*O. Scrivner, T. Gilmanov*

# SWIFT ALIGNER: A TOOL FOR THE VISUALIZATION AND CORRECTION OF WORD ALIGNMENT AND FOR CROSS LANGUAGE TRANSFER

**Abstract**. It is well known that parallel corpora are valuable linguistic resources. One of the benefits of such corpora is that they allow for the building an annotated corpus for resource-poor languages via cross-language transfer. That is, given accurate alignment between a word from a source language and its equivalent in a target language, some linguistic information, such as part-of-speech tags or syntactic annotation, can be projected to the aligned word. While there are several state-of-the-art word-aligners, such as GIZA++ and Berkeley, there is no simple visual tool that would enable correcting and editing aligned corpora of different formats. We have developed Swift Aligner, a free, open source, portable software written in Java that facilitates the visual representation of corpora, the correction of alignment and finally the transfer of morphological information and syntactic relations from an annotated source language into an unannotated target language, by means of word-alignment. In addition, this tool is flexible, as it imports corpora in various formats, such as GIZA++, Berkeley, and LIHLA. Finally, we have also shown that by using cross-language transfer, we would need only an estimated 30% of correction by human annotator, compared to 100% of manual annotation.

## 1. Introduction

In recent years parallel translated corpora have gained research attention as useful resource-light tools for building annotated corpora. It is well known that manual annotation is a costly and time consuming process and that only few languages have linguistically annotated large corpora. On the other hand, there is a great need for annotated data in many areas of research, such as machine translation and language studies, among others. Cross-language transfer has been introduced as a strategy for making use of the existing resource-rich languages in order to annotate resource-poor languages. Given the accurate alignment between a word from a source language and its equivalent in a target language, some linguistic information, such as part-of-speech tags or syntactic annotation, can be projected to the aligned word. Several studies have demonstrated the feasibility of such transfer not only for structurally similar languages, such as English and French, but also for structurally distant languages, such as English and Vietnamese, English and Hindi, and others. It has also been shown that morpho-syntactic transfer requires high precision of alignment between languages. While there are several stateof-art word-aligners, such as GIZA++ and Berkeley, there is no simple visual tool that enables the correcting and editing of aligned corpora of different formats. Thus, the goal of the present work is to develop a portable visual application for editing word-aligned parallel corpora produced by different aligners, and for providing initial steps for building resource-light corpora via cross-language transfer.

The remainder of the paper is organized as follows: section 2 reviews the concept of word-alignment and cross-language transfer in parallel corpora. Section 3 describes the architecture of Swift Aligner and demonstrates how the tool works. Finally, the conclusions and directions for further work are presented in Section 4.

## 2. Parallel Corpus

### *2.1. Word Alignment*

Traditionally, parallel corpus refers to a text written in two or more languages: the original text and its translation. Each corpus may consist of several hierarchical levels of alignment, namely paragraph, sentence and word. At word level the most common type of alignment is one-toone, where one word from a source language corresponds to only one word in a target language (see Fig. 1). The remaining types include one-to-many, many-to-one or unaligned words.

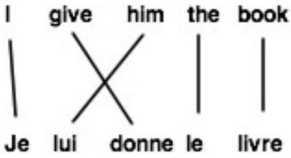

*Fig. 1.* One-to-one word alignment (English-French)

It is important to note that the accuracy of automatic alignment depends on many factors, such as the size of training data, genre, and quality of translation, among others. While there are many available aligners, not all of them are supported with visual tools for the manual editing and correcting of corpora. In addition, each alignment editor is restricted to a specific format of word alignment, for example Cairo (Giza) or Visual LIHLA.

### *2.2. Cross-Language Transfer*

The idea of transfer from one language into another by means of word-alignment is not new. Yarowsky and Ngai describe an experiment for morphological transfer from English into French. The results show that morphological annotation can be effectively transferred using a small core tagset. While the core tagset is limited to essential morphological information, for example, N (noun), J (adjective), some language specific morphological information can be resolved later by using a morphological analyzer specific for each language. The experiments also demonstrate that the accuracy of annotation can be further increased from 76% to 85% by simply correcting word alignment.

Similarly, syntactic information can be induced by means of word alignment. It has been shown in the literature that syntactic dependencies are more suitable than syntactic constituents for cross-language syntactic transfer. Fig. 2 illustrates the underlined hypothesis for many models of syntactic dependency transfer, namely the principle of direct correspondence assumption (DCA), that is, that syntactic relations between nodes of the source language hold for the corresponding aligned nodes of the target language. In Fig. 2 f1 and f5 are two nodes of a target language that correspond to two nodes *I* and *got* in a source language (English). Based on the DCA assumption, the syntactic relation of the type verb-subject *I got* can be transferred from the source language to nodes f1 and f5 in the target language via their alignment.

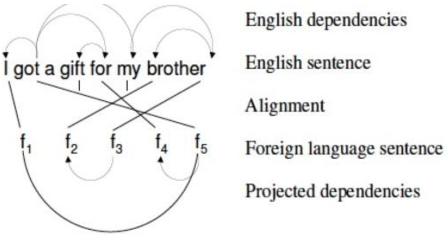
English dependencies
I got a gift for my brother
English sentence
Alignment
$f_1$ $f_2$ $f_3$ $f_4$ $f_5$
Foreign language sentence
Projected dependencies

*Fig. 2.* Dependency projections via word-alignment

To test the feasibility of projecting English syntactic information, Hwa et al. performed several experiments on languages with different word order, namely English, Spanish, and Chinese. While the direct projection from English yielded low unlabeled dependency F-scores (37% for Chinese and 38% for Spanish), the errors mostly occurred in cases where the target language required more projections than the source language (English). For instance, Chinese aspectual markers are not realized as a separate projection in English; therefore, they are left unlabeled during the transfer. The application of simple language-specific transformation, however, increased the accuracy to 68% for Chinese and 72% for Spanish transfer.

