BP2EP – Adaptation of Brazilian Portuguese texts to European Portuguese

Luís Marujo [1,2,3], **Nuno Grazina** [1], **Tiago Luís** [1,2], **Wang Ling** [1,2,3], **Luísa Coheur** [1,2], **Isabel Trancoso** [1,2]
[1] Spoken Language Laboratory / INESC-ID Lisboa, Portugal
[2] Instituto Superior Técnico, Lisboa, Portugal
[3] Language Technologies Institute / Carnegie Mellon University, Pittsburgh, USA
{luis.marujo,ngraz,tiago.luis,wang.ling,luisa.coheur,imt}@l2f.inesc-id.pt

Abstract

This paper describes a method to efficiently leverage Brazilian Portuguese resources as European Portuguese resources. Brazilian Portuguese and European Portuguese are two Portuguese varieties that are very close and usually mutually intelligible, but with several known differences, which are studied in this work. Based on this study, we derived a rule based system to translate Brazilian Portuguese resources. Some resources were enriched with multiword units retrieved semi-automatically from phrase tables created using statistical machine translation tools. Our experiments suggest that applying our translation step improves the translation quality between English and Portuguese, relatively to the same process without this adaptation step.

1 Introduction

Modern statistical machine translation (SMT) depends crucially on large parallel corpora and on the amount and specialization of the training data for a given domain. Depending on the domain, such resources may sometimes be available in only one of the language varieties. For instance, classroom lecture and talk transcriptions such as the ones that one can find in the MIT OpenCourseWare (OCW) website[1], and the TED talks website[2] (838 BP vs 308 EP) are examples of parallel corpora where Brazilian Portuguese (BP) translations can be found much more frequently than European Portuguese (EP) translations. This discrepancy has origin in the Brazilian population size that is near 20 times larger than the Portuguese population.

This paper describes the progressive development of a tool that transforms BP texts into EP, in order to increase the amount of EP parallel corpora available.

This paper is organized as follows: Section 2 describes some related work; Section 3 presents the main differences between EP and BP; the description of the BP2EP system is the focus of Section 4. The following section describes an algorithm for extracting Multiword Lexical Contrastive pairs from SMT Phrase-tables; Section 6 presents the results, and Section 7 concludes and suggests future work.

2 Related Work

Few papers are available on the topic of improving Machine Translation (MT) quality by exploring similarities in varieties, dialects and closely related languages.

Altintas (2002) states that developing an MT system between similar languages is much easier than the traditional approaches, and that by putting aside issues like word reordering and most of the semantics, which are probably very similar, it is possible to focus on more important features like grammar and the translation itself. This also allows the creation of domains of closely related languages which may be interchangeable and that, in this particular case, would allow, for instance, the development of MT systems between English and a set of Turkic languages instead of only Turkish. The system uses a set of rules, written in the XEROX Finite State Tools (XFST) syntax, which is based on Finite State Transducers (FST), to apply several morphological and grammatical adap-

[1] http://ocw.mit.edu/OcwWeb
[2] http://www.ted.com/talk

tations from Turkish to Crimean Tatar. Results showed that this approach does not cover all variations possible in these languages, and that in some cases, there is no way of adapting the text without an additional parser capable of determining if an adaptation not covered by the rules should be performed.

Other authors have developed very similar systems with identical approaches to translate from Czech to Slovak (Hajič et al., 2000), from Spanish to Catalan (Canals-Marote et al., 2001)(Navarro et al., 2004), and from Irish to Gaelic Scottish (Scannell, 2006). Looking at Scannel's system (2006) gives us a better understanding of the common system architecture. This architecture consists of a pipeline of components, such as a Part-of-Speech tagger (POS-tagger), a Naïve Bayes word sense disambiguator, and a set of lexical and grammatical transfer rules, based on bilingual contrastive lexical and grammatical differences.

