# Experimental Studies for Revealing Key Factors of Cross-Language Ontology Alignments

**Juliana Medeiros Destro**[1], **Julio Cesar dos Reis**[1], **Ariadne Maria Brito Rizzoni Carvalho**[1], **Ivan Luiz Marques Ricarte**[2]

[1] Institute of Computing, University of Campinas, Brazil

[2]School of Technology, University of Campinas, Brazil

{juliana.destro,jreis,ariadne}@ic.unicamp.br, ivan.ricarte@ft.unicamp.br

***Abstract.*** *Cross-language alignment between ontologies is relevant for the interoperability of systems in specific domains, such as in the life science domain. Although the literature has proposed techniques for the alignment of ontologies described in different languages, the influence of linguistic characteristics from domain-specific ontologies on such alignments has barely been appraised. This study proposes a series of experiments based on real-world mappings to understand the matching between ontologies in different languages. It investigates the role of a pivot-language related to the domain for the purpose of a fully automatic cross-language alignment. In particular, we analyse the influence of syntactic and semantic similarity methods and the structure of terms denoting concepts in ontologies. Experimental results, focused on the life science domain, indicate useful factors to take into account in the design of matching algorithms for domain-specific cross-language alignment.*

## 1. Introduction

Modern information systems explore several ontologies, described with different natural languages, in domain-specific contexts, such as in life sciences. Mapping between ontologies as the outcome of an alignment process plays a central role in data integration and other semantic-enabled analysis tasks. For instance, a set of mappings is required to explicitly interconnect concepts from different ontologies to allow the smooth interoperability among systems.

Although several contributions for ontology matching do exist [Shvaiko and Euzenat 2013], well-defined techniques to perform cross-language alignment between ontologies still deserve further studies. The issue of considering different languages poses additional difficulties for automatic matching, because straightforward string-based techniques may have limitations. Whereas some preliminary studies have investigated cross-language alignment between ontologies and multilingual matching [Trojahn et al. 2014], the literature lacks thorough empirical studies to understand the influence of linguistic characteristics for this task.

In this article, we conduct a series of experiments with several sets of cross-language ontology mappings. We systematically investigate underlying factors for the alignment between concepts from biomedical ontologies defined in different languages.

We aim at determining and understanding relevant properties that might allow for automatic cross-language matching in life sciences. In particular, we investigate the cross-language similarity between the concepts by using a translated version of the concepts and the original version of the other concept involved in the mapping. We are not concerned with similarity between the original concepts of mappings; instead, we investigate the degree of similarity between the translated and the original label of the interrelated concepts. In summary, this work makes the following contributions:

- Design thorough and original experiments to analyze key factors that might result in the correct mapping between biomedical concepts declared in different languages.
- Conduct extensive experiments by using real-world interconnected biomedical ontologies to obtain empirical evidences from the analyses that might be useful for the development of novel cross-lingual matching techniques.

We explore large biomedical ontologies defined in English and Spanish, and existing mapping sets between them available in open repositories. In our procedure, the interrelated concepts for a given mapping are translated to a pivot-language. We execute four distinct experiments with two analyses in each of them to examine linguistic, structural and similarity aspects between the original concept content, and its translated form. Results indicate that the choice of the pivot-language plays an important role in cross-language matching and the structure of concept elements affects the effectiveness of the semantic similarity, and that semantic similarity measure heavily depends on the domain corpus available in the target pivot-language.

The remaining of this article is organized as follows: Section 3 reports on the organization and description of the experiments; Section 4 describes the obtained results. Section 5 discusses the related work and our findings; and finally, Section 6 draws our conclusions and future work.

## 2. Related Work

There has been a number of investigations on specific aspects of cross-language for ontology matching. The work of Meilicke *et al.* [Meilicke et al. 2012] studies the performance of a set of matching systems based on a dataset defined to evaluate ontology alignment. Their results indicate the difficulties of traditional ontology matching algorithms for carrying out multilingual ontology alignment. Similarly, Trojahn *et al.* [Trojahn et al. 2014] describe an extensive survey of matching systems and strategies for accomplishing multilingual and cross-lingual ontology matching.

Several approaches explore the translation effects and the use of a third language in cross-language ontology alignment. In particular, Fu *et al.* [Fu et al. ] analysed the impact of automatic translations on multilingual ontology alignment, highlighting the translation's relevance for achieving adequate matching quality. Spohr *et al.*[Spohr et al. 2011] studied the translation of concept labels to a third language for matching two ontologies described in different languages.

