# An English-Portuguese parallel corpus of questions: translation guidelines and application in Statistical Machine Translation

**Ângela Costa$^{*\star}$, Tiago Luís$^{\star}$, Joana Ribeiro$^{\dagger\star}$, Ana Cristina Mendes$^{\dagger\star}$, Luísa Coheur$^{\dagger\star}$**

$^{*}$ Centro de Linguística da Universidade Nova de Lisboa
$^{\dagger}$Instituto Superior Técnico, Technical University of Lisbon
$^{\star}$Spoken Language Systems Laboratory - L$^2$F/INESC-ID
R. Alves Redol, 9 - 2º– 1000-029 Lisboa, Portugal
{first.last}@l2f.inesc-id.pt

**Abstract**

The task of Statistical Machine Translation depends on large amounts of training corpora. Despite the availability of several parallel corpora, these are typically composed of declarative sentences, which may not be appropriate when the goal is to translate other types of sentences, e.g., interrogatives. There have been efforts to create corpora of questions, specially in the context of the evaluation of Question-Answering systems. One of those corpora is the UIUC dataset, composed of nearly 6,000 questions, widely used in the task of Question Classification. In this work, we make available the Portuguese version of the UIUC dataset, which we manually translated, as well as the translation guidelines. We show the impact of this corpus in the performance of a state-of-the-art SMT system when translating questions. Finally, we present a taxonomy of translation errors, according to which we analyze the output of the automatic translation before and after using the corpus as training data.

**Keywords:** Questions Dataset, Translation Guidelines, Machine Translation

## 1. Introduction

The area of Statistical Machine Translation (SMT), like many others in NLP, heavily depends on the availability of corpora. There are several parallel corpora available for the SMT task, like Europarl (Koehn, 2005) and JRC-Acquis (Ralf et al., 2006), which constitute the foundations of the learning process. One of the characteristics shared by these corpora is that they are mainly composed of declarative sentences. Therefore, when we try to translate a different kind of sentence, like interrogatives, with an SMT system trained with these type of corpora, we may not get the desired translation quality due to the gap between them.

When it comes to the specific task of translating questions, there have been efforts to create multilingual corpora of questions (usually with their respective correct answers). These corpora were created in the context of the Cross Language Evaluation Forum (CLEF), an evaluation forum for Question-Answering (QA) systems, and include: the Multisix corpus (Magnini et al., 2003), which contains 200 English questions translated into 5 different languages, and the *DISEQuA* corpus, composed of 450 questions in Dutch, Italian, Spanish and English; the *Multieight-04 corpus* (Magnini et al., 2004), composed of 700 questions available in 7 different languages, including Portuguese; and, the Multi9-05 (Vallin et al., 2005) that contains 900 questions written in 9 languages, including Portuguese. Although of indisputable importance to QA, as it allows the benchmarking of systems, these corpora are clearly insufficient in any language when one wants to create (and train) a SMT system that efficiently translates natural language questions to/from European Portuguese.

A widely known monolingual corpus of questions is the one built by Li and Roth (2002), composed of nearly 6,000 questions, along with their semantic categories. This corpus – the University of Illinois at Urbana-Champaign (UIUC) dataset – has become a very valuable resource for training and testing machine learning models, since the authors have made it freely available on the Web. It is frequently used in QA, for the task of Question Classification (Silva et al., 2011; Huang et al., 2008; Zhang and Lee, 2003). More recently, this corpus was manually annotated with named entities of categories HUMAN, LOCATION and ORGANIZATION (Mendes et al., 2010).

The fact that the UIUC dataset corpus is broadly used by the research community was a strong motivation for us to translate it into Portuguese. Thus, in this paper we present the Portuguese version of the UIUC dataset.

This work contributes by providing a corpus of nearly 6,000 questions manually translated into Portuguese to the community, split into train and test sets. Two obvious applications of this corpus are the SMT of questions from/to English to/from Portuguese and Question Classification in Portuguese.

Also, this work gives a description of the translation guidelines and shows how this corpus can be used to improve an SMT system in the automatic translation of questions. In fact, there have been efforts to adapt SMT to domains with very limited data resources. For example, the work by Tiedemann (2009) focused on the optimization of an English to Dutch phrase-based systems to questions, and got substantial improvements over the baseline system, by building models with a mixture of the questions data and out-of-domain data. We will show how results improved with the new corpus and, finally, a taxonomy of translation errors is used to analyse the obtained results.

