of the obtained bitexts. These bitexts are injected into our generic training data in order to adapt the baseline MT system. Tables 5 and 6 present the BLEU scores obtained with the best bitext extracted from each multimodal corpus with *PhrExtract* and *PhrExtract\_LLRL* methods. The TER threshold is set to 50 for Euronews-LIUMand 60 for TED-LIUM.

Systems	devTED	tstTED
Baseline	22.93	23.96
PhrExtract (TER 60)	23.70	24.84
PhrExtract_LLRL	23.63	24.88

Table 5: BLEU scores on devTED and tstTED after adaptation of a baseline system with bitexts extracted from TED-LIUMcorpus.

Systems	devEuronews	tstEuronews
Baseline	25.19	22.12
PhrExtract (TER 50)	30.04	27.59
PhrExtract_LLRL	30.00	27.47

Table 6: BLEU scores on devEuronews and tstEuronews after adaptation of a baseline system with bitexts extracted from Euronews-LIUMcorpus.

In the experiment with TED data, we seek to adapt our baseline SMT system to a new domain. We can see in table 5 that our new system obtains similar results as the

*PhrExtract* method. This means that the extracted texts are useful for adaptation purpose.

The same behavior is observed on Euronews task (Table 6). The extracted text can be used to improve an existing SMT system already trained on the same kind of data.

This new extraction method bypass the use of the TER filtering which required many experiments in order to determine the best threshold for each task.

Moreover, looking at the extracted text sizes in Tables 3 and 4, we can observe that the LLR method generate much less data while obtaining equivalent performance. This suggests that only the most relevant data is extracted by this technique.

We can see in the example in Table 7, that adding the extracted phrases can have a positive effect on translation quality.

## 5. Related Work

There has been considerable amount of work on exploiting comparable corpora, although from a different perspective than the one taken in this paper.

(Zhao and Vogel, 2002) proposed an adaptive approach aims at mining parallel sentences from a bilingual comparable news collection collected from the web. A maximum likelihood criterion was used by combining sentence length models and lexicon-based models. The translation lexicon was iteratively updated using the mined parallel data to get better vocabulary coverage and translation probability estimation. In (Yang and Li, 2003), an alignment method at different levels (title, word and character) based on dynamic programming (DP) is presented. The goal is to identify the one-to-one title pairs in an English/Chinese corpus collected from the web, They applied longest common subsequence (LCS) to find the most reliable Chinese translation of an English word. (Resnik and Smith, 2003) propose a web-mining based system called STRAND and show that their approach is able to find large numbers of similar document pairs.

A cross-language information retrieval techniques is used by (Utiyama and Isahara, 2003) to extract sentences from an English/Japanese comparable corpus. They identify similar article pairs, and then, considering them as parallel texts, they align their sentences using a sentence pair similarity score and use DP to find the least-cost alignment over the document pair.

(Munteanu and Marcu, 2005) uses a bilingual lexicon to translate some of the words of the source sentence. These translations are then used to query the database to find matching translations using information retrieval (IR) techniques. (Abdul-Rauf and Schwenk, 2011) bypass the need of the bilingual dictionary by using their own SMT system. They also use simple measures like word error rate (WER) or translation edit rate (TER) in place of a maximum entropy classifier.

In (Munteanu and Marcu, 2006) a first attempt to extract parallel sub-sentential fragments (phrases) from comparable corpora is presented. They used a method based on a Log-Likelihood-Ratio lexicon and a smoothing filter. They showed the effectiveness of their method to improve an SMT system from a collection of a comparable sentences.

Source EN (ASR output)	for me it's a necessity to greece stays in the euro zone and that greece gets the chance to get back on track the problem
Baseline FR	<b>pour moi une nécessité pour la grèce</b> reste dans la zone euro et que la <b>grèce</b> aura la chance <b>de revenir sur la piste problème</b>
Adapted FR	<b>Je vois la nécessité que la Grèce</b> reste dans la zone euro et que la <b>Grèce</b> aura la chance <b>de se remettre sur pieds .</b>

Table 7: Example of translation quality improvements of the baseline MT system after adding parallel data extracted from Euronews-LIUMcorpus.

The second type of approach consist in extracting parallel phrases with an alignment-based approach (Quirk et al., 2007; Riesa and Marcu, 2012). These methods are promising, because (Cettolo et al., 2010) show that mining for parallel fragments is more effective than mining for parallel sentences, and that comparable in-domain texts can be more valuable than parallel out-of-domain texts. But the proposed method in (Quirk et al., 2007) do not significantly improve MT performance and model in (Riesa and Marcu, 2012) is designed for parallel data.

So, it's hard to say that this approach is actually effective for comparable data.

Since our method can increase the precision of the extraction method, it greatly expands the range of corpora which can be usefully exploited.

## 6. Conclusion

In this paper, we have presented a new multimodal corpus built to extract parallel data for SMT systems. We also presented a new system to extract parallel fragments from a multimodal comparable corpus. Experiments conducted on TED and Euronews data showed that our method significantly outperforms the existing approaches and improves MT performance both in situations of domain adaptation (TED data) and of in-domain improvement (Euronews). This is an encouraging result which do not require any threshold empirically determined comparing to TER filtering method. Our approach can be improved in several aspects. A parallel corpus is used to generate the LLR lexicon used for filtering. An alternative method could be to construct a large bilingual dictionary from comparable corpora, and use it in the filtering module. In this case, the lexicon would benefit from containing words specific to the targeted task (in the case of adaptation). Another interesting extension is to carefully select the comparable data to be used in the extraction framework. This selection could be based on a similarity measure computed before the extraction process, and would help to improve the system performances.

## 7. Acknowledgements

This work has been partially funded by the French Government under the project DEPART.

## 8. References

- S. Abdul-Rauf and H. Schwenk. 2011. Parallel sentence generation from comparable corpora for improved smt. *Machine Translation*.
- H. Afi, L. Barrault, and H. Schwenk. 2012. Parallel texts extraction from multimodal comparable corpora. In *Jap-TAL*, volume 7614 of *Lecture Notes in Computer Science*, pages 40–51. Springer.
- Haithem Afi, Loïc Barrault, and Holger Schwenk. 2013. Multimodal comparable corpora as resources for extracting parallel data: Parallel phrases extraction. *International Joint Conference on Natural Language Processing*, October.
- Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Comput. Linguist.*, 19:263–311, June.
- Mauro Cettolo, Marcello Federico, and Nicola Bertoldi. 2010. Mining parallel fragments from comparable texts. *Proceedings of the 7th International Workshop on Spoken Language Translation*.
- P. Deléglise, Y. Estève, S. Meignier, and T. Merlin. 2009. Improvements to the LIUM french ASR system based on CMU Sphinx: what helps to significantly reduce the word error rate? In *Interspeech 2009*, Brighton (United Kingdom), 6-10 september.
- Q. Gao and S. Vogel. 2008. Parallel implementations of word alignment tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, SETQA-NLP '08, pages 49–57.
- F. Grézl and P. Fousek. 2008. Optimizing bottle-neck features for LVCSR. In *2008 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 4729–4732. IEEE Signal Processing Society.
- P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180.
- D. S. Munteanu and D. Marcu. 2005. Improving Machine Translation Performance by Exploiting Non-Parallel Corpora. *Computational Linguistics*, 31(4):477–504.
- D. S. Munteanu and D. Marcu. 2006. Extracting parallel sub-sentential fragments from non-parallel corpora. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, ACL-44, pages 81–88.
- Franz J. Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st An-*

- nual Meeting on Association for Computational Linguistics - Volume 1, ACL '03*, pages 160–167, Stroudsburg, PA, USA. Association for Computational Linguistics.
- P. Ogilvie and J. Callan. 2001. Experiments using the lemur toolkit. *Proceeding of the Tenth Text Retrieval Conference (TREC-10)*.
- K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL '02*, pages 311–318.
- Q. Quirk, R. Udupa, and A. Menezes. 2007. Generative models of noisy translations with applications to parallel fragment extraction. In *In Proceedings of MT Summit XI, European Association for Machine Translation*.
- Philip Resnik and Noah A. Smith. 2003. The web as a parallel corpus. *Comput. Linguist.*, 29:349–380, September.
- J. Riesa and D. Marcu. 2012. Automatic parallel fragment extraction from noisy data. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT '12*, pages 538–542.
- Anthony Rousseau, Paul Deléglise, and Yannick Estève. 2012. Ted-lium: an automatic speech recognition dedicated corpus. *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*.
- S. Snover, B. Dorr, R. Schwartz, M. Micciulla, and J. Makhoul. 2006. A study of translation edit rate with targeted human annotation. *Proceedings of Association for Machine Translation in the Americas*, pages 223–231.
- A. Stolcke. 2002. SRILM - an extensible language modeling toolkit. In *International Conference on Spoken Language Processing*, pages 257–286, November.
- 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 - Volume 1, ACL '03*, pages 72–79.
- Christopher C. Yang and Kar Wing Li. 2003. Automatic construction of english/chinese parallel corpora. *J. Am. Soc. Inf. Sci. Technol.*, 54:730–742, June.
- 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, ICDM '02*, Washington, DC, USA. IEEE Computer Society.

# Extended Translation Memories for Multilingual Document Authoring

Jean-Luc Meunier, Marc Dymetman

Xerox Research Centre Europe

6 chemin de Maupertuis, Meylan, France

E-mail: jean-luc.meunier@xrce.xerox.com, marc.dymetman@xrce.xerox.com

## Abstract

This paper discusses multilingual document authoring, viewed as providing computer support for a user to author a document in some source language while automatically generating the same content in one or many target languages. A kind of unanticipated use of multilingual authoring appeared in the service sector, in situations where an employee is servicing customers by answering their requests, or helping them, via written electronic communication. Decoupling the employee's language from the customer's language may open up new perspectives and motivated this work, where we propose a small set of extensions to be made on a translation memory to support multilingual authoring more efficiently. We describe how an instance of such extended formalism can be conveniently created thanks to a domain specific language and describe how we implemented a full system. Finally, we report on the experiment we ran in a real business setting.

**Keywords:** Translation Memory, Writing assistance, Multilingual Authoring, Experimentation, Domain Specific Language

## 1. Introduction

The need for efficiently producing a document in multiple languages most probably appeared long time ago, and the Rosetta Stone is a famous example of this need. One conventional approach to the problem consists in an authoring step followed by a translation step. With the advent of computers and computer science, new tools emerged, and authoring support tools, translation memories and machine translation are particularly relevant with this respect. A new approach emerged in the 90s, which aimed at providing computer support for authoring a document in multiple languages, merging two steps into a single activity. One early publication from Hartley and Paris (1997) says it all in its title: "Multilingual document production from support for translating to support for authoring".

The work presented here contributes to this approach by extending translation memories for use in multilingual authoring support. We will first introduce a motivating business use that was probably not imagined in the 90s, before giving some background on an existing multilingual authoring tool. We will then describe how to extend a translation memory for multilingual authoring and report on the experiment we ran in a real business setting.

## 2. Motivation

A kind of unanticipated use of multilingual authoring appeared in the service sector, in situations where an employee is servicing customers by answering their requests, or helping them, via written electronic communication. This situation is very common in sectors like customer care, human resource, finance, etc. The customer, or more generally requestor, contacts the agent by email, or by filling in a web form. The agent uses dedicated tools, e.g. a knowledge base or some customer relationship management tool, in order to fulfill the request and provides the requestor with a written answer.

Some requests may need multiple cycles of communication, forming a conversation. So far, agents were grouped into language teams in one or several helpdesk centers and each team was sized to answer the peak load and cover for the opening hours of the customer service.

With the globalizing market, the number of serviced languages is increasing and finding agent speaking the required language(s) often becomes problematic. Since companies try to avoid opening one helpdesk per language/country they service but rather look for ways to centralize the helpdesks in one or a few helpdesk center(s), they often face the problem of finding in a certain country an agent speaking a language that is not generally spoken in that country. To accommodate with organizational issues, those agents are often also required to speak the language of the country or the company. Finding a person with the required technical and language skills can prove quite difficult and may require paying a premium to get the person onboard.

Breaking the language barrier and allowing an agent who does not speak the requestor's language to provide him/her with the required help is therefore attractive to companies operating in this business sector, even if the solution allows for handling only a portion of the total volume of requests.

Machine translation ideally should answer this need: a request could be automatically translated into the agent's language and vice-versa for the agent's answer. Practically, coping with translation errors is both critical and not easy. We distinguish two situations with different constraints: inbound and outbound correspondence.

For inbound, the request needs to be translated in the agent's language so that the agent understands the request and feels confident about his/her understanding. No need for a perfect translation quality. In usual quality evaluation terms, the fluency of the translation is of less importance than its adequacy, which can be critical.

For outbound correspondence, the translation quality that is required is much higher since the company is sending a

written answer to a customer. Both fluency and adequacy are important and the consequence of any translation errors must be carefully assessed before rolling out such a system. Although automatic confidence estimation (Blatz et al., 2004) of the translation could play a role, we have chosen a different approach based on multilingual authoring with the goal of allowing the agent to author a reply in both her/his language and in the customer's language. In term of reply's quality, the multilingual authoring tool will bring the language knowledge while the agent will bring the subject matter expertise. The goal is to create a high quality reply, both at language- and semantic-levels, so that it is not perceptible that the agent does not speak the customer's language.

In the rest of the paper, we will focus on the use of multilingual authoring for supporting the outgoing correspondence. More precisely, we focus on how to extend translation memories for setting up a multilingual authoring support system.

### 3. Background: the MDA Tool

Before introducing how a translation memory can be extended for supporting multilingual authoring, let us introduce here one pre-existing tool called MDA (Brun et al., 2000), which stands for Multilingual Document Authoring. This tool was conceived in the years 1998-2002. It allows a monolingual user to interactively produce a document in multiple languages, including a language s/he masters, following a document template that controls both the semantics and the realization of the document in multiple languages.

This section describes the MDA tool and its template inner working, by using excerpts of the publication "Document structure and multilingual authoring" by Brun, Dymetman and Lux (2000), so as to introduce the challenges one faces to support multilingual authoring.

In the next section, we will relate the extended translation memory formalism to this tool's template.

#### 3.1 Approach

First, the main requirement for such a tool is that the authoring process is monolingual, but the results are multilingual. At each point of the process the author can view in his/her own language the text s/he has authored so far. This is in line with the WYSIWYM (What You See Is What You Mean) editing method described in (Power & Scott, 1998). In MDA, the areas where the text still needs refinement are highlighted and menus for selecting a refinement are also presented to the author in his/her own language. Thus, the author is always overtly working in the language s/he knows, but is implicitly building a language-independent representation of the document content.

From this representation, the system builds multilingual texts in any of several languages simultaneously. This approach characterizes our system as belonging to the paradigm of "natural language authoring" (Hartley & Paris, 1997; Power & Scott, 1998), which is distinguished from natural language generation by the fact that the

semantic input is provided interactively by a person rather than by a program accessing digital knowledge representations.

Second, the system maintains strong control both over the semantics and the realizations of the document. At the semantic level, dependencies between different parts of the representation of the document content can be imposed: for instance the choice of a certain chemical at a certain point in a maintenance manual may lead to an obligatory warning at another point in the manual. At the realization level, which is not directly manipulated by the author, the system can impose terminological choices (e.g. company-specific nomenclature for a given concept) or stylistic choices (such as choosing between using the infinitive or the imperative mode in French to express an instruction to an operator).

Finally, the semantic representation underlying the authoring process is strongly document-centric and geared towards directly expressing the choices which uniquely characterize a given document in an homogeneous class of documents belonging to the same domain. The screenshot in figure 1 shows the MDA tool, with a document being authored.

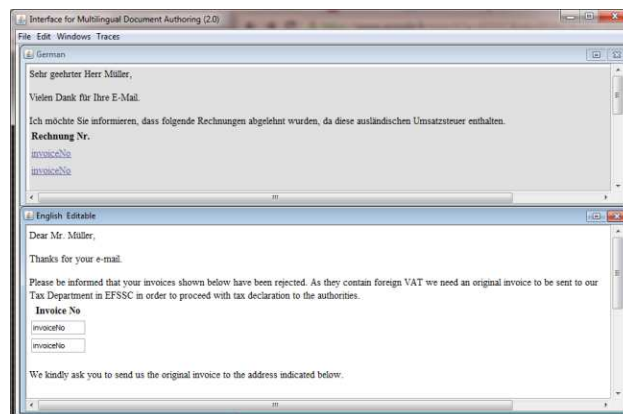


Figure 1: Screenshot of the MDA tool in use

#### 3.2 Interaction Grammars (IG)

Let us now give some details about the formalism of Interaction Grammars used by MDA. We start by explaining the notion of choice tree on the basis of a simple context-free grammar.

##### 3.2.1. Context-free grammars and choice trees

Let's consider the following Context-Free Grammar (CFG), ignoring for now the first column (italic text):

```

warnSympt warning --> "in case of", symptom, ",",
    action.
weak symptom --> "weakness".
conv symptom --> "convulsions".
hea symptom --> "headache".
rest action --> "get some rest".
consult action --> "call your doctor
    immediately".

```

What does it mean to author a "document" with such a CFG? It means that the author is iteratively presented with partial derivation trees relative to the grammar (partial in

the sense that leaves can be terminals or non-terminals) and at each given authoring step both selects a certain nonterminal to “refine”, and also a given rule to extend this non-terminal one step further; this action is repeated until the derivation tree is complete.

If one conventionally uses the identifier in *italic* in first column to name each rule, then the collection of choices made by the author during a session can be represented by a choice tree labelled with rule identifiers, also called combinators. An example of such a tree can be written `warnSymp(weak, rest)` reflecting the generation of the text “in case of weakness, get some rest”.

### 3.2.2. Making choice trees explicit

Choices trees are in MDA the central repository of document content and we want to manipulate them explicitly. Definite Clause Grammars (DCG) (Pereira & Warren, 1980) represent possibly the simplest extension of context-free grammars permitting such manipulation.

Our context-free grammar can be extended straightforwardly into the DCG<sup>1</sup>.

```
warning(warnSymp(S, A)) --> "in case of",
    symptom(S), ",", action(A).
symptom(weak) --> "weakness".
symptom(conv) --> "convulsions".
symptom(hea) --> "headache".
action(rest) --> "get some rest".
action(consult) --> "call your doctor
    immediately".
```

What these rules do is simply to construct choice trees recursively. Thus, the first rule says that if the author has chosen a symptom through the choice tree *S* and an action through the choice tree *A*, then the choice tree `warnSymp(S, A)` represents the description of a *warning*.

If now, in this DCG, we “forget” all the terminals, which are language-specific, by replacing them with the empty string, we obtain the following “abstract grammar”:

```
warning(warnSymp(S, A)) --> symptom(S),
    action(A).
symptom(weak) --> [].
symptom(conv) --> [].
symptom(hea) --> [].
action(rest) --> [].
action(consult) --> [].
```

This grammar is in fact equivalent to the definite clause program:

```
warning(warnSymp(S, A)) :- symptom(S),
    action(A).
symptom(weak) .
symptom(conv) .
symptom(hea) .
action(rest) .
action(consult) .
```

This abstract grammar (or, equivalently, this logic program), is language independent and recursively defines a set of well-formed choice trees of different categories, or types. Thus, the tree `warnSymp(weak, rest)`

<sup>1</sup> According to the usual logic programming conventions, lowercase letters denote predicates and functors, whereas up-percase letters denote metavariables that can be instantiated with terms

is well-formed “in” the type *warning*.

### 3.2.3. Dependent Types

In order to stress the type-related aspects of the previous tree specifications, we are actually using in our current implementation the following notation for the previous abstract grammar:

```
warnSymp(S, A)::warning --> S::symptom,
    A::action.
weak::symptom --> [].
conv::symptom --> [].
hea::symptom --> [].
rest::action --> [].
consult::action --> [].
```

The first rule is then read: “if *S* is a tree of type *symptom*, and *A* a tree of type *action*, then `warnSymp(S, A)` is a tree of type *warning*”, and similarly for the remaining rules.

The grammars we have given so far are deficient in one important respect: there is no dependency between the symptom and the action in the same warning, so that the tree is `warnSymp(weak, rest)` is well-formed in the type address. In order to remedy this problem, dependent types (Ranta, 2004) can be used. From our point of view, a dependent type is simply a type that can be parameterized by objects of other types. We write:

```
warnSymp(S, A)::warning -->
    S::symptom(Severity), A::action(Severity).
weak::symptom(mild) --> [].
conv::symptom(severe) --> [].
hea::symptom(severe) --> [].
rest::action(mild) --> [].
consult::action(severe) --> [].
```

We have introduced a *Severity* parameter that is shared by the two type *symptom* and *action* forcing certain associations between a given symptom and a given action.

### 3.2.4. Parallel Grammars and Semantics-driven Compositionality for Text Realization

We have just explained how abstract grammars can be used for specifying well-formed typed trees representing the content of a document.