## 3. Swift Aligner

Because, as we have seen from the previous section, the accuracy of word-alignment is very important for cross-language transfer, the first goal of our tool is to facilitate the editing of a corpus. Swift Aligner currently supports the following input formats: a) Giza (1); b) Berkeley (2); and c) LIHLA (3):

1) Je lui donne le livre .
NULL ({}) I ({ 1 })give ({ 3 })him ({2}) the ({ 4 })book ({ 5 }) . ({ 6 })

2) a. I give him the book .
b. Je lui donne le livre .

c. 1-1 2-3 3-2 4-4 5-5 6-6

3) a. <s snum=1>I give him the book .</s>
b. <s snum=1>Je lui donne le livre .</s>
c. <s snum=1>I:1 give:3 him:2 the:4 book:5 .:6</s>

Giza format in example 1 is an output file from GIZA++ aligner, where each word from a source file (second line) is assigned an alignment number, namely the relative position of its translation in a target language (first line). For instance, the source word *him* is aligned to a second token in a target sentence, which corresponds to the word *lui*. Berkeley format in example 2 is an output of Berkeley aligner. This format consists of three separate text files. One file contains a text from a source language, one sentence per line (2a), the second file contains a text from a target language, one sentence per line (2b), and the third file represents the word-alignment, where the first number is the source word position and the second number is the target word position. For example, 2-3 in example 2c describes an alignment between the second token in a source language (English) *give* and the third token, namely its translation equivalent, in a target language (French) *donne*. LIHLA is a word alignment produced by a lexical aligner LIHLA. This format consists of three files, similar to Berkeley: source text (3a), target text (3b), and alignment text (3c).

In a graphical user interface, the sentence pairs are displayed horizontally with a sentence from a source language placed on the top and a sentence from a target language placed on the bottom. Aligned words are connected with vertical lines. A Swift Aligner user can mouse-drag a line between two aligned words to correct the alignment. After the required corrections are performed visually, the edited version of an aligned corpus can be imported to the text representation. By default it is saved to the xml file produced during the initial conversion. Fig. 3 shows the GUI of a Swift Aligner. The drop down menus provide a number of utilities, such as various format import, file format conversion, and a few help topics relevant to the process of using the alignment software.

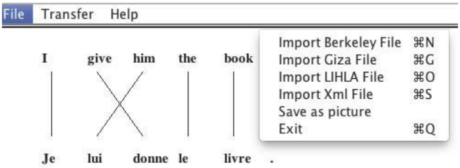

*Fig. 3* Swift Aligner GUI

At this stage our tool can process only one-to-one word alignment for cross-language transfer task. Morpho-syntactic transfer is performed via annotated source file in TNT format: each word and its pos tag are per line. Part-of-speech tags from a source language are copied to their aligned equivalent in a target language. By default, non-aligned words receive a tag «NA». The morphologically annotated target language is further saved as a text file in TNT format. Syntactic information is processed by means of an annotated source file in CONLL format. Syntactic dependencies among words in the source language are projected by finding the corresponding head elements of dependencies in the target language. The non-aligned tokens are assigned the tag «NA» and are attached to the root of a sentence. Table 1 illustrates the transfer method for the sentence pair *He loved the book* and *il aimait le livre* as an example.

*Table 1.* Dependency relation
from English (source) into French (target)

| **Relation** | **Head(source)** | **Dep(source)** | **Head(target)** | **Dep(target)** |
|---|---|---|---|---|
| **verb-subj** | loved | he | aimait | il |
| **verb-obj** | loved | book | aimait | livre |
| **noun-det** | book | the | livre | le |

As it is shown in Table 1, syntactic relations between the head and its dependent in a source language are carried over to a target language through their alignment. For example, the relation between *loved* and *he* is described as verb-subject relation, where *loved* is a head (verb) and *he* is its dependent (subject). This relation is transferred to their aligned word pairs *aimait* and *il*.

Finally, to illustrate and evaluate our tool we conducted two experiments on part of speech and syntactic transfer. The data for the experiment came from a small annotated parallel corpus of English-Latin. This corpus has one-to-one word alignment and is morpho-syntactically annotated using the same tagset (PennTreebank) and dependency relations in both bitexts. First, we randomly selected and manually corrected word-alignment in 10 sentences. The transfer of part-of-speech tags from English to Latin yielded 71% accuracy. We further projected syntactic dependencies from English into Latin and evaluated our results. The results yielded a 67% label accuracy score with punctuation and 77% without punctuation.

## 4. Conclusion and Future Work

This paper has introduced a multi-functional tool for parallel bilingual corpora. We felt a need for a simple visual tool that not only allows for editing word-alignment but also facilitates the building of new annotated corpora by means of cross-language transfer. We have shown that by using resource-light methods, we need human annotators to correct only an estimated 30%, compared to 100% of manual annotation. In addition, this tool is flexible to importing corpora in various formats, such as GIZA++, Berkeley, and LIHLA.

Presently, Swift Aligner is able for processing one-to-one word alignment. We continue its development in order to incorporate multiple types of alignments: one-to-many, many-to-one. Finally, we plan to increase the available formats for importing. The up-to-date version of the code and executables can be obtained by contacting the authors directly.

## 5. Acknowledgements

The authors would like to thank Dr. Adrian German for a valuable contribution to the project and Dr. Sandra Kübler for motivating discussions and help in writing this paper.