The remainder of this paper is organized as follows: in Section 2. we present the annotation guidelines; in Section 3. we describe how we used this corpus in the automatic translation of questions; and, in Section 4. we present an analysis of the errors in the translation of questions done by the automatic translator. The paper finishes in Section 5., where we conclude and point to future work directions.

## 2. Guidelines

In this section we will describe some issues that arose when manually translating the English corpus into Portuguese and the solutions we propose for them. These problems were mainly of two natures: semantic level issues and structure level issues.

### 2.1. Semantic level issues

On the semantic level of analyses we were able to find cases of expressions, words or phrases that could not be translated as they did not have a direct equivalent in Portuguese.

For instance, quotes from poems such as *What American poet wrote: "Good fences make good neighbors"?* should be translated keeping the quote in English *Que poeta americano escreveu: "Good fences make good neighbours"?*. By the same token, lines from commercial ads, for instance *Which company claimed to be "the world's biggest toy store"?* should also be kept in the source language *Que empresa diz ser "the world's biggest toy store"?*.

Through the corpus we were able to find other examples that could not be translated, as for instance, dates, names of historic events, brands, foods, names of institutions, laws and organizations, frozen expressions, just to mention a few. As we will see in Section 3., the fact that many expressions, words or phrases were not translated, affects the machine translation results.

Nevertheless, many general knowledge questions that refer to a shared knowledge could be translated or adapted into Portuguese. Examples of this are, for instance, *When was the slinky invented?*. Although slinky does not have the same name in Portugal it has an equivalent *Quando é que foi inventada a ondamania?*. Another example is *What "melts in your mouth , not in your hands"*. This is a commercial line of a product that also exists in Portuguese and that also uses the same lines for advertising in Portugal *O que é que se "derrete na boca, mas não nas mãos"?* .

The third issue worth mentioning is the case of questions that when translated into Portuguese resulted is a loss of meaning. Considering the following example, the answer is present on the question (*tubarão* is *shark* in Portuguese): *What animal occurs in Spielberg's "Jaws"?* a literal translation in Portuguese would sound like *What animal occurs in Spielberg's "Shark"?*, otherwise, the answer of the question would be given in the question itself.

### 2.2. Structure level issues

Some guidelines followed at structure level should also be mentioned. Of course the idiosyncrasies of each language should not be taken into consideration for these purposes. The following are three examples of deeper syntactic modifications that were done during the translation process.

The first issue regards some regular formed questions that turned into an imperative sentence with the pragmatic value of a question. This choice was made as the second formula was stylistically better in Portuguese, for instance *What's another word that means "knows all"?* was translated into *Diga um sinónimo de "knows all"*..

The second issue had to do with double questions. As we have previously mentioned, questions in the UIUC dataset were classified according to their semantic category, therefore double questions had to be divided into two questions, since the category of the question could be altered. For instance, the question *What company is being bought by Yahoo and how much is the deal worth?* was divided into two questions *Que empresa está a ser comprada pela Yahoo?* and *E por que valor?*, as two categories – entity and numeric – are the correct categorization.

Finally, also for stylistic purposes, some active sentences were turned into passive sentences, for instance *What city does McCarren Airport serve?* (active) becomes in Portuguese *Que cidade é servida pelo Aeroporto McCarren?* (passive).

## 3. Application to MT

This section shows the improvements that can be achieved by a phrase-based SMT system when using the parallel corpus of questions built during this work, as well as Europarl. Some details on the used corpora can be seen in Table 1.

| | Data | Lang. | Sentences | Words | Avg. Length |
|---|---|---|---|---|---|
| Train | europarl | pt | 1,302,000 | 28,041,534 | 21.5 |
| | | en | 1,302,000 | 27,471,864 | 21.0 |
| | questions | pt | 4,457 | 47,731 | 10.7 |
| | | en | 4,457 | 44,632 | 10.0 |
| Dev | questions | pt | 1,000 | 10,597 | 10.5 |
| | | en | 1,000 | 9,995 | 9.9 |
| Test | questions | pt | 500 | 4,182 | 8.3 |
| | | en | 500 | 3,734 | 7.4 |

Table 1: Data statistics of the datasets.

