Mining missing (French) Translations in DBpedia

Rifi and Fifi

Somewhere
some.where.org

Abstract. The present work addresses the issue of enriching DBpedia with information facilitating Semantic Web Annotation (SWA) in French. Each entry (concept) in DBpedia comes along a set of surface strings (property rdfs:label) which are possible realizations of the concept being described. Those are the strings that are typically used in SWA systems. Currently, only a fifth of the English DBpedia entries have a surface string in French. If we want SWA to perform similarly in other languages, and French in particular, we must find a way to automatically produce surface strings in other languages, for each concept in DBpedia. This is the problem we consider in this paper.

Keywords: Semantic Web, DBpedia, Translation Mining, Comparable Corpora

1 Introduction

The LOD (Linked Open Data) [Bizer, 2009] is conceived as a language independent resource in the sense that the information is represented by abstract concepts to which "human-readable" strings possibly in different languages are attached, *e.g.* the rdfs:label properties in DBpedia. For instance, we can access the abstract concept of *computer* by natural language queries such as **ordinateur** (rdfs:label@fr) in French or **computer** (rdfs:label@en) in English. Thanks to this, Semantic Web offers the advantage of having a truly multilingual World Wide Web [Gracia et al., 2012].

At the core of LOD, lies DBpedia [Jens Lehmann, 2014], the largest dataset that constitutes a hub to which most other LOD datasets are linked[1]. Since DBpedia is (automatically) generated from Wikipedia, which is multilingual, one would expect that each concept in DBpedia is labeled with a French surface string. This is for instance the case of the concept House of Commons of Canada[2] which is labelled in French as Chambre des communes du Canada. One problem however, is that most labels are currently in English [Gómez-Pérez et al., 2013].

Indeed, the majority of datasets in LOD are generated based on the extraction of anglophone resources. DBpedia, the endogenous RDF dataset of

[1] December 2014 - http://lod-cloud.net/

[2] http://dbpedia.org/page/House_of_Commons_of_Canada

Wikipedia is at no exception here, since it proposes labels in French (`rdfs:label@fr`) for only one fifth[3] of the concepts. For instance, the concept `School life expectancy`[4] has — at least at the time of writing — no label in French, while `Esprance de scolarisation` or `dure moyenne de scolarit` are valid translations in French. This comes from the fact that currently, a concept in `DBpedia` receives as its `rdfs:label` property in a given language the title of the Wikipedia article which is interlanguage linked to the (English) Wikipedia article associated to the `DBpedia` concept. And currently, only a fifth of the Wikipedia articles in English have an interlanguage link to a French Wikipedia article, the target language we are interested in here.

The lack of surface strings in a foreign language does not only reduce the usefulness of RDF indexing engines such as `sig.ma`[5], but also limits the deployment of Semantic Web Annotator (SWA) systems (*e.g.* [**?**,**?**]). This motivates the present study, which aims at automatically mining French labels of concepts in `DBpedia` that do not possess one yet.

The remaining of this article is organized as follows. We describe the approaches we tested in Section 2. Our experimental protocole is presented in Section 3. Section 4 reports the results we obtained. We conclude in Section 5.

2 Approches

In this section, we describe two families of approaches that can potentially solve the problem of finding the translations (into French) of an English surface string which corresponds to a concept in `DBpedia`, and for which no `rdfs:label` property is known in French. The first approach consists in mining external parallel corpora and is described in section 2.1. The other family encompasses three approaches we implemented which all make use of Wikipedia, seen as a comparable corpus. Those approches are described in section 2.2, section 2.3, and section 2.4.

2.1 Mining Parallel Corpora

One way to find translations is to mine a large parallel corpus, that is, a set of sentences in translation relation. This is the realm of statistical word alignment for which a decent accuracy is typically reported (in the order of 80% according to [Bourdaillet et al., 2010]). In this study, we assume the existence of a perfect translation spotter and we simply count the number of pair of sentences in which a given term and its reference translation are found in the sentence pairs of a given parallel corpus. We call this situation a *match*, an example of which is illustrated in Figure 1.

[3] `http://wiki.dbpedia.org/Datasets/DatasetStatistics`
[4] `http://dbpedia.org/page/School_life_expectancy`
[5] `http://sig.ma`

GigaWord^{en}	GigaWord^{fr}
21677278: *French and English* greeting *translation cards* [...]	21677278: *Des fiches de* salutation *en anglais et en franais* [...]

