Procesamiento del Lenguaje Natural, núm. 35 (2005), pp. 237-244 recibido 29-04-2005; aceptado 01-06-2005

Evaluating the LIHLA lexical aligner on Spanish, Brazilian Portuguese and Basque parallel texts *

Helena M. Caseli, Maria G. V. Nunes
NILC – ICMC – Univ. São Paulo
CP 668P, 13560-970 São Carlos–SP, Brazil
{helename,gracan}@icmc.usp.br

Mikel L. Forcada
Transducens – DLSI – Univ. d'Alacant
E-03071 Alacant, Spain
mlf@dlsi.ua.es

Resumen: El alineamiento de palabras y de unidades multipalabra desempeña un papel importante en muchas aplicaciones del procesamiento de lenguaje natural, tales como la traducción automática basada en ejemplos, la inducción de reglas de transferencia para la traducción automática, la lexicografía bilingüe, la desambiguación de la polisemia, etc. En esta comunicación describimos LIHLA, un alineador de palabras que utiliza léxicos probabilísticos bilingües generados por un paquete de herramientas libremente disponible (NATools) y heurísticas independientes del idioma para encontrar alineamientos entre palabras y unidades multipalabra en textos paralelos alineados por oraciones. El método ha alcanzado una precisión de un 92.44% y un 85.09% y una cobertura de un 91.13% y un 64.66% en textos paralelos escritos en portugués brasileño–español y español–euskera, respectivamente.
Palabras clave: Alineamiento de palabras, portugués, español, euskera

Abstract: Alignment of words and multiword units plays an important role in many natural language processing applications, such as example-based machine translation, transfer rule learning for machine translation, bilingual lexicography, word sense disambiguation, etc. In this paper we describe LIHLA, a lexical aligner which uses bilingual probabilistic lexicons generated by a freely available set of tools (NATools) and language-independent heuristics to find links between single words and multiword units in sentence-aligned parallel texts. The method has achieved a precision of 92.44% and 85.09% and a recall of 91.13% and 64.66% on Brazilian Portuguese–Spanish and Spanish–Basque parallel texts, respectively.
Keywords: Lexical alignment, Portuguese, Spanish, Basque

1 *Introduction*

Alignment of words and multiword units plays an important role in many natural language processing (NLP) applications, such as example-based machine translation (EBMT) (Somers, 1999) and statistical machine translation (SMT) (Ayan, Dorr, and Habash, 2004; Och and Ney, 2000), transfer rule learning (Carl, 2001; Menezes and Richardson, 2001), bilingual lexicography (Gómez Guinovart and Sacau Fontenla, 2004), and word sense disambiguation (Gale, Church, and Yarowsky, 1992), among others.

Aligning two (or more) texts means finding correspondences (translation equivalences) between segments (paragraphs, sentences, words, etc.) of the source text and segments of its translation (the target text). In this paper the focus is on lexical alignment, that is, alignment between single words and multiword units in Brazilian Portuguese (`pt`), Spanish (`es`) and Euskera (`eu`) parallel texts.

In the last years, several lexical alignment systems have been proposed in the literature among all of them, statistical systems are considered to be the state of the art (e.g., Hiemstra (1998) and Och and Ney (2000)). Although these systems provide quite satisfactory results they can not deal properly with syntactic differences between languages, such as non-consecutive phrasal information, long-range dependencies (Ayan, Dorr, and Habash, 2004) and alignments involving multiword units. These problems are very frequent in lexical alignment and unfortunately also very difficult to handle.

Following the same idea of many recently

* We thank FAPESP, CAPES, CNPq and the Spanish Ministry of Science & Technology (Project TIC2003-08681-C02-01) for financial support. We also thank Miriam A. G. Scalco for her help in the manual lexical alignment process.

ISSN: 1135-5948 © 2005 Sociedad Española para el Procesamiento del Lenguaje Natural

proposed approaches on lexical alignment (e.g., Wu and Wang (2004) and Ayan, Dorr, and Habash (2004)), the method described in this paper, LIHLA (Language-Independent Heuristics Lexical Aligner), tries to solve some of these problems by using statistical alignments between single words (defined in bilingual probabilistic lexicons) as a starting point, and by applying language-independent heuristics to them, aiming at finding the best alignments between words or multiword units.

