RESEARCH ARTICLE OPEN ACCESS

The First Parallel Multi-lingual Corpus of Amazigh

Fadoua ATAA ALLAH*, Nina MIFTAH**
**(CEISIC, Royal Institute of Amazigh Culture, Rabat, Morocco*
*** (Mohammed V University, Rabat, Morocco*
Corresponding Auther: Fadoua ATAA ALLAH

ABSTRACT
Amazigh language is the autochthon language of North Africa. However, until 2011 that it became a constitutionally official language in Morocco, after years of persecution. Amazigh language is still considered as one of the under resourced languages. This paper presents the development of a multi-lingual parallel corpus (Amazigh-English-French) aligned on the sentence level. The objective is to be used in linguistic research, teaching, and natural language processing application, primarily machine translation. The paper discusses this aspect, and presents the corpus encoding. A multi-lingual parallel corpus, which brings together Amazigh, English and French, is a new resource for the NLP community that completes the present panorama of parallel corpora. To the best of our knowledge, this corpus is the first Amazigh-English-French multi-lingual parallel corpus. The built corpus is sentence aligned, including 31864 sentences. The alignment was done automatically, while the evaluation was done manually. The evaluation results are satisfactory, achieving more than 90%.
Keywords –Amazigh, Encoding, Multi-lingual Corpus, Sentence-Alignment.

Date of Submission: 14-06-2018 Date of acceptance: 29-06-2018

I. INTRODUCTION

Amazigh language, as one of the Afro-asiatic languages, poses many challenges on natural language processing [1]. The Moroccan Amazigh language is considered as a prominent constituent of the Moroccan culture, due to its richness and originality. Morphology based on word formation process of roots and patterns, and the lack of linguistic corpora make computational approaches to Amazigh language challenging [1]. In the perspective to build an automatic translation system for the Amazigh language, it proves essential to equip the Amazigh language with rich and usable corpora essential for automatic processing. In this context, we are interested, in this work, to study multi-lingual corpus used in Machine Translation (MT).

Corpora are used in different machine translation approaches: expert approaches based on linguistic rules, statistical approaches, and hybrid ones. Theoretically, a corpus is able to represent potentially unlimited selections of texts [2]. Typically, corpora are characterized by the nature of the language. A corpus can represent one language or more, as it can contain text or multimedia. There are two types of multi-lingual corpora: parallel corpora and comparable ones. The parallel corpora are composed of pairs of documents that have mutual translations, while comparable corpora are composed of materials with common traits such as gender, time, field, etc., without translations.

In this paper, we are interested in contributing in a publicly available Amazigh corpus of written text to the natural language processing community. The rest of this paper is organized as follows: We devote Section 2 to the Amazigh language. While in Section 3, we present the multi-lingual corpora; furthermore, we introduce parallel and comparable ones. We turn, then, Section 4 to aligners, presenting traditional alignment methods and available tools in the literature. Section 5 introduces corpus standard encoding. In Section 6, we discuss our contribution, and we present our corpus format. Finally, in the conclusion section, we draw the conclusions, and present potential future research directions.

II. A NOTE ON THE AMAZIGH LANGUAGE

In linguistic terms, Amazigh language is characterized by the proliferation of dialects due to historical, geographical and sociolinguistic factors. In Morocco, this language is divided into three main regional varieties, depending on the area and the communities: Tarifite in the North, Tamazight in the Central Morocco and the Southeast, while Tachelhite in the South-West and the High Atlas.

Amazigh was essentially a language with an oral tradition. In order to preserve this language, it transited from orality to literacy. In Morocco, three competing graphic systems are used for writing Amazigh. These competing systems are:

- Arabic, widely, used for religion and rural poetry writing;
- Latin, supported by the International Phonetic Alphabet (IPA), used particularly by berberists since early works of missionaries;

- Tifinaghe, the ancestral writing system, which has been preserved by Touareg [3].

The Royal Institute of Amazigh Culture (IRCAM) has engaged to achieve a standardization process for Amazigh language in order to introduce it in the school system, as well as in the media. Over the last 16 years, IRCAM has published about 400 books related to the Amazigh language and culture, a number, which exceed the whole amount of Amazigh publications in the 20th century. However, in Natural Language Processing (NLP) terms, Amazigh, like most non-European languages, still suffers from the scarcity of language processing tools and resources.