Fig. 1. Example of a pair of parallel sentences in the corpus GigaWord where a match occurs between `greeting` and its french translation `salutation`.

2.2 Rapp

Identifying the translations of a term in a comparable corpus — two texts (one in each language of interest) that share similar topics without being in translation relation — is a challenge that has attracted many researchers. A popular idea that emerged for solving this problem is based on the assumption that the context of a term and its translation share similarities that can be used to rank translation candidates [**?**,Rapp, 1995]. Many variants of this idea have been implemented, see for instance [Bouamor, 2014].

In a nutshell, a term to translate is being represented by a so-called *context vector*, the set of words that cooccur in the source language more often than chance to the term. An association measure is typically used to score the strength of the correlation between the term and the context words. Each candidate translation (typically all the target words) is similarly represented in the target language. Thanks to a bilingual seed lexicon, the source context vector is projected into a target one[6], which is compared to the vectors of all the candidates by the means of a similarity measure.

The main parameters of this approach are a) the size of the window used to collect concurrents of the term to translate (we varied it), b) the association measure, c) the similarity used to compare context vectors, and d) the seed lexicon used for projecting source vectors. We followed the recommandation of [Laroche and Langlais, 2010] and used the discontinuous odds-ratio [**?**, p. 86] for measuring the degree of correlation between a term and its context words (see eq. 1) as well as the cosine measure for comparing vectors (see eq. 2).

$$oddsRatio_{disc} = \log \frac{(O_{11} + \frac{1}{2})(O_{22} + \frac{1}{2})}{(O_{12} + \frac{1}{2})(O_{21} + \frac{1}{2})} \tag{1}$$

where O_{ij} is given by the contingency table of two words w_1, w_2 where, for instance, O_{12} stands for the number of times w_2 occurs in a window, while w_1 does not:

	w_1	$\neg w_1$
w_2	O_{11}	O_{12}
$\neg w_2$	O_{21}	O_{22}

[6] When no translation is found for a source word, the latter is left as is. On the contrary, multiple translations are all added to the target context vector

$$cos(v_{src}, v_{trg}) = \frac{v_{src} \cdot v_{trg}}{\|v_{src}\| \cdot \|v_{trg}\|} \quad (2)$$

In the standard approach, the cooccurrent words are extracted from all the source documents of the comparable corpus in which the term to translate appears. We name this variant ALL hereafter. Since in our case, a term to translate is indeed the title of a Wikipedia article, a somehow natural way of populating the context vector of a term is to consider its only occurrences in this article; we call this variant RA (for reference article). This article may be small, thus limiting the number of words collected in the context vector. Therefore we investigated sets of articles in which the concurrents of a term are searched for, using *neighbourhood functions*, two of which come from Wikipedia article structure. Precisely, for an article a, we consider:

$\mathcal{N}_{LKI}(a)$, the set of articles that have a link to a (in-links),
$\mathcal{N}_{LKO}(a)$, the set of articles to which a points to (out-links),
$\mathcal{N}_{LKC}(a)$, the set of articles that are the most similar to a according to the *MoreLikeThis* method of the search engine Lucene[7] we used for indexing the collection of source texts available.

We note $\mathscr{D}_{src}$ the set of documents in which the term to translate is being searched in order to populate the context vector. We considered many combinations of those neighbourhood function. For example, $\mathscr{D}_{src} = \mathcal{N}_{LKI} + \mathcal{N}_{LKO} + \mathcal{N}_{LKC}$ means that the subset of Wikipedia articles considered is the union of all the articles returned by each function (*LKI*, *LKO* and *LKC*). We can also visit only a subset of the articles returned by a neighbourhood function, so $\mathcal{N}_{LKI50}$ indicates that only the first 50 titles returned by the $\mathcal{N}_{LKI}$ function (random order) are members of $\mathscr{D}_{src}$.

2.3 EsaB

We implemented the approach described in [Bouamor, 2014] which has been shown by the author to be more accurate than the aforementioned standard approach. The proposed method is an adaptation of the *Explicit Semantic Analysis*(ESA) approach described in [Gabrilovich and Markovitch, 2007] where a term to translate is represented by the titles of the articles in which it appears. The projection of the resulting context vector is obtained without a bilingual lexicon by simply following the interlanguage links of the related (English) articles, thus leading to a context vector of French titles.[8] The candidate translations are searched for among the words of the French articles obtained by projection (using a tf-idf score).