Although the most frequent alignment category is $1:1$ (in which one source word is translated exactly as one target word), other categories such as omissions ($1:0$ or $0:1$) or those involving multiword units ($n:m$, with n and/or $m \geq 1$) are also possible. An example of alignment involving a multiword unit is the $1:2$ alignment between `pt` word *dos* and `es` multiword unit *de los*.

This paper is organized as follows: section 2 presents an overview of bilingual lexicon generation and section 3 explains how LIHLA works. Section 4 describes some experiments carried out with LIHLA and their results. Finally, in section 5, some concluding remarks are presented.

2 Bilingual lexicon generation

As the first step, LIHLA uses alignments between single words defined in two statistical bilingual lexicons (source–target and target–source) generated from sentence-aligned parallel texts using NATools.[1]

So, given two sentence-aligned corpus files, the NATools word aligner —based on the Twenty-One system (Hiemstra, 1998)— counts the co-occurrences of words in all aligned sentence pairs and builds a sparse matrix of word-to-word probabilities using an iterative expectation-maximization algorithm. Finally, the elements with higher values in the matrix are chosen to compose two probabilistic bilingual lexicons (source–target and target–source) (Simões and Almeida, 2003). For each word in the corpus, each bilingual lexicon gives: the number of occurrences of that word in the corpus (its absolute frequency) and its most likely translations together with their probabilities.

[1]NATools is a set of tools developed to work with parallel corpora, which is freely available in `http://natura.di.uminho.pt/natura/natura/`.

Figure 1 shows an entry in the `pt`–`es` bilingual lexicon to the `pt` word *dos*. In this example, the best translation is *los* and the second one is *de*. It is due to the fact that the `pt` word *dos* can be translated to several combinations of other prepositions plus the definite article *los* or just the article. Also, the probability of omission of its translation (indicated by \(`null`\)) is specified which is higher than the probability of its translation as *dos* or *la*.

```
"dos" => {
    count => 2196,
    trans => {
       "los" => 0.74646669626236,
       "de" => 0.178675398230553,
       "\(null\)" => 0.0156443770974874,
       "dos" => 0.0111551126465201,
       "la" => 0.00522150145843625,
    },
},
```

Figure 1: Possible translations for `pt` word *dos* in the `pt-es` bilingual lexicon

3 How LIHLA works

Using two bilingual lexicons generated by NATools (in the previous step) and some language-independent heuristics, LIHLA tries to find the best alignment between source and target tokens (words, numbers, special characters, etc.) in a pair of parallel sentences by following the algorithm on Figure 2.[2] As its output, LIHLA produces a set A of alignments ($\alpha : \beta$) where α is a sequence of one or more source tokens (separated by '+'), and β is a similar sequence of target tokens.

For each source token s_j in a source sentence S, initially, LIHLA takes those source and target tokens in the parallel sentences with the same type (word or special character[3]) as s_j as possible translations of each other. Those source and target tokens are stored in source (C_S) and target (C_T) candidate sets, respectively. Then, the tokens are aligned according to their types using the `align_char` or `align_word` functions. This process is repeated while alignments can still

[2]A previous version of LIHLA, version 1.0 (Caseli, Nunes, and Forcada, accepted paper), aligns raw parallel texts in spite of sentence-aligned ones.