In order to produce actual multilingual documents from such specifications, a simple approach is to allow for parallel realization English, French, ... grammars, which all have the same underlying abstract grammar (program), but which introduce terminals specific to the language at hand. Thus the following French and English grammars are parallel to the previous abstract grammar<sup>2</sup>:

```
warnSymp(S, A)::warning --> "In case of",
    S::symptom(Severity), " , " ,
    A::action(Severity) , "." .
weak::symptom(mild) --> "weakness".
conv::symptom(severe) --> "convulsions".
hea::symptom(severe) --> "headache".
rest::action(mild) --> "get some rest".
```

<sup>2</sup> Because the order of goals in the right-hand side of an abstract grammar rule is irrelevant, the goals on the right-hand sides of rule in two parallel realization grammars can appear in a different order, which permits certain reorganizations of the linguistic material (situation not shown in the example).



```

consult::action(severe) --> "call your doctor".

warnSymp(S, A)::warning --> "En cas de",
  S::symptom(Severity), " , " ,
  A::action(Severity) , "." .
weak::symptom(mild) --> "fatigue".
conv::symptom(severe) --> "convulsions".
hea::symptom(severe) --> "maux de tête".
rest::action(mild) --> "prenez du repos".
consult::action(severe) --> "consultez votre
  médecin".

```

The logic programming representation of such a grammar has rules of the following form:

```

a1(B,C,...)::a(D,...)-english[X,Y, ...] -->
  B::b(E,...)-english[X, ...] ,
  ". . ." ,
  C::c(F,...)-english[Y, ...] ,
  ...
  {constraints(B,C,...,D,E,F,...)},
  {conditional_code(X, Y, ...)}.

```

Those rules are close to the grammar rules, with additional language-specific parameters to deal with constraints that are specific to one language.

As the reader can see, the creation of a MDA template was a complex task, requiring unusual skills, namely the knowledge of definite clause grammars and Prolog. On the other hand we were attracted by the power of the tool and chose to use it as target platform for our new formalism.

#### 4. Extending Translation Memories

While the interaction grammars (IG) presented above proved to apply well to the problem of modelling agents' replies, or more generally agents' language, their creation was somehow complex and requiring uncommon expertise. We therefore looked for some alternative formalism. In particular, we considered the structure of a translation memory, since it intrinsically captures the desired parallelism between one source language and some target(s) one(s). It however lacks of the power of a grammar to define or guide the agent's language. We have therefore defined a minimal set of mechanisms that should be added to a translation memory structure to support our goal.

The proposal consists in following a Translation Memory (TM) paradigm, with a set of extensions towards supporting the creation of document template for multilingual document authoring by a monolingual user. Our aim is to facilitate the design of document grammar for multilingual document authoring by non-experts. More precisely, where a translation memory stores document fragments together with the corresponding translation, our extension consists in adding the notion of fragment type, allowing a fragment to be generalized to a certain type of textual content; we also introduce the notion of global variable, allowing some textual contents to be shared across a document. Each fragment remains aligned with its counterpart(s) in the other language(s).

Additional mechanisms include constraints and conditional realization.

Without loss of generality, let's consider the case of generating some document in English and French.

We will call 'designer' the person in charge of designing a document grammar, which can then be used by a 'user' of the MDA tool

#### 4.1. A translation memory approach with Context Free Grammar power

Where a standard translation memory would be a two-columns table, with parallel segments in English and French, our extended TM will be a sequence of four-columns tables:

- Column 1 is the so-called **case**: it uniquely identifies, and labels, a specific row within a table.
- Column 2 is the so-called **wizard**: it is used to guide the interaction between the multilingual authoring tool, e.g. legacy MDA, tool and the user, when she/he authors a new document.
- Column 3 and 4 are the **English** and **French** columns: they contain the realizations (concrete realizations as character string) of the segment in the two languages.
- Each additional language would require one addition column.

Each such table is called a **type** and has a unique name as well. See the table named "MyType" in figure 2. Some common types such as **STRING**, **NUMBER** and **DATE** are pre-defined in the formalism and in the tool.

The underlying formalism has ties with Context Free Grammars (CFG), since a type can be seen as a CFG non-terminal, while the cases correspond to enumerating and naming the possible production rules for that non-terminal. More precisely, this formalism has ties with Synchronous Context Free Grammar (Chiang & Knight, 2006).

Let's consider a simple CFG grammar like:

```

Document -> Det Noun Adj ". ."
Det -> "one"
Det -> "two"
Noun -> ...
...

```

We would express such a CFG as the sequence of tables shown in figure 3.

We see that the wizard allows the template designer to associate a question with a given type. Typically, in the MDA tool (when a user authors a new document), the tool will display the question and propose (some or all of) the case names for that type as possible answers to the user.

The English and French columns of a case can refer (zero or multiple times) to the types listed in the **wizard** part of the case, in any order, and can interleave them with terminal strings. In the previous example, observe how the English and French realizations re-order the non-terminals.

We will call 'type call' a non-terminal in the Wizard, English and French columns, since it can be seen as



‘calling’ a type that is defined in its own extended TM table.

In addition, because the English and French refer to the wizard type calls, it may be necessary to distinguish multiple calls to the same type, e.g. for a rule like `Document -> Det Noun Verb Det Noun`.

So a type call may be named for further reference within the same case from the English or French realization, as for instance in figure 4.

This Translation Memory Grammar (TMG) approach makes one step towards supporting multilingual document authoring using parallel context-free grammars, but requires additional mechanism to be available, as we will see below.

## 4.2. A translation memory approach with Interaction Grammar power

We are here extending our TMG formalism to support dependencies between types as well as dealing with extra conditions on the realization in natural language. As explained in section 3.2, the existing MDA tool relies on the so-called Interactive Grammars (IG) formalism, which is a specialization of the Definite Clause Grammars (Pereira & Warren, 1980) inspired by the GF formalism (Ranta, 2004). Please refer to (Brun et al., 2000) for full details on this formalism.

We reproduce below the IG abstract grammar (which does not show terminals) of the drug warning example:

```
warnSymp(S,A)::warning -->
  S::symptom(SympClass),
  A::action(SympClass).
weak::symptom(mild) --> [].
conv::symptom(severe) --> [].
hea::symptom(severe) --> [].
rest::action(mild) --> [].
consult::action(severe) --> [].
```

We propose here a simple way to inject some key aspect of the IG formalism in our TM-based formalism to deal with dependencies among types.

For doing so, a type may have one or multiple attribute(s), the value of which can be constrained by an equality operator. The constraint can involve an attribute and a constant or two attributes. Note that the ‘=’ operator is asserting a constraint rather than expressing an assignment.

So the above example would be reflected as shown in figure 5.

Scoping: the attributes of a type are accessible from the type itself using the keyword `this`, or via a reference of a wizard’s type call within a case. An attribute set in the wizard column is visible in other columns, while if set in the ‘French’ column, it will only be visible from a ‘French’ column.

Moreover, it is common when designing a grammar to require access to certain information from several different places. Typically, when designing a template of a letter to a customer, the designer may need to access the customer name from several parts of the documents, which will typically correspond to accessing it from

several types of the TM-like template.

We therefore introduce one more mechanism allowing the designer to declare a so-called `global` by associating a (grammar-)unique name with a type. This name can then be used as reference in any case of any type.

Back to the drug warning, the designer could have for instance declared `DrugName` as a global of type `STRING` to conveniently insert the name of the drug in a realization. In addition, the designer could have declared a global `DrugForm` of type `pharm_form` (see in next section) to reflect the pharmaceutical form of the drug (tablet, capsule, syrup, eye drop).

## 4.3. Conditional Realization

We introduce the last mechanism to deal with fine realization issues. Typically, in French the noun ‘tablet’ has a genre which must be taken into account by a related adjective or past-participle (among others...).

We introduce conditional realization, where the designer can condition the realization by constraints on attributes. (The constraint is enforced locally to the case, unless it involves a global.)

The example in figure 6 below illustrates this.

The generated grammar also includes a catch-all mechanism so that if no condition is met, some error message is produced and shown to the user.

With such formalism, the interaction grammar example given in section 3.2 is shown in figure 7.

We believed this formalism to considerably alleviate the complexity of defining the resource required to support multilingual authoring and were interested in testing this belief, as described in next sections.

## 5. Implementation: dedicated tool suite for the TM Grammar

Editing such a TM grammar is not straightforward because of its structure as well as the multiple inner references to types, attributes, etc. We therefore decided to create some dedicated editing tool.

### 5.1. XML Lingua

First an XML representation was defined thanks to a RelaxNG (Clark & Murata, 2001) XML schema. Any TMG (translation-memory grammar) expressed in this XML language can then be displayed in the above tabular structure thanks to a CSS stylesheet.

We then explored the possible use of some off-the-shelf schema-aware XML editor, but none were supporting the CSS view in editing mode. So the use of an XML representation was both convenient and good engineering practice but was not appropriate for editing purpose.

### 5.2. Domain Specific Language

We therefore decided to design a Domain Specific Language (DSL) for our translation-memory grammars and implemented it using the Eclipse/Xtext/Xtend framework ([www.eclipse.org/org](http://www.eclipse.org/org)). Eclipse is an “an open

development platform comprised of extensible frameworks, tools and runtimes for building, deploying and managing software across the lifecycle“. Xtext is “a framework for development of programming languages and domain specific languages”. Xtend is “a flexible and expressive dialect of Java”.

The result is an editor with syntax coloring, content assistance, outline, validation and quick fix facilities integrated into the Eclipse IDE, which comes with rich functionalities for versioning etc, and able to generate the XML representation of a translation-memory grammar.

In Xtext, designing a DSL involves specifying a particular kind of BNF for the language to describe the concrete syntax and how it is mapped to an in-memory representation - the semantic model. This model will be produced by the parser on-the-fly when it consumes an input file. The full-fledged editor and required parser are automatically generated from the special BNF.

In Xtend, one can further enrich the editor, for instance to define the outline view appearing on the right panel in the screenshot below. But more importantly, we used Xtend to automatically generate the XML corresponding to a TMG being edited.

In order to generate the IG grammar required for the MDA tool given a TMG instance, we specifically developed a compiler from XML to IG.

Figure 8 shows the same Symptom/Action example created within this DSL.

## 6. Experiment

We experimented the MDA tool and the translation-memory grammar (TMG) with the help of colleagues from Xerox service who are running the Account-Payable office of a Xerox customer. In this office, Xerox agents are receiving emails from suppliers of the Xerox customer regarding invoices, payments, etc. The agents use the customer database and IT infrastructure to answer the requestors by email as well. The contractual language is German and this was requiring the agents to be fluent in German in addition to the job-specific skills.

Xerox service was interested in testing if combining machine translation and multilingual authoring would allow a monolingual English-speaking agent to work in this context where the business language, contracted by the customer, is German. More precisely, the goal was to evaluate the proportion of replies that could be handled by an English agent using MDA, assuming the machine translation of the request was satisfactory. Should the translation be unsatisfactory or MDA inappropriate to author a reply, then the request would be escalated to a German-speaking agent.

With the aim of handling the highest possible proportion of replies, the service team provided us with a typology of replies and selected the most frequent types for us to encode those types in a TMG. Given this list of pairs of (English, German) texts, we then devised a TMG. Looking at the regularities, we structured each reply as a sequence made of: greetings, thanks?, message+, ending (where ? denotes an optional item and + an item occurring one or more times). We identified 5 different forms of greetings and ending. The core of the reply could be

structured further into 6 sub-types, totalizing 90 cases, as they are called in TMG.

In order to jointly design the TMG with the Xerox service team, we exposed them to the TMG thanks to the tabular view created by use of the CSS on the XML file. Despite some of our colleagues were not IT expert, the tabular structure was easy to understand. So we ended up exchanging annotated document, namely MS-Word document in track change mode, so as to work jointly on the TMG. We show in figure 9 an excerpt of such a document sent back from our service colleagues who fixed the German side of the case “AP13\_Missing\_Invoice”.

Three rounds of tests were required to reach a satisfactory level, after a dozen of exchange of the TMG between the research and service teams. For each tests, the service team evaluated if a reply was both doable with MDA and of acceptable quality, on about 150 requests, by asking a monolingual English agent to answer a (machine-)translated request.

The table below summarizes the results:

Test results	Round 1	Round 2	Round 3
<b>Outbound unacceptable</b>	81%	42%	7%
<b>Outbound acceptable</b>	19%	58%	93%

The creation of the first version of TMG took about 4 days of work, while the following two next versions took 2 days each. The result obtained at round 3 is quite satisfactory. The use of a human-readable tabular structure proved to be valuable in this context where actors with different expertise, linguistic/business/IT, need to cooperate.

However, the TMG we created remains rather simple in the sense that only few semantic constraints and linguistic difficulties were to be handled. Actually, this relatively low complexity may also be characteristic of the domain of application because agents’ discourse often follows some company policy.

It remains unclear how well the TMG can scale to more sophisticated and advanced answer writing since the complexity of the grammar may become too high for handcrafting it. In 2000 Brun et al. chose a rather complex example involving pharmaceutical notices. We believe this example would be much easier to write with the TMG than with the 2000 original formalism. We are looking forward to new example of practical use to answer this important question.

## 7. Conclusion

In this paper we have presented a novel formalism for multilingual authoring so as to support a user in creating a document in “his” language while automatically generating the same content in some foreign language(s). The proposed formalism consists in a translation memory structure with a minimal set of additional mechanisms, to form what we call a Translation Memory Grammar (TMG).

To operationally implement it, we have relied on a pre-existing tool called MDA and on its underlying interactive grammars (IG), themselves implemented in a logic programming language. While logic programming was convenient, we believe there are alternative ways to

implement our proposed formalism.

To support the editing of the TMG, we have devised a domain specific language using modern software engineering techniques.

Since we introduced this tool in the context of a particular business need, we have described the experiment we did with our colleagues from the service arm of our company, in the context of a contracted provision of service to an external customer.

From the experiment, we draw the following conclusions:

- The tabular structure is valuable for supporting the necessary interaction between team members with different and complementary expertise: linguistic (source and target languages), business (Account payable here), IT (for creating the TMG).
- Basic linguistic phenomenon can be captured by simple syntactic encoding in the tabular structure, provided the IT person has rudimentary knowledge of both the source and target languages.
- The Eclipse/Xtext/Xtend framework allowed us to create a robust DSL.
- The Translation Memory grammar was powerful and expressive enough for answering these business needs.

Unfortunately, at the time of writing of this article we have no feedback from the field regarding the user acceptance of the tool and how the new practice compares to previous one in term of effort/resource. On the other hand, during test phases, no concern was raised regarding this matter, so we are optimistic.

We are now looking forward to experimenting with transferring the TMG editing tool suite to our service colleagues so as to validate the use of this formalism by non-specialists.

## 8. Acknowledgements

We thank Caroline Brun and Veronika Lux as their work and publications on MDA, jointly with Marc Dymetman, are central to the present work. We are also thankful to our colleagues from Xerox service for their participation in the experiment.

## 9. References

- Blatz, J., Fitzgerald, E., Foster, G., Gandrabur, S., Goutte, C., Kulesza, A., Sanchis, N., Ueffing, N. (2004, August). Confidence estimation for machine translation. In Proceedings of the 20th international conference on Computational Linguistics (p. 315). Association for Computational Linguistics.
- Brun, C., Dymetman, M., & Lux, V. (2000, June). Document structure and multilingual authoring. In Proceedings of the first international conference on Natural language generation-Volume 14 (pp. 24-31). Association for Computational Linguistics.
- Chiang, D., & Knight, K. (2006). An introduction to synchronous grammars. Tutorial available at <http://www.isi.edu/~chiang/papers/synchtut.pdf>.
- Clark, J., & Murata, M. RELAX NG Specification. Oasis, December 2001.
- Hartley, A., & Paris, C. (1997). Multilingual document production from support for translating to support for authoring. *Machine Translation*, 12(1-2), 109-129.

Pereira, F. C., & Warren, D. H. (1980). Definite clause grammars for language analysis—a survey of the formalism and a comparison with augmented transition networks. *Artificial intelligence*, 13(3), 231-278.

Power, R., & Scott, D. (1998, August). Multilingual authoring using feedback texts. In Proceedings of the 17th international conference on Computational linguistics-Volume 2 (pp. 1053-1059). Association for Computational Linguistics.

Ranta, A. (2004). Grammatical framework. *Journal of Functional Programming*, 14(2), 145-189.

## 10. Figures

MyType	(Wizard)	(English)	(French)
Case1-name	...	...	...
Case2-name	...	...	...
...			

Figure 2: a type named “MyType” in tabular view.

Document	(Wizard)	(English)	(French)
One-noun-phrase-document	“Choose a determiner:” <b>Det</b> “Choose a noun:” <b>Noun</b> “Choose an adjective:” <b>Adj</b>	<b>Det Adj Noun</b> “.”	<b>Det Noun Adj</b> “.”

Det	(Wizard)	(English)	(French)
Case one		“one”	“un”
Case two		“two”	“deux”

...

Figure 3: Example of CFG in the proposed formalism.

Document	(Wizard)	(English)	(French)
One-simple-sentence-document	“Choose a determiner:” <b>Det:d1</b> “Choose a noun:” <b>Noun:n1</b> “Choose a verb:” <b>Verb</b> “Choose a determiner:” <b>Det:d2</b> “Choose a noun:” <b>Noun:n2</b>	<b>d1 n1 Verb d2 n2</b> “.”	<b>d1 n1 Verb d2 n2</b> “.”

Figure 4: Reference to type calls

warning	(Wizard)	(English)	(French)
warnSymp	“Choose a symptom:” <b>symptom:S</b> “Choose an action:” <b>action:A</b> <b>S.severity =A.severity</b>	”In case of” <b>S</b> ”,” <b>A</b> “.”	”En cas de” <b>S</b> ”,” <b>A</b> “.”

symptom	(Wizard)	(English)	(French)
weak	<b>this.severity=mild</b>	“weakness”	...
conv	<b>this.severity=severe</b>	“convulsions”	...
hea	<b>this.severity=severe</b>	“headache”	...

action	(Wizard)	(English)	(French)
rest	<b>this.severity=mild</b>	”take some rest”	...
consult	<b>this.severity=severe</b>	“consult immediately”	...

Figure 5: An example of constraint

pharm_form	(Wizard)	(English)	(French)
tablet		“tablet”	”comprimé” <b>this.gender=m</b>
capsule		“capsule”	”gélule” <b>this.gender=f</b>

use	(Wizard)	(English)	(French)
swallow	"select a form:" pharm_form:F	"Swallow the" F "without crunching."	"Avaler" (F.gender=f "la"   F.gender=m "le") F "sans croquer."

Figure 6: Conditional Realization

Warning	(Wizard)	(English)	(French)
simple	"Choose a symptom:" Symptom:S "Choose an action:" Action:A S.severity=A.severity	"In case of" S ", " A "."	"En cas de" S ", " A "."
Symptom	(Wizard)	(English)	(French)
weak	this.severity=mild	"weakness"	"faiblesse"
conv	this.severity=severe	"convulsions"	"convulsions"
hea	this.severity=severe	"headache"	"maux de tête"
Action	(Wizard)	(English)	(French)
rest	this.severity=mild	"take some rest"	"prendre du repos"
consult	this.severity=severe	"consult immediately"	"consulter immédiatement"

Figure 7: Example 3.2.3 fully implemented

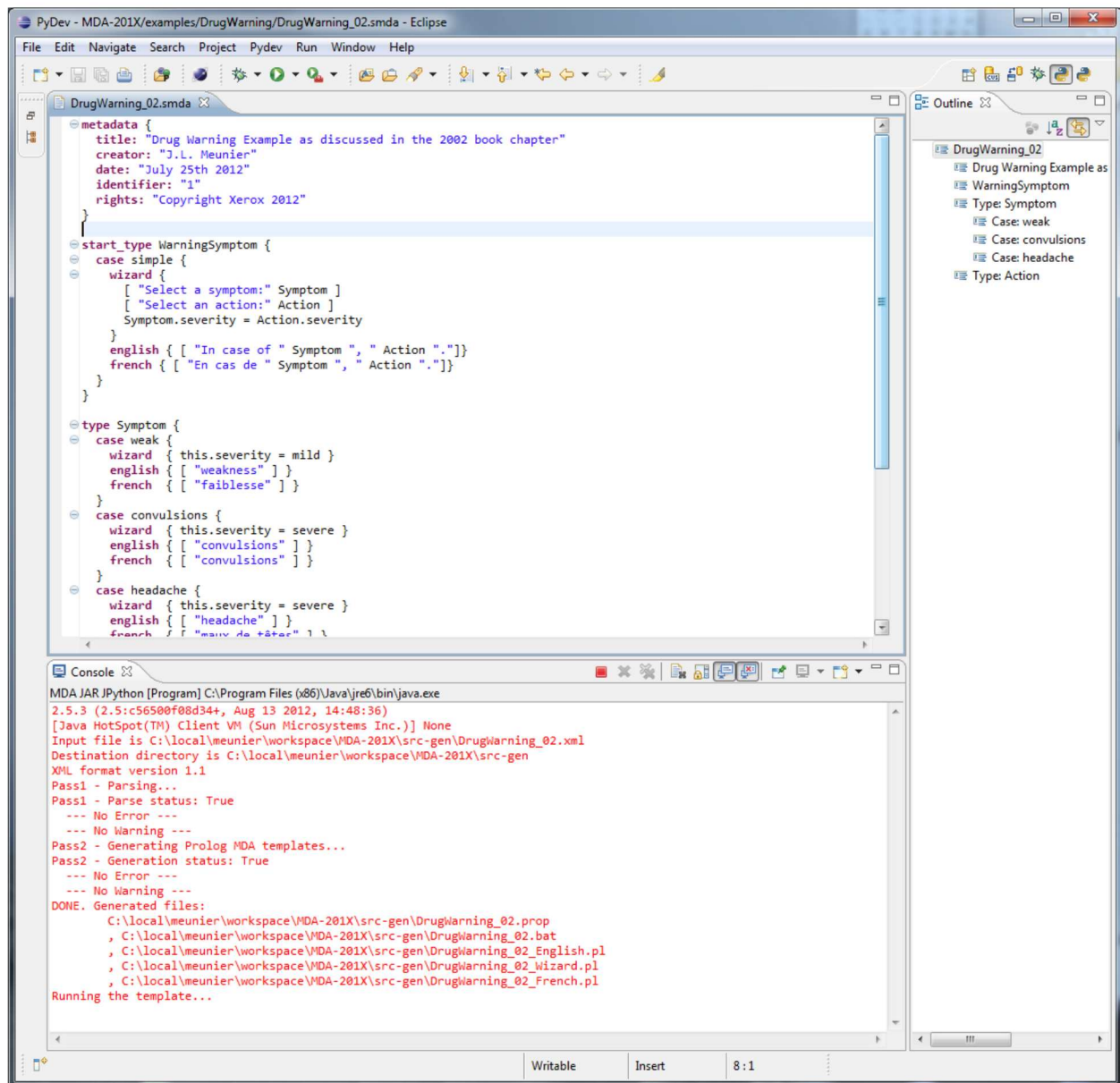


Figure 8: The DSL editor for translation-memory grammar in use, with a trace of the compiler producing the

corresponding MDA IG grammar

AP13_Missing_Invoice	
English	German
<p>Please be informed that your ((invoice   invoices)) No. <b>STRING:InvoiceNo</b> <b>TYPE:Further_List_Invoice</b> (( has   have)) not reached us. We kindly request that you send us the original ((invoice   invoices)) with all needed paperwork by post once again using the following details:</p> <p>The invoicing address to be indicated on your invoice is:</p> <p>Actual text not shown</p> <p>And the postal address where you have to send your invoice is:</p> <p>Actual text not shown</p>	<p>Leider haben wir <b>folgende</b> ((die unten aufgeführte Rechnung   die unten aufgeführten Rechnungen)) nicht erhalten: <b>STRING:InvoiceNo</b> <b>TYPE:Further_List_Invoice</b>. Bitte senden Sie uns Ihre Original- ((Rechnung   rRechnungen)) mit <b>allen</b> erforderlichen Anlagen erneut per Post an:</p> <p>Rechnungsanschrift (<b>auf der Rechnung selbst</b>):</p> <p>Actual text not shown</p> <p>Postanschrift (auf dem Umschlag):</p> <p>Actual text not shown</p>

Figure 9: MS-Word was used in track-change mode to interact with the service team. Note that conditional text, surrounded by double- red parentheses, was not an issue for them.