All experiments were performed using the phrase-based Moses decoder (Koehn et al., 2007). The directional word alignments were produced by GIZA++ (Och and Ney, 2003) using the IBM M4 model and combined using the *grow-diagonal-final* heuristic. The weights of the models were tuned with Minimum Error Rate Training (MERT) using the devel corpus. Results were evaluated using the BLEU (Papineni et al., 2002) metric.

We started by training baseline systems for the EN-PT and PT-EN directions, using only data from the Europarl parallel corpus. Next, we trained SMT models with the training set from the parallel corpus of questions and combined them (translation and language models) with the Europarl models. The SMT models are combined during decoding using a set of weights tuned with MERT. In this way, the Moses decoder tries to gather the translation hypothesis from the questions models, and collects additional options from the Europarl models. If the same translation hypothesis (in terms of identical input phrase and output phrase) is found in both models, separate translation hypothesis are created for each occurrence, but with different scores. Despite the huge difference in terms of size between the two datasets, the combination of the two systems yield significant improvements in the translation quality (Table 2 shows the attained BLEU scores when evaluated on the test set from our corpus). Using the bootstrap method (Koehn, 2004) we concluded that the improvements are statistically significant ($p < 0.01$).

| Direction | Model | BLEU |
|---|---|---|
| EN-PT | Europarl | 32.80 |
| | Europarl + Questions | **42.40** |
| PT-EN | Europarl | 36.96 |
| | Europarl + Questions | **44.45** |

Table 2: BLEU scores achieved by the SMT systems when evaluated on our test set (parallel corpus of questions).

The next section presents a detailed analysis of the differences between the two systems.

# 4. Error analysis

In this section we present a taxonomy of errors and an analysis of the errors found.

## 4.1. Error Taxonomy

We have selected the first 50 questions from the test corpus and we translated these sentences into Portuguese with Moses. Afterwards we did an analysis of the errors according to a simplified version of the taxonomy defined by (Vilar et al., 2006). For instance, we have removed the category Punctuation: as our corpus was constituted by questions, no errors of this type were present.
The taxonomy we have used is the following:

1. Missing Words

   When one or more words are missing in the translation, they can either be classified as missing filler words or missing content words.

   1.1 Missing Filler Words
   Original: *What is amitriptyline?*
   Translation: *O que é amitriptyline?*
   Correct Translation: *O que é **a** amitriptilina?*

   1.2 Missing Content Words
   Original: *What is the average weight of a Yellow Labrador?*
   Translation: *Qual é o peso de um Labrador amarelo?*
   Correct Translation: *Qual é o peso **médio** de um Labrador amarelo?*

2. Word Order

   This type of error occurs when the reordering model is unable to perform a reordering of the sentence, producing an odd sentence. Some taxonomies distinguish between short and long range ordering, but for our purposes we only consider word order.

   Original: *Who was the first American to walk in space?*
   Translation: *Quem foi o primeiro a andar **americano** no espaço?*
   Correct Translation: *Quem foi o primeiro **americano** a andar no espaço?*

3. Incorrect Words

   The type of error occurs when the translation engine is not able to correctly translate a word or expression, producing instead a wrong translation that severely affects the understandability of the sentence.

   3.1 Lexical Choice

   In this case, the translation engine chose the wrong translation candidate word.

   Original: *What hemisphere is the Philippines in?*
   Translation: *O que é que **lutam** as filipinas?*
   Correct Translation: *Em que **hemisfério** são as Filipinas?*

   3.2 Disambiguation

   In some situations, the system is not able to disambiguate the correct meaning of a source word in a given context. This happens when the source language word has more than one meaning on the target language.

   Original: *What is the temperature at the center of the Earth?*
   Translation: ***O que** é que a temperatura no centro da Terra?*
   Correct Translation: ***Qual** é a temperatura no centro da Terra?*

   3.3 Incorrect Form

   These errors occur when the translation engine, despite producing the word with the correct root, fails to produce the correct form of the word, usually incorrectly translating a verb form or not doing the correct gender or number transformations in noun, adjectives or pronouns.