[3]Each token in a source/target sentence is classified as a word if it contains at least one alphanumeric character or as a special character otherwise.

```
algorithm LIHLA
Input:
  a source sentence S = {s_1, ..., s_x} with x tokens
  a target sentence T = {t_1, ..., t_y} with y tokens
  a source–target bilingual lexicon B_S
  a target–source bilingual lexicon B_T
Output:
  a set A of alignments between tokens in S and T
Pseudo code:
A ← ∅
while alignments can still be produced
   and not maximum number of iterations do
 for j ← 1 to x
  if not_aligned(s_j)
  then
    C_S ← same_type(s_j, S)
    C_T ← same_type(s_j, T)
    if special_char(s_j)
    then A ← A ∪ {align_char(s_j, C_S, C_T)}
    else A ← A∪{align_word(s_j, C_S, C_T, B_S, B_T)}
  end_then
 end_for
end_while
if align_remained then
 for-each (s_j : t_i) ∈ A and (s_{j+k} : t_{i+l}) ∈ A
  with k, l > 1 do
  if (k = l)
  then
    for z ← 1 to (k − 1) A ← A ∪ {(s_{j+z} : t_{i+z})}
  end_then
  else
    M_S ← s_{j+1} + ... + s_{j+k−1}
    M_T ← t_{i+1} + ... + t_{i+l−1}
    A ← A ∪ {(M_S : M_T)}
  end_else
 end_for-each
end_then
return A
```

Figure 2: LIHLA algorithm (version 2.0)

```
function align_word
Input:
 the source word being aligned s_j
 a set of source candidate words C_S
 a set of target candidate words C_T
 a source–target bilingual lexicon B_S
 a target–source bilingual lexicon B_T
Output:
 an alignment (s_j : β) between s_j and β
Pseudo code:
1. if (∃ t_i ∈ C_T | t_i = s_j)
2. then return (s_j : t_i)
3. C'_T ← C_T ∩ look_for_translation(s_j, B_S)
4. if C'_T ≠ ∅
5. then
6.   continue ← true
7.   do
8.     t_i ← best_candidate(s_j, C'_T)
9.     if (t_i =NULL) then continue ← false #a
10.    else
11.      C'_S ← C_S∩ look_for_translation(t_i, B_T)
12.      s_k ← best_candidate(t_i, C'_S)
13.      if (s_k = s_j) or (s_k = NULL)
14.      then continue ←false #b
15.      else
16.        C''_T ← C_T∩ look_for_translation(s_k, B_S)
17.        t_l ← best_candidate(s_k, C''_T)
18.        if (t_i ≠ t_l) then continue ← false #c1
19.      end_else
20.    end_else
21.    if (continue = true) then remove(t_i, C'_T)
22.  until (continue = false) or (|C'_T| ≤ 0)
23.  M_S ← look_for_multiword(s_j, t_i, C_S, C_T)
24.  M_T ← look_for_multiword(t_i, s_j, C_T, C_S)
25.  return (M_S : M_T)
26. end_then
27. else
28.  C'_T ← look_for_cognate(s_j, C_T)
29.  if (C'_T ≠ ∅)
30.  then
31.    t_i ← best_cognate(C'_T)
32.    return (s_j : t_i)
33.  end_then
34. end_else
35. return (s_j : null)
```

Figure 3: Function `align_word`

be produced and a maximum number of iterations is not reached (in the experiments described in section 4 LIHLA has performed on average 4 iterations for each pair of parallel sentences). Furthermore, at the first iteration, all words with a frequency higher than a threshold are aligned only if they have a sure alignment to avoid erroneous alignments since all subsequent alignments are based on the previous ones.

In the last step (which is optional) LIHLA aligns the remaining unaligned source and target tokens between two pairs of already aligned tokens (in A) establishing several $1:1$ alignments when there are the same number of source and target tokens ($k = l$), or just one alignment involving all source and target tokens if they exist in different quantities. The decision of creating n $1:1$ alignments in spite of just one $n:n$ alignment when there is the same number of source and target tokens is due to the fact that a $1:1$ alignment is more likely to be found than a $n:n$.

To align words LIHLA uses the function `align_word` (see Figure 3). In this function some language-independent heuristics are applied to the words in C_S and the words in C_T aiming at finding the best possible lexical alignments between s_j (and maybe other words in C_S) and one or more words in C_T.

First of all, LIHLA priorizes a target word which is identical to s_j, to find exact matches, for instance, between proper names and numbers. If this word is found then a $1:1$ alignment is established (line 2); otherwise, LIHLA looks for possible translations in the source–target bilingual lexicon (B_S) and makes an intersection between them and the words in C_T.