Similar investigations also emphasized the use of a third language on a theoretical approach of indirect alignment between multilingual ontologies [Jung et al. 2009]. A noteworthy approach is explored by CroLOM (Cross-Lingual Ontology Matching System), which used translation together with a hybrid syntactic and semantic similarity

computation, increasing accuracy of the obtained mappings [Khiat 2016]. Even though CroLOM explored syntactic and semantic similarity measures to perform ontology matching, the approach did not shed light on how the different elements of the ontology concepts can impact the matching process.

Although these proposals have attempted to reach automatic cross-lingual ontology alignment, linguistic characteristics of the domain are not taken into account when choosing a pivot-language for translation. Our research aimed at empirically shedding light on key aspects of concept structure similarities involved in identification of cross-language mappings. To the best of our knowledge, this has not been done before. In addition, we studied the potential impact of choosing a linguistic pivot-language that is relevant for translating both ontologies to a target language.

## 3. Study Design

This study aims at describing the role played by similarity in cross-language ontology alignments using a set of real-world ontology mappings. Section 3.1 presents preliminary definitions. We describe the experimental setup in Section 3.2, which is followed by the description of the experiments (Section 3.3). Section 3.4 reports on the conducted analyses and used datasets.

### 3.1. Preliminaries

This work considers an ontology $O$ as a set of concepts interrelated by relationships, *e.g.*, "*is-a*", "*part-of*", "*related-to*" [Gruber 1995]. The set of concepts of an ontology $O_x$ is defined as $Concepts(O_x) = \{C_1, C_2, ..., C_n\}$. Each concept is characterized by a unique *identifier*, a *preferred label* and a set of *terms*.

Given a concept $C_k \in Concepts(O_x)$, $L(C_k)$ defines the value of the preferred label of $C_k$ expressing its local name denoted by a natural language string. For example, *"cardio vascular diseases"* describes the label of a concept. The labels can be defined by properties like *rdfs:label* and *skos:prefLabel*. We also define the set of terms (strings) to further characterize a concept $C_k$ as $T(C_k) = \{t_1, t_2, ..., t_n\}$. Terms provide additional information about the concept including its definition, a list of synonyms, *etc*. Each term has a particular semantics and may differ from one ontology to another. For instance, synonym terms define equivalent terms with respect to meanings, *e.g.*, the term "*hypotension*" is the synonym of "*low blood pressure*".

The translation of a concept $C_k$ is denoted by $C_k^T$. Given that the result of $L(C_k)$ and $T(C_k)$ is expressed in a language $\beta$, the label $L(C_k^T)$ and terms $T(C_k^T)$ of $C_k^T$ are expressed in $\alpha$ (pivot-language) as a different language.

A *mapping* $m_{ab}$ is established between two given concepts $C_a$ and $C_b$ from two different ontologies as $m_{ab} = (C_a, C_b, sim, \equiv)$, where $\equiv$ concerns the semantic relation connecting $C_a$ and $C_b$, $C_a \in Concepts(O_x)$ and $C_b \in Concepts(O_y)$. The $sim \in [0, 1]$ value represents the similarity measure between $C_a$ and $C_b$. In this work, we only consider equivalent concept-to-concept mappings. The $\mathcal{L}_{XY} = \{(m_{ab})_i | i \in \mathbb{N}\}$ consists of the set of different mappings between two ontologies $O_x$ and $O_y$ as the result of an alignment process.

### 3.2. Experimental Setup

The experiments investigate the similarity between the original and translated version of the concepts from a given mapping. Figure 2 presents a mapping with the involved concepts and their translation. The similarity function between two elements of a concept is given by $sim(el_1, el_2) \in [0, 1]$. The elements are strings representing a label or a synonym.

Figure 1 shows the elements denoting the concepts and the approach to examine the similarity between them. We study the similarity between labels only (*cf.* a in Figure 1), and the similarity between labels and synonyms of concepts involved in a mapping (*cf.* b in Figure 1).