# Using partly multilingual patents to support research on multilingual IR by building translation memories and MT systems

Lingxiao WANG<sup>1,2</sup>, Christian Boitet<sup>2</sup>, Valérie Bellynck<sup>2</sup>, Mathieu Mangeot<sup>2</sup>

<sup>1</sup> SAS Lingua et Machina, c/o Inria, Domaine de Voluceau, Rocquencourt, B.P. 105  
78153 Le Chesnay Cedex, France

<sup>2</sup>LIG-GETALP, Bâtiment IM<sup>2</sup>AG B, Laboratoire LIG, 41 rue des mathématiques  
38400 ST Martin d'Hères, France

E-mail: {Lingxiao.Wang, Christian.Boitet, Valerie.Bellynck, Mathieu.Mangeot}@imag.fr

## Abstract

In this paper, we describe the extraction of directional translation memories (TMs) from a partly multilingual corpus of patent documents, namely the CLEF-IP collection, and the subsequent production and gradual improvement of MT systems for the associated sublanguages (one for each language), the motivation being to support the work of researchers of the MUMIA community. First, we analysed the structure of patent documents in this collection, and extracted multilingual parallel segments (English-German, English-French, and French-German) from it, taking care to identify the source language, as well as monolingual segments. Then we used the extracted TMs to construct statistical machine translation systems (SMT). In order to get more parallel segments, we also imported monolingual segments into our post-editing system, and post-edited them with the help of SMT.

**Keywords:** Extraction of parallel segments, SMT, CLEF patent collection, translation memories, source language identification, support for CLIR

## 1. Introduction

Parallel corpora have an important role in the natural language processing (NLP), and are a valuable resource for many NLP applications, such as statistical machine translation (SMT), cross-lingual information retrieval and multilingual lexicography. Patent description documents, because they often contain multilingual translations of some segments, are also seen as an important source of parallel corpora. Much work has been done on this topic, such as (Utiyama and Isahara, 2007), (Lu et al., 2009), and (Wäschle and Riezler, 2012).

In this paper, we describe our method for extracting a multilingual parallel corpus from a patent corpus, namely the CLEF-IP collection<sup>1</sup>, and present how to use these data. From the extracted multilingual parallel segments (English-German, English-French, and French-German), we built a translation memory (TM) and added it into our iMAG/SECTra system (Wang and Boitet, 2013). We then produced several SMT systems from this MT. In order to contribute to WG2 of the MUMIA<sup>2</sup> community on infrastructure, we transformed the collection of patents in a website where each patent is monolingual, and can be accessed (and collaboratively) post-edited into any language, using the above desired MT system when applicable, and free MT Web servers otherwise (e.g., for access in Chinese).

## 2. The CLEF-IP Collection

The latest version of collection corpus is the same as the one used in the CLEF-IP 2011 lab (the data corpus used

in 2012 and 2013 is the same as the one used in 2011), so our work is based on the CLEF-IP 2011 collection. This collection comprises more than 117 GB of multilingual patent documents derived from European Patent Office (EPO) and World Intellectual Property Organization (WIPO) sources. The CLEF-IP 2011 collection is composed of about 3.5 M XML files containing the textual part (no images) of about 1.5 M partially multilingual patent documents, corresponding to over 1.5 million patents published until 2002.

A patent document of the CLEF-IP 2011 collection is an *application document*, a *search report*, or a *granted patent document*, which is stored as a XML file. Each patent document has a unique patent name (EP for the EPO, or WO for the WIPO, followed by a series of digits and a code A or B<sup>3</sup>, like EP-0071719-B1.xml). Different information and different content of the patent document are stored in various XML fields, such as *<bibliographic-data>*, *<invention-title>*, *<abstract>*, *<description>*, *<claims>*, *<copyright>*, etc., and the fields of some patent documents also have subfields. The content of the various XML fields can be in English, French, or German (official languages of the EPO). However, not all segments of patent documents have content in these fields.

Each XML patent document of CLEF-IP 2011 has an associated document language, which we can find it in the *<patent-document>* field. During our extraction process, we consider the document language as the source language. We analyzed patents with respect to the structure of their XML fields, and found that four main fields may have parallel segments: *<invention-title>*, *<abstract>*, *<description>*, and *<claims>*. Each field may have some subfields, for example, a field *<claims>* may contain 6 *<claim>* subfields in EP-0260000-B1.xml

<sup>1</sup> Cross-Language Experiment Forum (CLEF), <http://www.clef-campaign.org>,

and <http://www.ifs.tuwien.ac.at/~clef-ip/index.html>

<sup>2</sup> MUMIA (MUltilingual, multimodal, Multifaceted Information access) is a COST action (CE1002) of the UE. Many members of its network do research on CLIR in patents.

<sup>3</sup> List of patent document kind codes: <https://register.epo.org/help?topic=kindcodes> and [http://www.wipo.int/patentscope/en/wo\\_publication\\_informatio\\_n/kind\\_codes.html](http://www.wipo.int/patentscope/en/wo_publication_informatio_n/kind_codes.html)



(Figure 1). We begin with these fields, looking for fields that appear more than once in the patent document and each field with a different language attribute. For example, Figure 2 shows an `<invention-title>` field with 3 different language attributes (`lang="DE"`, `lang="EN"`, and `lang="FR"`). Each field also contains some content, in the language that corresponds to its attribute.

```

▼<claims load-source="patent-office" status="new" mxw-id="PCLM9874066" lang="DE">
  ><claim num="1">...</claim>
  ><claim num="2">...</claim>
  ><claim num="3">...</claim>
  ><claim num="4">...</claim>
  ><claim num="5">...</claim>
  ><claim num="6">...</claim>
</claims>

```

Figure 1: `<claims>` has 6 `<claim>` subfields in EP-0260000-B1.xml

```

▼<invention-title lang="DE" load-source="ep" status="new">
  Verfahren zur Herstellung von Polyimidestern der Trimellitsäure
</invention-title>
▼<invention-title lang="EN" load-source="ep" status="new">
  PROCESS FOR THE FABRICATION OF POLYIMIDE-ESTERS OF TRIMELLITIC ACID
</invention-title>
▼<invention-title lang="FR" load-source="ep" status="new">
  Procédé pour la fabrication de polyimide-esters de l'acide trimellitique
</invention-title>

```

Figure 2: Example of an `<invention-title>` field with 3 different language attributes and the corresponding contents in 3 different languages

### 3. Extraction of Parallel Data

We started from the 3.5 million XML files corresponding to 1.5 million patents. The first goal was to extract from them as many useful parallel segments as possible. First, we traverse every patent document. For each patent document, we select the source language from the `<patent-document>` field, according to the language attribute of this field. Second, we search the parallel segments contained in the four main fields (`<invention-title>`, `<abstract>`, `<description>`, and `<claims>`). Sometimes, some fields occur with different language attribute than the document language. For example, in *EP-0260700-B1.xml*, English is the document language, but `<claims>` segments do not exist in English, only German and French versions are available. Even though it is always desirable to collect as much text as possible, it is even more important to ensure the quality of the texts, so in this case, we do not store the German and French parts as a parallel segment.

All fields, which appear more than once in a patent document and have different language attributes, are treated as a collection. In general, an EPO patent document has a maximum of 3 languages (English, French, and German). We chose as source segment the segments whose language attribute is consistent with the source language, and then extract the target parallel segments from the other fields. For example, in EP-0301015-B1.xml, the source language is English, and the `<claims>` field appears 3 times. Hence, we use the English part of the claims fields as the source segments, and consider the French and German parts as the target segments. The source segment and the target segments are then stored separately into different files. In the above example, the source segment has been stored into *CLEF\_claims\_en-fr:en* and *CLEF\_claims\_en-de:en*, and the target segments in *CLEF\_claims\_en-fr:fr* and *CLEF\_claims\_en-de:de*, respectively. In order to reduce

the noise in the data, we keep only the extracted text, and remove all tags.

Not all the extracted data is fully suitable for direct use for NLP applications. We have to clean the extracted data and eliminate some noise. First, we split the text into sentences, and then remove useless whitespace, and duplicate sentences. For alignment, we use the LF Aligner<sup>4</sup>, an open-source tool based on Hunaligne (Varga et al., 2005), which has the widest linguistic backbone (a total of 32 languages), and permits the automatic generation of dictionaries in any combination of these languages. Aligned segments are prepared bilingually for 4 types (title, abstract, description, and claims), and all 6-language pairs (`de_en`, `de_fr`, `en_de`, `en_fr`, `fr_de`, `fr_en`).

## 4. Some Statistics About the Corpus

Table 1 shows the number of segments and words that are extracted from the title and claims fields on the source and the target after segment aligning. All extracted parallel sentences are saved in TMX and TXT formats, and can be found at <http://membres-liglab.imag.fr/wang/downloads>

## 5. Application in SMT

We used our extracted parallel corpus (the title and claims fields) to construct SMT systems with the Moses Toolkit (Koehn et al., 2007). First, preparing the development set and the test set, we extracted 2,000 sentences for training the feature weights of Moses, and extracted 1,000 sentences for testing. Then we use the rest to train translation models of Moses. We actually built SMT system for only 3 directions: `de-en`, `de-fr`, and `en-fr`.

The systems also include 5-gram language models trained on the target side of corresponding parallel texts using IRSTLM (Federico et al., 2008). The feature weights required by the Moses decoder were further determined with MERT (Och, 2003) by optimizing BLEU scores on the development set (1,000 sentences). The test sets were translated by the resulting systems and then used to evaluate the systems in terms of BLEU scores (Papineni et al., 2001), as shown in Table 2.

## 6. Post-editing Monolingual Sentences Pre-translated by SMT

When we extracted parallel sentences from the CLEF-IP collection, we also derived large amount of monolingual sentences, which are not translated in the patent documents. The language of patents, although having a large amount of vocabulary and richness of grammatical structure, can be considered as a specialized sub language, because its grammar is quite restricted compared to that of the whole language. Second, patents have attributes of domain, this has been proven in some works, for example, (Wäschle and Riezler, 2012). Third, recent experiments in specializing empirical MT systems have shown that remarkably good MT results can be obtained (Rubino et al., 2012). So we combine these features with framework iMAG/SECTra (Wang and Boitet, 2013).

<sup>4</sup> <http://sourceforge.net/projects/aligner/>



Language pairs		Title		Claims	
		Segments	Words	Segments	Words
de-en	de	311,298	2,038,785	1,696,498	62 M
	en		2,582,703		71 M
de-fr	de	311,184	2,036,112	1,661,419	79 M
	fr		2,482,257		86 M
en-de	en	884,759	6,661,481	5,218,024	332 M
	de		5,508,289		296 M
en-fr	en	884,727	6,661,322	5,373,452	330 M
	fr		8,538,012		380 M
fr-de	fr	106,211	963,508	572,356	36 M
	de		1,204,439		37 M
fr-en	fr	106,246	1,285,467	586,498	38 M
	en		1,048,374		37 M

Table 1: Number of extracted segments as source and target after segment aligning in the <title> and <claims> fields

Language pairs	Development set	Test Set
de-en	37.46	31.67
de-fr	35.41	28.72
en-de	43.16	36.01
en-fr	42.59	38.82
fr-en	44.12	42.61
fr-de	34.85	30.14

Table 2: BLEU scores of SMT systems



Figure 3: Interfaces of post-editing on SECTra

We store all monolingual sentences into html files, and add them into iMAG/SECTra. Pre-translation is provided by SMT systems, which are built with data extracted from the CLEF-IP 2011 collection. Figure 3 presents an example, where source sentences (de) are pre-translated (fr) by Moses and Google.

Figure 3 shows SECTra translation editor interface, similar to those of translation aids and commercial MT systems. It makes post-editing much faster than in the presentation context. Not yet post-edited segments can be selected, and global search-and-replace is available. All post-edited sentences are saved in a translation memory called CLEF-IP. When it becomes large enough after some period of using SECTra (about 10-15000 'good' bi-segments for the sublanguages of classical web sites), it can be used to build an empirical MT system for that sublanguage, and then to improve it incrementally as time goes and new segments are post-edited.

iMAG/SECTra also provides more languages

options for patent translation, such as Chinese, Hindi, or Arabic, using SMT or some online free web servers such as Google Translator, Systran, or Bing.

## 7. Support research on multilingual IR

Multilingual information search becomes important due to the growing amount of online information available in non-English languages and the rise of multilingual document collections. Query translation for CLIR became the most widely used technique to access documents in a different language from the query. For query translation, SMT is one way in which those powerful capabilities can be used (Oard, 1998). Our 3 SMT systems offer translation service by API. IR systems can use them directly. Due to robustness across domains and strong performance in translating named entities (like titles or short names), using SMT for CLIR can produce good results (Kürsten et al., 2009).

## 8. Conclusion and future work

In this paper, we gave an account of the extraction of a multilingual parallel corpus from the CLEF-IP 2011 collection. We first analyzed the structure of the patent documents of this collection and chose the fields to be extracted. To ensure the quality of parallel data, we cleaned them and aligned them with LF Aligner. The first version of the extracted patent parallel corpus consists of 3 languages, 6-language pairs, and is available in different formats (plain text files for Moses and TMX). This corpus is available to the research community. We also developed 3 specialized Moses-based SMT systems, from the TM resulting from the extraction process, and evaluated them, setting good BLEU scores on segments for which no translation was presented in the CLEF-IP 2011 files. We also transformed the initial collection of multilingual files into 3 collections of monolingual files, keeping only the source language text in each segment, and accessible in many languages using 3 dedicated iMAGs, and using the TM extracted from the original multilingual files. Multilingual access is provided by using our 3 Moses systems for the 3 corresponding language pairs, and other online free MT systems for the other language pairs.

One interesting perspective is the development of an infrastructure for the multilingual aspect of MUMIA-related research on patents. In the near future, we will setup a web service to support evaluation of the translation quality, both subjective (based on human judgments) and objective (task-related, such as post-editing time, or understanding time).

What has been done so far should enable researchers on CLIR applied to patents to include the multilingual aspect in their experiments. In future experiment, we plan to ask visitors of 3 websites to post-edit the MT "pre-translations". Interactive post-editing will transform the MT pre-translations of segments having no translation in the original CLEF-IP 2011 corpus into good translations, and the SMT systems will thus be incrementally improvable.

## 9. References

- Oard, Douglas W. (1998). A Comparative Study of Query and Document Translation for Cross-Language Information Retrieval, *Proceedings of the Third Conference of the Association for Machine Translation in the Americas on Machine Translation and the Information Soup*, p.472-483, October 28-31, 1998.
- Eisele, A., and Yu C. (2010). "MultiUN: A Multilingual Corpus from United Nation Documents". *Proceedings of the Seventh International Conference on Language Resources and Evaluation*.
- Federico, M., Bertoldi, N., and Cettolo, M. (2008), "IRSTLM: an Open Source Toolkit for Handling Large Scale Language Models", *Proceedings of Interspeech*, Brisbane, Australia, 2008.
- Kürsten, J., Wilhelm, T., and Eibl, M. (2009). The Xtrieval framework at CLEF 2008: domain-specific track. *Proceedings of CLEF*, pages 215–218, 2009.
- Koehn, P. (2005). "Europarl: A Parallel Corpus for Statistical Machine Translation". *Proceedings of Machine Translation Summit X*, Phuket, Thailand
- Koehn, P., Hoang, H., Birch, A., Callison-Birch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). "Moses: Open source toolkit for statistical machine translation". *Proceedings of the ACL 2007 Demo and Poster Sessions*, Prague, Czech Republic.
- Lu, B., Tsou, B.K., Zhu, J., Jiang, T. and Kwong, O.Y. (2009). "The construction of a Chinese-English patent parallel corpus". *Proceedings of the MT Summit XII*, Ottawa, Canada 2009.
- Och, F.J. (2003). Minimum error rate training in statistical machine translation, *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*. Sapporo, Japan, Volume 1, 160-167.
- Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. *In Proceedings of the Association of Computational Linguistics*, pp. 311–318.
- Ralf, S., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufis, D., and Varga, D. (2006). The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006)*. Genoa, Italy, 24-26 May 2006.
- Rubino, R., Huet, S., Lefèvre, F., and Linarès., G. (2012). Post-édition statistique pour l'adaptation aux domaines de spécialité en traduction automatique, *In Conférence en Traitement Automatique des Langues Naturelles*, pp. 527-534, Grenoble, France.
- Utiyama, M., and Isahara, H. (2007). A Japanese-English Patent Parallel Corpus. *Proceedings of MT Summit XI*.
- Varga, D., Halacsy, P., and et al. (2005). Parallel Corpora for Medium Density Languages. *RANLP 2005 Conference*.
- Wäschle, K., and Stefan, R. (2012). Analyzing Parallelism and Domain Similarities in the MAREC Patent Corpus. *Proceedings of the 5th Information Retrieval Facility Conference, IRFC 2012*, Vienna, Austria, July 2-3, 2012 12–27.
- Wang, L., and Boitet, C. (2013), Online production of HQ parallel corpora and permanent task-based evaluation of multiple MT systems: both can be obtained through iMAGs with no added cost. *Proceedings of MT Summit XIV, The 2nd Workshop on Post-Editing Technologies and Practice*. Nice, France 2 - 6 September 2013.

# Comparability of Corpora in Human and Machine Translation

Ekaterina Lapshinova-Koltunski & Santanu Pal

Saarland University  
Universität Campus A2.2,  
66123 Saarbrücken, Germany  
e.lapshinova@mx.uni-saarland.de, santanu.pal@uni-saarland.de

## Abstract

In this study, we demonstrate a negative result from a work on comparable corpora which forces us to address a problem of comparability in both human and machine translation. We state that it is not always defined similarly, and comparable corpora used in contrastive linguistics or human translation analysis cannot always be applied for statistical machine translation (SMT). So, we revise the definition of comparability and show that some notions from translatology, i.e. registerial features, should also be considered in machine translation (MT).

**Keywords:** comparable corpora, paraphrases, machine translation, register analysis, registerial features

## 1. Introduction

Numerous studies and applications in both linguistic and language engineering communities use comparable corpora as essential resources, e.g. to compare phenomena across languages or to acquire parallel resources for training in statistical Natural Language Processing (NLP) applications, e.g. statistical machine translation.

Due to the fact that parallel corpora remain a scarce resource (despite the creation of automated methods to collect them from the Web) and often cover restricted domains only (political speeches, legal texts, news, etc.), comparable corpora have been used as a valuable source of parallel components in SMT, e.g. as a source for parallel fragment of texts, paraphrases or sentences (Smith et al., 2010).