In this intersection, if no candidate word identical to those in B_S is found in C_T then, for each word in B_S, LIHLA tries to look for cognates for this word in C_T using the longest common subsequence ratio (LCSR).[4] The cognates which have been found are added to C'_T and the search follows with the next word in B_S until all words in B_S and/or C_T have already been processed. By doing this, LIHLA can deal with small changes in possible translations such as different forms of the same verb, changes in gender and/or number of nouns, adjectives, and so on. Furthermore, if a \(`null`\) is found in B_S it is added to C'_T to allow an omission alignment to be set.

If there is at least one candidate word ($C'_T \neq \emptyset$) LIHLA looks for the best translation (t_i) for s_j among all the target candidates in C'_T following three assumptions (**a**, **b** and **c** illustrated graphically below and indicated in Figure 3 preceded by a character #) and considering as the best candidate the one pointed out by the bilingual lexicon or that at the best position in relation to s_j. LIHLA has a parameter (set by the user) to define which of these criteria (bilingual lexicon or position) will be used when looking for the best candidate word. In the following examples, "NULL" indicates an omission alignment.

a. $s_j \rightarrow \text{NULL}\ (t_i = \text{NULL})$

b. $s_j \rightarrow t_i \rightarrow s_k$

  b$_1$. $s_j \leftrightarrow t_i\ (s_j = s_k)$

  b$_2$. $s_j \rightarrow t_i \rightarrow \text{NULL}\ (s_k = \text{NULL})$

c. $s_j \rightarrow t_i \rightarrow s_k \rightarrow t_l$

  c$_1$. $s_j \rightarrow t_i \rightarrow s_k \rightarrow t_l\ (t_i \neq t_l)$

  c$_2$. $s_j \rightarrow t_i \leftrightarrow s_k\ (t_i = t_l)$

Following these assumptions, if the best candidate for s_j is NULL (case **a**) it is taken as its the best translation and the search terminates (line 9). However, if $t_i \neq$ NULL, LIHLA looks for the best candidate for t_i, s_k, and if $s_k = s_j$ (**b$_1$**) or $s_k =$ NULL (**b$_2$**) t_i is taken as the best translation for s_j (line 14). Otherwise, LIHLA looks for the best candidate for s_k, t_l, and if $t_l \neq t_i$ (**c$_1$**) t_i is taken as the best candidate for s_j (line 18). In the last case (**c$_2$**) t_i has a better bidirectional alignment with another word different from s_j and, in this case, LIHLA removes this word from candidate set (C'_T) (line 21) and repeats the search with another word until there is not candidate words available in C'_T or the best translation is found.

After finding the best translation t_i to s_j, LIHLA looks for source and target multiword units involving them. A source (target) multiword unit M_S (M_T), in this case, is composed of words in C_S (C_T) that occur immediately before and/or after the source word s_j (target word t_i). Furthermore, each word in M_S (M_T) has to be a possible translation of at least one word in M_T (M_S) and not a possible translation of other words in C_T (C_S). M_S and M_T will contain at least s_j and t_i, respectively, and a $n : m$ alignment is established between them (line 25) according to the number of source (n) and target (m) words in M_S and M_T.

LIHLA can also deal with target words that do not occur in the source–target bilingual lexicon (B_S) and the set of target candidate words (C_T) at the same time by looking for cognate words using the LCSR and setting a $1 : 1$ alignment between s_j and its best cognate (line 32). An omission alignment (indicated by the special word `null`) for s_j can also be established if no target candidate word that satisfies the heuristics is available (line 35).

Some examples of `pt`–`es` lexical alignments produced by LIHLA and a brief description about how they were found (with indications of lines on Figure 3) are given as follows.

