http://dx.doi.org/10.5007/2175-7968.2015v35n1p225

FIRST STEPS TOWARDS A BILINGUAL PARALLEL CORPUS GEARED TO THE TREATMENT OF LEXICAL AMBIGUITY ON THE INTERFACE PORTUGUESE - LIBRAS

Jorge Bidarra*
Universidade Estadual do Paraná

Abstract: The process of translation from one language to another, either by humans or by machines, has been an ongoing challenge for researchers worldwide. Assuming that translation is not simply the replacement of a word in the source language with another word, supposedly similar to the original; they recognize that beyond the linguistic knowledge, the translator must consider the realization scenario in both languages. Herein, in a general way, we present some preliminary results obtained from the research we have been conducting around this subject. More specifically, with the focus on the phenomenon of lexical ambiguity, we bring in this article a brief discussion concerning to the problem. These results have served as an important support for the development of a bilingual parallel corpus that our work team is developing, so that in hereafter we will have the necessary conditions for the future implementation of an electronic bilingual dictionary, already in feasibility analysis phase for specification and modeling.
Keywords: Lexical Ambiguity. Parallel Corpus. Bilingual Electronic Dictionary. Portuguese–Brazilian Sign Language.

* Associated Professor and Researcher in State University of Paraná (UNIOESTE. Postdoctorate at Federal University of Santa Catarina/Brazil (Center for Sign Language Acquisition - NALS) and University of Sheffield/UK (Research Group in Natural Language Processing) - 2013-2014. Cascavel, Paraná, Brazil E-mail: bidarra@unioeste.br/jbidarra@pq.cnpq.br.

PRIMEIROS PASSOS EM BUSCA DE UM CORPUS PARALELO BILÍNGUE VOLTADO PARA O TRATAMENTO DA AMBIGUIDADE LEXICAL NA INTERFACE PORTUGUÊS - LIBRAS

Resumo: O processo de tradução de uma língua para outra, seja executada por humanos ou por máquinas, tem sido um desafio constante para pesquisadores em todo o mundo. Partindo do princípio de que a tradução não é simplesmente a substituição de uma palavra na língua de origem para outra língua, supostamente semelhante ao original, especialistas reconhecem que, além do conhecimento linguístico, o tradutor deve considerar o cenário de realização numa e noutra línguas. Apresentamos aqui alguns resultados preliminares obtidos a partir da investigação que vimos desenvolvendo em relação a esse assunto. Mais especificamente, com o foco no fenômeno de ambiguidade lexical, trazemos para debate uma breve discussão a respeito do problema. Tais resultados nos têm servido de base importante para a construção de um corpus paralelo bilíngue que se encontra em desenvolvimento por nossa equipe de trabalho, com vista à implementação futura de um dicionário bilíngue eletrônico, já em fase de análise e estudo de viabilidade para a especificação e modelagem
Palavras-chave: Ambiguidade lexical. Corpus paralelo. Dicionário Eletrônico bilíngue. Português-Língua Brasileira de Sinais.

Introduction

The most recent scientific research conducted on natural languages has shown important achievements, in terms of both theory and application. The advent of corpus linguistics, computer linguistics, and natural language processing (NLP), new and interesting studies, specifically in the fields of human or machine translation, have started to emerge. Even so, it is evident that a great deal of research on translation that is presented by the scientific community nearly always focuses on oral language. Investment in the field of machine translation, especially involving languages belonging to distinct modalities, are yet incipient. However, the

requirements of deaf communities who simultaneously address both oral and sign languages have started to attract attention from many researchers worldwide who work in this area. Herein, we address some partial results obtained from studies (which are still in development) on the description and discussion of linguistic and theoretical aspects relating to the Portuguese–Libras interface.

In section 2, we present the scope of our research and in section 3 we summarize some of the general aspects concerning the translation process. In section 4 we address a recurring issue relating to lexical ambiguity; a frequent phenomenon in human languages, and what has been done to solve it. In the final section (5), we present the material and methods used up to now, our general considerations, and acknowledgements and references.

1. The research scope: complexities of linguistic systems

The way lexical, syntactic, and semantic structures of languages are organized and articulated to give meaning to words, sentences, and phrases constitutes one of the most complex processes in the field of human cognition. How natural languages work and are processed has, over decades, challenged the curiosity of several specialists from different fields of knowledge. Despite all the remarkable research that has been successfully conducted on various languages and their representations, Frazier (1990) points out that we are still considerably far from a definite and thorough answer to all the existing puzzles and gaps identified by current theory. According to the author, the relationship between applied principles and lexical access, as a backup for syntactic analysis, is so complex that it implies the existence of not only one, but many subsystems. These subsystems are processes which are distinct among themselves, although they do complement each other to some extent.

Keeping this in mind, through linguistic theories (theoretical, corpus, and cognitive) and NLP, we have tried to contribute to the debates that have occurred, without losing sight of the

main challenges. More specifically, we are considering at least two central issues regarding interlanguage translation: (i) in the translation process, what are the necessary considerations when a translator faces the occurrence of ambiguous words? (ii) How can we guarantee that a nominated concept in a source language will find a semantic correspondent in the target language?