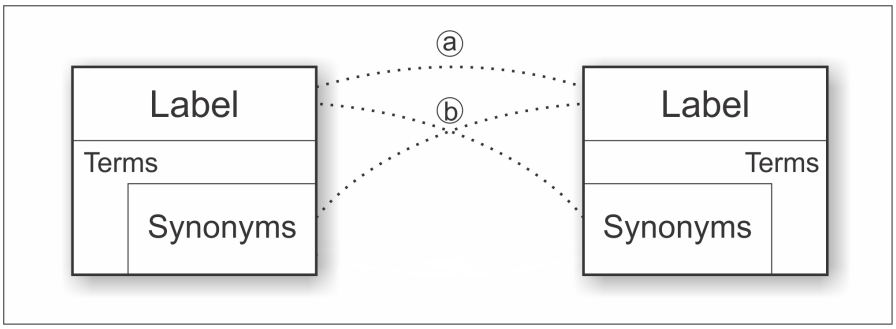


**Figure 1. Study of similarity between elements of cross-language concepts.**

In order to understand the role played by a pivot-language for matching ontologies in different languages, this work considers the similarity between the translated version of concepts involved in a mapping and its original content. To this end, given a mapping $m_{ab} \in \mathcal{L}_{XY}$, the first step was to translate the involved concepts. From the concepts $C_a$ and $C_b$ interrelated by the mapping, the translation outcome results in $C_a^T$ and $C_b^T$ (*cf.* Figure 2).

The translation is applied to label and terms of concepts to $\alpha$ (the Latin language), which differs from the original $\beta$ (English language) or $\gamma$ (Spanish language), in which the original concepts are described. We use *Google API* through Python module *TextBlob* at run-time to obtain the automatic translation of labels and terms of a given concept $C_x$ resulting in $C_x^T$. This method was chosen motivated by the fact that it can be used at run-time.

We use Latin ($\alpha$) as the pivot-language because it has the most prevalent etymology in the chosen domain [Charen 1951]. This means that a significant number of words in the domain have radicals originated from Latin. We assume that similarity advantage can be obtained when comparing strings from different languages.

### 3.3. Description of Experiments

We propose four distinct experiments exploring concept elements and their translation. Each experiment is applied to all datasets and the results are aggregated over all datasets.

**Experiment 1: similarity of translated labels.** This experiment (denoted TL in future references, standing for translated labels) investigates if there is a relevant similarity between labels translated to another language $\alpha$ of concepts involved in the mapping. Our

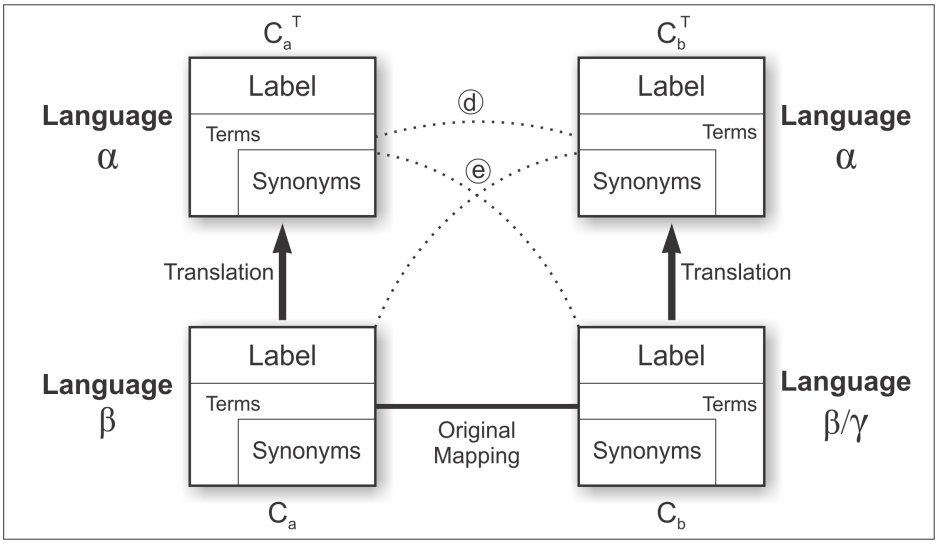


**Figure 2. Overview of the cross-language study setup. It shows the concepts involved in the mapping described in a Language $\beta$ or $\gamma$ and their translation to Language $\alpha$.**

motivation is to understand the role played by translated labels to language $\alpha$ for cross-language matching. The similarity function is applied to the translated label of concept $L(C_a)$ and the translated label of concept $L(C_b)$ of a mapping, *i.e.*, $sim(L(C_a^T), L(C_b^T))$.