On a different level, (Nakov and Ng, 2009) describes a way of building MT systems for less-resourced languages by exploring similarities with closely related and languages with much more resources. More than allowing translation for less-resourced languages, this work also aims at allowing translation from groups of similar languages to other groups of similar languages just like stated earlier. This method proposes the merging of bilingual texts and phrase-table combination in the training phase of the MT system. Merging bilingual texts from similar languages (on the source side), one with the less-resourced language and the other (much larger) with the extensive resource language, provides new contexts and new alignments for words existing in the smaller text, increases lexical coverage on the source side and reduces the number of unknown words at translation time. Also, words can be discarded from the larger corpus present in the phrase-table simply because the input will never match them (the input will be in the low resource language). This approach is heavily based on the existence of a high number of cognates between the related languages. Experiments performed when both approaches are combined between the similar languages show that extending Indonesian-English translation models with Malaysian texts yielded a gain of 1.35 points in BLEU. Similar results are obtained when improving Spanish-English translation with larger Portuguese texts, which improved the BLEU score by 2.86 points (Nakov and Ng, 2009).

3 Corpora

Several corpora were collected to use in our experiments:
– CETEMPublico corpus: collection of EP newspaper articles [3]
– CETEMFolha: collection of BP newspaper articles [4]
– 115 texts from the Zero Hora newspaper and 50 texts from the Folha Ciência da Folha de São Paulo newspaper. This corpus includes about 2,200 sentences, 62,000 words, and it was initially presented in (Caseli et al., 2009). It has 3 versions (original and 2 simplified versions). A parallel original corpus version in EP was also created to allow a fair evaluation using very close translations. The unavailability of EP text simplification corpora also motivated this choice.
– Ted Talks (TEDs): 761 TEDs in English were collected. From those, only 262 TEDs have a corresponding EP version and 749 TEDs have a BP version (Table 1).

Language	Nr. T.T.	Nr. Sentences	Nr. Words
EN	761	99,970	1,817,632
EP	262	29,284	512,233
BP	749	95,872	1,706,223

Table 1: Description of the gathered Ted Talks corpus.

4 Main Differences Between EP and BP Texts

The two Portuguese varieties involved in this work are very close and usually mutually intelligible, but there are several sociolinguistics, orthographic, morphologic, syntactic, and semantic differences (Mateus, 2003). Furthermore, there are also relevant phonetic differences that are beyond of the scope of this work.

4.1 Sociolinguistics Differences

Silva (2008) made the point that language varieties contribute to the sociolinguistic variations. As a matter of fact, such variations generate emotive meanings (e.g.: pejorative terms), strict sociolinguistic meanings (e.g.: erudite, popular, and regional terms/expressions), discursive meanings

[3] http://www.linguateca.pt/cetempublico
[4] http://www.linguateca.pt/cetenfolha/

(e.g.: interjections, discourse markers) and ways of addressing people (e.g.: *senhor*, *você*, and *tu*). In BP, *você* (*you*) is used as a personal pronoun when addressing someone, in the majority of situations, instead of *tu* (*you*) or its omission in EP.

4.2 Orthographic and Morphologic Differences

The recent introduction (2009) of the orthographic agreement of 1990 [5] mitigated some differences between the two varieties, but most existent linguistic resources were written using the pre-agreement version. The germane orthographic differences from EP to BP are: inclusion of muted consonants; abolition of umlaut; and different accentuation in Proparoxytone words (words with stress on third-to-last syllable), some Paroxytone words (stress on the penultimate syllable) ending in *-n*, *-r*, *-s* , *-x*; and words ending in *-eica*, and *-oo* (Teyssier, 1984). Examples:
– EN: project, water, tennis.
– BP: projeto, ágüa, tênis.
– EP: projecto, água, ténis.

4.3 Syntactic Differences

Both varieties differ in their preferences when expressing a progressive event. While in BP such an event is preferentially described using the gerund form of the verb, in EP, this is expressed using the verb's infinitive form. Examples:
– EN: He was running.
– BP: Ele estava correndo.
– EP: Ele estava a correr.

The placement of clitics also varies from one language to the other. In EP these are commonly joined with the verb and linked with a hyphen in affirmative sentences (proclitic position), and placed separately before the verb in negative sentences (enclitic position). In BP clitics are always placed separately from the verb and their relative positioning is dependent on the type of clitic, with several exceptions. For instance, third person clitics are placed after the verb and pronominal clitics are placed before. Examples:
– EN: He saw me on the street.
– BP: Ele me viu na rua.
– EP: Ele viu-me na rua.

The handling of articles is also distinct in some cases. In BP, articles that are followed by possessive pronouns are frequently omitted, while this is not correct in EP. Examples:
– EN: I sold my car.
– BP: Vendi meu carro.
– EP: Vendi o meu carro.