- (*vida : vida*)
  A $1 : 1$ alignment between two identical words established at line 2.
- (*atmosfera : atmósfera*)
  A $1 : 1$ alignment between two cognate words, in the case that they were not found in the bilingual lexicons and, in this case, the LCSR had to be used

[4] The LCSR of two words is computed by dividing the length of their longest common subsequence by the length of the longer word. For example, the LCSR of `pt` word *alinhamento* and `es` word *alineamiento* is $\frac{10}{12} \simeq 0.83$ as their longest common subsequence is *a-l-i-n-a-m-e-n-t-o*.

to find the best target cognate word for the given source word. Here, the LCSR(*atmosfera*,*atmósfera*) $= 8/9 \cong 0.89$ is greater than LIHLA's default threshold of 0.75, hence a 1 : 1 alignment between these cognate words is set in line 32.

- (*apelo : llamado*)
  A 1 : 1 alignment found at line 14 since the best translation for the `pt` word *apelo* according to the bilingual lexicon is the `es` word *llamado* and vice-versa. In this case no multiword unit involving source and target words was found, so, a 1 : 1 alignment was established at line 25.
- (*dos : de+los*)
  A 1 : 2 alignment in which the best translation for the `pt` word *dos*, the `es` word *los* (see Figure 1), was found at line 14 and a target multiword unit was found at line 24 since *de* is also a possible translation for the `pt` word *dos*. Therefore, a 1 : 2 alignment was established at line 25.
- (*ou+seja : es+decir*)
  A 2 : 2 alignment in which the best translations for the `pt` word *seja* (in this context) are the `es` words *decir* and *es*, in this order. Since the `pt` word *ou* is a possible translation for both `es` words (*es* and *decir*), source and target multiword units were found at lines 23 and 24, respectively, and a 2 : 2 alignment was established at line 25.
- (*por :* `null`)
  A 1 : 0 alignment between the `pt` word *por* and the special word `null` established at line 35 since LIHLA did not find a target word t_i for the given source word s_j during the alignment process.
- (*tão : tan*)
  (bons : halagüeños)
  (*que : que*)
  A 1 : 1 alignment between `pt` word *bons* and `es` word *halagüeños* generated at the alignment of remained unaligned tokens since these words are between two previously aligned pairs (*tão:tan*) and (*que:que*) (see Figure 2).

4 Evaluation and results

Alignments produced by LIHLA were evaluated using the well-known precision, recall and alignment error rate (AER) metrics.

Let R be the set of reference alignments, A the set of alignments proposed by the method; $|A \cap' R|$ stands for the number of source and target tokens found in reference (R) and proposed (A) alignments at the same time, splitting the tokens in reference alignment between more than one proposed alignment if needed. Precision (1), recall (2) and AER (3) (the complement of the F-measure, a combination of precision and recall metrics) are shown below. In these experiments, AER was calculated considering all alignments as sure links[5] —as in (Wu and Wang, 2004)— and not as possible and sure links— as done in (Och and Ney, 2000).

$$\text{Precision} = \frac{|A \cap' R|}{|A|} \tag{1}$$

$$\text{Recall} = \frac{|A \cap' R|}{|R|} \tag{2}$$

$$\text{AER} = 1 - 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

The following sections describe some experiments carried out with LIHLA on `pt`–`es` (section 4.1) and `es`–`eu` (section 4.2) sentence-aligned parallel corpora.

4.1 Experiments with `pt-es`

The `pt`–`es` parallel corpus (`CorpusFAPESP`) used in these experiments is composed of articles from the online version of the Brazilian scientific magazine *Pesquisa FAPESP*.[6]

In the experiments described here, the 646 parallel articles on `CorpusFAPESP` were sentence-aligned by a version of Translation Corpus Aligner (TCA) (Hofland, 1996), but any other sentence alignment method proposed in the literature could be similarly used, such as the well-known method of Gale and Church (1991).[7]

The 15,192 aligned sentences composed of 798,641 words (381,656 in `pt` and 416,985

[5]A sure link is an unambiguous alignment while a possible link is an alignment that might or might not be established since there is not a straight correspondence between source and target tokens.

[6]The *Pesquisa FAPESP* magazine is available at `http://revistapesquisa.fapesp.br` with parallel texts written in Brazilian Portuguese (original), English (version) and Spanish (version).