[7] `http://www.lucene.org`
[8] Articles with interlanguage links are simply ignored.

2.4 EsaT

This approach takes advantage of the fact that the terms to translate are titles of articles in Wikipedia. For each of them, a context vector is constructed which contains the titles of the articles identified by the neighbourhood functions described in section 2.2 ($\mathcal{N}_{LKI}$, $\mathcal{N}_{LKO}$ and $\mathcal{N}_{LKC}$). The titles of a context vector are translated similarly to EsaB, by picking the title of the article identified by the (French) interlanguage link associated to the source title. After translation, the target context vector is compared to the vectors computed for all the candidate (French) titles.[9].

It is important to note that this approach can only propose candidate translations that are titles in the French Wikipedia. This can still be useful for identifying missing interlanguage links, a real concern since those links are manually populated [**?**].

3 Experimental Protocole

In this section, we present our experimental protocole.

3.1 Reference List

To evaluate our different approaches, we built a test set of single word English terms for which selected as a gold standard translation the title of the French article to which the interlanguage link points. We initially built a list of 1000 terms, but realized in the curse of investigations that many of them where actually named entities, for which translation is not required. In order to avoid dealing with this problem at first, we intersected the list with 3 lexicons[10]. Some examples of pairs are displayed in table 1. In the end, we ended up with 125 terms.

t_s	t_c (Reference Translation)
`Efficiency`	`Efficience`
`Greeting`	`Salutation`
`Joystick`	`Joystick`
`Junk food`	`Malbouffe`
`Woodcut`	`Xylographie`

Table 1. Examples of pairs of terms and its sanctioned translation.

[9] 79 152 Wikipedia^{fr} titles in our case.

[10] Ergane: 29387 entries - `http://download.travlang.com`; Starbucks: 187749 entries - In house lexicon; Freelang: 38877 entries - `http://www.freelang.net`

Evaluation Except for the translation mined from parallel corpora, we ask our approach to produce a ranked list of (at mots) 20 candidates for each source term. The reference translation t_c is searched in the list of translation candidates and success is possible from 1 to r (variable r is called *rank* and its maximum is 20 in our case). We compute Mean Average Precision at rank 20 (MAP_{20}) [Manning et al., 2008].

$$MAP_{rmax}(L) = \frac{1}{|L|} \sum_{t_s=1}^{|L|} \frac{1}{ntr_{t_s}} \sum_{k=1}^{ntr_{t_s}} Precision(C_{t_s}) \tag{3}$$

$$Precision(C_{t_s}) = \begin{cases} \frac{1}{r} \text{ if } t_c \text{ (de } t_s) \in C_{t_s} \text{ at rank } r(<= rmax) \\ 0 \text{ else} \end{cases}$$

where $|L|$ is the number of terms t_s from our list L (125 in our case), ntr_{t_s} is the number of reference translation for the t_s^{th} term (always 1 in our case) and $Precision(C_{t_s})$ is 0 if the reference translation isn't found for the t_s^{th} term or $\frac{1}{r}$ if its is (r is the rank of the reference translation t_c in the translation candidates list C_{t_s}).

3.2 Corpora

We evaluated the usefulness of both parallel and comparable corpora to the task of mining missing translations into French of concepts in `DBpedia`.

Parallel Corpora The first approach we considered exploits bitexts, that is, pairs of documents that are in translation relation (also called a parallel corpus), and where sentences that are in translation relation are identified. We have selected four bitexts, which main characteristics are detailed in table 2 where Hansards refers to Canadian parliament debates published in French and English between 1986 and 2007; Europarl stands for the English-French bitext typically used for training statistical engines[11]; Health gathers over 800k sentence pairs mined from the websites of *Health Canada*[12] and the *Public Health Agency of Canada*[13], two websites that are particularly well organized from a bilingual point of view; and GigaWord[14], as its name stands, gathers over 10^9 words coming from institutional texts downloaded over the Web.