In addition, word expansions in BP are commonly word contractions in EP (Caseli et al., 2009). The list of contractions was extracted from (Abreu and Murteira, 1994) and it is also available at Wikipedia[6]. Examples:
– EN: He lived in that house.
– BP: Ele vivia em aquela casa.
– EP: Ele vivia naquela casa.

4.4 Lexical and Semantic Differences

In the same manner that *color* and *colour*, or *gas* and *petrol* are examples of the differences between American and British dialects, there are also correlative lexical differences between BP and EP. At this level, there are innumerable differences between this varieties. The origin of these differences is also very varied, ranging from the influence of other languages, cultural differences and historical reasons. As a result, many words with the same meaning are written differently and some words that are written equally have different meanings. For instance, the word "sentença" means both sentence and verdict in BP, but it only means verdict in EP.

5 BP2EP

In a preliminary experiment we trained a SMT system (Moses) using BP and EP aligned Ted Talks. However, the results were relatively low with approximately 30 BLEU points. Hence, we developed a RBMT BP2EP system. It follows a pipes and filters architecture (Fig. 1). The very first step in the conversion of BP text to EP is to take care of words that are incongruous with the orthographic agreement. We use the Bigorna (Almeida et al., 2010) system to handle this transformation. The following step is the application of the in-house POS tagger (Ribeiro et al., 2003), which provides the syntactic information necessary to trigger several transformation rules from BP to EP. These rules will be described next.

5.1 Handling of Sociolinguistics Differences

When the word *você* followed by a verb is detected, the sentence is transformed in order to

[5] http://www.portaldalinguaportuguesa.org/index.php?action=acordo&version=1990

[6] http://pt.wikipedia.org/wiki/Lista_de_contracoes_na_lingua_portuguesa

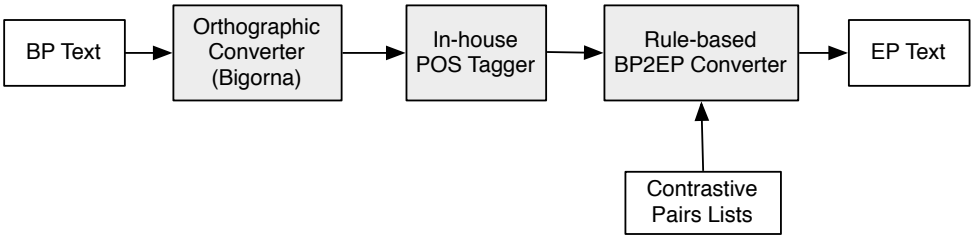

Figure 1: System architecture of BP2EP

match an EP construction using the pronoun *tu* instead. So, *você* is replaced by *tu* and the verb is changed from the third person singular to the second person form, with the help of the verb list. The following general rule is applied:

"você" + (adverb) +verb (3^{rd} person) → "tu" + (adverb) +verb (2^{nd} person)

In some situations, the sentence obtained by applying the general rule is perfectly acceptable in EP, but addressing the listener/reader as *tu* would be disrespectful. On the other hand, omitting the *tu* or *você* makes native speakers question the grammaticality of the sentence.

These transformations are of great value for colloquial and oral text, but seldom used in other type of text, namely, journalistic texts.

In addition, we have also compiled a manually filtered (to remove ambiguity) list of idiomatic expressions, containing about 380 expressions, based on Wikipedia [7] to resolve popular and regional expressions. Example:
EN: Booze, make a mistake
BP: Enfiar o pé na jaca
EP: Embriagar-se, cometer um erro

5.2 Handling of Orthographic, Morphologic, Lexical, and Semantic Differences

Using the orthographic agreement conversor, Bigorna, is the initial step to resolve orthographic differences. In order to address the remaining conversions, a list of 2,200 entries was compiled automatically based on word occurrence in CETEM-Publico and CETEMFolha. Then, the list was enriched with entries from another Wikipedia article: list of lexical differences between versions of Portuguese Language [8]. All those entries were manually checked for accuracy. Entries in BP with more than one meaning in EP were excluded, e.g.: *bala* in BP represents either candy or bullet, while in EP it is only used as bullet. In addition, we developed an algorithm to extract automatically entries from an SMT Phrase-Table (Section 6). The final list has around 3,000 entries.