[7]For more information on sentence alignment methods see PESA (Portuguese-English Sentence Alignment) project home-page: `http://www.nilc.icmc.usp.br/projects/PESA.html`.

in es) derived from this process were used to generate the bilingual lexicons (see section 2). The automatically aligned sentences were not post-processed for correction of misalignments because we believe that a few misaligned sentences will not significantly degrade the translation probabilities of all words in the corpus.

A manual reference alignment has been built with 591 aligned sentences (4%) randomly selected from the whole set. The 31,471 tokens (14,756 in pt and 16,719 in es) in the reference corpus were manually aligned by two bilingual annotators following the guidelines established in (Caseli, Scalco, and Nunes, 2005)[8] and the observed inter-annotator agreement rate of 95% indicates that the annotations are reasonably reliable. As expected, most of the alignments on the pt-es reference corpus as annotated by the human annotators are 1 : 1 (83.85%), but other categories such as omissions (6.60%) or those involving multiword units (9.55%) can also be found.

Table 1 shows the metric values per alignment category in pt–es parallel sentences. In this alignment process, the best candidate word (see algorithm on Figure 3) was chosen based on its position as opposed to the best translations found in the bilingual lexicons (as in es–eu sentences, see section 4.2) since the word order on pt and es sentences does not change much. The alignment of the remaining unaligned tokens was performed (see algorithm on Figure 2) since it has improved the overall performance. As can be noticed from Table 1, the worst AER is on the omission category (60.24%) and the AER for all categories except omissions (all – omissions) is 8.22%.

The current version of LIHLA has improved the results of the previous version (1.0) (Caseli, Nunes, and Forcada, accepted paper) in more than 4% AER in the alignment of multiword units. This improvement was already expected since in the current version LIHLA performs a more elaborate search for the best translation (see section 3) in which a $n : m$ alignment can be established in any iteration rather than just in the alignment of the remained unaligned tokens as done before. The improvement on the overall performance was of 1.43% AER.

Category	Precision	Recall	AER
1 : 1	82.71%	89.17%	14.18%
1 : 1-omissions	88.50%	92.33%	9.63%
omissions	33.35%	49.21%	60.24%
multiword	81.34%	70.53%	24.45%
all	86.73%	88.28%	12.50%
all - omissions	**92.44%**	**91.13%**	**8.22%**

Table 1: Evaluation of LIHLA per alignment category on pt–es parallel sentences

In order to compare LIHLA with another lexical aligner, the pt–es parallel texts were aligned by GIZA++ (Och and Ney, 2000) achieving better results as shown on Table 2.[9] However, LIHLA had a little better performance on multiword and omission categories. Furthermore, while LIHLA lasted 9 minutes (5 minutes to generate the lexicons and 4 minutes to align the test corpus) the training and alignment performed by GIZA++ lasted almost three times more (35 minutes).

Category	Precision	Recall	AER
1 : 1	84.22%	90.78%	12.62%
1 : 1-omissions	90.84%	94.93%	7.46%
omissions	31.45%	48.97%	61.70%
multiword	75.32%	71.25%	26.77%
all	90.53%	91.72%	8.88%
all - omissions	**97.38%**	**94.88%**	**3.89%**

Table 2: Evaluation of GIZA++ per alignment category on pt–es parallel sentences

4.2 Experiments with es-eu

The Spanish–Basque (es–eu) parallel corpus, in turn, is composed of 7,007 sentences from a translation memory avaliable on Internet,[10] that is, an already aligned set of parallel sentences. These aligned sentences were not post-processed for correction of possible misalignments for the same reason as pt–es sentences were not (see section 4.1).

A manual reference alignment has been built with 51 aligned sentences (0.73%) randomly selected from the whole set. The 1,715 tokens (1,012 in es and 703 in eu) in the reference corpus were manually aligned by a bilingual annotator.

[8]The guidelines defined in (Caseli, Scalco, and Nunes, 2005) are based on those defined for ARCADE (Véronis and Langlais, 2000) and Blinker (Melamed, 1998) projects.