With the aim to integrate Amazigh into information and communication technologies, many scientific studies have been undertaken, in different fields, to improve its current situation:

- Graphic encoding [4].
- Optical character recognition [5] [6].
- Speech processing [7] [8].
- Basic NLP tools, such as converters [9] [10] and concordancer [11].
- Textual Corpora: raw corpus [12], tagged corpus [13] and bi-lingual corpus [14].
- Enhanced NLP tools, such as morphological analyzer/generator [15] [16], named entities extractor [17], and machine translation system [18].

From the overview of the existed studies, it is remarked that the Amazigh language still suffer from the lack of multi-lingual corpus. To help overcome this situation, we have been interested in building the first Amazigh multi-lingual parallel corpus.

III. MULTI-LINGUAL CORPUS

To define corpora types, we can use different criteria (the number of languages involved, the content of the corpus and the form of the corpus). A corpus can be classified as mono-lingual or multi-lingual in regard to the number of treated languages. Mono-lingual corpus contains text in one language, while multi-lingual one is composed of documents in several languages. These documents are selected according to the same criteria [19].

The remaining of this section is devoted to the description of various multi-lingual corpora and presentation of the characteristics of each one. In the literature, we distinguish between two types of multi-lingual corpora, which are the parallel corpus and the comparable corpus.

3.1 Parallel corpus

Parallel texts exist since a long time, even before Naturel Language Processing (NLP) startups. The anthropological research found, in North of Africa, objects (stelae) containing parallel texts, including Punic and Amazigh with inscriptions in Tifinaghe.

In fact, a parallel corpus consists of a set of document pairs such that, for a couple, one of the documents is the translation of the other. The most known of multi-lingual parallel corpus are the Europarl corpus [20] and the Hansards corpus[1]. In a parallel corpus, textual information, only, is not sufficient for most NLP applications. The mapping between equivalent parts to a sufficient level of granularity - such paragraphs or sentences - is essential to implement useful applications. This mapping information is called alignment.

In 1980, parallel texts began systematically to be exploited in the context of language processing. Among the parallel corpus reference, we can mention [21]:

- Hansards corpus created in 2001. This corpus is composed of English and French texts from the Canadian Parliament debates transcriptions from 1970 to 1988.
- Europral corpus, which brings together the European Parliament texts in 11 languages, with more than 20 million words per language.
- Bible corpus built in 1999 that was hold in 13 languages.
- JRC-ACQUIS corpus, which is composed of European Union texts acquired in 20 official languages of the European Union (EU).

The parallel corpus is very important for building robust bilingual dictionaries, machine translation and cross-language information systems. Nevertheless, these corpora, including specialized areas, are inherently scarce resources, and difficult to build.

3.2 Comparable corpus

Because of the various problems affecting the parallel corpus, specially their rarity, it is necessary to find an alternative approach. Thus, in 1995, methods for aligning non-parallel corpus (noisy parallel corpus and comparable corpus) were appeared [22]. The research community considers that comparable corpus contains documents, which are not translations of each other, but “closely linked by the same content” to “different levels of parallelism, such as words, strings of words, phrases, etc.” [22].

Bowker and Pearson consider that “[...] comparable corpora consist of sets of text in different languages that are not translations of each other. We use the word ‘comparable’ to indicate that the text in different languages have been selected because they have some characteristics or features in common” [23]. There are two sets of texts selected

[1] U. Germann (ed.), Aligned Hansards of the 36th Parliament of Canada - Release 2001-1a. http://www.isi.edu/natural-language/download/hansard/ (Viewed 06/10/2017)

by the same criteria, the difference is only the language, but there are not mutual translations. The comparable term indicates that these texts share certain characteristics or certain traits, such as, in the majority of the time, the subject, the time or the degree of technicality, etc.

Two characteristics of comparable corpus are specifically discussed in the literature of bilingual alignment: the size and the degree of comparability of a comparable corpus [24].

The corpus size: The size of a corpus is a balance between the purpose of the study and the time that was available. A corpus does not have to be huge, what matters is that it must properly represent the target language. Obviously, a small corpus does not, sufficiently, represent all the linguistic features of a language.