5.3 Handling of Syntactic Differences

The first syntactic rule adds an article before a possessive pronoun when necessary. The two following rules transform clitics from proclitic into enclitic position or replace a pronoun by an enclitic:

verb + pronoun → verb + "-" + pronominal form

pronominal form + verb → verb + "-" + pronominal form

The system also deals with several exceptions to the general rule of changing clitics from proclitic to enclitic position listed in Abreu (1994): negative sentences (i.e. sentences containing *não*, *ninguém*, *nunca*, and/or *jamais*); sentences with subordinate clauses (i.e. sentence containing *quando*, *até*, and/or *que*), questions (i.e. sentences containing a question mark and at least one of the following words: *que*, *quem*, *qual*, *quanto*, *como*, *onde*, *por que*, *porquê*, *porque*, and *para que*); sentences containing undefined pronouns (*alguém*, *ninguém*, *nenhum*, *nenhuma*, *nenhuns*, *nenhumas*, *todo*, *todas*, *qualquer*, *quaisquer*, *nada*, *tudo*, and *ambos*); and sentences containing adverbs (*apenas*, *só*, *até*, *mesmo*, *também*, *já*, *talvez*, and *sempre*); and exclamative sentences (!).

The transformation from proclitic to mesoclitic was not handled, because it is often optional (Mateus, 2003)(Montenegro, 2005), and because of mesoclitics being relatively rare (0.01% of the total number of words in an EP newspaper corpus of 148 Million words).

The gerund is another syntactic structure which frequently needs transformation. The in-house

[7] http://pt.wikipedia.org/wiki/Anexo:Lista_de_expressoes_idiomaticas
[8] http://pt.wikipedia.org/wiki/Anexo:Lista_de_diferencas_lexicais_entre_versoes_da_lingua_portuguesa

POS tagger identifies not only the POS of a word, but it also returns the infinite form of a verb identified in its gerund form.

gerund → "a" + infinite

The processing of gerunds also captures the typical exception to this rule, i.e. when the gerund is after a comma or semi-comma, and therefore should not be transformed.

Additionally, a rule to handle word contractions was created based on a list from Wikipedia.

6 Extraction of Multiword Lexical Contrastive pairs from SMT Phrase-Tables

Handling of Orthographic, Morphologic, Lexical, and Semantic differences is based on lists of contrastive pairs. The available lists, described in 5.2, have very limited coverage, namely on multiword lexical pairs. Table 2 contains examples. Such fact motivated the development of a generic extraction algorithm that extracts a list of words equally applicable to translate from both EP to BP and BP to EP. The algorithm involves 2 steps. Firstly, we modified the phrase extraction algorithm used in MT to eliminate phrase pairs with dangling words, i.e . words in the source sentence that are not aligned to any word in the target sentence and vice versa. Figure 2 illustrates a phrase pair with dangling words. In the default phrase extraction algorithm, phrase pairs with dangling words are extracted. For instance, in the sentence pair *no comboio* (train) to *trem* (train), where the translation of *comboio* is *trem* and the proposition *no* (in/by) is not aligned to any words, the phrase pair is *comboio* and *trem* is extracted, but the phrase pair *o comboio* and *trem* is also extracted. For the purposes of our work, we altered the algorithm to limit the dangling words that can be extracted using Geppetto (Ling et al., 2010).

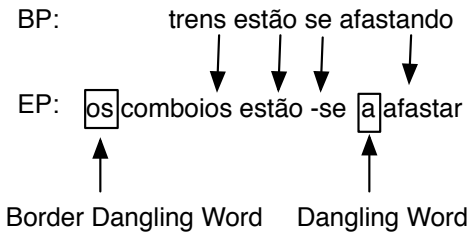

Figure 2: Example of tokenized phrase pair containing dangling words.

Two approaches were tested. The first approach does not allow any dangling word to be extracted. The second one only allows dangling words if they are not in the border of either side of the phrase pair. In the example given in Figure 2, according to the first criteria, the phrase pair would be discarded, since it contains dangling words. However, the phrase pair with the source *trens estão -se* and the target *comboios estão se* would be accepted. The second criteria would also accept the phrase pair the source *trens estão se afastando* and the target *comboios estão se a afastar*, because the dangling word is not in the border.