In contrast to parallel corpora, which contain originals and their translations, comparable corpora can contain originals only, or translations only, and can thus be defined as a collection of texts with the same sampling frame and similar representativeness (McEnery, 2003). For example, they may contain the same proportions of the texts belonging to the same genres, or the same domains in a range of different languages.

However, the concept of 'comparable corpora' may differ depending on which measure is taken into account (register or domain), and what are the purposes of the analysis. In this paper, we present an experiment which demonstrates that comparability in human translation studies does not always coincide with what is understood under comparability in machine translation.

The remainder of the paper is structured as follows. In section 2., we outline the aims and the motivation of the present study. Section 3. presents related work on comparable corpora, the clarification of the notions of domain and register, as well as their definition applied in this work. Section 4. describes the resources at hand and the applied methodology. Here, we describe the resources at hand, and the methods used. In section 5., we show the results, and discuss the problems we face.

## 2. Aims and Motivation

The original aim of our experiment was to enhance the resources available for machine translation with the help of

a paraphrase extraction from both parallel and comparable corpora at hand. The extracted paraphrases can then be used to improve statistical machine translation, as it was done in our previous studies. For example, in (Pal et al., 2013), multi-word expressions (MWE) were extracted from comparable corpora aligned on document level. These were aligned and used for the improvement in English-Bengali Phrase-Based SMT (PB-SMT) by incorporating them directly and indirectly into the phrase table. In another study, n-gram overlapping parallel fragment of texts were extracted from comparable corpora to serve as an additional resource to improve a baseline PB-SMT system, see (Gupta et al., 2013). Another possible application of such paraphrases is acquisition of parallel and comparable data from the web, which can also be used for MT enhancement.

For this experiment, we decide for English-German resources consisting of two parts: a baseline created for a PB-SMT system, and an existing comparable corpus, which was originally compiled to serve human translation tasks. Hence, comparability of its texts was stated according to criteria used in translatology, see sections 3.2. and 4.1. below.

The texts of the corpus belong to two genres – political speeches and popular science. The choice of these datasets for our experiment is motivated by the difference in the availability of resources. Whereas extensive parallel resources are available for political speeches, it is difficult to find parallel resources for popular-scientific texts. Therefore, we decide to apply procedures for both datasets, as on the one hand, we hope to enhance the resources available (improving machine translation with paraphrases), and on the other hand, we want to test how our procedures work on a dataset different to what is commonly used, e.g. news articles or political speeches.

Moreover, these two datasets are different not only in the amount of parallel resources available. They also differ in the correlation of the notions of domain vs. genre/register. In political speeches, the notion of domain correlates more with that of register, whereas in popular scientific texts, it doesn't. Therefore, we observed different results in the application of our procedures, which make us address the problem of corpus comparability in translation.

### 3. Related Work and Theoretical Issues

#### 3.1. Comparable corpora

**Comparable corpora in MT** As already mentioned above, comparable corpora have become widely used in NLP, contrastive language analysis and translatology. In NLP, they found application in the development of bilingual lexicons or terminology databases, e.g. in (Chiao and Zweigenbaum, 2002; Fung and Cheung, 2004) or (Gaussier et al., 2004) and in cross-language information research, see e.g. (Grefenstette, 1998) or (Chen and Nie, 2000), as well as MT improvement, e.g. (Munteanu and Marcu, 2005) or (Eisele and Xu, 2010).

The methods used in these approaches are mostly based on context similarity: the same concept tends to appear with the same context words in both languages, the hypothesis that is also used for the identification of synonyms. Several earlier studies have shown that there is a correlation between the co-occurrences of words which are translations of each other in any language (Rapp, 1999) and that the associations between a word and its context seed words are preserved in comparable texts of different languages, cf. (Fung and Yee, 1998).

In most cases, the starting point is a list of bilingual “seed expressions” required to build context vectors of all words in both languages. This is either provided by an external bilingual dictionaries or databases, as in (Déjean et al., 2002), or is extracted from a parallel corpus, as in (Otero, 2007). We also start with a list of “seed expressions”, which are paraphrases in our case. They are extracted from a bilingual parallel corpus, and enhanced with paraphrases from a comparable corpus.

There are similar works with the application for automatic extraction of terms, e.g. in (Chiao and Zweigenbaum, 2002) and (Saralegi et al., 2008). The authors used specialised comparable corpora, e.g. English-French corpora in medical domain, or English-Basque corpora in popular science, for automatic extraction of bilingual terms. In both cases, comparability is accounted for by the distribution of topics (or also publication dates).

**Comparable corpora and comparability** In most works, comparability is correlated with the comparability of potential word equivalents and their contexts or collocates, which is reasonable for bilingual terminology extraction task. Although these criteria might be sufficient for creation of multilingual lexicons or terminology databases, translation of whole texts involve more influencing factors, as more levels of description, i.e. conventions of a register a text belongs to are at play. In translation studies, which are concerned with human translations, as well as human translator training, these aspects take on an important role. While translating a text from one language into another, a translator must consider the conventions of the text type to be translated.

In existing MT studies these conventions (specific register features) have not been taken into account so far. Describing comparable data collected for training, authors consider solely domains, i.e. topics described in the collected texts, ignoring the genre or the register of these texts. We claim that register features should also be considered in the defi-

nition of a comparable corpus in MT, as they are in human translation.

In the following, we define the notions of genre, register and domain, as well as their role in the definition of comparability in our analysis.

#### 3.2. Genre, Register and Domain

We consider multilingual corpora comparable if they contain texts which belong to the same register.

In our analysis, we use the term *register*, and not *genre*, although they represent two different points of view covering the same ground, see e.g. (Lee, 2001). However, we refer to genre when speaking about a text as a member of a cultural category, about a register when we view a text as language, its lexico-grammatical characterisations, conventionalisation and functional configuration of a language which are determined by a context use situation, variety of language means according to this situation. Different situations require different configurations of a language.

This kind of register definition is used in human translation studies, e.g. corpus-based approaches as in (Teich, 2003; Steiner, 2004; Hansen Schirra et al., 2013; Neumann, 2013), and coincides with the one formulated in register theory, e.g. in (Quirk et al., 1985; Halliday and Hasan, 1989; Biber, 1995). In their terms, registers are manifested linguistically by particular distributions of lexico-grammatical patterns, which are situation-dependent. The canonical view is that situations can be characterised by the parameters of *field*, *tenor* and *mode* of discourse. Field of discourse relates to processes and participants (e.g., Actor, Goal, Medium), as well as circumstantials (Time, Place, Manner etc.) and is realised in lexico-grammar in lexis and colligation (e.g. argument structure). Tenor of discourse relates to roles and attitudes of participants, author-reader relationship, which are reflected in stance expressions or modality. Mode of discourse relates to the role of the language in the interaction and is linguistically reflected at the grammatical level in Theme-Rheme constellations, as well as cohesive relations at the textual level. So, the contextual parameters of registers correspond to sets of specific lexico-grammatical features, and different registers vary in the distribution of these features.

The definition of domain is also present in register analysis. Here, it is referred to as *experiential domain*, or what a text is about, its topic. Experiential domain is a part of the context parameter of field, which is realised in lexis, as already mentioned above. However, it also includes colligation, in which also grammatical categories are involved. So, domain is just one of the parameter features a register can have. Some NLP studies, e.g. those using web resources, do claim the importance of register or genre conventions, see e.g. (Santini et al., 2010). However, to our knowledge, register or genre features remain out of the focus in machine translation. Whereas there exist some works on domain adaptation, e.g. adding bilingual data to the training material of SMT systems, as in (Eck et al., 2004), or (Wu et al., 2008) and others, register features are mostly ignored. In human translator training, on the contrary, the knowledge on lexico-grammatical preferences of registers plays an important role. A human translator learns to analyse

texts according to the register parameters both in a source and in a target language.

## 4. Resources and Methodology

### 4.1. Resources at hand

In our experiment, we use two types of dataset: (1) a big English-German parallel training corpus; (2) a small English-German comparable corpus. The first one is based on the English-German component of EUROPARL<sup>1</sup> (Koehn, 2005), used to build the baseline system and to create the initial paraphrase table, see section 4.3. below. The other dataset (2) is used for the enhancement of this paraphrase table. This dataset was extracted from the multilingual corpus CroCo (Hansen Schirra et al., 2013), which contains English and German texts, belonging to the same register. As already mentioned above, we decide for the registers of political speeches (SPEECH) and popular science (POPSCI), see section 1.

**Data selection** The texts in the corpus are selected according to the criteria of register analysis as defined in 3.2. above. According to the general register analysis, SPEECH belongs to the communication of an 'expert to expert' in a formal social distance, whereas the latter is rather 'expert to layperson' in a causal social distance. Both express an equal social and a constitutive language role. For popular-scientific texts in both languages, it is essential that texts are perceived as pleasurable, and not only informative reading. This means that author-reader relationship (the contextual parameter of tenor) is very important in this register, see (Kranich et al., 2012).

English originals (EO) in SPEECH are collected from the US public diplomacy and embassy web services, whereas German texts (GO) originate from German governmental, ministry and president websites. Both EO and GO texts have 'exposition', 'persuasion' and 'argumentation' as goal orientation, 'expert to expert' as agentive role, and include information on economic development, human security and other issues in both internal, foreign or global perspective. Both EO and GO texts in POPSCI originate from popular-scientific articles, which have 'exposition' as goal orientation, 'expert to layperson' as agentive role. The information in the articles are on psychotherapy, biology, chemistry and others.

Although no attention was paid to the parallelity of topics discussed in both corpora (which could mean that their domains do not necessarily coincide), English and German registers are comparable along other features. Moreover, they have a number of commonalities in English and German. For example, popular-scientific texts show preference for particular process types, e.g. relation processes (expressed by transitivity), underspecified Agent (expressed by extensive use of passive constructions), and others in both languages (Teich, 2003).

**Data processing** We used Stanford Parser, see (Socher et al., 2013; Rafferty and Manning, 2008), and Stanford NER<sup>2</sup> for parsing and named entity tagging for the EO

and GO texts. The experiments were carried out with the help of the standard log-linear PB-SMT model as baseline: GIZA++ implementation of IBM word alignment model 4, phrase-extraction heuristics as described in (Koehn et al., 2003), minimum-error-rate training (Och, 2003) on a held-out development set, target language model trained with the SRILM toolkit (Stolcke, 2002) with Kneser-Ney smoothing (Kneser and Ney, 1995) and the Moses decoder (Koehn et al., 2007).

### 4.2. Paraphrase extraction

We start our experiment with the identification of paraphrases from the English-German parallel training corpus, (1) in section 4.1. above.

Paraphrase is a phrase or an idea that can be represented or expressed in different ways in the same language by preserving the meaning of that phrase or idea. Paraphrases can be collected from parallel corpora as well as from comparable corpora. Extraction of parallel fragments of texts, sentences and paraphrases from comparable corpora is particularly useful for any corpus-based approaches to MT, especially for SMT (Gupta et al., 2013). Paraphrases can be used to alleviate the sparseness of training data (Callison-Burch et al., 2006), to handle Out Of Vocabulary (OOV) words, as well as to expand the reference translations in automatic MT evaluation (Denoual and Lepage, 2005; Kauchak and Barzilay, 2006). Moreover, in SMT, the size of the parallel corpus plays a crucial role in the SMT performance. However, large volume of parallel data is not available for all language pairs or all text types (see section 1.).

A significant number of works have been carried out on paraphrasing. A full-sentence paraphrasing technique was introduced by (Madnani et al., 2007). They demonstrated that the resulting paraphrases can be used to drastically reduce the number of human reference translations needed for parameter tuning without a significant decrease in translation quality. (Fujita and Carpuat, 2013) describe a system that was built using baseline PB-SMT system. They augmented the phrase table with novel translation pairs generated by combining paraphrases where these translation pairs were learned directly from the bilingual training data. They investigated two methods for phrase table augmentation: source-side augmentation and target-side augmentation. (Aziz and Specia, 2013) report the mining of sense-disambiguated paraphrases by pivoting through multiple languages. (Barzilay and McKeown, 2001) proposed an unsupervised learning algorithm for identification of paraphrases from a corpus of multiple English translations of the same source text. A new and unique paraphrase resource was reported by (Xu et al., 2013), which contains meaning-preserving transformations between informal user-generated texts. Sentential paraphrases are extracted from a comparable corpus of (temporally and topically related) messages in Twitter which often express semantically identical information through distinct surface forms. A novel paraphrase fragment pair extraction method was proposed by (Wang and Callison-Burch, 2011) in which the authors used a monolingual comparable corpus containing different articles about the same topics or events.

<sup>1</sup>the 7th Release v7 of EUROPARL

<sup>2</sup><http://nlp.stanford.edu/software/CRF-NER.shtml>

The procedure consisted of document, sentence and fragment pair extraction.

Our approach is similar to the identification technique used by (Bannard and Callison-Burch, 2005). In our study, identification of paraphrases has been carried out by pivoting through phrases from the bilingual parallel corpus (1). We consider all phrases in the phrase table as potential candidates for paraphrasing.

After extraction of potential paraphrase pairs, we compute the likelihood of them being paraphrases. For a potential paraphrase pair  $(e_1, e_2)$  we have defined a paraphrase probability  $p(e_2|e_1)$  in terms of the translation model probabilities  $p(f|e_1)$ , that the original English phrase  $e_1$  is translated as a particular target language phrase  $f$ , and  $p(e_2|f)$ , that the candidate paraphrase  $e_2$  is translated as the same foreign language phrase  $f$ . Since  $e_1$  can be translated to multiple foreign language phrases, we sum over all such foreign language phrases. Thus the equation reduces to as follows:

$$\hat{e}_2 = \arg \max_{e_2 \neq e_1} P(e_2|e_1) \quad (1)$$

$$= \arg \max_{e_2 \neq e_1} \sum_f P(f|e_1)P(e_2|f) \quad (2)$$

We compute translation model probabilities using standard formulation from PB-SMT. So, the probability  $p(e|f)$  is calculated by counting how often the phrases  $e$  and  $f$  were aligned in the parallel corpus as follows :

$$p(e|f) = \frac{\text{count}(e, f)}{\sum_f \text{count}(e, f)} \quad (3)$$

Using the equation (2) and (3) we calculate paraphrase probabilities from the phrase table.

### 4.3. Incorporation of paraphrases into PB-SMT System

The next step is to create additional training material using these extracted paraphrases. We initially found and marked the paraphrases in the source English sentences within the training data and then replaced each English paraphrase with all of its other variants, gradually creating more training instances. For example, consider the English phrase “throughout the year” and its two paraphrases “all year round” and “all around the year”. Now we consider following sentences from our training data for each of these phrase and paraphrases.

- (1) a. Events, parties and festivals occur throughout the year and across the country.
- b. Weather on all of the Hawaiian islands is very consistent, with only moderate changes in temperature all year round.
- c. There is an intense agenda all around the year and the city itself is a collection of art and history.

In example (1), the first sentence, the phrase “throughout the year” is replaced by its two paraphrases “all year round” and “all around the year” to create two additional sentences to be added to the existing training data. Similarly “all year

round” and “all around the year” are replaced by the remaining two variants for the second and third sentence, respectively.

In this way, for these three training sentences, we can create six additional sentences from all combinations of replacement. Combining these additional resources with the existing training data, we enhance the existing baseline of the PB-SMT system.

We decode English original (EO) sentences from both SPEECH and POPSCI through our enhanced English-German PB-SMT system. The density of population of words for GO with respect to EO are measured through the decoded output provided by the enhanced system. The population measure is defined as how many translated German word words are corresponding to the GO words by measuring distance between them. For this, we use the following distance measure techniques: *Minimum Edit Distance Ratio* (MEDR) and *Longest Common Subsequence Ratio* (LCSR). Let,  $|W|$  be the length of the string  $W$  and  $ED$  is the minimum edit distance or levenshtein distance calculated as the minimum number of edit operations such as insert, replace, delete – needed to transform  $W_1$  into  $W_2$ .

The definition of the Minimum Edit Distance Ratio is given in (4), and the definition of Longest Common Subsequence Ratio in (5).

$$MEDR(W_1, W_2) = 1 - \frac{|ED(W_1, W_2)|}{\max(|W_1|, |W_2|)} \quad (4)$$

$$LCSR(W_1, W_2) = \frac{|LCS(W_1, W_2)|}{\max(|W_1|, |W_2|)} \quad (5)$$

The training corpus was filtered with the maximum allowable sentence length of 100 words and sentence length ratio of 1:2 (either way). In the end, the training corpus contained 1,902,223 sentences. In addition to the target, side monolingual German corpus containing 2,176,537 sentences from EUROPARL was used for building the target language model. We experimented with different n-gram settings for the language model and the maximum phrase length and found that a 5-gram language model and a maximum phrase length of 7 produced the optimum baseline result.

This baseline is now to be enhanced with additional paraphrases from comparable corpora at hand, which we describe in the following section.

### 4.4. Analysis of comparable corpora

To expand the paraphrase table, we first perform manual comparison of each corresponding comparable file in terms of token and part-of-speech (POS) alignment.

Then, we analyse density with the help of named entities (NE). Named entities are identified on both EO and GO sentences separately with the help of English and German Stanford NER. So, using NEs we prove the comparability between the comparable parts of the corpus, i.e we check whether NEs are present on both its sides (English and German). We follow the same word similarity technique: MEDR and LCSR, as described in section 4.3. above. The comparability has been measured according to the population density (how many NEs correspond between the EO and GO) on both side of the comparable corpus.

## 5. Experiment Results

### 5.1. Comparison results

In tables 1 and 2, we present the results of the comparison for texts from the analysed corpus, including the total number of tokens (token) and NEs, as well as their population (pop) and population density (pop.dens) calculated as populated tokens/AVG (the sum of total EO and GO tokens), see section 4.3. for details.

	EO	GO	pop	pop.dens
<b>token</b>	13906	14598	5729	0.40
<b>NE</b>	369	263	8	0.02

Table 1: Similarities between EO and GO in POPSCI

	EO	GO	pop	pop.dens
<b>token</b>	9753	7094	3969	0.47
<b>NE</b>	387	297	149	0.43

Table 2: Similarities between EO and GO in SPEECH

Our results show that token alignment in SPEECH is much more reliable than that in POPSCI. The same results are obtained on the POS level: the total number of nouns are more probably matching between the comparable files in SPEECH. Moreover, we found more population density in the SPEECH data, if compared with the data in POPSCI.

This means that whereas we can prove the comparability of EO and GO in SPEECH using these measuring techniques, we are not able to do the same for POPSCI. Hence, we cannot extract paraphrases from the comparable corpus of POPSCI texts at hand. This shows that our method of paraphrase enhancement with the data from comparable corpora does not work with all types comparable corpora.

The reason for it is the nature of the comparable data. On the one hand, English and German texts are comparable in POPSCI if register settings in both languages are considered. On the other hand, they are not necessarily comparable in their domains. At the same time, SPEECH, which was also set up under same conditions of register analysis, seem to be comparable in both aspects. We assume that the notion of domain in SPEECH correlates with that of register, whereas in popular science it doesn't.

### 5.2. Discussion

Facing the negative results of our experiment, we decide to revise the notion of comparability, which does not always correspond in machine translation and in human translation. Defining comparability criteria for corpora, these scientific communities have often two different things in mind: (1) register in human translation (register-oriented perspective), (2) domain in machine translation (domain-oriented perspective). We assume that the relation between these two perspectives is inclusive: domain definition is implied in the register analysis as a part of 'experiential domain definition'. This is confirmed by the results of our experiment which demonstrates that in some cases, the definition of domain and register coincide. For instance, in political speeches, experiential domain is not that diverse as in popular-scientific texts, and thus, the texts identified as

comparable according to the register-oriented perspective, are also comparable in terms of the domain-oriented perspective.

At the same time, if we define corpora as being comparable along the domain-oriented criterion only, they would not necessarily be comparable from the register-oriented perspective. For instance, for human translation, news reporting on certain political topics cannot be comparable with political speeches discussing the same topics as in the news texts. The latter would lack 'persuasion' and 'argumentation' in their as goal orientation, as well as 'expert to expert' as agentive role, which would be reflected in their lexicogrammatical features.

We believe that both perspectives are important for translation (both human and machine). The first one has an impact on the lexical level, e.g. terminology or general vocabulary used in a translated text. The other is important for lexicogrammar, i.e. morpho-syntactic preferences of registers and their textual properties, e.g. cohesive phenomena and information structure. Therefore, we claim that there is a need to define new measures of corpus comparability in translation, which can be measured e.g. by homogeneity<sup>3</sup>, and would consider both domain and further registerial features.

In MT studies this problem has not been addressed so far. To our knowledge, none of the existing MT studies integrate register features. As a result, machine-translated texts would (not) have features characteristic for the register they belong to. For example, German popular-scientific texts can be characterised by a high number of passive constructions, see section 3.2. above. We calculate the ratio of passive constructions<sup>4</sup> in German originals and compare it to the passive ratio in German translations from English, considering human (HU) and a statistical machine translation (SMT)<sup>5</sup>. Whereas human translations demonstrate a similar proportion of passives as in comparable originals, machine translations seem to underuse this verb construction type.

corpus	ratio
<b>GO</b>	6.62
<b>HU</b>	6.98
<b>SMT</b>	3.10

Table 3: Passive verb constructions in POPSCI

Undoubtedly, we need to test more features to come to the final conclusion about the impact of registerial features on the translation output. However, it was not the original aim of the present paper. Moreover, we need to expand the parallel training corpus with additional genre to show possible differences in the resulting models. For future work, we also plan to experiment with another approach on MT enhancement, e.g. the one described in (Munteanu and Marcu, 2005).