**Experiment 2: cross-language labels similarity.** This experiment (denoted XL, standing for cross-language labels) checks if there is a relevant similarity between the original label and the translated version of the labels to another language $\alpha$. It explores the translated label of concept $L(C_a^T)$, the original label of concept $L(C_b)$ and vice-versa. For computing the similarity, it takes the maximum value between $sim(L(C_a), L(C_b^T))$ and $sim(L(C_a^T), L(C_b))$.

**Experiment 3: similarity of translated labels and synonyms.** This experiment (denoted TLS, standing for translated labels and synonyms) aims to study the behaviour of similarity calculated between the translated label and synonyms from the original concepts of mapping. It indicates whether exploring the matching between labels and synonyms is relevant for cross-language alignment. Given a mapping and its translated concepts, this experiment explores the translated label of concept $L(C_a^T)$ and the translated set of synonyms from $C_b^T$. It also considers the inverted possibility, taking into account the translated label of concept $L(C_b^T)$ and the set of translated synonyms from $L(C_a^T)$. For each mapping, the procedure retains the maximum value of similarity calculated between all comparisons made.

**Experiment 4: similarity of cross-language labels and synonyms.** This experiment (denoted XLS, standing for cross-language labels and synonyms) is similar to experiment XL, but at this stage we aim at experimenting the behaviour of similarity calculated between the translated label and the original version of synonyms of the other concept involved in mapping (*i.e.*, cross-language). The goal is to comprehend how this configuration might be useful for cross-language alignment. Given the translated concepts of a mapping, it explores the translated label of concept $L(C_a^T)$ and the original set of synonyms from $C_b$. It also considers the inverted option, taking the translated label of

concept $L(C_b^T)$ and the set of synonyms from $C_a$ into account.

### 3.4. Analyses and Datasets

For each experiment, we perform specific analysis transversely to examine the influence of the string elements composing labels and terms (Analysis 1). We also investigate different aspects concerning the type of similarity functions in the matching between concepts denoted in different languages (Analysis 2).

**Analysis 1: the influence of the organization of the string elements.** The first analysis of the experiments (Analysis Org, standing for organization) performs the calculation of similarity considering the string of elements as a whole. The description of labels and terms may be organized differently between two different ontologies, $O_a$ and $O_b$. For instance, the concept label *"cardio vascular diseases"* in $O_a$ may be described as *"diseases of the heart"* in $O_b$. This aspect may have an impact on concept matching. Therefore, we wanted to further analyze whether the similarity measures are affected by the organization of the strings denoting labels and synonyms. To this end, for each concept element (a label or a synonym), we compared the similarity values obtained when the element is considered a single string, and when we split each concept element into tokens, divided by empty spaces, and removed stop-words (*e.g.*, of, for, and). This results in an array of tokens for each concept element. Figure 3 depicts a representation of a concept considering the structure of string. This shows the way the similarity measure is calculated in this analysis, between labels (*cf.* f in figure 3) and between labels and synonyms (*cf.* g in figure 3)

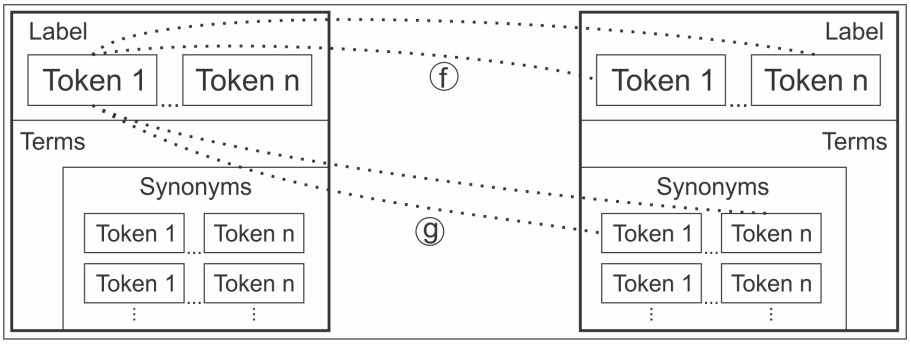


**Figure 3. Analysis of the structure of textual strings denoting concepts' elements.**

In the experiments TL and XL, given the array of tokens of the label $L(C_a^T)$, we calculate the similarity between each token with the label $L(C_b^T)$ and $L(C_b)$. For each experiment, it retains the maximum similarity value computed and keeps it into an array. This is performed for each token of the label $L(C_a^T)$. Afterwards, since no weight is given to each token, the output result remains the average of similarity values stored in the array of similarity.