In the second step, the phrase table produced by the algorithm is filtered to eliminate spurious phrase pairs. The following filters are applied:

- Punctuation Removal: Punctuation is generally translated one to one, e.g.: "eu irei tentar . " →"vou tentar ." because the same entries without punctuaction also exists.
- Multiple Source Removal: If more than one possible translation exists, only the best phrase pair is chosen. To do this, we sort the phrase table source entries and we select the entry that has higher probability and scores, e.g.: "Eu achei" → "Eu pensei", "Eu achei" → "Pensei", "Eu achei" → "Pensava".
- Number based Removal: Numbers are generally translated one to one e.g.: "609 bilhões em 2008" → "609 biliões em 2008"
- Identical Translation Removal : Many words in EP are translated equally to BP, which is done by default. Furthermore, if any word in the source phrase is contained in the target one or vice versa, the phrase pair is also removed, e.g.: "Eu"→ "E Eu".
- Confidence based Entries Filtering: Removes phrase pairs with low confidence based on their features, which are used in (Koehn et al., 2003). We remove entries having probabilities lower than 1 (to trim ambiguous entries) and the respective weights are lower than 0.5 (empiric threshold). We plan in future work to improve this filter by using a linear combination of the 4 parameters.
- Lexicon based Filtering: Removes phrase pairs where the source or the target contain words that are not present in the lexicon of the respective language.
- Number of Words Filtering: Removes phrase pairs where the number of words in the source

is different from the number of words in the target. In our experiments, phrase pairs were limited to 2 words maximum, as shown in Table 2, allowing the extraction of some multiword units.

BP	EP
geladeiras	frigoríficos
bagdá	bagdade
antropólogos	antropologistas
ônibus	autocarro
astronômica internacional	astronómica internacional
deusa nicaraguense	deusa nicaraguana
tecnologia projeta	tecnologia projecta
neocortex é	neocórtex é
papel milimetrado	papel milimétrico

Table 2: Examples of extracted phrase pairs translations.

7 Results

In order to evaluate the BP2EP system, we made several types of evaluation. The first one consisted of evaluating the phrase table extraction algorithm. Our second experiment was formulated to analyze how the system can translate from BP to EP. It was used the manually parallel corpora of EP and BP.

Our third experiment evaluated the usage of the BP2EP output in SMT. Our goal was to determine whether it is observed translation quality gains when adding the BP2EP output, created from the BP texts, to the EP models. The parallel corpora used in the SMT evaluation was created from TED talks. Since the audio transcriptions and translations available at the TED website are not aligned at the sentence level, we used the Bilingual Sentence Aligner (Moore, 2002) to accomplish this task. Table 5 shows some details about the EP-EN, BP-EN, BP2EP-EN, PT-&-BP-EN and PT&-BP2EP-EN. The BP2EP corpus corresponds to the output of the BP2EP system with the BP corpus as input. The EP-&-BP and EP-&-BP2EP corpus is the concatenation of the EP corpus with the BP and BP2EP corpus, respectively.

7.1 Evaluation of the extraction of Multiword Lexical Contrastive pairs from SMT Phrase-Tables

We run the phrase table extraction algorithm for pairs of translations containing 1 and 2 words. The corpus was retrieved from the set of TED talks, that had both the Brazilian Portuguese and European Portuguese translated transcriptions. Table 3 illustrates the dimension of the corpus in each language. The initial phrase table had 668,276 entries. The inclusion of the results from this evaluation in the BP to EP improved the BLEU Score (0.2) of translation of "BP to EP" evaluation described below. Table 2 provides some examples of well extracted phrases and Table 4 contains the number of pairs extracted.

Lang. Pair	Sentences	Words
EP	23812	396763
BP	23812	402983

Table 3: Description of the corpus used to create the translation table for Lexical Contrastive pairs.

Nr. Words	Total Nr. Entries	Useful Entries
1	634	320 (51 %)
2	499	219 (44 %)

Table 4: Phrase Table Extraction Results.

7.2 "Translation" of BP to EP

Using EP as reference, the impact of our system in the translation of BP texts to EP was tested. The BLEU score between the original BP text and our manually created reference was 70.92. The BLEU score between the BP text processed with BP2EP and our reference was 75.84. This experiment showed an improvement of about 5 BLEU points.

7.3 Impact of BP2EP Output in SMT

In order to measure the impact of the BP2EP parallel corpus on the EP translation, we made several experiments. Using the EP→EN and EN→EP models as baseline, we compared them with BP and BP2EP models, and also EP-&-BP and EP-&-BP2EP.