2. Translating process: general aspects

Translation is not merely the substitution of a word in the source language for another one that corresponds semantically or similarly in the target language. A much more elaborate linguistic-cognitive level of processing is required, which demands of the translator not only mastery of the involved languages but also the use of adequate strategies, which, among others, include linguistic and cultural knowledge of both systems. Through the research, we became aware that having a sort of "worldly" knowledge is an important part of the translating process. The following extract by Campos elaborates:

> One does not translate from one language to another, but from one culture to another; translation requires that the translator, a repository of general knowledge, of culture in general, that each professional will enlarge and improve according to the interests of the field to which his / her paper aims at. (Campos, 1986, p. 27-28 – our translation).

As far as we know and according to Campos (1986), NLP presents tasks that require the development of different modules that go from phonetic-phonological and lexical treatment to syntactic-semantic-conceptual levels. Although definitely not in every case, but at least in great part, the context tends to show itself enough to solve many phenomena manifested by the languages, within the languages themselves or when compared with other languages.

There have been plenty of studies in the field of translation, mainly from the 1950s onwards; (Vinay and Dabernelt (1977), Nida (1964), Catford (1965), Vázquez-Ayora (1977), Newmark (1981), and Jakobson (1975). More recently, Mona Baker (1999, 2001) introduced modern concepts based on the theory of Corpus linguistics; linking her approach to the field of linguistic studies, within the study of translation. The author brought the factor of political bias in translation into discussion. According to Baker, it is this bias that determines the desired and given way a translator expresses, in the target language. Regarding all discussions in this area, one of the problems that interested us the most in our research was the matter of similarity between meanings given in one language compared to another. By this we mean the identification of possible semantic correspondents between the involved languages in the translation process.

3. Recurring problems found in the translation process and the search for solutions

Regardless of the nature or modality of the languages, there are many obstacles to be faced throughout the translation process. Among them, two central issues call more attention in our word: (1) Lexical ambiguity phenomenon and (2) the problem of similarity (or lack of similarity) in meaning between the lexical units of the source and target languages.

3.1 The phenomenon of lexical ambiguity

Characterized by the plurality of meanings, and/or senses[1], that certain words of a language may bear within a context, the occurrence of lexical ambiguity have become one of the biggest issues facing the scientific community. Traditionally, the lexical ambiguity of languages has been dealt with based on two kinds of manifestation: polysemy and homonymy. Generally, disambiguation involving homonymy does not impose extensive difficulties, neither for the

speaker nor within linguistic theoretical studies. Context is usually enough to determine the meaning that such a word bears in that moment. Regarding polysemy, the situation becomes harder. There are plenty of interesting and published papers on this subject (Bidarra, 2004; Silva, 2006; Ullmann, 1979; Gibbs et al., (1994); Kilgarriff (1997), Pustejovsky, 1995; Lakoff, 1987).

In fact, the occurrence of lexical ambiguity in languages, as well as the necessity to guarantee a correspondence of meanings between a source language statement (be it oral or not) and the target language statement, constitute two crucial points; one, for the profession of translation, and two, for automatic systems of translation.

3.2 Lexical-semantic concept *and* correspondence

Common sense dictates that words would only exist with the goal of naming things and concepts that are inherent in the world. However, by attempting to understand linguistic processing through literal translation, one of the most critical problems faced by translators is the need to ensure that the original information in a given language will transfer to another language without any loss of meaning or context. To what extent is this possible? Herein lies an interesting question.

Although it would be ideal if each word in the source language should semantically correspondent in the target one, some occurrences such as mentioned in the sequence can happen. Although there may be one word (or more) in a language to designated to a concept or object in the world, it is not rare to observe the lack of a corresponding word that is able to fully cover the concept or object in another language. To designate to a concept, in many languages one word is sufficient, while in others it would take a more complex arrangement to readily garner the same meaning. As we can note ´concept´ is the central issue, but what does it mean? Generally speaking, “concept” is a mental description or a shared idea by the speakers that may refer not

only to objects, but also to events or phenomena of the real or fictional world. (Taylor, 1985; Rosch, 1973; Cruse, 1986, 2004; Aitchinson, 1994). Through it, native speakers of a language can establish some ontology of common sense which is necessary to successfully achieve communication via their own language. According to Di Felippo and Dias-da-Silva (2009) this is the reason why, "linguistic communities tend to show different conceptual repertoires that reveal different categorizations or perspectives of the world." . Cruse (2004), Vossen *et al.* (1998), Alonge *et al.* (1998) and Bentivogli *et al.* (2000) claim that this variation is susceptible to cases of divergence; the topic to which our research has been dedicated. According to these researchers, language is subject to four kinds of divergence: conceptual, denotative, connotative, and pragmatic. Briefly, conceptual divergence occurs when a given lexicalized concept in the source language does not have a semantic correspondent in the target language. Denotative divergence happens when a lexicalized concept in a language does have one or more similar concepts in the target language. Connotative divergence, although similar to denotative, differs in that the lexicalized concept(s) in the target language would not have connotative meaning in the source language. Lastly, pragmatic divergences occur when a lexicalized concept in one language is not lexicalized in the other, but accomplished by free expression[2].

4. Material and methods

Remembering that we are particularly interested in the issue of lexical ambiguity, for our present debate, we have selected 3 out of the 19 ambiguous words in Brazilian Portuguese (to be referred to as Portuguese from this point onwards): *banco* (bank), *brilhante* (bright), and *processo* (process). We have divided this section into 5 subsections, starting with the following adopted procedures, to select and prepare the ambiguous words and sentence for analysis.