However, the negative results of our experiments made us raise the questions about (1) comparability, and (2) ad-

<sup>3</sup>see work on homogeneity measure by (Kilgarriff, 2001).

<sup>4</sup>We calculate the ratio of passives in all final verb constructions.

<sup>5</sup>the translations are available in VARTRA, see (Lapshinova-Koltunski, 2013).

ditional features which could have impact on translation, which we address to both communities and aim to raise a discussion in these issues.

## Acknowledgments

The research leading to these results has received funding from the EU project EXPERT – the People Programme (Marie Curie Actions) of the European Union’s Seventh Framework Programme FP7/2007-2013 under REA grant agreement no. [317471]. The resources available were provided within the project VARTRA supported by a grant from Forschungsausschuß of Saarland University.

## 6. References

- W. Aziz and L. Specia. 2013. Multilingual WSD-like Constraints for Paraphrase Extraction. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 202–211, Sofia, Bulgaria.
- C. Bannard and C. Callison-Burch. 2005. Paraphrasing with Bilingual Parallel Corpora. In *Proceedings of ACL-2005*, pages 597–604.
- R. Barzilay and K.R. McKeown. 2001. Extracting Paraphrases from a Parallel Corpus. In *Proceedings of ACL-2001*, pages 50–57.
- D. Biber. 1995. *Dimensions of Register Variation. A Cross-linguistic Comparison*. Cambridge University Press, Cambridge.
- C. Callison-Burch, P. Koehn, and M. Osborne. 2006. Improved Statistical Machine Translation Using Paraphrases. In *Proceedings of the Main Conference on HLT-NAACL-2006*, pages 17–24.
- J. Chen and J.-Y. Nie. 2000. Parallel web text mining for cross-language Ir. In *Proceedings of RIAO 2000: Content-Based Multimedia Information Access*, volume 1, pages 62–78, Paris.
- Y. Chiao and P. Zweigenbaum. 2002. Looking for Candidate Translational Equivalents in Specialized, Comparable Corpora. In *Proceedings of the 19th International Conference on Computational Linguistics - Volume 2, COLING-02*, pages 1–5.
- H. Déjean, É. Gaussier, and F. Sadat. 2002. Bilingual Terminology Extraction: An Approach Based on a Multilingual Thesaurus Applicable Comparable Corpora. In *Proceedings of the 19th International Conference on Computational Linguistics, COLING-02*.
- E. Denoual and Y. Lepage. 2005. Bleu in characters: towards automatic Mt evaluation in languages without word delimiters. In *The Second International Joint Conference on Natural Language Processing*, pages 81–86.
- M. Eck, S. Vogel, and A. Waibel. 2004. Improving statistical machine translation in the medical domain using the unified medical language system. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING-2004)*, pages 792–798, Geneva, Switzerland.
- A. Eisele and J. Xu. 2010. Improving Machine Translation Performance Using Comparable Corpora. In *Proceedings of the 3rd Workshop on Building and Using Comparable Corpora*, pages 35–41, Malta. LREC-2010.
- A. Fujita and M. Carpuat. 2013. Fun-nrc: Paraphrase-augmented Phrase-based Smt Systems for Ntcir-10 Patentmt. In *The 10th NTCIR Conference*, Tokyo, Japan.
- P. Fung and P. Cheung. 2004. Mining Verynon-parallel Corpora: Parallel Sentence and Lexicon Extraction via Bootstrapping and Em. In *Proceedings of EMNLP*, pages 57–63.
- P. Fung and L.Y. Yee. 1998. An IR Approach for Translating New Words from Nonparallel, Comparable Texts. In *Proceedings of the 17th International Conference on Computational Linguistics*, volume volume 1 of *COLING-98*, pages 414–420.
- E. Gaussier, J.-M. Renders, I. Matveeva, C. Goutte, and H. Djean. 2004. A geometric view on bilingual lexicon extraction from comparable corpora. In *Proceedings of ACL-04*, pages 527–534.
- G. Grefenstette. 1998. *Cross-Language Information Retrieval*. Kluwer Academic Publishers, London.
- R. Gupta, S. Pal, and S. Bandyopadhyay. 2013. Improving Mt System Using Extracted Parallel Fragments of Text from Comparable Corpora. In *Proceedings of the 7th Workshop on Building and Using Comparable Corpora*, pages 69–76, Sofia, Bulgaria.
- MAK Halliday and R. Hasan. 1989. *Language, context and text: Aspects of language in a social semiotic perspective*. Oxford University Press.
- S. Hansen Schirra, S. Neumann, and E. Steiner. 2013. *Cross-linguistic Corpora for the Study of Translations. Insights from the Language Pair English-German*. de Gruyter, Berlin, New York.
- D. Kauchak and R. Barzilay. 2006. Paraphrasing for Automatic Evaluation. In *Proceedings of the Main Conference on HLT-NAACL-2006*, pages 455–462.
- A. Kilgariff. 2001. Comparing corpora. *International Journal of Corpus Linguistics*, 6(1):1–37.
- R. Kneser and H. Ney. 1995. Improved backing-off for n-gram language modeling. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume I, pages 181–184, Detroit, Michigan.
- P. Koehn, F. J. Och, and D. Marcu. 2003. Statistical phrase-based translation. In *Proceedings of NAACL-2003*, volume 1, pages 48–54.
- P. Koehn, H. Hoang, A. Birch, C. Callison Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of ACL-2007*, pages 177–180.
- P. Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT.
- S. Kranich, J. House, and V. Becher. 2012. Changing conventions in English-german translations of popular scientific texts. In Kurt Braunmüller and Christoph Gabriel, editors, *Multilingual Individuals and Multilingual Societies*, volume 13 of *Hamburg Studies on Multilingualism*, pages 315–334. John Benjamins.
- E. Lapshinova-Koltunski. 2013. VARTRA: A Compar-



- ble Corpus for Analysis of Translation Variation. In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*, pages 77–86, Sofia, Bulgaria. Association for Computational Linguistics.
- D. Y. Lee. 2001. Genres, registers, text types, domains and styles: clarifying the concepts and navigating a path through the bnc jungle. *Technology*, 5:37–72.
- N. Madnani, N.F. Ayan, P. Resnik, and B.J. Dorr. 2007. Using Paraphrases for Parameter Tuning in Statistical Machine Translation. In *Proceedings of the Second Workshop on StatMT*, pages 120–127.
- T. McEnery. 2003. Corpus Linguistics. In Ruslan Mitkov, editor, *Oxford Handbook of Computational Linguistics*, Oxford Handbooks in Linguistics, pages 448–463. Oxford University Press, Oxford.
- D. S. Munteanu and D. Marcu. 2005. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31:477–504.
- S. Neumann. 2013. *Contrastive Register Variation: A Quantitative Approach to the Comparison of English and German*. Trends in Linguistics. Studies and Monographs [Tilsm]. Walter de Gruyter.
- P. G. Otero. 2007. Learning Bilingual Lexicons from Comparable English and Spanish Corpora. In *Proceedings of MT Summit XI*, pages 191–198.
- S. Pal, S. K. Naskar, and S. Bandyopadhyay. 2013. MWE Alignment in Phrase Based Statistical Machine Translation. In *Proceedings of the Machine Translation Summit XIV*, pages 61–68, Nice, France.
- R. Quirk, S. Greenbaum, G. Leech, and J. Svartvik. 1985. *A Comprehensive Grammar of the English Language*. Longman, London.
- Anna Rafferty and Christopher D. Manning. 2008. Parsing Three German Treebanks: Lexicalized and Unlexicalized Baselines. In *ACL Workshop on Parsing German*.
- R. Rapp. 1999. Automatic Identification of Word Translations from Unrelated English and German Corpora. In *Proceedings of the 37th ACL*.
- M. Santini, A. Mehler, and S. Sharoff. 2010. Riding the rough waves of genre on the web. In A. Mehler, S. Sharoff, and M. Santini, editors, *Genres on the Web: Computational Models and Empirical Studies*, pages 3–30. Springer.
- X. Saralegi, I. S. Vicente, and A. Gurrutxaga. 2008. Automatic extraction of bilingual terms from comparable corpora in a popular science domain. In *Proceedings of the 1st Workshop on Building and Using Comparable Corpora*, Marrakesh. LREC-2008.
- J. R. Smith, C. Quirk, and K. Toutanova. 2010. Extracting parallel sentences from comparable corpora using document level alignment. In *Proceedings of Human Language Technologies: The 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT-10)*, pages 403–411.
- R. Socher, J. Bauer, C.D. Manning, and A.Y. Ng. 2013. Parsing With Compositional Vector Grammars. In *Proceedings of ACL-2013*, Sofia, Bulgaria.
- E. Steiner. 2004. *Translated texts: Properties, Variants, Evaluations*. Peter Lang, Frankfurt a. Main.
- A. Stolcke. 2002. Srilm—an extensible language modeling toolkit. In *Proceedings International Conference on Spoken Language Processing*, pages 257–286.
- E. Teich. 2003. *Cross-linguistic Variation in System and Text. A Methodology for the Investigation of Translations and Comparable Texts*. Mouton de Gruyter, Berlin and New York.
- R. Wang and C. Callison-Burch. 2011. Paraphrase Fragment Extraction from Monolingual Comparable Corpora. In *4th Workshop on Building and Using Comparable Corpora*, Portland, Oregon.
- H. Wu, H. Wang, and C. Zong. 2008. Domain adaptation for statistical machine translation with domain dictionary and monolingual corpora. In D. Scott and H. Uszkoreit, editors, *Proceedings of the 22nd International Conference on Computational Linguistics (COLING-2008)*, pages 993–1000, Manchester, UK.
- W. Xu, A. Ritter, and R. Grishman. 2013. Gathering and Generating Paraphrases from Twitter with Application to Normalization. In *7th Workshop on Building and Using Comparable Corpora*, Sofia, Bulgaria.

# Identifying Japanese-Chinese Bilingual Synonymous Technical Terms from Patent Families

Zi Long<sup>†</sup> Lijuan Dong<sup>†</sup> Takehito Utsuro<sup>†</sup> Tomoharu Mitsuhashi<sup>‡</sup> Mikio Yamamoto<sup>†</sup>

<sup>†</sup>Graduate School of Systems and Information Engineering, University of Tsukuba, Tsukuba, 305-8573, Japan

<sup>‡</sup>Japan Patent Information Organization, 4-1-7, Toyo, Koto-ku, Tokyo, 135-0016, Japan

## Abstract

In the task of acquiring Japanese-Chinese technical term translation equivalent pairs from parallel patent documents, this paper considers situations where a technical term is observed in many parallel patent sentences and is translated into many translation equivalents and studies the issue of identifying synonymous translation equivalent pairs. First, we collect candidates of synonymous translation equivalent pairs from parallel patent sentences. Then, we apply the Support Vector Machines (SVMs) to the task of identifying bilingual synonymous technical terms, and achieve the performance of over 85% precision and over 60% F-measure. We further examine two types of segmentation of Chinese sentences, i.e., by characters and by morphemes, and integrate those two types of segmentation in the form of the intersection of SVM judgments, which achieved over 90% precision.

**Keywords:** synonymous technical terms, patent families, technical term translation

## 1. Introduction

For both high quality machine and human translation, a large scale and high quality bilingual lexicon is the most important key resource. Since manual compilation of bilingual lexicon requires plenty of time and huge manual labor, in the research area of knowledge acquisition from natural language text, automatic bilingual lexicon compilation have been studied. Techniques invented so far include translation term pair acquisition based on statistical co-occurrence measure from parallel sentences (Matsumoto and Utsuro, 2000), translation term pair acquisition from comparable corpora (Fung and Yee, 1998), compositional translation generation based on an existing bilingual lexicon for human use (Tonoike et al., 2006), and translation term pair acquisition by collecting partially bilingual texts through the search engine (Huang et al., 2005).

Among those efforts of acquiring bilingual lexicon from text, Morishita et al. (2008) studied to acquire Japanese-English technical term translation lexicon from phrase tables, which are trained by a phrase-based SMT model with parallel sentences automatically extracted from parallel patent documents. In more recent studies, they require the acquired technical term translation equivalents to be consistent with word alignment in parallel sentences and achieved 91.9% precision with almost 70% recall. Furthermore, based on the achievement above, Liang et al. (2011a) considered situations where a technical term is observed in many parallel patent sentences and is translated into many translation equivalents. More specifically, in the task of acquiring Japanese-English technical term translation equivalent pairs, Liang et al. (2011a) studied the issue of identifying Japanese-English synonymous translation equivalent pairs. First, they collect candidates of synonymous translation equivalent pairs from parallel patent sentences. Then, they apply the Support Vector Machines (SVMs) (Vapnik, 1998) to the task of identifying bilingual synonymous technical terms.

Based on the technique and the results of identifying Japanese-English synonymous translation equivalent pairs

in Liang et al. (2011a), we aim at identifying Japanese-Chinese synonymous translation equivalent pairs from Japanese-Chinese patent families. We especially examine two types of segmentation of Chinese sentences, namely, by characters and by morphemes. Although both types of segmentation achieved almost similar performance around 95~97% (in recall / precision / f-measure) in the task of acquiring Japanese-Chinese technical term translation pairs, they have different types of errors. Also in the task of identifying Japanese-Chinese synonymous technical terms, both types of segmentation achieved almost similar performance, while they have different types of errors. Thus, we integrate those two types of segmentation in the form of the intersection of SVM judgments, and show that this achieves over 90% precision.

## 2. Japanese-Chinese Parallel Patent Documents

Japanese-Chinese parallel patent documents are collected from the Japanese patent documents published by the Japanese Patent Office (JPO) in 2004-2012 and the Chinese patent documents published by State Intellectual Property Office of the People's Republic of China (SIPO) in 2005-2010. From them, we extract 312,492 patent families, and the method of Utiyama and Isahara (2007) is applied<sup>1</sup> to the text of those patent families, and Japanese and Chinese sentences are aligned. In this paper, we use 3.6M parallel patent sentences with the highest scores of sentence alignment.

## 3. Phrase Table of an SMT Model

As a toolkit of a phrase-based SMT model, we use Moses (Koehn et al., 2007) and apply it to the whole 3.6M parallel patent sentences. Before applying Moses, Japanese sentences are segmented into a sequence of morphemes by the Japanese morphological analyzer MeCab<sup>2</sup> with the

<sup>1</sup>Here, we used a Japanese-Chinese translation lexicon consisting of about 170,000 Chinese head words.

<sup>2</sup><http://mecab.sourceforge.net/>

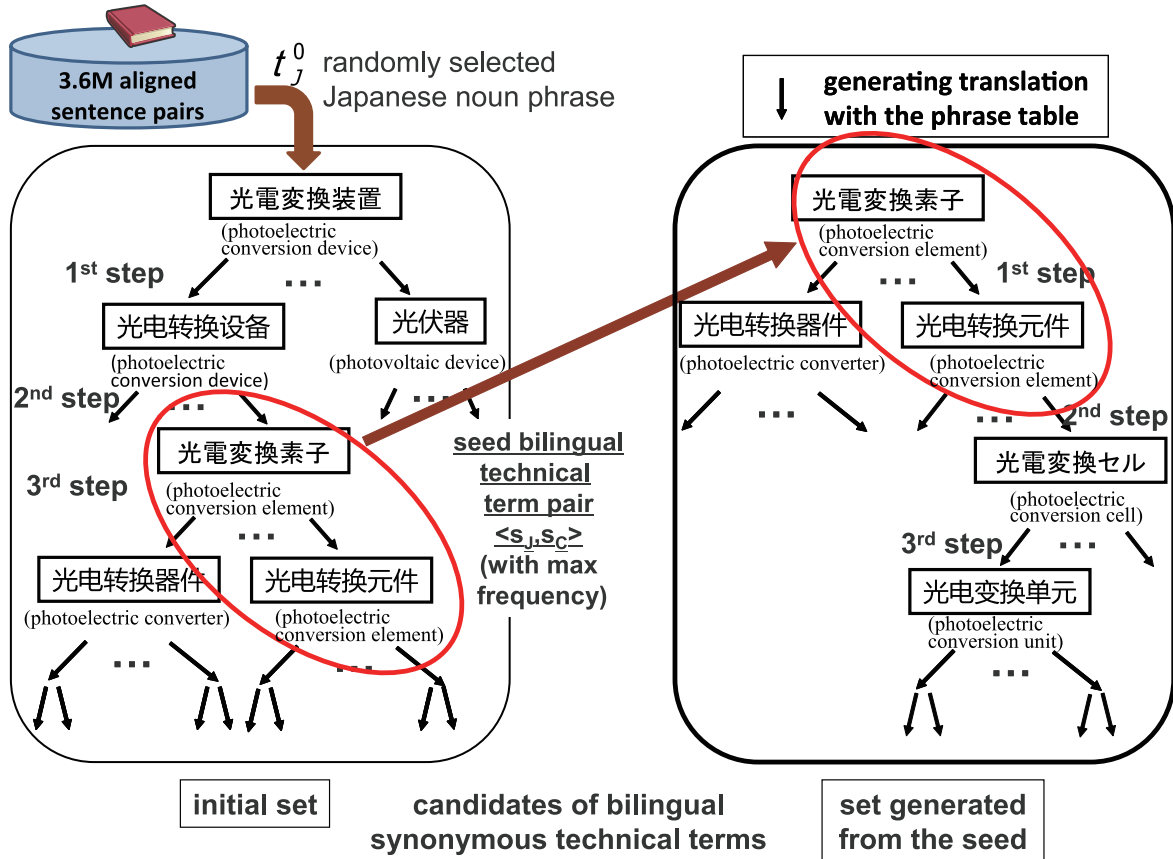


Figure 1: Developing a Reference Set of Bilingual Synonymous Technical Terms

morpheme lexicon IPAdic<sup>3</sup>. For Chinese sentences, we examine two types of segmentation, i.e., segmentation by characters<sup>4</sup> and segmentation by morphemes<sup>5</sup>.

As the result of applying Moses, we have a phrase table in the direction of Japanese to Chinese translation, and another one in the opposite direction of Chinese to Japanese translation. In the direction of Japanese to Chinese translation, we finally obtain 108M (Chinese sentences segmented by morphemes) / 274M (Chinese sentences segmented by characters) translation pairs with 75M / 197M unique Japanese phrases with Japanese to Chinese phrase translation probabilities  $P(p_C | p_J)$  of translating a Japanese phrase  $p_J$  into a Chinese phrase  $p_C$ . For each Japanese phrase, those multiple translation candidates in the phrase table are ranked in descending order of Japanese to Chinese phrase translation probabilities. In the similar way, in the phrase table in the opposite direction of Chinese to Japanese translation, for each Chinese phrase, multiple Japanese translation candidates are ranked in descending order of Chinese to Japanese phrase translation probabilities.

Those two phrase tables are then referred to when identifying a bilingual technical term pair, given a parallel sen-

tence pair  $\langle S_J, S_C \rangle$  and a Japanese technical term  $t_J$ , or a Chinese technical term  $t_C$ . In the direction of Japanese to Chinese, given a parallel sentence pair  $\langle S_J, S_C \rangle$  containing a Japanese technical term  $t_J$ , Chinese translation candidates collected from the Japanese to Chinese phrase table are matched against the Chinese sentence  $S_C$  of the parallel sentence pair. Among those found in  $S_C$ ,  $\hat{t}_C$  with the largest translation probability  $P(\hat{t}_C | t_J)$  is selected and the bilingual technical term pair  $\langle t_J, \hat{t}_C \rangle$  is identified. Similarly, in the opposite direction of Chinese to Japanese, given a parallel sentence pair  $\langle S_J, S_C \rangle$  containing a Chinese technical term  $t_C$ , the Chinese to Japanese phrase table is referred to when identifying a bilingual technical term pair.

#### 4. Developing a Reference Set of Bilingual Synonymous Technical Terms

When developing a reference set of bilingual synonymous technical terms (detailed procedure to be found in Liang et al. (2011a)), starting from a seed bilingual term pair  $s_{JC} = \langle s_J, s_C \rangle$ , we repeat the translation estimation procedure of the previous section six times and generate the set  $CBP(s_J)$  of candidates of bilingual synonymous technical term pairs. Figure 1 illustrates the whole procedure.

Then, we manually divide the set  $CBP(s_J)$  into  $SBP(s_{JC})$ , those of which are synonymous with  $s_{JC}$ , and the remaining  $NSBP(s_{JC})$ . As in Table 1, we collect 114 seeds, where the number of bilingual technical terms included in  $SBP(s_{JC})$  in total for all of the 114 seed bilin-

<sup>3</sup><http://sourceforge.jp/projects/ipadic/>

<sup>4</sup>A consecutive sequence of numbers as well as a consecutive sequence of alphabetical characters are segmented into a token.