In the experiments TLS and XLS, which explores the similarity between the $L(C_a^T)$ and the synonyms of $C_b^T$ and of $C_b$, the comparison performed in experiments TL and XL is repeated, but for each synonym of $C_b^T$ and $C_b$. Finally, for each experiment, it is returned the maximum value from the set of stored average values of similarity (*i.e.*, the maximum medium).

**Analysis 2: the impact of syntactic and semantic similarity.** This analysis (Analysis SynSem, standing for syntactic and semantic) aims to inquire the influence of similarity methods in the conducted experiments. We examine the difference in the obtained results when calculating the similarity by exploring syntactic and semantic techniques. The syntactic measure ($Sim_{sy}$) explores the traditional edit-distance technique (Levenshtein distance) [Levenshtein 1966]. This technique relies on the number of single character edits (*i.e.,* insertions, deletions, substitutions) required to change one word into another.

The semantic measure ($Sim_{sm}$) explores the *Weighted Overlap* method applied to NASARI semantic vectors [José Camacho-Collados and Navigli 2015]. This method makes cross-language similarity measurement possible by using vectors in a unified language independent space of concepts from semantic representations in *BabelNet* [Navigli and Ponzetto 2012]. Formally,

$$Sim_{sm}(el_1, el_2) = WO(v_1, v_2), \tag{1}$$

where $v_1$ and $v_2$ refer to the word-based vector representation of the string elements $el_1$ and $el_2$, respectively. The similarity is computed by comparing the corresponding vectors, which results in similarity scores. The measure $WO$ computes the weighed average of the two similarity scores resulting in a normalized value $0 \leq x \leq 1$.

The mapping datasets in the experiments were collected from two different sources, the *BioPortal*[1] repository and the *Unified Medical Language System* (UMLS)[2]. The study explored mappings between the *Systematized Nomenclature of Medicine-Clinical Terms* (SNOMEDCT) with several other ontologies. Table 1 describes the set of mappings.

**Table 1. Mappings between biomedical ontologies: mapping sets, source and number of mappings and percentage of mappings with exact match label (#Exact match).**

| Mapping set $\mathcal{L}$ | Source | #Mappings | #Exact match |
|---|---|---|---|
| SNOMEDCT-LOINC | BioPortal | 29 676 | 69% |
| SNOMEDCT-NCIT | BioPortal | 16 746 | 97% |
| SNOMEDCT-SNMI | BioPortal | 166 932 | 73% |
| SNOMEDCT-MSHSPA | UMLS | 17 859 | $< 1\%$ |
| SNOMEDCT-MDRSPA | UMLS | 20 623 | $< 1\%$ |

## 4. Results

Figure 4 presents the results obtained with experiment TL. This shows the distribution of the computed similarity values organized in three groups of similarity ranges. Analysis OrgW refers to the similarity results considering the content of textual string concept elements as a whole. Analysis OrgT presents the similarity results considering the organization (tokens) of textual concept elements. Whereas Figure 4 (a) presents the results

[1] bioportal.bioontology.org

[2] www.nlm.nih.gov/research/umls/

with syntactic similarity measure, Figure 4 (b) shows the results with semantic similarity measure.

We statistically analyse the results obtained with the *t-test* to indicate the significance of findings with 95% of confidence. We denote by (*) the series presenting the higher averages from the statistical test. Results of the remaining experiments follow the same presentation approach.