4.1 Selection and preparation of the ambiguous words and sentences for analysis

The selection of the words was based on the frequency in which they appeared in academic texts, from different fields of knowledge. Whilst completing our research, it must be noted that all matters were within a university environment. For each word, different sentences were collected, as per Table 1.

Table 1 –Ambiguous words in Portuguese Selected for the Study and number of Sentences per word

Ambiguous Words[3]	Amount of Sentences (per word)
Abatimento (*reduction*)	527
Anestesiar (*anesthetize*)	509
Arquivo (*file*)	413
Artigo (*article*)	226
Assar (bake)	151
Banco (*bank*)	225
Brilhante (*bright*)	452
Cabeça (*head*)	496
Capital (*capital*)	503
Coração (*heart*)	398
Educação (*education*)	119
Estado (*state, status*)	774
Exame (*exam*)	346
Explorar (*explore*)	150
Holocausto (*holocaust*)	631
Livro (*book*)	549

Pensão (*pension*)	222
Processo (*process*)	455
Relaxar (*relax*)	500
TOTAL OF SENTENCES	**7,646**

4.2 From the dictionaries to the definition of meanings of the words

Since it was necessary to disambiguate words before starting translations, the use of dictionaries in Portuguese, as well as in Libras, was vital. For this study, 5 dictionaries of Portuguese language were used [Ferreira, A.B.H. Novo Aurélio Século XXI: The Dictionary of Portuguese Language. RJ: Nova Fronteira, 1999; http://www.dicio.com.br; www.dicionarioweb.com.br; http://www.priberam.pt. We also used The Synonyms Dictionary available in Microsoft WORD (topic of the menu: Review) and 2 of Libras [Capovilla, F.C.; Raphael, W.D. Dicionário enciclopédico ilustrado trilíngue da língua de sinais brasileira. SP: Edusp, 2001 (new edition in 2009); www.acessobrasil.org.br/libras/].

The first, and perhaps the greatest, problem we face with regard to dictionaries was the lack of uniformity in their contents. Recurrently, for the same word, the dictionaries provided fewer or more meanings, without offering any explanation as to what had led the lexicographers to produce such records. This situation forced us to make a serious decision regarding the project. That is, for each of the 19 ambiguous words, we had to select a minimum number of set of meanings which best covered all definitions found in the dictionaries used.

It was equally important to consider the words where some consistency was found in these dictionaries, but this was not always the case. Just as an example, let us see what happened to the word “bright.” Even if all the consulted dictionaries registered

"intelligent" as a possible meaning for "bright," we thought it better not to include it in the set we created. The reason is that "intelligent" is also an ambiguous word, just like brilhante (*bright*). Our solution was to replace "intelligent" with two other meanings: notável (*remarkable*) and talentoso (*talented*).

The next step was to identify the linguistic environment in which each one tended to manifest. For this, we adopted a set (quantified in Table 1) of sentences for each word. Prior to analysis, the sentences underwent an extensive process of preparation, the details of which will be outlined later on in the research.

4.3 Preparation, analysis and translation of sentences to glosses Libras

The sentences to support lexical analyzes were taken from the texts of books (didactic, non-didactic, technical and non-technical), technical and scientific articles, newspaper and magazine articles, Bible texts, and different websites (random sentences by writers, thinkers, artists, scientists, politicians, etc.). For this selection we initially used the tool Wordsmith v.4, and later, v.6, a more updated version (*www.lexically.net/wordsmith/*). The greatest concern regarding this selection were both to try and guarantee that the sentences would not be restricted to only one field of knowledge. We understood that the more expansive the subjects expressed in these sentences was, the better. In other words, this extensive search prioritized the criterion of generality of subjects and fields. The reason for this was that the ambiguous words would comprise a vast universe of possibilities, which was the ideal situation for our goals.

Two actions had to be taken prior to analysis. The first was to dismiss the sentences strongly identified as metaphors. This needed to be done because metaphorical sentences bring along comparative and figurative elements to express reality situations which goes beyond the scope of this research. The second, by using wordsmith, some original sentences had to be truncated due to the lack of important excerpts in the beginning or at the end. This could cause

difficulties in determining the exact meanings of the ambiguous words. For this reason, we chose to complement the sentences so that they would start making more sense (we must remember that in our research, lexical-semantic is a central matter). The problem noted with the use of wordsmith was that the cutting line of the extracted excerpts was not determined through punctuation, but through the size of the window defined by the user.

We tried to recover greater contexts, hoping that the information we needed would arise. However, it was soon revealed to be a hard and innocuous task, and therefore, unnecessary. After all, we were not interested in analyzing the syntactic structures of the sentences, but simply in having a favorable syntactic environment so that the meanings of the words could be sought and determined by local contexts.

This way, the criterion adopted was to reduce the linguistic environment in which the ambiguous words were appearing. By "reduced environment" we understand that the sentences were to be composed of one, or, at the most, two clauses, and without major changes. For instance, we transformed sentences which were in the passive voice to their correspondent in the active voice. We also tried to adjust sentences which did not fit the canonical syntactic structure of Portuguese language – subject verb [verbal complement] (optional).