Comparable Corpus While a parallel corpus is a valuable ressource for mining translations, it is widely agreed that a comparable corpus — that is a pair of documents in two languages that share some similarities (*e.g.* are about the same topic) — is a more widely available resource. We considered Wikipedia

[11] `http://www.statmt.org/europarl/`
[12] `www.hc-sc.gc.ca`
[13] `http://www.phac-aspc.gc.ca/`
[14] `http://www.statmt.org/wmt13/translation-task.html`

Corpus	$\#Sent.^{pairs}$	$\#Words^{en}$	$\#Words^{fr}$
	(in thousands)		
HANSARDS	8 802	242	253
EUROPARL	2 007	132	142
HEALTH	809	133	142
GIGAWORD	22 520	3 002	2 775

Table 2. Characteristics in numbers of words and sentences of each selected corpus.

as a natural comparable corpus, since DBpedia is extracted from Wikipedia [Jens Lehmann, 2014]. It is generally accepted that the articles *paired with an interlanguage link* provide a translation relation between their respective title, while the content of the articles are in comparable relationship [Hovy et al., 2013]. Thus, the textual content of these pairs of articles are the comparable resources we considered.

4 Results

4.1 Mining Parallel Corpora

If we consider that more frequent terms are more likely to be correctly aligned by a statistical word aligner, and taking as a conservative threshold that matches should have a frequency of at least 5 in a parallel corpus, we measured that GIGAWORD — the larget parallel corpus we considered — could be used to retrieve 76.8% of the translations of our test set (96 terms). Gluing the 4 parallel corpora we considered only slightly increases the performance (99 translations). The remaining terms are either absent from the parallel material, or the translations available are different than the sanctioned one.

In [Bourdaillet et al., 2010], the authors report that their alignment model is characterized by a precision of 78.3%. Therefore a gross estimate is that nearly 60% of the translations of our test set could be identified perfectly thanks to our parallel data, a not too bad figure.

4.2 Rapp

Table 3 summarizes the experiments we conducted with the standard approach using Wikipedia as a source of comparable corpora. It deserves several comments. First, we observe that increasing the size of window in which we collect the context words leads to noise. The optimal window size seems to be almost 3, that is the cooccurrent words should be taken in the immediate vicinity of the term to translate. This corroborates the study in [**?**]. For this reason, we stick to this value for subsequent experiments. Second, and somehow disappointedly, we observe that the MAP measure is far higher when the term to translate is considered in all the documents of Wikipedia(0.75). Considering the occurrences of the term

in the content of the article only is drastically impacting the performance (0.19). Among the neighbourhood functions, $\mathcal{N}_{LC}$ seems the less appropriate. None of their combinations lead to increased performance, even if this does increase the number of documents inspected. We do not have a clear explanation for this. Last, all our variants outperform one where random documents are sampled when building the context vector, which to some extent demonstrates that our neighbourhood functions do capture useful material.

		Window Size			
		1	3	7	15
Documents Set $\mathscr{D}_{src}$	ALL	0.72	**0.75**	0.62	0.55
	RA	0.10	0.19	0.12	0.09
	$\mathcal{N}_{LKI50}$	**0.30**	**0.33**	0.12	0.06
	$\mathcal{N}_{LKO50}$	0.22	**0.34**	0.13	0.05
	$\mathcal{N}_{LKC50}$	0.22	0.27	0.11	0.05
	$\mathcal{N}_{LKI50} + RA$	0.26	0.27	0.10	0.05
	$\mathcal{N}_{LKI50} + \mathcal{N}_{LKO50}$	0.34	0.32	0.13	0.05
	$\mathcal{N}_{LKI100}$	0.31	0.31	0.11	0.06
	$\mathcal{N}_{LKI1000}$	0.29	0.25	0.12	0.08
	$\mathcal{N}_{RD50}$	0.00	0.00	0.0016	0.0026

Table 3. MAP_{20} of some variants of the standard approach (RAPP). $\mathcal{N}_{RD50}$ stands for a neighbourhood function which random picks 50 (English) articles.

4.3 EsaB

The performance of the ESAB is 0.46, which is significantly lower than the standard approach which uses all the Wikipedia articles for populating the context vector (0.75), but higher than the variants we tested which use a subset of the articles only (0.34). This result contradicts the better performance of ESAB observed in [Bouamor, 2014]. On reason could be that she tested the translation of domain specific terms while we are translating all kinds of terms.

We inspected the terms for which our implementation failed and observed that often, the approach is actually victim of some sort of *semantic drift*. For instance, while translating the term `tears`, the context vector is populated with articles related to music albums that contain this term in their title, while the associated French article (when available) almost never contains the translation. We also noted that in our implementation, we set the size of the context vector to 100 which often exacerbates this issue. Since the titles in the context vector are sorted by their association strength (tf-idf) to the term to translate, the first articles are more likely to be relevant. We mesure this in table 4 where we observe that keeping 20 titles only seems to deliver the best performance.

nb. of article titles	MAP$_{20}$
5	0.45
10	0.55
20	**0.57**
30	0.54
50	0.51
100	0.46

Table 4. MAP$_{20}$ of EsaB as a function of the context vector dimension.