All experiments were performed using the Moses decoder [9]. Before decoding the test set (shown in Table 5), we tune the weights of the phrase table using Minimum Error Rate Training (MERT) using the devel corpus shown in Table 5. The devel and test set are in EP and EN and are the same among the several experiments. The language model was created only with EP texts. The results were evaluated using the BLEU (Papineni et al., 2002) and METEOR (Lavie and Denkowski, 2009) metrics.

[9] http://www.statmt.org/moses/

Tables 6 and 7 shows the results for the EP/BP/BP2EP→EN and EN→EP/BP/BP2EP models, respectively. We observed that BP models generate better results than using only the EP ones. The larger amount of parallel data for BP explains these differences. The BP2EP results were systematically better than using BP models. This shows that our hypothesis of converting BP to EP using this approach led to consistent improvements to the translation between EN and EP.

Data	Lang. Pairs	Sentences	Words
Train	EP	24500	487267
	EN	24500	508146
	BP	84500	1683639
	EN	84500	1775634
	BP2EP	84500	1689143
	EN	84500	1775634
Devel	EP	500	10269
	EN	500	10949
Test	EP	438	10181
	EN	438	11417

Table 5: Data statistics for the TED Talks parallel corpora used in BP2EP evaluation.

Model	BLEU	METEOR
EP→EN	37.57	58.40
BP→EN	38.29	58.27
BP2EP→EN	**38.55**	**58.50**
EP-&-BP→EN	40.91	60.12
EP-&-BP2EP→EN	**41.07**	**60.30**

Table 6: Results of the EP/BP/BP2EP→EN models on the EP-EN test set

Model	BLEU	METEOR
EN→EP	33.13	52.60
EN→BP	34.46	54.07
EN→BP2EP	**34.48**	**54.29**
EN→EP-&-BP	35.90	54.86
EN→EP-&-BP2EP	**36.57**	**55.47**

Table 7: Results of the EN→EP/BP/BP2EP models on the EN-EP test set

8 Conclusions and Future work

In this paper, we show that the algorithm of extraction of Multiword Lexical Contrastive pairs from SMT phrase tables was useful to the BP2EP system, because it gave us several lexical / morphological entries used to resolve differences between BP to EP. We have used this to extract the lexical rules, where one entity in BP is written differently in EP. As the translation lexicon created automatically has a margin of error, it is necessary to manually filter spurious entries. If the language pair of varieties and/or dialects is well covered by the workers of crowdsourcing systems such as Amazon's Mechanical turk (AMT), it is a viable option to avoid manually filtering these entries (Callison-Burch, 2009). Another possible improvement to extraction of multiword lexical contrastive pairs algorithm is the inclusion of the translational entropy to help to identify idiomatic multiword expressions (Moirón and Tiedemann, 2006).

We have also got encouraging results (about 5 BLEU points) when we evaluated the BP2EP output against manually created EP corpora. However, the system still needs some improvements to handle particular cases. The incomplete lexical pair coverage is one of the reasons. But there are exceptions to the rules that are hard to capture by rules.

Also, the usage of BP2EP system to translate BP text to EP yields better results in an EN to EP and EP to EN translation task, in comparison with the use of BP texts.

In the future, we plan to pursue the automatically generation of syntactic rules. This will allows us to manage low frequent syntactic occurrence, e.g.: clitic reordering exceptions. Furthermore, our system only extracts lexical rules for unambiguous entities. Thus, we plan to incorporate the contextual information in the disambiguation process. Also as future work, we intent to pursue an automatic way of handling the way of addressing, i.e., creating a machine learning classifier that indicates whether the word *você* could be left in place or simply removed instead of being replaced by *tu*.

Acknowledgements

The authors would like to thank Nuno Mamede, Amália Mendes, and the anonymous reviewers for many helpful comments. Support for this research by FCT through the Carnegie Mellon Portugal Program under FCT grant SFRH/BD/33769/2009, FCT grant SFRH/BD/51157/2010, FCT grant SFRH/BD/62151/2009, and also through

projects CMU-PT/HuMach/0039/2008, CMU-PT/0005/2007. This work was supported by FCT (INESC-ID multiannual funding) through the PIDDAC Program funds.

References

Abreu, M. H. and R. B. Murteira. 1994. *Gramatica Del Portoghese Moderno*. Zanucheli Editore, 5 edition.