The need to apply these criteria over the sentences we had already chosen was an arduous process. However, it was from this process that we could produce a significantly large set of sentences for the future analyzes. Since this paper resulted in 7,646 different sentences, it was unfeasible to go further on the analysis and translation of all of them within the timeframe available. Therefore, what we have shown here corresponds only to a sample of all the material researched in the study.

For each of the three ambiguous words cited above, we selected 100 sentences. This amount can be taken without any impairment because we noticed that with this number we were able to cover the different meanings that these words took, successfully

reaching our set targets. For this task, dictionaries were used, as previously mentioned, as well as assumptions as per the theories of Lexical Semantic and Corpus Linguistics. The main goal in this phase was, from the listed senses, to determine which one would manifest in each of the sentences. For that, not only ambiguous words were observed, but also the influence the words that co-occured[10] with them had over the manifested meaning. According to the guidance proposed by Corpus Linguistics, most of the so-called stop words (prepositions, articles, conjunctions, pronouns, etc.) were discarded, to concentrate on the words of open class or those referred to as "words with content" (nouns, verbs, adjectives, etc.).

Given that the degree of influence the co-occurring words had over determining the meaning of ambiguous words was often different, they were separated into two major groups: strong co-occurring and not-so-strong co-occurring. We referred strong co-occurring to the ones that, in local context, could assure that a given meaning, and not another, prevailed in that context. The words that, although contributing to this determination, had a less impressive performance were defined as not-so-strong co-occurring (the way they were identified was the fact that, by themselves, these words did not bear the same strength as the strong co-occurring ones). The criterion used to identify which were co-occurring was that they belonged to "conceptual organization of an activity or knowledge field" (for instance, lexicon of fishing, lexicon of music, soccer, etc.) – terminologies – or, that they were part of the same semantic field of the ambiguous word.

Once the meaning was determined, the information was attached to the ambiguous word in each sentence. In order to set the difference among the three kinds of information mentioned, we decided that those underlined would indicate the strong co-occurring ones and those in italics as the not-so-strong co-occurring ones and the notation <...> to explain the meaning taken by the word in each of the different contexts. Below are some examples; the ambiguous words are shaded in yellow.

- Banco

(1) Na *praça*, na *igreja*, num banco <assento> de *jardim*, é bom sentar para *descansar*.
[In the *square*, at the *church*, on a bench <seat> of a *garden*, it is good to sit in order to rest.]

(2) Acessamos os bancos <Lugar para armazenamento de dados ou informações> de dados de cada *cliente*.
[We have accessed the databases <Place to store data or information> of each client]

- Brilhante

(3) Conheça a estrela mais brilhante <cintilante> do *planeta*.
[Know the brightest star <shining> of the planet.]

(4) *Ele* foi um médico brilhante <talentoso>.
[He was a bright <talented> doctor.]

- Processo

(5) A brincadeira é um dos mais importantes processos <atividade> para fins de *ensino* e *aprendizagem*.
[Playing is one of the most important processes <activity> in order to teach and learn.]

(6) O processo <evolução, transformação> de amadurecimento acontece *ao longo de sua vida escolar*.
[Maturing process <evolution, transformation> happens throughout school life.]

- Translating process Portuguese - glosses Libras

The glosses shown here resulted from a long and careful work involving four interpreters and teachers, all of them proficient in Brazilian Sign Language and Portuguese, 1 deaf and 3 listeners. Each sentence has been translated by each of them separately and at the end the results compared and adjusted by themselves, a joint work. To facilitate the visualization of connections established between words/expressions in Portuguese

and glosses in Libras, we have added a sub-indexed numeric note to the compositions of the representations. The sub-indexes correspond to the terms (words and/or expressions) which would be the semantic equivalent, in the sentence in Portuguese to the sentences translated to Glosses Libras, as shown in the example sentences provided below.

(1) Na *praça*$_1$, na *igreja*$_2$, num banco$_3$ <assento> de *jardim*$_4$, é$_5$ bom$_6$ sentar$_7$ para *descansar*$_8$.
[In the square$_1$, at the church$_2$, on a bench$_3$ <seat> of a garden$_4$ it is$_5$ good$_6$ to sit$_7$ in order to rest$_8$.]

PRAÇA$_1$ IGREJA$_2$ JARDIM$_4$ BANCO$_3$ SER$_5$ BOM$_6$ SENTAR$_7$ DESCANSAR$_8$
[SQUARE$_1$ CHURCH$_2$ GARDEN$_4$ BENCH$_3$ BE$_5$ GOOD$_6$ TO SIT$_7$ TO REST$_8$]

(2) Acessamos$_1$ os bancos$_2$ <Lugar para armazenamento de dados ou informações> de dados$_3$ de cada$_4$ *cliente*$_5$.
[We have accessed$_1$ the databases$_{2\text{-}3}$ <a place to store data or information> of each$_4$ *client*$_5$.]