4.4 EsaT

Table 5 summarizes the performances of the EsaT variants we tested. It is striking that most variants outperform the Rapp approach, with a MAP$_{20}$ performance sometimes above 80%, which outperforms all the approaches reported so far. One must recall however that the translations ranked by EsaT are titles of French articles. In many case of interest, there is simply no French title available in Wikipedia which corresponds to the English title we seek to translation to.

In order to measure this, we translated with EsaT121 article titles in English that do not have a French counterpart (no interlanguage link) in Wikipedia. We manually rated each translation produced as relevant or not. Only 26 terms (21.4%) had a relevant translation in the top 20 ones proposed, and only 19 terms (15.7%) had their first translation rated relevant.

		Subset of $Wikipedia^{fr}$ titles						
		$\mathcal{N}_{LKI}$	$\mathcal{N}_{LKO}$	$\mathcal{N}_{LKC}$	$\mathcal{N}_{LKI}+\mathcal{N}_{LKO}$	$\mathcal{N}_{LKI}+\mathcal{N}_{LKC}$	$\mathcal{N}_{LKO}+\mathcal{N}_{LKC}$	$\mathcal{N}_{LKI}+\mathcal{N}_{LKO}+\mathcal{N}_{LKC}$
$\mathscr{D}_{src}$	$\mathcal{N}_{LKI}$	*0.76*	0.49	0.33	**0.79**	0.53	**0.79**	**0.80**
	$\mathcal{N}_{LKO}$	0.34	*0.57*	0.28	0.48	0.58	0.45	0.56
	$\mathcal{N}_{LKC}$	0.28	0.34	*0.31*	0.31	0.40	0.35	0.36
	$\mathcal{N}_{LKI}+\mathcal{N}_{LKO}$	0.75	0.64	0.34	**0.81**	0.63	**0.78**	**0.82**
	$\mathcal{N}_{LKI}+\mathcal{N}_{LKC}$	0.74	0.53	0.34	0.77	0.54	0.75	0.77
	$\mathcal{N}_{LKO}+\mathcal{N}_{LKC}$	0.41	0.59	0.31	0.54	0.58	0.47	0.58
	$\mathcal{N}_{LKI}+\mathcal{N}_{LKO}+\mathcal{N}_{LKC}$	0.73	0.63	0.33	**0.80**	0.61	0.74	**0.79**

Table 5. MAP$_{20}$ of some EsaT variants.

4.5 Combination

In this section we evaluate the complementarity of the three approaches we tested that are making use of Wikipedia. This is done by implementing a rather crude combination of them, which basically aggregates the candidates produced by each approach, rewarding candidates produced by several approaches. This is reported in Table 6. Our crude combination outperforms each approach alone, which indicates that the approaches are complementary and the fact that several propose a given term is a useful information. While this is encouraging, we must note our small test set.

Approach	MAP$_{20}$	MAP$_1$
ESAT ($\mathcal{N}_{LKI} + \mathcal{N}_{LKO}$)	0.82	0.79
RAPP (ALL)	0.75	0.73
ESAB (20 articles)	0.57	0.43
Combination	**0.97**	0.96

Table 6. Combination of some variants.

4.6 BabelNet

Identifying the translations of Wikipedia articles in English is partially solved in the BabelNet project [Navigli and Ponzetto, 2012]. The translation of concepts in Wikipedia that are not interlanguage linked are taken care of by applying machine translation on (minimum 3 and maximum 10) sentences extracted from Wikipedia that contain a link to the article which title they seek to translate. The most frequent translation is finally selected. We roughly evaluated the coverage and the quality of translations proposed by BabelNet[15].

Among the 3 573 789 English Wikipedia articles, 2 816 502 do not have an interlanguage link to a French Wikipedia article (78, 80%). We removed from those terms the named entities which are numerous in Wikipedia [Hovy et al., 2013], and for which no translation is required.[16]. It remains 521 895 entries. Only 66 242 (13%) of them have a translation in BabelNet. We randomly sampled 50 of those translations and evaluated their quality. This was a consuming tasks for which we had to read the Wikipedia articles, and use available resources such as *Google Translation*[17] and Linguee[18]. We observed that 38 terms (76%) have a correct translation in BabelNet. Clearly, this suggests that the projection of a resource such as `DBpedia`into French is not a solved problem.