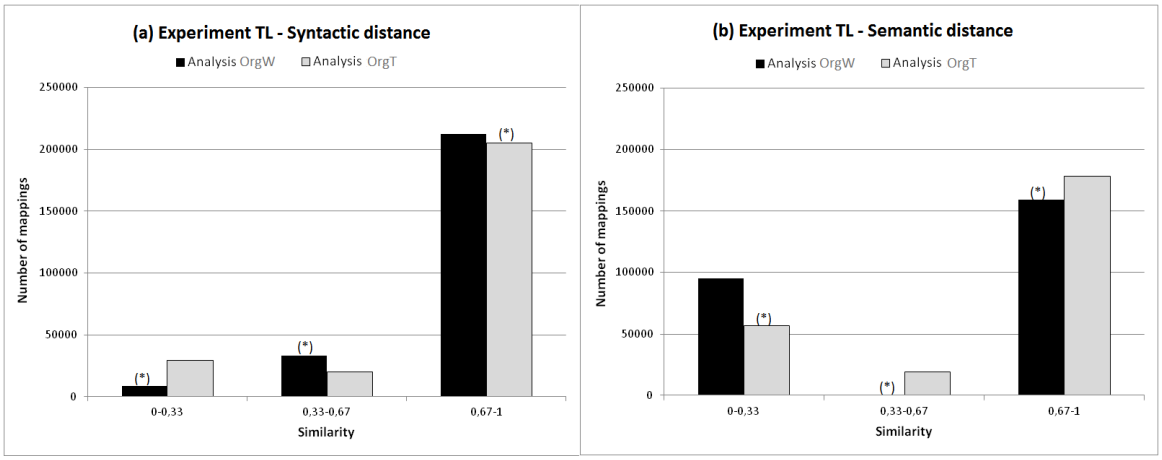


**Figure 4. Results of experiment TL (translated labels). (a) - results for syntactic similarity measure; (b) - results for semantic similarity measure. The x-axis presents the ranges of similarity value and the y-axis shows the number of mappings. The bars compare Analysis OrgW (whole string) and OrgT (tokens).**

Figure 4 (a) shows that values found in $Sim_{sy}$ are concentrated in the range with the highest ratio. Figure 4 (b) reveals a similar behaviour for $Sim_{sm}$. A possible explanation for this behaviour can be the high number of labels having an exact match.

Figure 5 (a) presents the results obtained with experiment XL for cross-language similarity comparisons of labels ($Sim_{sy}$). We notice a high accuracy on the syntactic distance since the $\alpha$ language used is closely related to the domain. Furthermore, our findings indicate that Analysis OrgT (with token) performs better when the semantic measure is applied (for both experiments TL and XL). Spliting the strings into tokens favors the performance of the semantic measure as complex labels are split into smaller strings (*e.g.*, "*Congenital anomaly of thyroid cartilage*" does not have a direct match in NASARI, but a match is found in NASARI for its separated terms "*Congenital*","*anomaly*","*thyroid*" and "*cartilage*") .

Figure 6 presents the results obtained with experiment TLS (translated labels and synonyms). The impact in Analysis OrgW is clear when comparing labels and synonyms in the similarity calculation of $Sim_{sm}$ (*cf.* Figure 6 - b). The same results are not observed with the syntactic measure due to the difficulties in calculating $Sim_{sm}$ with the entire string, because labels and synonyms are represented by long and complex strings. Note that Analysis OrgT (*cf.* Figure 6 - b) keeps most mappings in the highest range of similarity. The separation into tokens improves the number of isolated terms found in the background knowledge.

Figure 7 presents the results achieved with experiment XLS. Since the $\alpha$ language

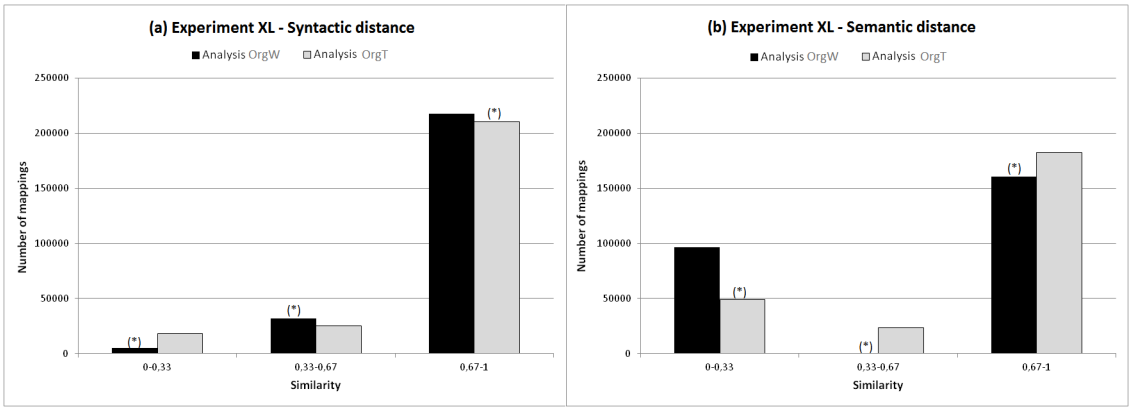


**Figure 5. Results of experiment XL (cross-language labels). (a) - results for syntactic similarity measure; (b) - results for semantic similarity measure. The x-axis presents the ranges of similarity value and the y-axis shows the number of mappings. The bars compare Analysis OrgW (whole string) and OrgT (tokens).**