NÓS ACESSAR$_1$ BANCO-DADOS$_{2\text{-}3}$ CADA-UM@$_4$ PESSOA^PAGAR$_5$
[WE ACCESS$_1$ DATABASE$_{2\text{-}3}$ EACH ONE@$_4$ PERSON^PAY$_5$]

(3) Conheça$_1$ a estrela$_2$ mais brilhante$_3$ <cintilante> do *planeta*$_4$.
[Know$_1$ the brightest$_3$ star$_2$ <shining> of the *planet*$_4$]

CONHECER$_1$ ESTRELA$_2$ BRILHO+++$_3$ MUNDO$_4$
[KNOW$_1$ STAR$_2$ BRIGHT+++$_3$ WORLD$_4$]

(4) *Ele*$_1$ foi$_2$ um médico$_3$ brilhante$_4$ <talentoso>.
[He$_1$ was$_2$ a bright$_4$ <talented> doctor$_3$.]

HOMEM^EL@$_{1}$ PASSADO SER2 MÉDICO3 TALENTO4
[MAN^HE$_{1}$ PAST TO BE$_{2}$ DOCTOR$_{3}$ TALENTED$_{4}$]

(5) A brincadeira$_{1}$ é um dos$_{2}$ mais importantes$_{3}$ processos$_{4}$ <atividade> para fins de$_{5}$ *ensino*$_{6}$ e$_{7}$ *aprendizagem*$_{8}$.
[Playing$_{1}$ is one of the$_{2}$ most important$_{3}$ processes$_{4}$ <activity> in order to$_{5}$ teach$_{6}$ and$_{7}$ learn$_{8}$.]

BRINCADEIRA$_{1}$ SER$_{2}$ TRABALHO-DESENVOLVER$_{4}$ IMPORTANTE+++$_{3}$
OBJETIVO$_{5}$ ENSINAR$_{6}$ TAMBÉM$_{7}$ APRENDER$_{8}$.
[PLAYING$_{1}$ TO BE$_{2}$ WORK-DEVELOP$_{3}$ IMPORTANT+++$_{3}$ OBJECTIVE$_{5}$ TO TEACH $_{6}$ ALSO$_{7}$ TO LEARN$_{8}$.]

(6) O processo <evolução, transformação> de$_{1}$ amadurecimento$_{2}$ acontece$_{3}$ *ao longo de*$_{4}$ *sua vida*$_{5}$ *escolar*$_{6}$
[Maturing$_{2}$ process <evolution, transformation> happens$_{3}$ *throughout*$_{4}$ *your school*$_{6}$ *life*$_{5}$.]

FREQUENTAR ESCOLA DESDE-PEQUEN@$_{4-5-6}$
ACONTECER$_{3}$ AMADURECIMENTO$_{2}$
TRANSFORMAÇÃO$_{1}$.
[TO ATTEND SCHOOL SINCE-CHILDHOOD$_{4-5-6}$ TO HAPPEN$_{3}$ MATURING$_{2}$ TRANSFORMATION$_{1}$]

For each word/expression, we also organized tables containing all the terms and their correspondent in glosses, per sentence, creating sets of information such as illustrations provided linearly in the sequence. The tables were needed, due to our requirement to visualize the data as a whole, and not only segments, as had been happening with the adopted notations. It was possible to develop bilingual lexical/thesaurus/dictionaries with the adopted notations by use of indexing methods through different keys, according to Portuguese, Glosses, and, in some future work, via *sign writing*.

(1) Na *praça*$_1$, na *igreja*$_2$, num banco$_3$ <assento> de *jardim*$_4$, é bom$_5$ sentar$_7$ para *descansar*$_8$.
[In the square$_1$, at the church$_2$, on a bench$_3$ <seat> of a garden$_4$, it is$_5$ good$_6$ to sit$_7$ to rest$_8$.]

PRAÇA$_1$ IGREJA$_2$ JARDIM$_4$ BANCO$_3$ SER$_5$ BOM$_6$ SENTAR$_7$ DESCANSAR$_8$
[SQUARE$_1$ CHURCH$_2$ GARDEN$_4$ BENCH$_3$ TO BE$_5$ GOOD$_6$ TO SIT$_7$ TO REST$_8$]

Words/Expression in Portuguese ≡ Words/Expression in Glosses:

- [na] praça ≡ PRAÇA [SQUARE]
- [na] igreja ≡ IGREJA [CHURCH]
- [num] banco ≡ BANCO [BENCH]
- [de] jardim ≡ JARDIM [GARDEN]
- É ≡ SER [TO BE]
- Bom ≡ BOM [GOOD]
- Sentar ≡ SENTAR [TO SIT]
- [para] descansar ≡ DESCANSAR [TO REST]

(2) Acessamos$_1$ os bancos$_2$ <Lugar para armazenamento de dados ou informações> de dados$_3$ de cada$_4$ *cliente*$_5$.
[We have accessed$_1$ the databases$_{2\text{-}3}$ <a place to store data or information> of each$_4$ *client*$_5$.]