<sup>5</sup>Chinese sentences are segmented into a sequence of morphemes by the Chinese morphological analyzer Stanford Word Segment (Tseng et al., 2005) trained with Chinese Penn Treebank.

Table 1: Number of Bilingual Technical Terms: Candidates and Reference of Synonyms

(a) With the Phrase Table based on Chinese Sentences Segmented by Characters

		# of bilingual technical terms for the total 114 seeds		average per seed	
Candidates of Synonyms $\bigcup_{s_J} CBP(s_J)$	included only in the set (a)	8,816	22,563	77.3	197.92
	included in the intersection of the sets (a) and (b)	13,747		120.6	
Reference of Synonyms $\bigcup_{s_{JC}} SBP(s_{JC})$	included only in the set (a)	309	2,496	2.7	21.9
	included in the intersection of the sets (a) and (b)	2,187		19.2	

(b) With the Phrase Table based on Chinese Sentences Segmented by Morphemes

		# of bilingual technical terms for the total 114 seeds		average per seed	
Candidates of Synonyms $\bigcup_{s_J} CBP(s_J)$	included only in the set (b)	14,161	28,948	124.2	253.9
	included in the intersection of the sets (a) and (b)	14,787		129.7	
Reference of Synonyms $\bigcup_{s_{JC}} SBP(s_{JC})$	included only in the set (b)	180	2,604	1.6	22.8
	included in the intersection of the sets (a) and (b)	2,424		21.3	

gual technical term pairs is around 2,500 to 2,600, which amounts to around 22 per seed on average. It can be also seen from Table 1 that although about 90% of reference of synonymous technical terms are shared by the two types of segmentation (by characters and by morphemes), only about 40% to 50% of candidates of synonymous technical terms are shared by the two types of segmentation.

## 5. Identifying Bilingual Synonymous Technical Terms by Machine Learning

In this section, we apply the SVMs to the task of identifying bilingual synonymous technical terms. In this paper, we model the task of identifying bilingual synonymous technical terms by the SVMs as that of judging whether or not the input bilingual term pair  $\langle t_J, t_C \rangle$  is synonymous with the seed bilingual technical term pair  $s_{JC} = \langle s_J, s_C \rangle$ .

### 5.1. The Procedure

First, let  $CBP$  be the union of the sets  $CBP(s_J)$  of candidates of bilingual synonymous technical term pairs for all of the 114 seed bilingual technical term pairs. In the training and testing of the classifier for identifying bilingual synonymous technical terms, we first divide the set of 114 seed bilingual technical term pairs into 10 subsets. Here, for each  $i$ -th subset ( $i = 1, \dots, 10$ ), we construct the union  $CBP_i$  of the sets  $CBP(s_J)$  of candidates of bilingual synonymous technical term pairs, where  $CBP_1, \dots, CBP_{10}$  are 10 disjoint subsets<sup>6</sup> of  $CBP$ .

<sup>6</sup>Here, we divide the set of 114 seed bilingual technical term pairs into 10 subsets so that the numbers of positive (i.e., syn-

As a tool for learning SVMs, we use TinySVM (<http://chasen.org/~taku/software/TinySVM/>). As the kernel function, we use the polynomial (1st order) kernel<sup>7</sup>. In the testing of a SVMs classifier, we regard the distance from the separating hyperplane to each test instance as a confidence measure, and return test instances satisfying confidence measures over a certain lower bound only as positive samples (i.e., synonymous with the seed). In the training of SVMs, we use 8 subsets out of the whole 10 subsets  $CBP_1, \dots, CBP_{10}$ . Then, we tune the lower bound of the confidence measure with one of the remaining two subsets. With this subset, we also tune the parameter of TinySVM for trade-off between training error and margin. Finally, we test the trained classifier against another one of the remaining two subsets. We repeat this procedure of training / tuning / testing 10 times, and average the 10 results of test performance.

### 5.2. Features

Table 2 lists all the features used for training and testing of SVMs for identifying bilingual synonymous technical terms. Features are roughly divided into two types: those of the first type  $f_1, \dots, f_6$  simply represent various characteristics of the input bilingual technical term  $\langle t_J, t_C \rangle$ , while those of the second type  $f_7, \dots, f_{16}$  represent relation of the input bilingual technical term  $\langle t_J, t_C \rangle$  and the

onymous with the seed) / negative (i.e., not synonymous with the seed) samples in each  $CBP_i$  ( $i = 1, \dots, 10$ ) are comparative among the 10 subsets.

<sup>7</sup>We compare the performance of the 1st order and 2nd order kernels, where we have almost comparative performance.

Table 2: Features for Identifying Bilingual Synonymous Technical Terms by Machine Learning

class	feature	definition ( where $X$ denotes $J$ or $C$ , and $\langle s_J, s_C \rangle$ denotes the seed bilingual technical term pair )
features for bilingual technical terms $\langle t_J, t_C \rangle$	$f_1$ : frequency	log of the frequency of $\langle t_J, t_C \rangle$ within the whole parallel patent sentences
	$f_2$ : rank of the Chinese term	given $t_J$ , log of the rank of $t_C$ with respect to the descending order of the conditional translation probability $P(t_C   t_J)$
	$f_3$ : rank of the Japanese term	given $t_C$ , log of the rank of $t_J$ with respect to the descending order of the conditional translation probability $P(t_J   t_C)$
	$f_4$ : number of Japanese characters	number of characters in $t_J$
	$f_5$ : number of Chinese characters	number of characters in $t_C$
	$f_6$ : number of times generating translation by applying the phrase tables	the number of times repeating the procedure of generating translation by applying the phrase tables until generating $t_C$ or $t_J$ from $s_J$ , as in $s_C \rightarrow \dots \rightarrow t_J \rightarrow t_C$ , or, $s_J \rightarrow \dots \rightarrow t_C \rightarrow t_J$
features for the relation of bilingual technical terms $\langle t_J, t_C \rangle$ and the seed $\langle s_J, s_C \rangle$	$f_7$ : identity of Japanese terms	returns 1 when $t_J = s_J$
	$f_8$ : identity of Chinese terms	returns 1 when $t_C = s_C$
	$f_9$ : edit distance similarity of monolingual terms	$f_9(t_X, s_X) = 1 - \frac{ED(t_X, s_X)}{\max( t_X ,  s_X )}$ (where $ED$ is the edit distance of $t_X$ and $s_X$ , and $ t $ denotes the number of characters of $t$ .)
	$f_{10}$ : character bigram similarity of monolingual terms	$f_{10}(t_X, s_X) = \frac{ bigram(t_X) \cap bigram(s_X) }{\max( t_X ,  s_X ) - 1}$ (where $bigram(t)$ is the set of character bigrams of the term $t$ .)
	$f_{11}$ : rate of identical morphemes (for Japanese terms)	$f_{11}(t_J, s_J) = \frac{ const(t_J) \cap const(s_J) }{\max( const(t_J) ,  const(s_J) )}$ (where $const(t)$ is the set of morphemes in the Japanese term $t$ .)
	$f_{12}$ : rate of identical characters (for Chinese terms)	$f_{11}(t_C, s_C) = \frac{ const(t_C) \cap const(s_C) }{\max( const(t_C) ,  const(s_C) )}$ (where $const(t)$ is the set of Characters in the Chinese term $t$ .)
	$f_{13}$ : subsumption relation of strings / variants relation of surface forms (for Japanese terms)	returns 1 when the difference of $t_J$ and $s_J$ is only in their suffixes, or only whether or not having the prolonged sound “—”, or only in their hiragana parts.
	$f_{14}$ : identical stem (for Chinese terms)	returns 1 when the difference of $t_C$ and $s_C$ is only whether or not haing the word “的” which is not the prefix or suffix.
	$f_{15}$ : rate of intersection in translation by the phrase table	$f_{15}(t_X, s_X) = \frac{ trans(t_X) \cap trans(s_X) }{\max( trans(t_X) ,  trans(s_X) )}$ (where $trans(t)$ is the set of translation of term $t$ from the phrase table.)
	$f_{16}$ : translation by the phrase table	returns 1 when $s_J$ can be generated by translating $t_E$ with the phrase table, or, $s_E$ can be generated by translating $t_J$ with the phrase table.

seed bilingual technical term pair  $s_{JC} = \langle s_J, s_C \rangle$ .

Among the features of the first type are the frequency ( $f_1$ ), ranks of terms with respect to the conditional translation probabilities ( $f_2$  and  $f_3$ ), length of terms ( $f_4$  and  $f_5$ ), and the number of times repeating the procedure of generating translation with the phrase tables until generating input terms  $t_J$  and  $t_C$  from the Japanese seed term  $s_J$  ( $f_6$ ).

Among the features of the second type are identity of monolingual terms ( $f_7$  and  $f_8$ ), edit distance of monolingual terms ( $f_9$ ), character bigram similarity of monolingual terms ( $f_{10}$ ), rate of identical morphemes (in Japanese,  $f_{11}$ ) / characters (in Chinese,  $f_{12}$ ), string subsumption and variants for Japanese ( $f_{13}$ ), identical stem for Chinese ( $f_{14}$ ), rate of intersection in translation by the phrase table ( $f_{15}$ ), and translation by the phrase tables ( $f_{16}$ ).

### 5.3. Evaluation Results

Table 3 shows the evaluation results for a baseline as well as for SVMs. As the baseline, we simply judge the input bilingual term pair  $\langle t_J, t_C \rangle$  as synonymous with the seed bilingual technical term pair  $s_{JC} = \langle s_J, s_C \rangle$  when  $t_J$  and  $s_J$  are identical, or,  $t_C$  and  $s_C$  are identical. When training / testing a SVMs classifier, we tune the lower bound of the confidence measure of the distance from the separating hyperplane in two ways: i.e., for maximizing precision and for maximizing F-measure. When maximizing precision, we achieve almost 87% precision where F-measure is over 40%. When maximizing F-measure, we achieve over 60% F-measure with around 71% precision and over 52% recall. As shown in Figure 2, the two types of segmentation of Chinese sentences, namely, by characters and by morphemes, tend to have different types of errors. So, we integrate those two types of segmentation in the form of the intersection of

Table 3: Evaluation Results (%)

		segmented by characters			segmented by morphemes			intersection		
		precision	recall	f-measure	precision	recall	f-measure	precision	recall	f-measure
baseline ( $t_J$ and $s_J$ are identical, or, $t_C$ and $s_C$ are identical.)		71.5	39.4	50.8	69.1	40.0	50.7	77.3	33.1	46.3
SVM	maximum precision	<b>86.9</b>	26.0	40.0	84.3	24.5	38.0	<b>90.0</b>	25.1	39.2
	maximum f-measure	71.0	52.8	60.6	68.6	54.4	<b>60.7</b>	—	—	—

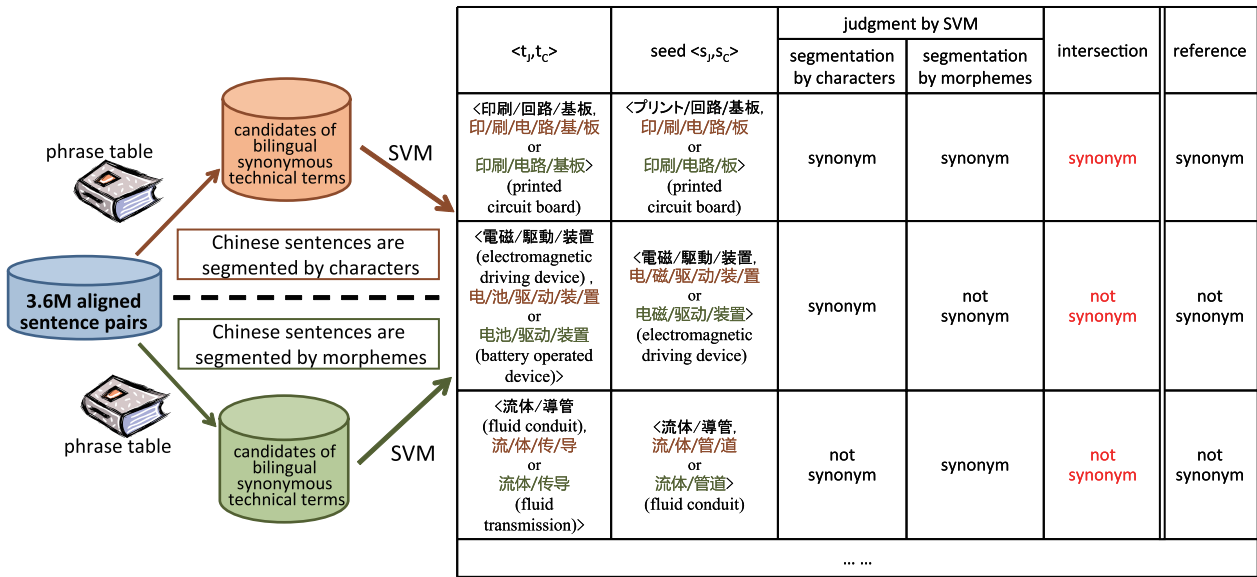


Figure 2: Evaluating Intersection of Judgments by SVM based on Character/Morpheme based Segmentation of Chinese Sentences

SVM judgments, where, for both types of segmentation, we tune the lower bound of the confidence measure of the distance from the separating hyperplane. We maximize precision while keeping recall over 25% with held-out data, and this achieves over 90% precision as shown in Table 3.

## 6. Related Work

Among related works on acquiring bilingual lexicon from text, Itagaki et al. (2007) focused on automatic validation of translation pairs available in the phrase table trained by an SMT model. Lu and Tsou (2009) and Yasuda and Sumita (2013) also studied to extract bilingual terms from comparable patents, where, they first extract parallel sentences from comparable patents, and then extract bilingual terms from parallel sentences. Those studies differ from this paper in that those studies did not address the issue of acquiring bilingual synonymous technical terms. Tsunakawa and Tsujii (2008) is mostly related to our study, in that they also proposed to apply machine learning technique to the task of identifying bilingual synonymous technical terms. However, Tsunakawa and Tsujii (2008) studied the issue of identifying bilingual synonymous technical terms only within manually compiled bilingual technical

term lexicon and thus are quite limited in its applicability. Our approach, on the other hand, is quite advantageous in that we start from parallel patent documents which continue to be published every year and then, that we can generate candidates of bilingual synonymous technical terms automatically.

Our study in this paper is also different from previous works on identifying synonyms based on bilingual and monolingual resources (e.g. Lin and Zhao (2003)) in that we learn bilingual synonymous technical terms from phrase tables of a phrase-based SMT model trained with very large parallel sentences. Also in the context of SMT between Japanese and Chinese, Sun and Lepage (2012) pointed out that character-based segmentation of sentences contributed to improving machine translation performance compared to morpheme-based segmentation of sentences.

## 7. Conclusion

In the task of acquiring Japanese-Chinese technical term translation equivalent pairs from parallel patent documents, this paper considered situations where a technical term is observed in many parallel patent sentences and is translated into many translation equivalents and studied the is-

sue of identifying synonymous translation equivalent pairs. We especially examined two types of segmentation of Chinese sentences, i.e., by characters and by morphemes, and integrated those two types of segmentation in the form of the intersection of SVM judgments, which achieved over 90% precision. One of the most important future works is definitely to improve recall. To do this, we plan to apply the semi-automatic framework (Liang et al., 2011b) which have been invented in the task of identifying Japanese-English synonymous translation equivalent pairs and have been proven to be effective in improving recall. We plan to examine whether this semi-automatic framework is also effective in the task of identifying Japanese-Chinese synonymous translation equivalent pairs.

## 8. References

- P. Fung and L. Y. Yee. 1998. An IR approach for translating new words from nonparallel, comparable texts. In *Proc. 17th COLING and 36th ACL*, pages 414–420.
- F. Huang, Y. Zhang, and S. Vogel. 2005. Mining key phrase translations from Web corpora. In *Proc. HLT/EMNLP*, pages 483–490.
- M. Itagaki, T. Aikawa, and X. He. 2007. Automatic validation of terminology translation consistency with statistical method. In *Proc. MT Summit XI*, pages 269–274.
- P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proc. 45th ACL, Companion Volume*, pages 177–180.
- B. Liang, T. Utsuro, and M. Yamamoto. 2011a. Identifying bilingual synonymous technical terms from phrase tables and parallel patent sentences. *Procedia - Social and Behavioral Sciences*, 27:50–60.
- B. Liang, T. Utsuro, and M. Yamamoto. 2011b. Semi-automatic identification of bilingual synonymous technical terms from phrase tables and parallel patent sentences. In *Proc. 25th PACLIC*, pages 196–205.
- D. Lin and S. Zhao. 2003. Identifying synonyms among distributionally similar words. In *Proc. 18th IJCAI*, pages 1492–1493.
- B. Lu and B. K. Tsou. 2009. Towards bilingual term extraction in comparable patents. In *Proc. 23rd PACLIC*, pages 755–762.
- Y. Matsumoto and T. Utsuro. 2000. Lexical knowledge acquisition. In R. Dale, H. Moisl, and H. Somers, editors, *Handbook of Natural Language Processing*, chapter 24, pages 563–610. Marcel Dekker Inc.
- Y. Morishita, T. Utsuro, and M. Yamamoto. 2008. Integrating a phrase-based SMT model and a bilingual lexicon for human in semi-automatic acquisition of technical term translation lexicon. In *Proc. 8th AMTA*, pages 153–162.
- J. Sun and Y. Lepage. 2012. Can word segmentation be considered harmful for statistical machine translation tasks between Japanese and Chinese? In *Proc. 26th PACLIC*, pages 351–360.
- M. Tonoike, M. Kida, T. Takagi, Y. Sasaki, T. Utsuro, and S. Sato. 2006. A comparative study on compositional translation estimation using a domain/topic-specific corpus collected from the web. In *Proc. 2nd Intl. Workshop on Web as Corpus*, pages 11–18.
- H. Tseng, P. Chang, G. Andrew, D. Jurafsky, and C. Manning. 2005. A conditional random field word segmenter for Sighan bakeoff 2005. In *Proc. 4th SIGHAN Workshop on Chinese Language Processing*, pages 168–171.
- T. Tsunakawa and J. Tsujii. 2008. Bilingual synonym identification with spelling variations. In *Proc. 3rd IJCNLP*, pages 457–464.
- M. Utiyama and H. Isahara. 2007. A Japanese-English patent parallel corpus. In *Proc. MT Summit XI*, pages 475–482.
- V. N. Vapnik. 1998. *Statistical Learning Theory*. Wiley-Interscience.
- K. Yasuda and E. Sumita. 2013. Building a bilingual dictionary from a Japanese-Chinese patent corpus. In *Computational Linguistics and Intelligent Text Processing*, volume 7817 of *LNCS*, pages 276–284. Springer.

# Revisiting comparable corpora in connected space

Pierre Zweigenbaum

CNRS, UPR 3251, LIMSI  
91403 Orsay, France  
pz@limsi.fr

## Abstract

Bilingual lexicon extraction from comparable corpora is generally addressed through two monolingual distributional spaces of context vectors connected through a (partial) bilingual lexicon. We sketch here an abstract view of the task where these two spaces are embedded into one common bilingual space, and the two comparable corpora are merged into one bilingual corpus. We show how this paradigm accounts for a variety of models proposed so far, and where a set of topics addressed so far take place in this framework: degree of comparability, ambiguity in the bilingual lexicon, where parallel corpora stand with respect to this view, e.g., to replace the bilingual lexicon. A first experiment, using comparable corpora built from parallel corpora, illustrates one way to put this framework into practice. We also outline how this paradigm suggests directions for future investigations. We finally discuss the current limitations of the model and directions to solve them.

## 1. Introduction

The standard approach to bilingual dictionary extraction from comparable corpora (Rapp, 1995; Fung and McKeown, 1997) proposes to perform monolingual distributional analysis in each of the two comparable corpora. It represents source and target words with context vectors, and a transformation of source context words into target context words through a dictionary. Previous work has investigated variations on context vector construction (context nature and size, association scores, e.g., (Laroche and Langlais, 2010; Gamallo and Bordag, 2011)) and on the seed-dictionary-based transformation: origin and coverage of the dictionary, e.g., (Chiao and Zweigenbaum, 2003; Hazem and Morin, 2012), complementary transformations (Gaussier et al., 2004), disambiguation of dictionary entries (Morin and Prochasson, 2011; Apidianaki et al., 2013; Bouamor et al., 2013b), acquisition of the dictionary from parallel corpora (Morin and Prochasson, 2011; Apidianaki et al., 2013).

Here we want to emphasize the overall space which is created by this construction. Previous work has hinted at this overall space (e.g., (Gaussier et al., 2004)) or used it explicitly (Peirsman and Padó, 2010) but has not to our knowledge investigated further the view that it can provide on the task and its related issues. The goal of this paper is to draft a model of this space and to point at the avenues it opens for further research. Therefore this paper is a rather abstract, first stab at a description of this model, and leaves both a precise formalization and concrete experiments for further work. It also leaves for future work the handling of multi-word expressions. This type of exposition may incur risks of “hand waiving”, which we have tried to minimize. Its main contributions (and outline) are the following:

- The description of a unified space embedding the context vectors of the two comparable corpora;
- The description of a connected, bilingual corpus generated from the two comparable corpora;
- A reformulation of some topics in bilingual lexicon extraction from comparable corpora;

- Suggestions for future research spawned by this unified space.