Almeida, J. J., A. Santos, and A. Simões. 2010. Bigorna–a toolkit for orthography migration challenges. In *Seventh International Conference on Language Resources and Evaluation (LREC2010)*, Valletta, Malta.

Altintas, K. 2002. A machine translation system between a pair of closely related languages. In *Seventeenth International Symposium On Computer and Information Sciences*.

Callison-Burch, C. 2009. Fast, cheap, and creative: evaluating translation quality using amazon's mechanical turk. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1*, EMNLP '09, pages 286–295, Morristown, NJ, USA. Association for Computational Linguistics.

Canals-Marote, R., A. Esteve-Guillén, A. Garrido-Alenda, M. Guardiola-Savall, A. Iturraspe-Bellver, S. Montserrat-Buendia, S. Ortiz-Rojas, H. Pastor-Pina, P.M. Pérez-Antón, and M.L. Forcada. 2001. The spanish-catalan machine translation system internostrum. *0922-6567 - Machine Translation*, VIII:73–76.

Caseli, H. M., T. F. Pereira, L. Specia, T. A. S. Pardo, C. Gasperin, and S. M. Aluísio. 2009. Building a brazilian portuguese parallel corpus of original and simplified texts. In *Advances in Computational Linguistics, Research in Computer Science - 10th Conference on Intelligent Text Processing and Computational Linguistics - CICLing*, volume 41, pages 59–70.

Hajič, J., J. Hric, and V. Kuboň. 2000. Machine translation of very close languages. In *Proceedings of the sixth conference on Applied natural language processing*, pages 7–12, Morristown, NJ, USA. ACL.

Koehn, P., F. J. Och, and D. Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 48–54. ACL.

Lavie, Alon and Michael Denkowski. 2009. The meteor metric for automatic evaluation of machine translation. *Machine Translation*, 23:105–115.

Ling, W., T. Luís, J. Graça, L. Coheur, and I. Trancoso. 2010. Towards a general and extensible phrase-extraction algorithm. In *IWSLT '10*, pages 313–320, Paris, France.

Mateus, M. H. M., 2003. *Gramática da Língua Portuguesa*, chapter Português europeu e português brasileiro: duas variedades nacionais da língua portuguesa, pages 45–51. Caminho, Lisbon, 5ª edition.

Moirón, B.V. and J. Tiedemann. 2006. Identifying idiomatic expressions using automatic word-alignment. In *Proceedings of the EACL 2006 Workshop on Multiword Expressions in a Multilingual Context*, Trento, Italy.

Montenegro, H. M., 2005. *Português para todos – A Gramática na Comunicação*. Mirandela: João Azevedo.

Moore, Robert C. 2002. Fast and accurate sentence alignment of bilingual corpora. In *Proceedings of the 5th Conference of the Association for Machine Translation in the Americas on Machine Translation: From Research to Real Users*, AMTA '02, pages 135–144, London, UK. Springer-Verlag.

Nakov, P. and H. T. Ng. 2009. Improved statistical machine translation for resource-poor languages using related resource-rich languages. In *EMNLP '09*, pages 1358–1367, Morristown, NJ, USA. Association for Computational Linguistics.

Navarro, J. R., J. González, D. Picó, F. Casacuberta, J. M. de Val, F. Fabregat, F. Pla, and J. Tomás. 2004. Sishitra : A hybrid machine translation system from spanish to catalan. In *Advances in Natural Language Processing*, volume 3230 of *Lecture Notes in Computer Science*, pages 349–359. Springer Berlin.

Papineni, Kishore, Salim Roukos, Todd Wardand, and Wei-Jing 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.

Ribeiro, Ricardo, Nuno J. Mamede, and Isabel Trancoso. 2003. Using Morphossyntactic Information in TTS Systems: comparing strategies for European Portuguese. In *PROPOR 2003*, volume 2721 of *Lecture Notes in Computer Science*. Springer.

Scannell, Kevin P. 2006. Machine translation for closely related language pairs. In *In Proceedings of the LREC2006 Workshop on Strategies for developing machine translation for minority languages*, Paris. ELRA.

Silva, A.S. 2008. Integrando a variação social e métodos quantitativos na investigação sobre linguagem e cognição: para uma sociolinguística cognitiva do português europeu e brasileiro. *Revista de Estudos da Linguagem*, 16(1):49–81.

Teyssier, P. 1984. *Manuel de Langue Portugaise*. Paris.