NÓS ACESSAR$_1$ BANCO-DADOS$_{2\text{-}3}$ CADA-UM@$_4$ PESSOA^PAGAR$_5$
[WE ACCESS$_1$ DATABASE$_{2\text{-}3}$ EACH ONE@$_4$ PERSON^PAY$_5$]

Words/Expression in Portuguese ≡ Words/Expression in Glosses:

- acessamos ≡ NÓS ACESSAR [We have accessed]
- [os] bancos ≡
- de dados ≡ } BANCO-DADOS [the databases]

- [de] cada ≡ CADA-UM@ [of each]
- cliente ≡ PESSOA^PAGAR [*client*]

(3) Conheça$_1$ a estrela$_2$ mais brilhante$_{3\,<\text{cintilante}>}$ do *planeta*$_4$.
[Know$_1$ the brightest$_3$ star$_{2\,<\text{shining}>}$ of the *planet*$_4$]

CONHECER$_1$ ESTRELA$_2$ BRILHO+++$_3$ MUNDO$_4$
[KNOW$_1$ STAR$_2$ BRIGHT+++$_3$ WORLD$_4$]

Words/Expression in Portuguese ≡ Words/Expression in Glosses:

- conheça ≡ CONHECER [KNOW]
- [a] estrela ≡ ESTRELA [STAR]
- mais brilhante ≡ BRILHO+++ [BRIGHT+++]
- [do] planeta ≡ MUNDO [WORLD]

(4) *Ele*$_1$ foi$_2$ um médico$_3$ brilhante$_{4\,<\text{talentoso}>}$.
[He$_1$ was$_2$ a bright$_{4\,<\text{talented}>}$ doctor$_3$.]

HOMEM^EL@$_1$ PASSADO SER2 MÉDICO3 TALENTO4
[MAN^HE$_1$ PAST TO BE$_2$ DOCTOR$_3$ TALENTED$_4$]

Words/Expression in Portuguese ≡ Words/Expression in Glosses:

- Ele ≡ HOMEM^EL@ [[MAN^HE]
- foi ≡ PASSADO SER [PAST TO BE]
- [um] médico ≡ MÉDICO [DOCTOR]
- brilhante ≡ TALENTO [TALENTED]

(5) A brincadeira$_1$ é um dos$_2$ mais importantes$_3$ processos$_4$ <atividade> para fins de$_5$ ensino$_6$ e$_7$ aprendizagem$_8$.
[Playing$_1$ is one of the$_2$ most important$_3$ processes$_4$ <activity> in order to$_5$ teach$_6$ and$_7$ learn$_8$.]

BRINCADEIRA$_1$ SER$_2$ TRABALHO-DESENVOLVER$_4$ IMPORTANTE+++$_3$
OBJETIVO$_5$ ENSINAR$_6$ TAMBÉM$_7$ APRENDER$_8$.
[PLAYING$_1$ TO BE$_2$ WORK-DEVELOP$_3$ IMPORTANT+++$_3$ OBJECTIVE$_5$ TO TEACH $_6$ ALSO$_7$
TO LEARN$_8$.]

Words/Expression in Portuguese ≡ Words/Expression in Glosses:

- [a] brincadeira ≡ BRINCADEIRA [PLAYING]
- é [um dos] ≡ SER [TO BE]
- mais importantes ≡ IMPORTANTE+++ [IMPORTANT+++]
- processos ≡ TRABALHO-DESENVOLVER [WORK-DEVELOP]
- para fins de ≡ OBJETIVO [OBJECTIVE]
- ensino ≡ ENSINAR [TO TEACH]
- e ≡ TAMBÉM [ALSO]
- aprendizagem ≡ APRENDER [TO LEARN]

(6) O processo <evolução, transformação> de$_1$ amadurecimento$_2$ acontece$_3$ *ao longo de$_4$ sua vida$_5$ escolar$_6$*
[Maturing$_2$ process <evolution, transformation> happens$_3$ *throughout$_4$ your school$_6$ life$_5$*.]

FREQUENTAR ESCOLA DESDE-PEQUEN@$_{4\text{-}5\text{-}6}$
ACONTECER$_3$ AMADURECIMENTO$_2$
TRANSFORMAÇÃO$_1$.
[ATTENDING SCHOOL FROM AN EARLY AGE@$_{4\text{-}5\text{-}6}$ TO HAPPEN$_3$ MATURING$_2$ TRANSFORMATION$_1$]

Words/Expression in Portuguese ≡ Words/Expression in Glosses:

- [o] processo [of] ≡ TRANSFORMAÇÃO [TRANSFORMATION]
- amadurecimento AMADURECIMENTO [MATURING]
- acontece ≡ ACONTECER [TO HAPPEN]
- ao longo de
- sua
- vida
- escolar

} ≡ FREQUENTAR ESCOLA DESDE-PEQUENO@ [ATTENDING SCHOOL FROM AN EARLY AGE@]

These procedures were not developed in a single step. Adaptations were made in the corpus which ended up producing different versions of the material. Initially, we had decided to highlight only the ambiguous words in Portuguese, their co-occurring and the semantically corresponding terms in glosses, and all of them together with their respective sub-indexes. However, as we went on with the notes, we noticed that, not only the translation of other terms of the sentences, but also their indexes would be useful for the work as a whole. This decision to use the indexes, although taking more time, was important because it allowed us to create more complete and demonstrative tables of all the possible semantic correspondents established among the terms of both languages. These more complete tables were the impetus for new research: the building of a bilingual lexicon.

The search for terms which showed similarity in Portuguese and Libras was neither an easy task, nor a trivial one. In this process, we faced situations, besides those previously mentioned (such as the exclusion of *stop words*), whereby a term in Portuguese did not have a similar corresponding word in Glosses Libras. In such circumstances, we used markers, such as INF (INFormation) used, in this case, to introduce a particular concept not materialized in Libras in the sentence in Libras.