[9]GIZA++ was trained using its default configuration and the same corpus used to generate the bilingual lexicons.

[10]The es–eu translation memory is available at `http://130.206.101.53/LegeBi/Botha/botha1992-1994.tmx`

The bilingual lexicons generated from the 7,007 es–eu parallel sentences composed of 277,664 words (163,096 in es and 114,568 in eu) were used to align the 591 sentences on test corpus and evaluation results are shown on Table 3.

In this alignment process, the best candidate word (see algorithm on Figure 3) was chosen based on the best translation according to the bilingual lexicons as opposed to their positions (as in pt–es sentences, see section 4.1) since the word order on es and eu sentences tends to change a lot; and the alignment of the remaining unaligned tokens was not performed for es–eu parallel sentences since it has not improved overall performance. As can be noticed from Table 3, the worst AER is on the omission category (85.60%) and the AER for all categories except omissions is 26.52%.

Category	Precision	Recall	AER
1 : 1	24.48%	81.21%	60.06%
1 : 1-omissions	47.07%	82.92%	39.95%
omissions	7.99%	72.58%	85.60%
multiword	55.13%	44.89%	50.51%
all	45.29%	65.58%	46.42%
all - omissions	**85.09%**	**64.66%**	**26.52%**

Table 3: Evaluation of LIHLA per alignment category on es–eu parallel sentences

The es–eu parallel texts were also aligned by GIZA++ achieving results worse than LIHLA's as shown on Table 4. Once again, LIHLA was faster lasting less than 2 minutes (1 minute and 17 seconds to generate the lexicons and 24 seconds to align the test corpus) while GIZA++ lasted 16 minutes to perform training and alignment.

Category	Precision	Recall	AER
1 : 1	23.14%	58.38%	66.86%
1 : 1-omissions	47.35%	57.30%	48.15%
omissions	7.17%	61.29%	87.16%
multiword	54.51%	40.77%	53.35%
all	40.30%	56.51%	52.95%
all - omissions	**80.05%**	**54.52%**	**35.14%**

Table 4: Evaluation of GIZA++ per alignment category on es–eu parallel sentences

5 *Concluding remarks*

This paper has presented a lexical alignment method, LIHLA, which aligns words and multiword units based on initial statistical word-to-word correspondences and language-independent heuristics. LIHLA has been evaluated on pt–es and es–eu parallel sentences and has achieved, respectively: 92.44% and 85.09% of precision, 91.13% and 64.66% of recall and 8.22% and 26.52% of AER. The results achieved by LIHLA on pt–es parallel sentences are worse (about 4% on AER) than those achieved by the state of the art statistical alignment system, GIZA++, but better on es–eu parallel sentences (about 8% on AER) .

So, based on these results it is possible to notice that LIHLA had a better performance than GIZA++ on some possible weak points of statistical alignment systems: non-consecutive phrasal information, long-range dependencies and multiword units — frequent problems, mainly in es–eu parallel sentences. The lower precision and recall values of LIHLA for multiword units alignment on es–eu parallel sentences, that is 55.13% and 44.89% respectively, may be explained if we consider the agglutinative nature of eu which leads to produce many alignments involving es multiword units.

Furthermore, LIHLA has some advantages when compared to other lexical alignment methods: it does not need to be trained for a new pair of languages (as in Och and Ney (2000))[11] and neither does it require preprocessing steps (apart from tokenization) to handle texts (as in Gómez Guinovart and Sacau Fontenla (2004)) or a large parallel corpus since it has achieved interesting results even with a very small amount of data.

Finally, the best contribution of LIHLA (apart from its speed) is that it is based on language-independent heuristics and, therefore, it can be applied to a new pair of languages without any modification (as has been done with pt–es and es–eu). As future work, we aim at investigating better ways to tokenize eu sentences in a language-independent way as well as using additional linguistic information (such as part-of-speech tags) to try to improve alignment results. As a long-term goal, LIHLA will be part of a system to learn transfer rules to machine translation from sequences of aligned words.

