# Grammarless Bracketing in an Aligned Bilingual Corpus 

Jorge Kinoshita<br>Escola Politécnica da Universidade de São Paulo - Brazil<br>jkinoshi@pcs.usp.br

Abstract
We propose a simple grammarless procedure to extract phrasal translation examples from aligned parallel texts. It is based on the


#### Abstract

difference of word sequence in two languages.

\section*{1. Introduction}

We propose a simple procedure to extract some phrases from an aligned bilingual corpus observing differences in the sequence of words between two aligned sentences. For example: given two sentences: "The black cat runs" and its translation to Portuguese: "O gato preto corre" and the alignment: $(O=$ The, black $=$ preto, cat $=$ gato, runs $=$ corre $)$; then it is possible to detect an inversion between the words "black" and "cat" in the Portuguese translation. Based on this inversion we extract the phrase: "black cat" = "gato preto".


## 2. Procedure

The idea is to observe an inversion in the sequence of words (ex: black cat $=$ gato preto) in two sentences and extract three pairs of phrases: (P1-, P2+), (P1+, P2-), (P1, P2):

First phrase:
P1-: Phrase in Language 1 to the left of the inversion (P1-) (ex: black).

P2+: Phrase in Language 2 to the right of the inversion. (ex: preto).
$\mathrm{P} 2+$ is the translation of $\mathrm{P} 1-$.

## Second phrase:

P1-: Phrase in Language 1 to the right of the inversion (P1+) (ex: cat).

P2-: Phrase in Language 2 to the left of the inversion. (ex: gato).

P2- is the translation of $\mathrm{P} 1+$.

## Third phrase:

We can join P1- and P1+, as well as P2- and P2+ in the following way:

P1 = P1- , J, P1+
$\mathbf{P 2}=\mathrm{P} 2-, \mathrm{K}, \mathrm{P} 2+$
where:
J is a string that appears between P 1 - and $\mathrm{P} 1+$ and that is not translated to Language 2. Usually it is an empty string.

K is a string that appears between P2- and P2+ and that is not translated to Language 1 . Usually it is an empty string.

Two aligned sentences can be represented in a graph (Melamed, 1996). For instance, two statements:

## The black cat drinks the white milk <br> O gato preto bebe o leite branco

were aligned in figure 1. A word is assigned to each position of the horizontal axis and vertical axis.

A point in this graph is generically given by ( X , $\operatorname{Pos}(\mathrm{X})$ ). Language 1 (ex: English) refers to the first argument and Language 2 (ex: Portuguese) to the second.
$\operatorname{Pos}(\mathrm{X})$ is defined if there is some translation (word or string) to the word in $\mathrm{X} . \operatorname{Pos}(\mathrm{X})$ is undefined if there is no translation to the word in X . A point ( $\mathrm{X}, \mathrm{Y}$ ) exists when $\operatorname{Pos}(\mathrm{X})$ is defined.

Given two points $(\mathrm{X}, \mathrm{Y})$ and $(\mathrm{A}, \mathrm{B})$ where $\mathrm{X}<\mathrm{A}$, they are adjacent if there is no other point $(\mathrm{J}, \mathrm{K})$ for $\mathrm{X}<\mathrm{J}<\mathrm{A}$.

Usually, for two adjacent points ( $\mathrm{X}, \mathrm{Y}$ ) and ( $\mathrm{A}, \mathrm{B}$ ), if $\mathrm{X}<\mathrm{A}$ then $\mathrm{Y}<\mathrm{B}$.

An inversion occurs when $\mathrm{X}<\mathrm{A}$ and $\mathrm{Y}>\mathrm{B}$.
Now we can redefine $\mathrm{P} 1-, \mathrm{P} 2+, \mathrm{P} 1+, \mathrm{P} 2-$ in a more precise way. Theses phrases are next to an inversion between adjacent points $(X, Y)$ and $(A, B)$ where $X<A$.
$\mathbf{P 1}$ - is the sequence of adjacent words to the left of X with positions: X, X-1, ..., X-i, ... X-n where $\operatorname{Pos}(\mathrm{X}-\mathrm{i})>\mathrm{B}$ or $\operatorname{Pos}(X-i)$ is undefined.