Even with all the information shown in our examples and attachments, we understand that, because the data is associated to each sentence in isolation, visualization was neither clear not functional enough to present the idea as a whole. We therefore decided to build tables of semantic correspondents, where all the relationships found throughout the analysis of the sentences could be presented, for each of the ambiguous words studied. The aim was, in the end, to have an extensive database, similar to those we have produced in relation to the parallel corpus, only with all the words we've been working on, with their semantic correspondents in Libras, organized and structured together.

The final step was to match the translations from glosses to practical Libras. Although this was not included in the initial scope of the research, we believe its availability will be extremely useful because it would provide us with better conditions for ascertaining whether or not the translations are compatible with practical Libras. This stage has begun and we can view its result, although it is still pretty incipient. The translations already performed have so far indicated interesting results for Libras.

Final considerations

We have shown here only some steps we have been executed along our work aiming at developing an electronic bilingual Parallel Corpus specifically geared to the implementation of a computational treatment of lexical ambiguity on the Interface Portuguese. While in sign language such as American, English and Spanish, to name a few, it is possible to access different types of parallel corpus, in Brazil this investment begins only happen more recently. One example has been the researches and works that are being developed by the research group that integrates NALS (http://nals.cce.ufsc.br/)/Federal University of Santa Catarina. Despite conducting plenty of research, in an attempt to find a Portuguese – Libras corpus with the necessary characteristics to meet the ones

required by our work, we did not succeed. Considering, however, that having a corpus was and is a main prerequisite here, we decided early on to construct one. Given a corpus constitutes a set of linguistic data, systematized through determined criteria, so as to be processed by computer, the design of one was not an easy task. This was due to challenges with the main goal; to provide plenty of useful results for description and linguistic analyzes, the process of which involves two different languages (Sinclair, 1991)

Based on all the challenges in the activities performed so far, the results from this project have been encouraging. We know, however, that reviews still need to be conducted, with regards to the project's content and other pieces of information relating to the databases and analyses we have been performing. These aspects of Stage I, along with future requirements such as diversified analyzes, expansion of the corpus, and formalization of a representation of the linguistic data obtained, will lead to establishing the basic and fundamental conditions for Stage II of the project, according to what has been shown in this text.

We are fully aware that several points of this research need to be highlighted. Choosing to work with Glosses, despite not being considered a language, was extremely important. Through them, we were able to easily identify the problems, as well as various phenomena, which were highlighted throughout the translating process. Although we do not have conditions nor space to hold this debate at present, it is worth mentioning 2 occurrences that called more our attention, namely: (i) some words that are verbs in Portuguese, become part of nominal classes in Libras, and (ii) words from closed classes, such as articles, pronouns, prepositions, conjunctions, etc., disappear in translation. Finally, reinforcing what we have already mentioned before, this report is only an initial part of what we are developing in our research group. We know that the challenges ahead are enormous, but we are confident of reaching satisfactory results that will be useful for our future demands.

Acknowledgements

To the researchers Ronice Müller de Quadros, Professor and Center for Sign Language Acquisition (NALS) Coordinator at Federal University of Santa Catarina, and Lucia Specia, Professor and Member of the Natural Language Processing Group at the University of Sheffield/UK Computer Science, with which I had the great opportunity to work, my supervisors in postdoctoral. Thanks to the teachers and translators/interpreters, hearing and deaf, whose assistance was invaluable in the development of this work. Special thanks to the teachers Tania Aparecida Martins/ UNIOESTE (campus of Marechal C. Rondon), Leidiani Reis and the undergraduate students of Language and Scientific Research. To Professor Mirna F. de Oliveira, UNIOESTE (campus of Foz do Iguaçu), for adding to the discussion and for review of the produced material. Finally, to the Araucaria Foundation (Support Scientific and Technological Development of Parana/BR) and Coordination for the Improvement of Higher Education (CAPES/ BR) for scholarships.

Notes

1. We make a distinction between sense and meaning. For such we took the theoretical view claimed by Vygotsky (1986), for whom the sense regards all the psychological facts immersed in our consciousness (zones of meanings) and the meaning would be part of these sense zones that the word has taken, within the context of some speech together with other ones. In other words, the meaning would be the most basic and unchanged semantic information that a word has, regardless of the different contexts in which it manifests itself, while the sense is given to the word according to the context manifested.

2. Conceptual and Pragmatic divergences are also known as lexical gaps.

3. The words given in English are for reference purposes only. Any discussion of these words in English is outside the scope of this research.

4. The writing of a text in the structure of Brazilian sign language, through the use of words from Portuguese, forming an intermediate language (glosses Libras), has become a useful resource, not only for translators, but also for the deaf (Quadros and Souza, 2008), even after considering their limitations. In the case of this project, using this resource represented an undeniable benefit, both for our better understanding of the operation of Libras, and as an interface to develop computer tools.

5. In fact, there are 4 words and not 3, because, besides the ones mentioned here, we also worked with the sentences with the word "state". These analyzes resulted in a master thesis supervised by Jorge Bidarra, Keli Vidarenko da Rosa, the title of which is: The impact of the Occurrence of Ambiguous Words in Portuguese in the Translating Process to Libras, through Glosses: The case of the word "State", with public defense on 03/11/2014, at UNIOESTE, campus Cascavel/Paraná/Brazil.