## 2. Related work

The introduction has shortly enumerated several dimensions of research on bilingual lexicon extraction from comparable corpora. The work closest to what we develop here is that of (Gaussier et al., 2004). A core component of the geometric view of (Gaussier et al., 2004) is the space defined by (source, target) word pairs in the bilingual dictionary. Among other things, (Gaussier et al., 2004) propose to represent words of both the source and target corpora in this common space, effectively creating a unified space. We propose below to extend this space and to study the view it gives of the joined comparable corpora.

Joint bilingual representations have been proposed in the past in various settings. Dual-language documents have been proposed by (Dumais et al., 1996), where a document and its translation are merged into a bilingual document; Latent Semantic Indexing is then performed on the collection of dual-language documents. Since we work with comparable corpora, we extend this concept to that of a dual-language corpus.

Translation pairs, i.e., bilingual dictionary entries, are used by (Jagarlamudi and Daumé III, 2010) as a substitute for ‘concepts’ to create cross-language topics. We also use translation pairs as basic units for cross-language representation; in our setting they are used in context vectors and in the above-mentioned dual-language corpus.

The notion of a bilingual vector space for comparable corpora, labeled with translation pairs, has already been proposed by (Peirsman and Padó, 2010). To avoid the need for a bilingual dictionary, they bootstrap translation pairs with “frequent cognates, words that are shared between two languages” (Peirsman and Padó, 2010). This creates a bilingual space in which words of each language are represented by context vectors in which context words are translation pairs. Both source and target words can be compared according to the similarity of their context vectors. Given a source word  $s$ , its nearest neighbor  $t$  in the target language is a candidate translation. (Peirsman and Padó, 2010) select



$$\begin{array}{c}
\cdots \\
\cdots \\
\text{pregnant} \\
\cdots \\
\cdots
\end{array}
\begin{pmatrix}
\vdots \\
\vdots \\
4.394197 \\
\vdots \\
\vdots
\end{pmatrix}
\begin{array}{c}
\text{women} \\
\vdots \\
\vdots \\
\vdots \\
\vdots
\end{array}
E =
\begin{array}{c}
e_1 \\
\vdots \\
e_i \\
\vdots \\
e_m
\end{array}
\begin{pmatrix}
\ddots & a(e_1, e_j) & & \\
& \vdots & & \\
& a(e_i, e_j) & & \\
& \vdots & & \\
& a(e_m, e_j) & & \ddots
\end{pmatrix}
\begin{array}{c}
f_1 \\
\vdots \\
f_k \\
\vdots \\
f_n
\end{array}
\begin{pmatrix}
\ddots & a(f_1, f_l) & & \\
& \vdots & & \\
& a(f_k, f_l) & & \\
& \vdots & & \\
& a(f_n, f_l) & & \ddots
\end{pmatrix}
\begin{array}{c}
f_1 \\
f_l \\
\cdots \\
f_n
\end{array}$$

Figure 1: Context vectors in source and target corpora: the column for  $e_j$  (resp.  $f_k$ ) represents its context vector, and  $a(e_i, e_j)$  (resp.  $a(f_k, f_l)$ ) is the association strength of  $e_i$  and  $e_j$  (resp.  $f_k$  and  $f_l$ ).

candidate pairs  $(s, t)$  where  $t$  is the nearest target neighbor of  $s$  and  $s$  is the nearest source neighbor of  $t$ . Iterating this process extends the initial set of seed bilingual pairs into a larger bilingual lexicon. This notion of a bilingual vector space was only a means to an end in (Peirsman and Padó, 2010). We explore it further in the present paper.

### 3. Reformulating the standard approach to bilingual lexicon extraction from comparable corpora

#### 3.1. Monolingual distributional analysis of source and target corpora

The distributional hypothesis characterizes the meaning of a word by the distribution of its usages in a language sample: a corpus. The original formulation by Harris (see details in (Habert and Zweigenbaum, 2002), citing (Harris, 1991)) relies on relations between operators and arguments. A common approximation consists in representing word usage through co-occurrence with other words in the corpus. Whatever the choice, given the vocabulary  $V$ , this associates to a given word  $e_i \in V$  a vector of words  $e_j \in V$  to which it is syntagmatically associated, and which is usually called its *context vector*. For example, context words (e.g., *pregnant*) in Sentence (1) contributes to the characterization of the context vector for *women* (see Figure 1, left):

- (1) information for pregnant women and children

$$\begin{array}{c}
\cdots \\
\cdots \\
[\text{pregnant} \sim \text{enceintes}] \\
\cdots \\
\cdots
\end{array}
\begin{pmatrix}
\vdots \\
\vdots \\
4.394197 \\
\vdots \\
\vdots
\end{pmatrix}
\begin{array}{c}
\text{women} \\
\vdots \\
\vdots \\
\vdots \\
\vdots
\end{array}$$

Figure 2: A context vector of the source corpus, with entries translated into the target language.

Overall, this creates a word $\times$ word matrix  $E$  of dimension  $|V| \times |V|$  in which  $E_i^j = a(e_i, e_j)$  is the association strength of  $e_i$  and  $e_j$ . Mutual information, log-likelihood ratio, and odds-ratio, among others, are common values for this association strength (see e.g. (Evert, 2005; Laroche and Langlais, 2010) for more association scores).

Given two corpora  $S$  and  $T$  (typically, here, two comparable corpora in two different languages), composed of vocabularies  $V$  and  $W$ , we can build word $\times$ word association matrices  $E$  and  $F$  of dimensions  $|V| \times |V|$  and  $|W| \times |W|$  (see Figure 1, center and right).

#### 3.2. How (unambiguous) bilingual links connect source wband target spaces

The standard approach additionally relies on a bilingual dictionary  $D = \{[s_i \sim t_j]\}$ , i.e., a set of [source $\sim$ target] word pairs. Its fundamental hypothesis is that word distribution reflects meaning and that meaning is preserved through translation, from which it assumes that the distribution of source words in the source corpus is similar to the distribution of their translations in the target corpus.<sup>1</sup> To simplify the exposition, we assume here that the dictionary introduces no ambiguity: it provides exactly one translation for the input source words that it contains (and reciprocally for target words). We do not assume that it has full coverage of the source or target corpus, otherwise there would remain no unknown word to translate.

Let us start from the context vector representation  $(a(e_i, e_j))_{i=1}^m$  of a source word  $e_j$  in the source corpus, where  $a(e_i, e_j)$  is the value of the vector on the axis provided by word  $e_i$ . The dictionary  $D$  is used to translate the entries in this context vector: based on translation pairs  $[e_i \sim f_k] \in D$ , i.e., where  $f_k$  is a translation of  $e_i$  through the dictionary, it produces a representation  $(a(f_k, e_j))_{k=1}^n$  of the source word  $e_j$  in the target corpus (see Figure (2)). In this representation, the same value  $a(f_k, e_j) = a(e_i, e_j) = a([e_i \sim f_k], e_j)$  is assumed to represent the association that the source word  $e_j$  would have with the target word  $f_k$  translated from  $e_i$  if  $e_j$  were occurring in the target corpus. This creates a representation of the position of  $e_j$  in the target space: target words  $f_l$  whose positions are close to it are candidates to translate  $e_j$ .

<sup>1</sup>Note that (Harris, 1988, viii) considers that this applies to the language of a given subsience (see again (Habert and Zweigenbaum, 2002)) rather than to the whole language.

$$E_t = \begin{matrix} e_1 \\ \vdots \\ e_{m-p} \\ [e_{m-p+1} \sim f_1] \\ \vdots \\ [e_m \sim f_p] \end{matrix} \begin{pmatrix} e_1 & e_j & e_m \\ \ddots & a(e_1, e_j) & \\ & \vdots & \\ & a(e_{m-p}, e_j) & \\ [e_{m-p+1} \sim f_1] & a([e_{m-p+1} \sim f_1], e_j) & \\ & \vdots & \\ [e_m \sim f_p] & a([e_m \sim f_p], e_j) & \ddots \end{pmatrix} \quad
F_t = \begin{matrix} [e_{m-p+1} \sim f_1] \\ \vdots \\ [e_m \sim f_p] \\ f_{p+1} \\ \vdots \\ f_n \end{matrix} \begin{pmatrix} f_1 & f_l & f_n \\ \ddots & a([e_{m-p+1} \sim f_1], f_l) & \\ & a([e_m \sim f_p], f_l) & \\ f_{p+1} & a(f_{p+1}, f_l) & \\ \vdots & & \\ f_n & a(f_n, f_l) & \ddots \end{pmatrix}$$

Figure 3: Translated context vectors in source ( $E_t$ ) and target ( $F_t$ ) corpora.  $[e_{m-p+d} \sim f_d]_{d \in \{1 \dots p\}}$  are translation pairs in the dictionary. Instead of discarding the non-translated contexts of the vectors, we keep them untouched.

$$G = \begin{matrix} e_1 \\ \vdots \\ e_i \\ [e_{m-p+1} \sim f_1] \\ \vdots \\ [e_m \sim f_p] \\ f_{p+1} \\ \vdots \\ f_n \end{matrix} \begin{pmatrix} e_1 & e_j & \dots & [e_{m-p+d} \sim f_d] & f_l & \dots & f_n \\ \vdots & a(e_1, e_j) & & \vdots & & & \\ \vdots & \vdots & & \vdots & & & \mathbf{0} \\ e_i & a(e_i, e_j) & & \vdots & & & \\ [e_{m-p+1} \sim f_1] & a([e_{m-p+1} \sim f_1], e_j) & & \vdots & a([e_{m-p+1} \sim f_1], f_l) & \vdots & \vdots \\ \vdots & \vdots & & \vdots & \vdots & \vdots & \vdots \\ [e_m \sim f_p] & a([e_m \sim f_p], e_j) & & \vdots & a([e_m \sim f_p], f_l) & \vdots & \vdots \\ f_{p+1} & & & \vdots & a(f_{p+1}, f_l) & \vdots & \vdots \\ & \mathbf{0} & & \vdots & \vdots & \vdots & \vdots \\ f_n & & & \vdots & a(f_n, f_l) & \vdots & \vdots \end{pmatrix}$$

Figure 4: Translated context vectors  $G$  in source and target corpora, embedded in unified context space.  $[e_{m-p+d} \sim f_d]_{d \in \{1 \dots p\}}$  are translation pairs in the dictionary.

Since generally not all source and target words belong to the dictionary, only a part of a source context vector (say  $p$  entries) goes through this translation, while the rest is ignored. Let us assume for ease of exposition that we re-order the rows (and columns) of  $E$  (resp.  $F$ ) with the  $p$  in-dictionary entries last (resp. first). The translated version  $E_t$  of the source (resp.  $F_t$  of the target) context vectors can then be schematized as shown in Figure 3 (we keep the out-of-dictionary part of the vectors though). This reveals the common representation subspace created by the dictionary entries ( $[e_{m-p+1} \sim f_1] \dots [e_m \sim f_p]$ , in red in Figure 3).

### 3.3. Embedding bilingual corpora into a unified space

This common subspace provides a basis on which to merge the two sets of context vectors. Of the  $m$  dimensions of  $E$  and of the  $n$  dimensions of  $F$ ,  $p$  are common to both. These vectors can thus be extended to dimension  $q = m + n - p$ : vectors of  $E_t$  are extended with  $n - p$  zeros at their end, and vectors of  $F_t$  are extended with  $m - p$  zeros at their beginning.<sup>2</sup> Besides, to highlight some properties of the

obtained representation, we re-order the context vectors so that the columns for source and target words in the dictionary are next to each other. This is schematized on Figure 4, where the common subspace is shown in red, zero extensions are shown in blue, two in-dictionary context vectors are grouped under each  $[e_{m-p+d} \sim f_d]$  header (in violet), and black shows the corpus-specific contexts. Note that only the red parts are used in the standard approach.

These in-dictionary context vectors have another interpretation at the text level. Substituting source (resp. target) words with translation pairs amounts to actually *replacing in the texts the source (resp. target) words present in the dictionary with concatenated bi-words*. For instance, depending on the dictionary, the English Sentence (1) may become as in Sentence (2 a) (the dictionary has no entry for *information* and *women*). Similarly, in the reverse direction, the French sentence *une forte proportion de femmes enceintes* may give rise to Sentence (2 b):

- (2) (a) information|for|intention pregnant|enceintes  
women|and|et children|enfants  
(b) a|une high|forte proportion|proportion of|de  
|femmes pregnant|enceintes

Figure 5 displays the same examples graphically, with En-

<sup>2</sup>Note again that we do not discard the non-translated contexts of these vectors. This contrasts to the standard approach where only the in-dictionary contexts are kept and then compared. We return to this point below.

information for pregnant women and children  
 intention enceintes et enfants  
a high proportion of pregnant  
une forte proportion de femmes enceintes

Figure 5: Bilingual corpus: an English sentence and a French sentence. In this example, *information*, *women*, and *femmes* are out-of-dictionary words.

English words on top and French words at the bottom. Blue color marks the source sentence. Once transformed this way, the two comparable corpora can be merged into one bilingual corpus. To avoid confusion between source and target cognates, all out-of-dictionary words in the source and target corpora are marked with their language.<sup>3</sup>

The representation of words in this corpus can follow the standard distributional practice outlined in Section 3.1. Since source corpus words outside the dictionary never co-occur with target corpus words outside the dictionary, the two corresponding quadrants of the matrix in Figure 4 are filled with zeros. This should make the contribution of out-of-dictionary contexts minimal in the computation of vector similarity.

More precisely, if the dot product is used to compare context vectors, the representation in Figure 4 leads to the same results as truncating context vectors to their dictionary part, as is performed in the standard approach. However, if the similarity of two vectors is instead computed through a formula which takes into account all components of both vectors (e.g., cosine similarity normalizes the dot product by dividing it by the norms of the two vectors, and the Jaccard index divides the common features by the union of all features of the two context vectors), the formulation in Figure 4 should lead to reduced similarity values for each word with a strong association with out-of-dictionary words. If we consider that for a given word, the stronger its associations with out-of-dictionary words, the poorer the fidelity of its context vector, reducing its similarity to other context vectors might not be a bad move. This suggests a direction for new investigations.

Note also that for each  $d \in \{1 \dots p\}$ , the context vectors of translation pair items  $e_{m-p+d}$  and  $f_d$  are expected to be more similar to each other than to any other context vector. These pairs of in-dictionary context vectors might thus provide a training set to tune some parameters or to train supervised methods. However, replacing  $e_{m-p+d}$  and  $f_d$  with a concatenated bi-word in the corpus replaces their two context vectors with a single one (not shown in Figure 4). This forces a single distribution on the resulting bi-word. Such merged context vectors are the only ones that may have non-zero out-of-dictionary context words in both

<sup>3</sup>For instance by prefixing them with *lang\_*, e.g. *en\_* and *fr\_*. In our experiments we adopted a simpler convention where a translation pair  $[e \sim m]$  is noted  $e | f$ , and source or target out-of-dictionary words are noted respectively  $e |$  and  $| f$ , as seen in Example (2 a) for *information* and *women*.

the source and target subspaces of the corpus.<sup>4</sup> To summarize, we have proposed here:

1. A unified context matrix which embeds context vectors of both source and target corpora; and
2. An associated merged bilingual corpus, some of whose “words” are bilingual word pairs.

The merged bilingual corpus has only been sketched. While computations are performed on the unified context matrix, the main intention of the merged bilingual corpus is to produce a concrete object which can support human observation and reasoning, and thereby complement the more abstract artifact of context vectors in unified context space. It is defined as a corpus whose contexts produce the unified context matrix. If the bilingual dictionary is not ambiguous (i.e., it only contains one-to-one mappings between source and target words), the merged corpus can be defined by simple substitution as in the present section. If the bilingual dictionary is ambiguous (see Section 4.3. below), creating the bilingual corpus requires a more complex management of individual contexts which goes beyond the present paper. This difficulty in building the bilingual corpus may be taken as a clue that ambiguous dictionary entries create a problem for bilingual lexicon extraction from comparable corpora, and should thus be resolved before bilingual lexicon extraction.

## 4. Revisiting common topics in bilingual lexicon extraction

### 4.1. Bilingual lexicon extraction as “a-lingual” distributional analysis and similarity

The unified context vector space contains both source and target context vectors. Similarity in this space can therefore be used to compare source and target context vectors directly, hence to look for word translations. Moreover, clustering in this space results in clusters which can contain at the same time source and target context vectors, which are similar either in source space (monolingual distributional similarity), in target space (same), or across the two (cross-lingual distributional similarity, aimed at spotting translations).

Having one unified space might be thought at first sight to help reduce the common propensity to use directional methods, which then need to be symmetrized a posteriori as in (Chiao et al., 2004). This is however not necessarily the case: even within unified space, (Peirsman and Padó, 2010) still opt to enforce symmetric conditions to select similar words.

### 4.2. Degree of comparability

(Déjean and Gaussier, 2002) consider that two corpora are comparable if a non-negligible subpart of the vocabulary  $V$

<sup>4</sup>We might also keep the original individual context vectors of  $e_{m-p+d}$  and  $f_d$ , and add to them, instead of substituting for them, their merged context vector. This amounts to duplicating the sentences (or more precisely the contexts) in which words  $e_{m-p+d}$  or  $f_d$  occur: keeping the original sentence and creating a copy where occurrences of  $e_{m-p+d}$  or  $f_d$  are replaced with  $[e_{m-p+d} \sim f_d]$ .

of the source corpus has a translation in the target vocabulary  $W$  and reciprocally. (Li and Gaussier, 2010) base their measure of comparability of two corpora on the proportion of words in  $V$  (resp.  $W$ ) whose translations are found in  $W$  (resp.  $V$ ). This proportion corresponds to the proportion of rows in the  $E_t$  or  $F_t$  matrix which could be covered by a complete dictionary—or which an oracle method could map to a correct translation in the corpus. In contrast, comparability measures which use features other than simple words translations (Su and Babych, 2012) do not have a simple counterpart in these matrices.

### 4.3. Ambiguity in the bilingual lexicon

The proposed construction emphasizes the importance of disambiguating dictionary word translations, which recent work (Apidianaki et al., 2013; Bouamor et al., 2013b) has shown to be able to bring substantial improvements in bilingual lexicon extraction from comparable corpora. However, if multiple translations remain for source dictionary words (e.g.,  $[e_{m-p+d} \sim f_{d_1}], \dots [e_{m-p+d} \sim f_{d_t}]$ ), the context vector view presented in Section 3.3. should be adapted.

One way to handle this would be to create additional rows (and columns) in matrix  $G$  for the additional translation pairs. This amounts to duplicating the sentences (more precisely, contexts) in which the source word  $e_{m-p+d}$  occurs: each resulting sentence  $S_i$  would replace occurrences of  $e_{m-p+d}$  with  $[e_{m-p+d} \sim f_{d_i}]$ . However, if several source words  $e_a, e_b, \dots$  map to the same target word  $f_d$ , this results in distinct representations  $[e_a \sim f_d], [e_b \sim f_d], \dots$  of the same target word  $f_d$  which split the distribution of this target word into several parts. This could be a reasonable option if this separates distinct senses of  $f_d$ .

Another way would be to assume a less constrained mapping (typically, a linear transformation) through the dictionary from source words to target words. This can be defined by a transformation matrix  $M$  (see, e.g., (Gaussier et al., 2004)) whose row indexes are the source words that have an entry in the dictionary, whose column indexes are the target words which the dictionary proposes for at least one source word, and where  $M_{ij} = 1$  (or some given positive weight, for instance such that  $\sum_j M_{ij} = 1$  to encode a distribution of word translation probabilities) iff  $[e_i \sim f_j]$  is in the dictionary and  $M_{ij} = 0$  otherwise. As announced in Section 3.3., this method makes it more difficult to design an associated merged corpus. A direction to consider to create this merged corpus would be to include in this corpus not only full sentences, but also isolated phrases embodying elementary contexts.

All in all, the present discussion emphasizes that disambiguating source (and target) words helps obtain a better-defined model and could help design a more natural merged corpus. The methods adopted by (Apidianaki et al., 2013) look particularly relevant for this purpose since they induce clusters of translations which create sense clusters in the target corpus, hence seem compatible with the first above-mentioned way to handle ambiguity.

### 4.4. Parallel corpora in connected space

Parallel corpora<sup>5</sup> are often considered to be an ideal version of comparable corpora: they maximize comparability inasmuch as most source words can be aligned to a target word, and reciprocally. Indeed, parallel corpora also have drawbacks, the main one being that they are subject to translation bias: at least one of the two parallel corpora has been obtained by translating from a source language, and may contain calques, so the parallel corpus is a less good sample of that language. However, as in most work on parallel corpora, we shall ignore this property here.

We can represent two parallel corpora in the same way as comparable corpora in Section 3.1.: each corpus is subjected to distributional analysis to build context vectors. Then, instead of using an external bilingual dictionary, we can take advantage of word alignments to connect the two corpora. An advantage of word alignments (assuming they are correct) over using an external dictionary is that no disambiguation is necessary: each word translation is precisely valid in the context where it is found. Another advantage is that as mentioned above, most source words are aligned with some target word.

What is the use of considering parallel corpora under this view? Indeed, since most words can find translations through alignment, which is much more precise than distributional similarity, handling them as comparable corpora is not directly relevant for bilingual lexicon acquisition. However, let us examine their representation more closely.