P2+ Corresponds to $(\operatorname{Pos}(\mathrm{X}), \mathrm{J} 1, \operatorname{Pos}(\mathrm{X}-1), \mathrm{J} 2, \ldots$ $\operatorname{Pos}(\mathrm{X}-\mathrm{n})$ ) rearranged in crescent order. Ji corresponds to words or an empty string between $\operatorname{Pos}(\mathrm{X}-\mathrm{i})$ and $\operatorname{Pos}(\mathrm{X}-\mathrm{i}-1)$ with no translation.
$\mathbf{P 1 +}$ is the sequence of adjacent words to the right of A with positions: $\mathrm{A}, \mathrm{A}+1, \mathrm{~A}+2, \ldots, \mathrm{~A}+\mathrm{i}, \ldots \mathrm{A}+\mathrm{n}$ where $\operatorname{Pos}(\mathrm{A}+\mathrm{i})<\mathrm{Y}$ or $\operatorname{Pos}(\mathrm{A}+\mathrm{i})$ is undefined.

P2- corresponds to $(\operatorname{Pos}(\mathrm{A}), \mathrm{J} 1, \operatorname{Pos}(\mathrm{~A}+1), \mathrm{J} 2$, $\operatorname{Pos}(\mathrm{A}+2), \ldots \operatorname{Pos}(\mathrm{A}+\mathrm{n}))$ rearranged in crescent order. Ji corresponds to words or an empty string between $\operatorname{Pos}(\mathrm{A}+\mathrm{i})$ and $\operatorname{Pos}(\mathrm{A}+\mathrm{i}+1)$ with no translation.

In figure 1, we observe two inversions. The first one yields:

P1- = black; P2+ = preto
P1+ = cat; P2- = gato
P1 = black cat
$\mathrm{P} 2=$ gato preto
We decided to make an experiment with Japanese.
In figure 2, we show the alignment between an English and the Japanese sentence: "kuroi neko ha shiroi miruko wo nomimasu", yielding the following aligned phrases:
drinks = nomimasu
shiroi miruko $=$ white milk
drinks the white milk = shiroi muruko wo nomimasu.
In figure 3, we repeated the process for Portuguese and Japanese, yielding the following phrases.
gato preto $=$ kuroi neko
leite branco $=$ shiroi miruko
bebe o leite branco $=$ shiroi miruko wo nomimasu
We observe that this simple procedure detects some phrases in a robust way.

Figure 1. Graphical representation of the English Portuguese alignment


Figure 2. Representation of the English Japanese alignment


Figure 3. Representation of the Portuguese Japanese alignment


## 3. Application

James Strong aligned two versions of the New Testment (in the Bible): in Greek and in English. He numbered each Greek word (each stem with a number) and added the same numbers to the English version (Bible, 1998). For instance, Jhon 1:1 in Greek and English was annotated as:

John 1:1 en <1722> arch <746> hn <2258> (5713) o <3588> logoj <3056> kai <2532> o <3588> logoj <3056> hn <2258> (5713) proj <4314> ton <3588> qeon <2316> kai <2532> qeoj <2316> hn <2258> (5713) o <3588> logoj <3056>

John 1:1 II In <1722> the beginning <746> was <2258> (5713) the Word <3056>, and <2532> the Word <3056> was <2258> (5713) with <4314> God <2316>, and <2532> the Word <3056> was <2258> (5713) God <2316>.

There are many other versions in other languages annotated in the same way. For instance, the version in Portuguese is

João 1:1 If No <1722> princípio <746> era <2258> (5713) o Verbo <3056>, e <2532> o Verbo <3056> estava <2258> (5713) com <4314> Deus <2316>, e <2532> o Verbo <3056> era <2258> (5713) Deus <2316>.

We created a Perl script to extract some phrases from texts given in the Strong's annotation. This script:

1) aligns two texts
2) searches for the inversions in the graph and displays the aligned phrases.
The alignment is made according to the numbers. Our idea is very simple: to align the first occurrence of a number in Greek to the first occurrence of the same number in English, and the second occurrence in Greek to the second occurrence in English, etc. This simple idea does not always work correctly. An example of an error is given in John 1:3.