6. Local Context is understood as the fact that co-occurring words are next to ambiguous words in proximity. The notion of closeness, although it could imply a direct connection between some, is less rigid. Generally, we consider close words the ones which appear in the same clause or, at most, in two.

7. We took the concepts of semantic fields and terminologies as described by Lehrer (1974).

8. The proposed numbers to the word Sent., in brackets and at the end of each sentence, indicate the number of the sentence in each of the attached files.

References

Aitchinson, J. 1994. *Words in the Mind: an introduction to the mental lexicon*. Blackwell.

Alonge, A.; Calzolari, N.; Vossen, P.; Bloksma, L.; Castellon, I.; Antonia, M.; Marti, M. A.; Peters, W. 1998. The linguistic design of the EuroWorNet database. In *Computers and the Humanities*, 32:91-115. Dordrecht: Kluwer Academic Publishers.

Aristotle. 1984. *Rethoric and Poetics of Aristotle* (Modern Library College). McGraw-Hill.

Baker, M. 1996. Linguistic and Cultural Studies. Complementary or Competing Paradigms in Translation Studies? A. Lauer *et al*. (eds.), *Ubersetzungswissenschaft im Umbruch,* Tubingen: Gunter Narr.

Baker, M.; Malmkjaer, K. 2001. *Routledge Encyclopedia of Translation Studies*. London, New York: Routledge.

Bentivogli, L; Pianta, E. Pianesi, F. 2000. Coping with lexical gaps when building aligned multilingual wordnets. In: *International Conference on Language Resources and Evaluation - LREC, 2, 2000, Athens*. Proceedings. (http://multiwordnet.itc.it/english/pubb.php).

Bidarra, J. 2004. *O léxico no processamento da Linguagem Natural*. Cascavel: Edunioeste.

Campos, G. 1986. *O que é tradução*. Coleção Primeiros Passos – 166 Leituras Afins. SP: Editora Brasiliense.

Catford, J.C. 1965. *A Linguistic Theory of Translation: An essay in Applied Linguistics*. London: Oxford Press.

Cruse, D.A. 2004. Meaning in language: an introduction to semantics and pragmatics. Oxford: Oxford University Press.

_____. 1986. Lexical Semantics. NY: Cambridge University Press.

Di Felippo, A. e Dias-da-Silva, B.C. 2009. *A interlíngua da base lexical bilíngue REBECA*. Veredas On-Line: Linguística de Corpus e Computacional, 2: 50-67, PPG Lingüística/UFJF.

Frazier, L. 1990. Exploring the Architecture of the Language-Processing System. G.T.M. Altman (Ed.). *Cognitive Models of Speech Processing*, 409-433. Cambridge: The MIT Press.

Gibbs, R.W.; Beitel, D.; Harrington, M. & Sanders, P. 1994. The Poetics of Mind. Figurative Thought, Language, and Understanding. Cambridge: Cambridge University Press.

Jakobson, R. 1975. Linguistica e Comunicação. São Paulo: Cultrix.

Lakoff, G. 1987. Women, Fire, and Dangerous Things: What Categories Reveal about the Mind. Chicago: The University of Chicago Press.

Kilgarriff, 1997. A. I don´t believe in word senses. *Computers and the Humanities,* 31 (2): 91-113.

Lehrer, A. 1974. Semantic Fields and Lexical Structure. North-Holland Linguistic Series (11). American Elsevier.

Nida, E. 1964. Toward a Science of Translation. Leiden: Brill.

Newmark, P.1981. Approaches to Translation (Language Teaching Methodology). Elsevier Science & Technology.

Perini, M.A. 2003. Sofrendo a gramática. 3ª ed., São Paulo: Ática.

Pustejovsky, J. 1995. The Generative Lexicon. Cambridge: The MIT Press.

Quadros, R.M.; Souza, S.X. 2008. Aspectos da tradução/encenação na Língua de Sinais Brasileira para um ambiente virtual de ensino: práticas tradutórias do curso de Letras Libras. In: R.M. Quadros (ed.). *Estudos Surdos III*. Petrópolis: Arara Azul.

Rosch, E. 1973. Natural categories. *Cognitive Psychology*, 4: 328-350.

Silva, A.S. 2006. *O Mundo dos Sentidos em Português: Polissemia, Semântica e Cognição*. Coimbra: Edições Almedina.

Sinclair, J. 1991. *Corpus, Concordance, Collocation*. Oxford University Press.

Taylor, J.R. 1985. *Linguistic Categorization: prototypes in linguistic theory*. Oxford: Clarendon Press.

Ullmann, S. 1979. *Semantics: An Introduction to the Science of Meaning*. Rowman&Littlefield.

Vinay, J.P. and Darbelnet, J. 1977. *Stylistique Comparée du Français et de l'Anglais*. Paris: Didier.

Vygotsky, L.S. 1986. *Thought and Language*. Newly revised by Alex Lozulin. The Massachusetts Institute Technology.

Vossen, P. Marinai, E; Peters, C; Castellon, I; Marti, A. Rigau, G. 1998. Compatibility in interpretation of relations in EuroWordNet. *Computers and the Humanities*, 32:153-184. Dordrecht: Kluwer Academic Publishers.

Recebido em: 17/01/2015
Aceito em: 08/03/2015