   Original: *When did John F. Kennedy get elected as President?*
   Translation: *Quando é que John F. Kennedy **ser eleitos** como presidente?*
   Correct Translation: *Quando é que John F. Kennedy **foi eleito** como presidente?*

   3.4 Extra Words

   This type of error refers to the cases where the translation engine generates sentences containing words, most commonly filler words, that should be removed in order to obtain a correct sentence.

   Original: *Where is John Wayne airport?*
   Translation: *Onde está o **senhor deputado** Wayne aeroporto?*
   Correct Translation: *Onde é o aeroporto John Wayne?*

   3.5 Idiomatic Expressions

   This category of errors refers to expressions that should have not been translated literally. In these situations, the translation will express a literal meaning that it is not the correct one.

Original: *Which mountain range in North America stretches from Maine to Georgia?*
Translation: *Que variam de montanha na América do Norte se estende desde o Maine até a Georgia?*
Correct Translation: *Que* **cordilheira** *na América do Norte se estende desde o Maine até a Georgia?*

4. Unknown Words

   Unknown words or expressions are the ones for which the translation engine could not find any translation candidate and for that reason were kept in the source language and copied to the translation output.

   Original: *What is the life expectancy for crickets?*
   Translation: *O que está a esperança de vida para* ***crickets****?*
   Correct Translation: *Qual é a esperança de vida dos* ***grilos****?*

### 4.2. Discussion

Figure 1 represents the number of errors we were able to find before and after the SMT models were trained with the training set from the parallel corpus of questions.

As we can see, the number of missing words and problems with word order were slightly the same before and after. The major changes occurred on the number of incorrect words level of analyses, as the number of disambiguations decreased. Analyzing the results in detail, we can see that with this corpus we manage to overcome a common problem when translating questions: how to translate the wh-words? For instance, *what* can be translated into Portuguese as *O que* but also as *Qual*, *O quê*, *Quais*, *A que*. This was the cause of the majority of the errors of type 3.2 and, after the adaptation, and now, this type of errors decreased significantly.

However, we should also mention the increase of the number of unknown words that were not translated after adaptation. This can be explained with the nature of our corpus. This corpus of 6,000 questions was built for Question-Answering systems and Question Classification and for these tasks, and accordingly to our guidelines, some words and expressions should be kept in English in order to facilitate the task of finding the correct answer to a given question. In the sentence *Who 's the founder and editor of The National Review?*, the name of the newspaper should be left untranslated. Thus, particularly on this case, the word *national* was no longer translated after the train. By the same token, in the question *Why does the moon turn orange?*, the word *orange* was correctly translated before the adaptation, but after it was one of the words listed as untranslated. This is due to the fact that the word *orange* appears in the training set in expressions such as *Orange Bowl* and *Orange County*. The first one is the name of a Stadium and the second one the name of a county and both have no equivalent in Portuguese, thus, were not translated. In this way, the system "learned" that the word *orange* should not be translated.

This leads to an interesting question: should the translation guidelines be adapted to the use that is going to be given to a corpus?

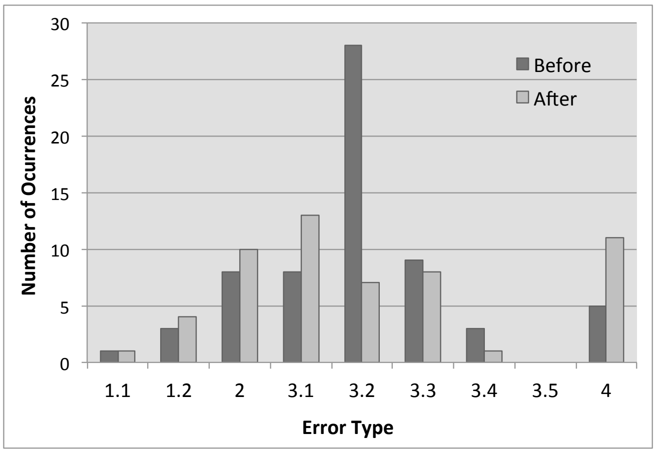

Figure 1: Comparison between the number of errors found per type before and after adding the questions to the training set.

## 5. Conclusions and Future Work

With this work, we made available a corpus of nearly 6,000 questions manually translated into Portuguese, split into train and test sets, with application in SMT from/to English to/from Portuguese and Question Classification in Portuguese. In addition, we described the translation guidelines.