Applying our procedure to John 1:1, we got the following phrases:

P1-, P2+: < qeoj :: God >
P1+, P2-: < hn o logoj :: the Word was >
P1, P2: qeoj hn o logoj :: the Word was God
Applying it until John 1:5, we have:
1- John 1:2 outoj 〈3778> hn <2258> (5713) en <1722> arch <746> proj <4314> ton <3588> qeon <2316>

2- John 1:2 The same <3778> was <2258> (5713) in <1722> the beginning <746> with <4314> God <2316>.

Note: There was no inversion in these verses.

1- John 1:3 panta <3956> di <1223> autou <846> egeneto <1096> (5633) kai <2532> cwrij <5565> autou <846> egeneto <1096> (5633) oude <3761> en <1520> o <3739> gegonen <1096> (5754)

2- John 1:3 All things <3956> were made <1096> (5633) by <1223> him <846>; and <2532> without <5565> him <846> was <1096> <0> not <3761> any thing <1520> made <1096> (5633) that <3739> was made <1096> (5754).
< o :: that > < gegonen :: made >
o gegonen :: made that
< di autou :: by him > < egeneto :: were made >
di autou egeneto :: were made by him
Note: The phrase "made that" is wrong due to the alignment phase. The number 1096 appeared 3 times in Greek and 4 times in English. The third occurrence in Greek was wrongly assigned to the third occurrence in English.

1- John 1:4 en <1722> autw <846> zwh <2222> hn <2258> (5713) kai <2532> h <3588> zwh <2222> hn <2258> (5713) to <3588> fwj <5457> twn <3588> anqrwpwn <444>

2- John 1:4 In <1722> him <846> was <2258> (5713) life <2222>; and <2532> the life <2222> was <2258> (5713) the light <5457> of men <444>.
< zwh :: life > < hn :: was >
zwh hn :: was life

1- John 1:5 kai <2532> to <3588> fwj <5457> en <1722> th <3588> skotia <4653> fainei <5316> (5719) kai <2532> h <3588> skotia <4653> auto <846> ou <3756> katelaben <2638> (5627)

2- John 1:5 II And <2532> the light <5457> shineth <5316> (5719) in <1722> darkness <4653>; and <2532> the darkness <4653> comprehended <2638> (5627) it <846> not <3756>.
<en th skotia :: in darkness > < fainei :: shineth > en th skotia fainei :: shineth in darkness
< auto ou :: it not > < katelaben :: comprehended > auto ou katelaben :: comprehended it not

The same procedure is now applied to Greek and Portuguese.

1- John 1:1 en <1722> arch <746> hn <2258> (5713) o <3588> logoj <3056> kai <2532> o <3588> logoj <3056> hn <2258> (5713) proj <4314> ton <3588> qeon <2316> kai <2532> qeoj <2316> hn <2258> (5713) o <3588> logoj <3056>

2- João 1:1 II No <1722> princípio <746> era <2258> (5713) o Verbo <3056>, e <2532> o Verbo <3056> estava
<2258> (5713) com <4314> Deus <2316>, e <2532> o Verbo <3056> era <2258> (5713) Deus <2316>.
< qeoj :: Deus > < hn o logoj :: o Verbo era > qeoj hn o logoj :: o Verbo era Deus

1- John 1:2 outoj 〈3778> hn <2258> (5713) en <1722> arch <746> proj <4314> ton <3588> qeon <2316>

2- João 1:2 Ele <3778> estava <2258> (5713) no <1722> princípio <746> com <4314> Deus <2316>.