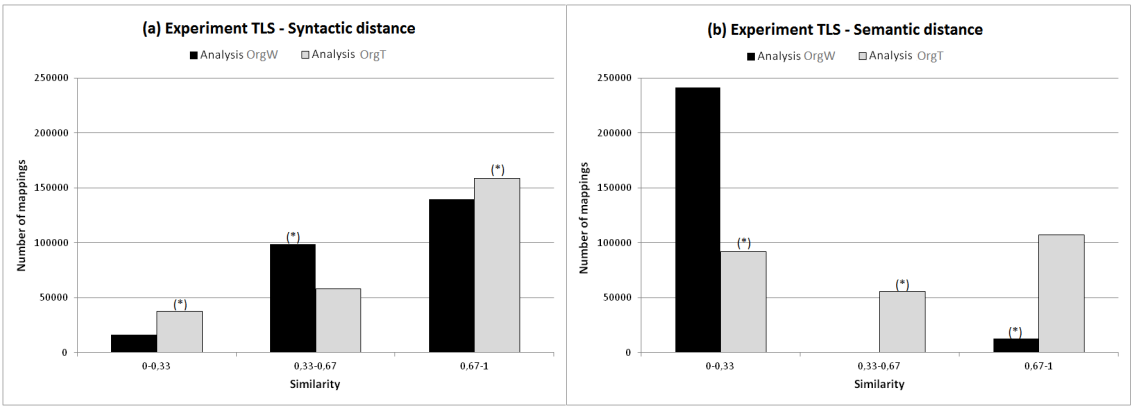


**Figure 6. Results of experiment TLS (translated labels and synonyms). (a) - results for syntactic similarity measure; (b) - results for semantic similarity measure. The x-axis presents the ranges of similarity value and the y-axis shows the number of mappings. The bars compare Analysis OrgW (whole string) and OrgT (tokens).**

is etymologically related to the domain, results of experiment XLS remain similar to the findings in experiment XLS. The analysis of $Sim_{sy}$ shows that cross-language labels and synonyms presents improved similarity values. For example, a higher similarity value is obtained when comparing "*Intravascular injection*" with "*Iniectio de sanguine vas*" (translated synonym of "*Injection of blood vessel*"), than when comparing its translation "*intravascular iniectio*" with "*Iniectio de sanguine vas*". Also, we observe a better result in Analysis OrgT when compared to experiment TLS, revealing that the organization of labels and synonyms affects positively the similarity values of cross-language comparisons.

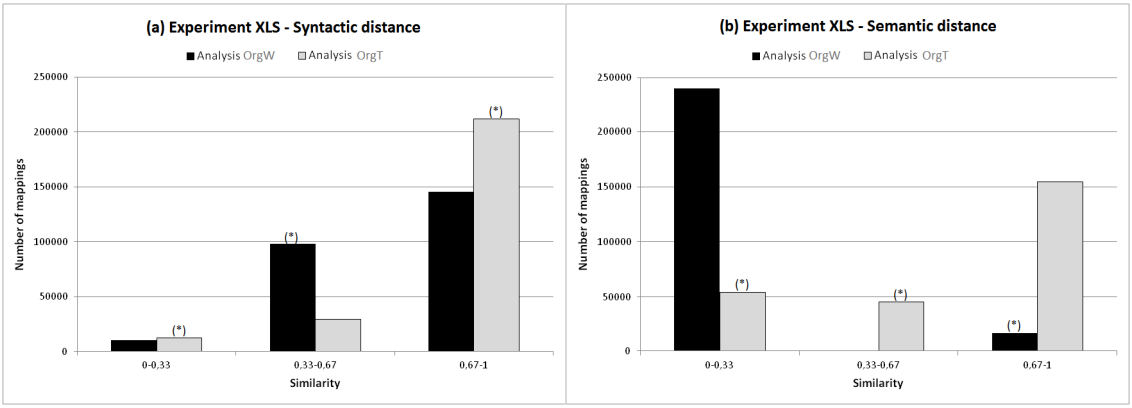


**Figure 7. Results of experiment XLS (cross-language labels and synonyms). (a) - results for syntactic similarity measure; (b) - results for semantic similarity measure. The x-axis presents the ranges of similarity value and the y-axis shows the number of mappings. The bars compare Analysis OrgW (whole string) and OrgT (tokens).**

## 5. Discussion

This work contributed with a set of experiments to reveal the relevant aspects to be considered in cross-language matching. Furthermore, it determined the influence of the type of similarity function for multilingual matching algorithms. It can be particularly useful to understand and select the adequate features to be used by machine learning approaches for ontology alignment.