[15] Version 2.0.1 - Mars 2014
[16] We filtered named entities thanks to heuristics used by BabelNet
[17] `https://translate.google.ca/`
[18] `http://www.linguee.com/english-french/`

5 Conclusion

We have proposed, implemented and tested a number of variants for mining translations in Wikipedia. We have tested it on a small test set (125 terms) for which we knew the reference translation (the one sanctioned by Wikipedia). Our results suggest that the standard projective approach behaves rather well (over 75% MAP_{20}) and better than some variants proposed recently by others. We also verified that 60% of the translation could be found in parallel material, an approach which should not be neglected at least for the English-French language pair for which large parallel corpora are available.

This work suggests a number of issues that should be addressed before deploying our technology to all the terms that require to be translated in `DBpedia`. First, control tests should be rerun on a larger test set. Second, we measured that among the non named entity terms in `DBpedia` that do not have a translation, only a third are single terms, the case we studied here. Multiword expressions in natural language processing [Sag et al., 2002] and their translation in particular is a problem that deserves further investigations. Also, all the approach we described are sensitive to the frequency of the term being translated in the collection mined, and we should look at this specifically.

Acknowledgments

This version of the paper clearly lacks proofreading. We are sorry the inconvenience it may have caused to the reader. The final version of this article will be thoroughly revised.

References

[Bizer, 2009] Bizer. 2009. Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems*, 4(2):1–22.

[Jens Lehmann, 2014] Jens Lehmann, R. I. (2014). DBpedia - a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web Journal*.

[Gracia et al., 2012] Gracia, J., Montiel-Ponsoda, E., Cimiano, P., Gmez-Pérez, A., Buitelaar, P., and McCrae, J. (2012). Challenges for the multilingual web of data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 11:63–71.

[Gómez-Pérez et al., 2013] Gmez-Pérez, A., Vila-Suero, D., Montiel-Ponsoda, E., Gracia, J., and Aguado-de Cea, G. (2013). Guidelines for multilingual linked data. In *Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics*, WIMS '13, pages 3:1–3:12, New York, NY, USA. ACM.

[Navigli and Ponzetto, 2012] Navigli, R. and Ponzetto, S. P. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

[Hovy et al., 2013] Hovy, E., Navigli, R., and Ponzetto, S. P. (2013). Collaboratively built semi-structured content and artificial intelligence: The story so far. *Artificial Intelligence*, 194:2–27.

[Manning et al., 2008] Manning, C. D., Raghavan, P., and Schtze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

[Bourdaillet et al., 2010] Bourdaillet, J., Huet, S., Langlais, P., and Lapalme, G. (2010). TransSearch: from a bilingual concordancer to a translation finder. *Machine Translation*, 24(3-4):241–271.

[Rapp, 1995] Rapp, R. (1995). Identifying word translations in non-parallel texts. In *Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics*, ACL '95, pages 320–322, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Bouamor, 2014] Bouamor, D. (2014). *Constitution de ressources linguistiques multilingues à partir de corpus de textes parallèles et comparables*. PhD thesis, Université Paris Sud - Paris XI.

[Laroche and Langlais, 2010] Laroche, A. and Langlais, P. (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.

[Gabrilovich and Markovitch, 2007] Gabrilovich, E. and Markovitch, S. (2007). Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, IJCAI'07, pages 1606–1611, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

[Penta et al., 2012] Penta, A., Quercini, G., Reynaud, C., Shadbolt, N., and others (2012). Discovering cross-language links in wikipedia through semantic relatedness. In *ECAI*, pages 642–647.

[Sag et al., 2002] Sag, I. A., Baldwin, T., Bond, F., Copestake, A. A., and Flickinger, D. (2002). Multiword expressions: A pain in the neck for NLP. In *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing '02, pages 1–15, London, UK, UK. Springer-Verlag.

[Sharoff, 2004] Sharoff, S. (2004). What is <u>at stake</u>: A case study of russian expressions starting with a preposition. In *Proceedings of the Workshop on Multiword Expressions: Integrating Processing*, MWE '04, pages 17–23, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Prochasson and Fung, 2011] Prochasson, E. and Fung, P. (2011). Rare word translation extraction from aligned comparable documents. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 1327–1335, Stroudsburg, PA, USA. Association for Computational Linguistics.