1- John 1:3 panta <3956> di <1223> autou <846> egeneto <1096> (5633) kai <2532> cwrij <5565> autou <846> egeneto <1096> (5633) oude <3761> en <1520> o <3739> gegonen <1096> (5754)

2- João 1:3 Todas as coisas <3956> foram feitas <1096> (5633) por intermédio <1223> dele <846>, e <2532>, sem <5565> ele <846>, nada <3761> do que <1520> foi feito <1096> (5633) se fez <1096> (5754).
< di autou :: por intermédio dele > < egeneto :: foram feitas >
di autou egeneto $::$ foram feitas por intermédio dele
< egeneto :: foi feito > < oude en :: nada do que >
egeneto oude en :: nada do que foi feito

1- John 1:4 en <1722> autw <846> zwh <2222> hn <2258> (5713) kai <2532> h <3588> zwh <2222> hn <2258> (5713) to <3588> fwj <5457> twn <3588> anqrwpwn <444>

2- João 1:4 A vida <2222> estava <2258> (5713) nele <1722> <846> e <2532> a vida <2222> era <2258> (5713) a luz <5457> dos homens <444>.
<en autw :: nele > < zwh hn :: A vida estava > en autw zwh hn :: A vida estava nele

1- John 1:5 kai <2532> to <3588> fwj <5457> en <1722> th <3588> skotia <4653> fainei <5316> (5719) kai <2532> h <3588> skotia <4653> auto <846> ou <3756> katelaben <2638> (5627)

2- João 1:5 ๆI A luz <5457> resplandece <5316> (5719) nas <1722> trevas <4653>, e <2532> as trevas <4653> não <3756> prevaleceram <2638> (5627) contra ela <846>.
<en th skotia :: nas trevas > < fainei :: resplandece
$>$
en th skotia fainei :: resplandece nas trevas
< auto :: contra ela > < ou katelaben :: não prevaleceram >
auto ou katelaben :: não prevaleceram contra ela
< kai to :: e > < fwj en th skotia fainei :: A luz resplandece nas trevas >
kai to fwj en th skotia fainei :: A luz resplandece nas trevas e

## 4. Conclusion

This procedure must be applied in bilingual prealigned corpus. If the bilingual corpus is not aligned, it is possible to use alignment methods such as (Melamed, 1996), (Melamed, 1997).
(Wu, 1995 ${ }^{\text {a }}$ ), (Wu, 1095b) proposes a method to extract phrases from bilingual corpus. In ( $\mathrm{Wu}, 1995^{\text {a }}$ ), words do not have to be aligned. His method uses a very simple grammar (but very ambiguous) and a very precise dictionary. Analysing the procedure in (Wu, 1995), we concluded that his method works better when there are inversions in word sequence. This observation encouraged us to search for another way of phrase extraction focused in word sequence inversions.

When dealing with multilingual texts, it is possible to extract more phrases. For instance, in verse John 1:4, the phrase "In him, was life" can not be extracted when comparing just Greek and English, but it can be extracted when observing the Portuguese phrase that was given in a different order: "A vida estava nele".

Some weak points to our ideas are:

1) There are many phrases that can not be collected because our approach is based only in order inversion. For instance: in John 1:2 there was not any phrase extracted.
2) The quality of the phrase extracted (it is not numerically measured) is doubtful. For instance: in John 1:4, "a vida estava" was considered a phrase but a normal parser would never get the phrase "a vida estava" from the statement "a vida estava nele".

We hope that depending on the application, these weak points are not relevant. We intend to apply these ideas in a Example Based Machine Translation system. Now we are trying to solve the question "how can we foresee an inversion in the translated sentence after collecting many phrases?"
acknowledgement: this research is supported by FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo; www.fapesp.br).

## 5. References

Bible, 1998. Bíblia Online 2.0. CD-ROM. Sociedade Bíblica do Brasil. www.sbb.org.br
Melamed, I. Dan, 1996. A Geometric Approach to Mapping Bitext Correspondence, IRCS Technical Report\#96-22, a revised version of the paper presented at the First Conference on Empirical Methods in Natural Language Processing (EMNLP'96), Philadelphia, PA.
Melamed, Dan, 1997. A Portable Algorithm for Mapping Bitext Correspondence, 35th Conference of the Association for Computational Linguistics (ACL'97), Madrid, Spain.
Wu, Dekai, 1995a. Grammarless extraction of phrasal translation examples from parallel texts. TMI-95, Sixth International Conference on Theoretical and

Methodological Issues in Machine Translation, v2, 354372. Leuven, Belgium.

Wu, Dekai, 1995b. An algorithm for simultaneously bracketing parallel texts by aligning words. ACL-95: 33rd Annual Meeting of the Assoc. for Computational Linguistics, 244-251. Cambridge, MA.