Results show that when using an $\alpha$ language related to the domain, the syntactic distance provides a reliable measurement of similarity. It was clear that when exploring labels with synonyms, their textual string structure can play a relevant role. This became even more evident when exploring the cross-language computation with the semantic measure. Although influenced by the background knowledge, results obtained with semantic measure were similar to those achieved with syntactic distance. The experiments point out that semantic measure performance is boosted when strings are explored with separate tokens.

Although our findings are relevant, the results are only applicable to languages within the same alphabetical universe. The advantage of using a pivot-language related to the domain is to increase the accuracy of syntactic distance measurements, but such benefit can be lost when the set of characters differs.

Further investigations involve thoroughly examine semantic similarity considering the influence of the corpus and other measure approaches. We plan future experiments to investigate the role of neighbour concepts.

## 6. Conclusion

Cross-language alignment of ontologies requires adequate techniques relying on similarity measures to overcome the difficulties on the matching task. This article contributed with empirical studies to thoroughly unveil relevant aspects to be considered in the definition of matching algorithms applied to the alignment of ontologies in different languages. We have shown that the use of a pivot-language related to the domain in the cross-language alignment is beneficial for automatic matching algorithms. In additon, we have shown that, in this context, the performance of syntactic and semantic similarity measures slightly differs. Future work encompass the design of an original cross-language matching algorithm for aligning biomedical ontologies.

## Acknowledgments

This work is supported by the São Paulo Research Foundation (FAPESP) (Grant #2014/14890-0).

## References

[Charen 1951] Charen, T. (1951). The etymology of medicine. *Bulletin of the Medical Library Association*, 39.

[Fu et al. ] Fu, B., Brennan, R., and O 'sullivan, D. Cross-lingual ontology mapping - an investigation of the impact of machine translation. In *The Semantic Web (ASWC'09)*.

[Gruber 1995] Gruber, T. R. (1995). Toward principles for the design of ontologies used for knowledge sharing. *International Journal of Human-Computer Studies*, 43.

[José Camacho-Collados and Navigli 2015] José Camacho-Collados, M. T. P. and Navigli, R. (2015). a novel approach to a semantically-aware representation of items. In *North American Chapter of the Association of Computational Linguistics (NAACL 2015)*, Denver,USA.

[Jung et al. 2009] Jung, J., Håkansson, A., and Hartung, R. (2009). Indirect alignment between multilingual ontologies: A case study of korean and swedish ontologies. In *Agent and Multi-Agent Systems: Technologies and Applications*, volume 5559. Springer.

[Khiat 2016] Khiat, A. (2016). Crolom: Cross-lingual ontology matching system. In *Proceedings of the 15th International Semantic Web Conference (ISWC 2016)*, pages 146–152.

[Levenshtein 1966] Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10.

[Meilicke et al. 2012] Meilicke, C., Trojahn, C., Šváb-Zamazal, O., and Ritze, D. (2012). Multilingual ontology matching evaluation–a first report on using multifarm. In *The Semantic Web: ESWC 2012 Satellite Events*. Springer Berlin Heidelberg.

[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.

[Shvaiko and Euzenat 2013] Shvaiko, P. and Euzenat, J. (2013). Ontology matching: State of the art and future challenges. *Knowledge and Data Engineering, IEEE Transactions on*, 25(1).

[Spohr et al. 2011] Spohr, D., Hollink, L., and Cimiano, P. (2011). A machine learning approach to multilingual and cross-lingual ontology matching. In *ISWC 2011*. Springer.

[Trojahn et al. 2014] Trojahn, C., Fu, B., Zamazal, O., and Ritze, D. (2014). *Towards the Multilingual Semantic Web: Principles, Methods and Applications*, chapter State-of-the-Art in Multilingual and Cross-Lingual Ontology Matching. Springer Berlin Heidelberg.