[11] The same bilingual lexicons can be used by LIHLA to align new sentence-aligned parallel texts in a much faster way.

References

Ayan, N. F., B. J. Dorr, and N. Habash. 2004. Multi-Align: Combining linguistic and statistical techniques to improve alignments for adaptable MT. In R. E. Frederking and K. B. Taylor, editors, *Proceedings of the 6th Conference of the AMTA (AMTA-2004)*, number 3265 in Lecture Notes in Artificial Inteligence (LNAI), pages 17–26. Springer-Verlag Berlin Heidelberg.

Carl, M. 2001. Inducing probabilistic invertible translation grammars from aligned texts. In *Proceedings of CoNLL-2001*, pages 145–151, Toulouse, France.

Caseli, H. M., M. G. V. Nunes, and M. L. Forcada. accepted paper. LIHLA: A lexical aligner based on language-independent heuristics. In *Proceedings of the V Encontro Nacional de Inteligência Artificial (ENIA05)*, São Leopoldo, RS, Brazil.

Caseli, H. M., M. A. G. Scalco, and M. G. V. Nunes. 2005. Manual para anotação de alinhamentos lexicais. Série de Relatórios do NILC NILC-TR-05-09, NILC, `http://www.nilc.icmc.usp.br/nilc/download/NILC-TR-05-09.zip`.

Gale, W. A. and K. W. Church. 1991. A program for aligning sentences in bilingual corpora. In *Proceedings of the 29th Annual Meeting of the ACL (ACL-1991)*, pages 177–184, Berkley.

Gale, W. A., K. W. Church, and D. Yarowsky. 1992. Using bilingual materials to develop word sense disambiguation methods. In *Proceedings of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI 1992)*, pages 101–112, Montreal, Canada, June.

Gómez Guinovart, X. and E. Sacau Fontenla. 2004. Métodos de optimización de la extracción de léxico bilingüe a partir de corpus paralelos. *Procesamiento del Lenguaje Natural*, 33:133–140.

Hiemstra, D. 1998. Multilingual domain modeling in Twenty-One: automatic creation of a bi-directional translation lexicon from a parallel corpus. In Peter Arno Coppen, Hans van Halteren, and Lisanne Teunissen, editors, *Proceedings of the 8th CLIN meeting*, pages 41–58.

Hofland, K. 1996. A program for aligning English and Norwegian sentences. In S. Hockey, N. Ide, and G. Perissinotto, editors, *Research in Humanities Computing*, pages 165–178, Oxford. Oxford University Press.

Melamed, I. D. 1998. Annotation style guide for the Blinker project, version 1.0.4. Technical Report 98-06, Institute for Research in Cognitive Science, University of Pennsylvania, Philadelphia, PA.

Menezes, A. and S. D. Richardson. 2001. A best-first alignment algorithm for automatic extraction of transfer mappings from bilingual corpora. In *Proceedings of the Workshop on Data-driven Machine Translation at 39th Annual Meeting of the ACL (ACL-2001)*, pages 39–46, Toulouse, France.

Och, F. J. and H. Ney. 2000. Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting of the ACL (ACL-2000)*, pages 440–447, Hong Kong, China, October.

Simões, A. M. and J. J. Almeida. 2003. NATools – A statistical word aligner workbench. *Processamiento del Lenguaje Natural*, 31:217–224.

Somers, H. 1999. Review article: Example-based machine translation. *Machine Translation*, 14(2):113–157.

Véronis, J. and P. Langlais. 2000. Evaluation of parallel text alignment systems: The ARCADE project. In J. Véronis, editor, *Parallel text processing: Alignment and use of translation corpora*, pages 369–388. Kluwer Academic Publishers.

Wu, H. and H. Wang. 2004. Improving domain-specific word alignment with a general bilingual corpus. In R. E. Frederking and K. B. Taylor, editors, *Proceedings of the 6th Conference of the AMTA (AMTA-2004)*, number 3265 in Lecture Notes in Artificial Inteligence (LNAI), pages 262–271. Springer-Verlag Berlin Heidelberg.