We used the translated questions to train a state-of-the-art phrase-based SMT system, and observed an improvement in the translation quality when compared to the baseline system. We have seen a significant decrease of the number of incorrectly translated words; however, and due to the nature of the corpus, the number of words that were not translated increased.

Regarding future work directions, we intend to use this corpus to train an hierarchical and/or a syntax-based SMT system. Moreover, we plan to train a question classifier for Portuguese and use it within a QA system.

## Acknowledgements

This work was supported by Fundação para a Ciência e a Tecnologia (INESC-ID multiannual funding) through the PIDDAC Program funds and by the EC CIP METANET4U project #270893. The Fundação para a Ciência e a Tecnologia also supports Ângela Costa and Tiago Luís scholarships through project PT-STAR (CMU-PT/HuMach/0039/2008). Ana Cristina Mendes is also supported by a PhD fellowship from Fundação para a Ciência e a Tecnologia (SFRH/BD/43487/2008).

## 6. References

Zhiheng Huang, Marcus Thint, and Zengchang Qin. 2008. Question Classification using Head Words and their Hypernyms. In *EMNLP*, pages 927–936.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan 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, Stroudsburg, PA, USA. Association for Computational Linguistics.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing , EMNLP 2004, A meeting of SIGDAT, a Special Interest Group of the ACL, held in conjunction with ACL 2004, 25-26 July 2004, Barcelona, Spain*, pages 388–395. ACL.

P. Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5.

Xin Li and Dan Roth. 2002. Learning question classifiers. In *Proc. 19th Int Conf. Computational linguistics*, pages 1–7. ACL.

Bernardo Magnini, Simone Romagnoli, Alessandro Vallin, Jesús Herrera, Anselmo Peñas, Víctor Peinado, Felisa Verdejo, and Maarten de Rijke. 2003. The Multiple Language Question Answering Track at CLEF 2003. In *CLEF*, pages 471–486.

Bernardo Magnini, Alessandro Vallin, Christelle Ayache, Gregor Erbach, Anselmo Peñas, Maarten de Rijke, Paulo Rocha, Kiril Ivanov Simov, and Richard F. E. Sutcliffe. 2004. Overview of the CLEF 2004 Multilingual Question Answering Track. In Carol Peters, Paul Clough, Julio Gonzalo, Gareth J. F. Jones, Michael Kluck, and Bernardo Magnini, editors, *CLEF*, volume 3491 of *Lecture Notes in Computer Science*, pages 371–391. Springer.

Ana Cristina Mendes, Luísa Coheur, and Paula Vaz Lobo. 2010. Named Entity Recognition in Questions: Towards a Golden Collection. In *LREC*. ELRA.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Comput. Linguist.*, 29:19–51, March.

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*, pages 311–318. Association for Computational Linguistics.

S. Ralf, B. Pouliquen, A. Widiger, C. Ignat, T. Erjavec, D. Tufiş, and D. Varga. 2006. The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pages 2142–2147.

João Silva, Luísa Coheur, Ana Mendes, and Andreas Wichert. 2011. From symbolic to sub-symbolic information in question classification. *Artificial Intelligence Review*, 35:137–154.

Jörg Tiedemann. 2009. Translating questions for cross-lingual qa. In Lluís Marqués and Harold Somers, editors, *Proceedings of 13th Annual Conference of the European Association for Machine Translation (EAMT'09)*, pages 112 – 119, Barcelona, Spain, May.

Alessandro Vallin, Bernardo Magnini, Danilo Giampiccolo, Lili Aunimo, Christelle Ayache, Petya Osenova, Anselmo Peñas, Maarten de Rijke, Bogdan Sacaleanu, Diana Santos, and Richard Sutcliffe. 2005. Overview of the CLEF 2005 Multilingual Question Answering Track. In *Proc. CLEF*.

David Vilar, Jia Xu, Luis Fernando DHaro, and Hermann Ney, 2006. *Error analysis of statistical machine translation output*, pages 697–702. Citeseer.

Dell Zhang and Wee Sun Lee. 2003. Question classification using support vector machines. In *SIGIR 2003*, pages 26–32, New York, NY, USA. ACM.