A direct equivalent of a dictionary translation pair in parallel corpora is a pair of aligned  $[e \sim f]$  words. However, a given source word may be translated as one among a set of variant words, and a set of different source words may obtain the same translation (which is useful to collect paraphrases (Barzilay and McKeown, 2001)). It may thus be beneficial to identify, among the possible translations of a given source word, those that are equivalent or closely related (Apidianaki, 2008) and those that are different (see also (Yao et al., 2012) for statistics on synonymy [equivalence] and polysemy [difference] in this context). Such sense clusters may provide a more relevant basis for translation pairs than individually aligned words in context vectors: by making (language-sensitive) word senses explicit, they should on the one hand lead to better generalization than individual words, while on the other hand differentiating different senses, thus potentially leading to better discrimination. Examining parallel corpora in the framework of unified context vector space thus naturally leads to considering questions and directions that have proved fruitful in the parallel corpus literature.

Another interest of representing parallel corpora in unified context space is that they can then be used in lieu of a dictionary to connect comparable corpora: this is the topic of the next section.

<sup>5</sup>In this paper we use the plural term ‘parallel corpora’ to refer to a pair of aligned corpora, to make it easier to refer to each corpus individually as the ‘source corpus’ and the ‘target corpus’. This departs from common usage where a parallel corpus (singular) refers to a corpus of bitexts.

#### 4.5. Substituting the bilingual dictionary with a parallel corpus

Replacing the bilingual dictionary with one obtained from a pair of parallel corpora has been proposed by (Morin and Prochasson, 2011; Apidianaki et al., 2013). As explained in the previous section, parallel corpora have an advantage over a dictionary: their word alignments are found in the context of a sentence, so that the translation they show for a given (possibly ambiguous) source word in a source sentence is a correct translation of that source word in that source context, displayed in the context of the target sentence in which it occurs. In other words, parallel corpora directly implement the substitution introduced in Section 3.3. Therefore, an ideal situation when using parallel corpora would be to add them to the comparable corpora, thereby directly connecting the source and target corpora. For consistency, the parallel corpora should be in-domain, i.e., the source (resp. target) parallel corpus should be comparable to the source (resp. target) comparable corpus.

However, (Morin and Prochasson, 2011) and (Apidianaki et al., 2013) kept their parallel corpora separate from the comparable corpora. (Morin and Prochasson, 2011) used in-domain parallel corpora but discarded them after obtaining a dictionary of aligned words. (Apidianaki et al., 2013) used out-domain parallel corpora, induced word senses from them, and used these sense clusters plus information from the parallel corpora to disambiguate translations. This makes better use of the observed word distributions in the parallel corpora. Still, a step further in this direction would consist in extending the latter method by using in-domain parallel corpora: applying (Apidianaki et al., 2013)’s method to induce word senses and to translate context vectors, passing to unified context space, and adding the parallel corpora to unified context space as explained in Section 4.4.

When in-domain parallel corpora are scarce, they can be generated by machine translation from a part of the comparable corpus (Abdul-Rauf and Schwenk, 2009). Assuming that the machine translation system used to do so has been trained on a large pair of parallel corpora for the considered language pair, this creates a chain of steps which propagate translation pairs: (i) translation pairs are learned from large (out-domain) parallel corpora into the phrase table; (ii) they are used to produce (artificial) (in-domain) parallel corpora by translating existing sentences of the comparable corpora (note that this can be done in both directions); (iii) translation pairs instantiated in the artificial parallel corpora link the two comparable corpora; (iv) distributional analysis and similarity in the comparable corpora suggest new translation pairs. Some amount of loss is to be expected at each stage: as in many other directions listed in this paper, experiments will be useful to know to which extent this impedes the outlined method.

### 5. A preliminary experiment

As a preliminary, controlled experiment, we performed translation spotting in unified space in a pair of comparable corpora. We created these comparable corpora in such a way that many of their words come with tailored, low-ambiguity translations. We started from English-French

parallel corpora obtained from the *Health Canada* bilingual Web site (Deléger et al., 2009) and re-used by (Ben Abacha et al., 2013) for cross-language entity detection. The corpus was word-aligned with Fast Align (Dyer et al., 2013) in forward and reverse directions, then symmetrized with atools with the grow-diag-final option. It was then split into two halves in the order of the files (hence the topics covered by the two halves are expected to show some differences). The first half was used as an English source corpus (with French translation), and the second half as a French source corpus (with English translation).

When a source word was aligned to multiple target words, a more selective word alignment was obtained by computing an association score (discounted log odds ratio) over the word alignment links and keeping the link with the most associated target word. Links under a threshold were also discarded (we selected a threshold of 1 based on initial experiments). The target word selected this way was considered to be the translation of the source word and was pasted to it to create a bi-word as per the notations showed in Sentences 2 a and 2 b in Section 3.3. (see also Figure 5). This created two artificial comparable corpora. In each of these two corpora, some source words were mapped to target words as though through a dictionary—actually thanks to the word alignment process.

We then simulated out-of-dictionary words by surgically removing some of these translations. Given a translation pair  $[e \sim f]$ , in the English corpus we modified all bi-words  $e|*$  into  $e|$  and all bi-words  $*|f$  into  $|f$ ; in the French corpus we did the same in the opposite order. The examples cited in Section 3.3. were actually extracted from this corpus; they were obtained by removing the translation pairs  $[women \sim femmes]$  and  $[information \sim information]$  from the two parts of the corpus. We did this for several series of translation pairs: 31 among the most frequent ones, 54 at rank 1000, 45 at rank 5000, 48 at rank 10000, and 49 at rank 15000, for a total of 227 translation pairs. After this operation, the two halves of the corpus were pasted together, thus producing one bilingual corpus with  $2 \times 227$  additional out-of-dictionary words (slightly less actually since our sample of translation pairs happened to include a few common source or target words). This corpus contains 2.1 million words.

We then performed distributional analysis of this corpus in unified space: we built context vectors for each (bi)word in the corpus (minimum 5 occurrences, stop-word removal in both languages, window of 5 words left and right, discounted log odds-ratio as in (Laroche and Langlais, 2010)). Context vectors were truncated to the 1000 most associated context words. Vector similarity was computed by taking the cosine of the two vectors (we also tested the dot product).

We performed the translation spotting task by taking as source words the above 227 pairs of artificial out-of-dictionary words. For each source word, we retrieved the corresponding context vector, computed its similarity to all other context vectors, and ranked them in descending similarity order (we kept up to 500 most similar context vectors). We evaluated the results by checking whether the word with the closest context vector was the refer-

	sim	dir	F-measure
success@1	cos	f→e	0.3982
		e→f	0.4398
	dot	f→e	0.5113
		e→f	0.4213
success@o1	cos	f→e	0.6833
		e→f	0.7083
	dot	f→e	0.6606
		e→f	0.6806

Table 1: Translation spotting in unified space. N=227 test pairs in either direction; sim = similarity; cos = cosine, dot = dot product; dir = direction of translation.

ence translation (the other word of the translation pair), e.g. whether starting from *women*], the closest context vector was that for *femme* (*success@1*). Sometimes the closest context vector may represent a word of the same language. Therefore we also performed the same check restricted to out-of-dictionary words of the other language (*success@o1*, where *o* stands for out-of-dictionary and also for other). This second measure can be seen as more realistic since we have this knowledge and can use it anyway in a translation spotting task. However, out-of-dictionary words include on the one hand natural OOD words which could not be aligned reliably when preparing the corpus, and on the other hand artificial OOD words which can have a different distribution. This may bias their recognition and lead to an optimistic evaluation. Hence our trying to reduce this bias by selecting words in a variety of frequency ranges.

Table 1 displays the obtained results. A detailed analysis of this first experiment is beyond the scope of this paper; we may observe nevertheless that *success@1*, between 0.40 and 0.51, would be rather high for comparable corpora, and that *success@o1*, between 0.66 and 0.71, is as expected much higher but probably optimistic. The important point is that this exemplifies distributional analysis in unified space, where the translation links which create bi-words are obtained from parallel corpora instead of a pre-existing dictionary. The extension of this experiment by adding non-parallel texts to a parallel kernel is left for future work.

## 6. Embedding space suggests directions for future investigations

Presenting the unified context space and the connected bilingual corpus led us to mention several topics about bilingual lexicon acquisition from comparable corpora which deserve investigation. Among others we mentioned keeping whole context vectors in similarity computation instead of truncating their out-of-dictionary part; performing similarity computation directly on unified context space; performing cross-language clustering on unified context space; whether or not to merge the context vectors of in-dictionary words, and its consequence on bilingual lexicon extraction; connecting parallel corpora to unified context space; exploring the relevance of creating them through

machine translation.

The handling of the context vectors of in-dictionary words, with a source view (see the violet  $e_{m-p+d}$  column in Figure 4), a target view (violet  $f_d$  column), and possibly a merged view (not shown on the figure), is reminiscent of the feature augmentation proposed by (Daumé III, 2007) to help domain adaptation. The parallel here would be that the merged context vectors of in-dictionary words could help connect word distributions in the two “domains” (here languages), for instance when computing cross-language word clusters on unified context space.

As an application, bilingual word classes obtained through cross-language clustering can provide additional data for methods such as (Täckström et al., 2012) which aim at direct transfer of NLP components from one language to another.

How to create a merged bilingual corpus when multiple translations are provided for some words in the dictionary has been left undetermined in the above sections. A word lattice representation (more exactly, a directed acyclic graph) encoding alternative words could help solve the problem. The translation pair representation adopted in this paper would then be extended to pairs of disjunctions of words. However, this is likely to amount to merging the target (resp. source) word distributions for all alternate translations, which should be separated at least into sense clusters (see Sections 4.4. and 4.5. above).

## 7. Relation to non-standard methods of bilingual lexicon extraction from comparable corpora

The present work focuses on the above-mentioned ‘standard approach’ to bilingual lexicon extraction from comparable corpora. (Déjean et al., 2002) have proposed to extend this method by representing words through their distributional similarity to the terms of a bilingual thesaurus. That is, instead of using context vectors to represent words directly, they use context vectors to compare words to the entries of a bilingual dictionary (more precisely a thesaurus of the domain), itself represented by the context vectors of its terms as computed in the corpus. Words are thus represented by vectors of similarity values to the dictionary. The source and target parts of their comparable corpora are still used to compute context vectors, but in this method they are used as intermediate representations to obtain the similarity vectors. Since this extended method also relies on a bilingual dictionary used to translate terms occurring in the corpus, it is also a possible candidate to submit to the reformulation that we propose below. However, its bilingual dictionary is actually a thesaurus where multiword terms are a majority, and (Déjean et al., 2002)’s method does not require these multiword terms to occur as a unit: this is an obstacle to the reformulation we proposed for the standard method.

Instead of using distributional similarity in local contexts and a bilingual dictionary, some bilingual lexicon extraction methods use bilingual pairs of documents. This is the case of (Bouamor et al., 2013a) who, following (Gabrilovich and Markovitch, 2007)’s Explicit Semantic Analysis (ESA) method, represent a word by the vector of

Wikipedia pages in which it occurs. Inter-language links identify pairs of pages which describe the same entry in different languages. (Bouamor et al., 2013a) follow these links to ‘translate’ source ESA vectors into target ESA vectors, and then to identify candidate translations of the source word. Wikipedia is arguably a comparable corpus, but knowledge of the comparability (and often the translation) of document pairs is used here as a replacement for the bilingual dictionary; the method does not rely on an external pair of comparable corpora. And since translation takes place at the level of whole documents (the Wikipedia pages) rather than at the level of individual words in the texts, it seems difficult to submit it to our reformulation. Beyond bilingual extraction from comparable corpora, a reference set of parallel documents (called “anchor texts”) is also used by (Forsyth and Sharoff, 2014): it serves as a base to compute the vector of similarities (a similarity profile) of a text to every document in the set. Having translations of each base document enables the authors to use the same device as in bilingual lexicon extraction through a bilingual dictionary: the similarity profile of a text in a source language can be ‘translated’ to a target language and compared to similarity profiles of texts in the target language, hence computing inter-text similarities across languages. Again we find here the principle of multilingual linkage at the level of whole documents.

## 8. Current limitations and future work

As announced in the introduction, this paper is a first sketch of a renewed framework for studying bilingual lexicon extraction from comparable corpora. It takes a simple form when a one-to-one dictionary is used, which is the case in a large subset of the comparable corpora literature, where often the first or most frequent translation is used alone. However, when multiple translations are taken into account, we have seen that details of the representation need to be worked out.

The main limitation of the present paper is its double lack of a precise formalization and of experiments, which are left for further work. We believe it may be productive however to give early exposure of the above principles to public scrutiny, rather than deliver them piecewise with accompanying formalization and experiments. The first experiment presented in this paper, using comparable corpora built from parallel corpora, illustrates one way to put this framework into practice.

We plan to continue oracle experiments with controlled corpora, to better study the properties of the unified context space and of the merged bilingual corpus. For instance, even more constrained than the experiment of Section 5. with parallel corpora, two pseudo-comparable corpora can be built by splitting a monolingual corpus into two halves and tagging each token in each half to mark its language (say *source* and *target* as in Section 5.). This creates two comparable corpora in two ‘distinct’ languages. Then a varying proportion of the words  $w_d$  can play the role of in-dictionary words by entering the pairs  $[source] \sim [target]$  into the dictionary, while the rest of the words are kept distinct.<sup>6</sup> The ability to spot pseudo-translations in various

settings can then be evaluated, without interfering with issues linked to multiple dictionary translations.

## 9. References

- Sadaf Abdul-Rauf and Holger Schwenk. 2009. Exploiting comparable corpora with TER and TERp. In *Proceedings of the 2nd Workshop on Building and Using Comparable Corpora: From Parallel to Non-parallel Corpora*, BUCC '09, pages 46–54, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Marianna Apidianaki, Nikola Ljubešić, and Darja Fišer. 2013. Vector disambiguation for translation extraction from comparable corpora. *Informatika (Slovenia)*, 37(2):193–201.
- Marianna Apidianaki. 2008. Translation-oriented word sense induction based on parallel corpora. In Nicoletta Calzolari, Bente Maegaard Khalid Choukri, Joseph Mariani, Jan Odijk, Stelios Piperidis, and Daniel Tapias, editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May. European Language Resources Association (ELRA). <http://www.lrec-conf.org/proceedings/lrec2008/>.
- Regina Barzilay and Kathleen R. McKeown. 2001. Extracting paraphrases from a parallel corpus. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, ACL '01, pages 50–57, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Asma Ben Abacha, Pierre Zweigenbaum, and Aurélien Max. 2013. Automatic information extraction in the medical domain by cross-lingual projection. In *Proceedings IEEE International Conference on Healthcare Informatics 2013 (ICHI 2013)*, Philadelphia, USA, September. IEEE.
- Dhouha Bouamor, Adrian Popescu, Nasredine Semmar, and Pierre Zweigenbaum. 2013a. Building specialized bilingual lexicons using large scale background knowledge. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 479–489, Seattle, Washington, USA, October. Association for Computational Linguistics.
- Dhouha Bouamor, Nasredine Semmar, and Pierre Zweigenbaum. 2013b. Context vector disambiguation for bilingual lexicon extraction from comparable corpora. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 759–764, Sofia, Bulgaria, August. Association for Computational Linguistics.
- Yun-Chuang Chiao and Pierre Zweigenbaum. 2003. The effect of a general lexicon in corpus-based identification of French-English medical word translations. In Robert Baud, Marius Fieschi, Pierre Le Beux, and Patrick Ruch, editors, *Proceedings Medical Informatics Europe*, volume 95 of *Studies in Health Technology and Informatics*, pages 397–402, Amsterdam. IOS Press.

pseudo-words used in word sense disambiguation (Gale et al., 1992), which concatenate two existing words in the same language then expect a system to separate the distributions of the two original words.

<sup>6</sup>This creation of pseudo-translations is the reverse of the



- Yun-Chuang Chiao, Jean-David Sta, and Pierre Zweigenbaum. 2004. A novel approach to improve word translations extraction from non-parallel, comparable corpora. In *Proceedings International Joint Conference on Natural Language Processing*, Hainan, China. AFNLP.
- Hal Daumé III. 2007. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263, Prague, Czech Republic, June. Association for Computational Linguistics.
- Louise Deléger, Magnus Merkel, and Pierre Zweigenbaum. 2009. Translating medical terminologies through word alignment in parallel text corpora. *Journal of Biomedical Informatics*, 42(4):692–701. Epub 2009 Mar 9.
- Susan T. Dumais, Thomas K. Landauer, and Michael L. Littman. 1996. Automatic cross-linguistic information retrieval using latent semantic indexing. In *Working Notes of the Workshop on Cross-Linguistic Information Retrieval, SIGIR*, pages 16–23, Zurich, Switzerland. ACM.
- Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM Model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia, June. Association for Computational Linguistics.
- Hervé Déjean and Éric Gaussier. 2002. Une nouvelle approche à l'extraction de lexiques bilingues à partir de corpus comparables. *Lexicometrica*. Numéro spécial Alignement lexical dans les corpus multilingues, resp. Jean Véronis.
- Hervé Déjean, Éric Gaussier, and Fatia Sadat. 2002. An approach based on multilingual thesauri and model combination for bilingual lexicon extraction. In *Proceedings of the 19th COLING*, Taipei, Taiwan, 24 August–1 September.
- Stefan Evert. 2005. *The Statistics of Word Cooccurrences. Word Pairs and Collocations*. Ph.D. thesis, Universität Stuttgart.
- Richard S. Forsyth and Serge Sharoff. 2014. Document dissimilarity within and across languages: A benchmarking study. *Literary and Linguistic Computing*, 29(1):6–22.
- Pascale Fung and Kathleen McKeown. 1997. Finding terminology translations from parallel corpora. In *Proceedings Fifth Annual Workshop on Very Large Corpora*, pages 192–202. ACL.
- Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *Proceedings of the 20th international joint conference on Artificial intelligence, IJCAI'07*, pages 1606–1611, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.
- William A Gale, Kenneth W Church, and David Yarowsky. 1992. Work on statistical methods for word sense disambiguation. In *Working Notes of the AAAI Fall Symposium on Probabilistic Approaches to Natural Language*, pages 54–60.
- Pablo Gamallo and Stefan Bordag. 2011. Is singular value decomposition useful for word similarity extraction? *Language Resources and Evaluation*, 45(2):95–119.
- Éric Gaussier, J.M. Renders, I. Matveeva, Cyril Goutte, and Hervé Déjean. 2004. A geometric view on bilingual lexicon extraction from comparable corpora. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume*, pages 526–533, Barcelona, Spain, July.
- Benoît Habert and Pierre Zweigenbaum. 2002. Contextual acquisition of information categories: what has been done and what can be done automatically? In Bruce E. Nevin and Stephen M. Johnson, editors, *The Legacy of Zellig Harris: Language and information into the 21st Century – Vol. 2. Mathematics and computability of language*, pages 203–231. John Benjamins, Amsterdam.
- Zellig Sabbetai Harris. 1988. *Language and information*. Columbia University Press, New York.
- Zellig Sabbetai Harris. 1991. *A theory of language and information. A mathematical approach*. Oxford University Press, Oxford.
- Amir Hazem and Emmanuel Morin. 2012. Adaptive dictionary for bilingual lexicon extraction from comparable corpora. In *LREC 2012, Eighth International Conference on Language Resources and Evaluation*, pages 288–292, Istanbul, Turkey. ELRA.
- Jagadeesh Jagarlamudi and Hal Daumé III. 2010. Extracting multilingual topics from unaligned comparable corpora. In *Proceedings of the 32nd European Conference on Advances in Information Retrieval, ECIR'2010*, pages 444–456, Berlin, Heidelberg. Springer-Verlag.
- Audrey Laroche and Philippe Langlais. 2010. Revisiting context-based projection methods for term-translation spotting in comparable corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics, COLING '10*, pages 617–625, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Bo Li and Éric Gaussier. 2010. Improving corpus comparability for bilingual lexicon extraction from comparable corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 644–652, Beijing, China, August. Coling 2010 Organizing Committee.
- Emmanuel Morin and Emmanuel Prochasson. 2011. Bilingual lexicon extraction from comparable corpora enhanced with parallel corpora. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, pages 27–34, Portland, Oregon, June. Association for Computational Linguistics.
- Yves Peirsman and Sebastian Padó. 2010. Cross-lingual induction of selectional preferences with bilingual vector spaces. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 921–929, Los Angeles, California, June. Association for Computational Linguistics.
- Reinhard Rapp. 1995. Identifying word translation in non-



- parallel texts. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics, student session*, volume 1, pages 321–322, Boston, Mass.
- Fangzhong Su and Bogdan Babych. 2012. Development and application of a cross-language document comparability metric. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May. European Language Resources Association (ELRA).
- Oscar Täckström, Ryan McDonald, and Jakob Uszkoreit. 2012. Cross-lingual word clusters for direct transfer of linguistic structure. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT '12*, pages 477–487, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Xuchen Yao, Benjamin Van Durme, and Chris Callison-Burch. 2012. Expectations of word sense in parallel corpora. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT '12*, pages 621–625, Stroudsburg, PA, USA. Association for Computational Linguistics.