The corpus comparability: To quantify the extent to which texts of a corpus are comparable, we rely on the notion of its degree of comparability. A corpus has a high degree of comparability if it contains texts of many characteristics (eg. release dates, areas, themes, genres, etc.). The comparability of a corpus varies according to the set of selected features or criteria.

3.3 Comparable corpus versus parallel corpus

Depending of the task to be performed, one of the corpus types may be called for. On table 1, we compare between parallel and comparable corpora.

TABLE 1: COMPARISON BETWEEN MULTILINGUAL CORPORA

	Comparable corpus	Parallel corpus
1	Words have multiple meanings in the same corpus.	A word has only one meaning in the corpus.
2	Multiple translations can be associated with a word.	A single translation is associated with each word.
3	The translations may not exist in the target document.	There are no missing translations between a source and a target corpus.
4	The positions and frequencies of words are incomparable.	The positions and frequencies of words in translation relationship are comparable.

The ideal corpus and the fundamental resource for translator and the most machine translation systems is a parallel one, which source text is aligned with their translations in target languages. This type of corpus can provide necessary terminology and phraseology for translation, as well as examples of translation strategies [25].

Generally, the parallel corpora structure is based on an alignment at the sentence's level. The difficulty of having this resource, especially in specialized fields, made emerge comparable corpus, which became an alternative to the parallel corpus. The ability of comparable corpus, to capture closer information of the translation framework can avoid literal translation errors, has led some researchers to jointly exploit the parallel and comparable corpora to take advantages of both of these resources [26].

However, the comparable corpora are not ideal for under-resourced languages. Since it is difficult to guarantee the comparability of texts in different languages. Furthermore, a comparable corpus gives a picture less clear of the correspondences of lexical items than a parallel corpus does [27].

The parallel corpus is one of the most valuable resources in the development of several Natural Language Processing (NLP) applications. Essentially, it provides indispensable training data for the statistical machine translation systems. Moreover, it is also useful for other applications like multilingual text retrieval [28] and automatic bilingual lexical acquisition [29].

IV. ALIGNMENT OF PARALLEL CORPUS

Text alignment is an important process of different machine translation systems. According to Rappazzo alignment can be defined as: "Aligning consists in finding correspondences, in parallel bilingual corpora, between textual segments that are translation equivalents".

Various alignment levels can be set in a parallel corpus. Each level corresponds to a structural level in the corpus. For example, for a literary work and its translation, we can define an alignment between chapters, between paragraphs, or sentences of these paragraphs, and even, between words inside sentences [30].

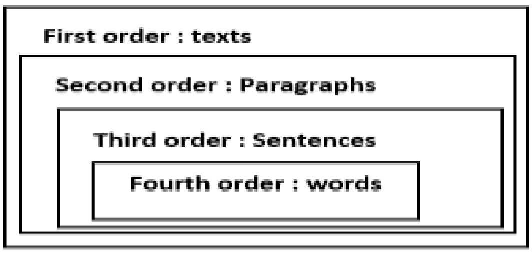

Fig. 1. Alignment levels

4.1 Alignment by sentences

Sentence alignment is the process to extract a parallel sentence pairs that are translations of each other, from parallel documents. We present, in this section, the main sentence alignment methods.

In 1984, the first automatic alignment method named "The lexical anchor" was appeared. It is based on the distribution of words, without any use of outside information source. It is, just, based on observing word co-occurrence within probably corresponding areas [31].

In 1991, a new alignment method has emerged. It is a purely statistical method, based on the length of sentences [32]. The idea is that long sentences in the source text tend to be translated into long sentences in the target text, and the short sentences are translated into short ones.

Later in 1992, the cognates' algorithm was appeared. Cognates are pairs of tokens of different languages that share "obvious" phonological or orthographic and semantic properties, with the result that they are likely to be used as mutual translations [33].

4.2 Alignment by words

This type of alignment is still a challenge to overcome. However, in the context of the sentence alignment, the alignment of the words is not the first goal. Grammatical words are sources of problems: their correspondence is even less necessary as between full words [34]. Nevertheless, it is not possible to ignore them completely as they can be part of an expression to spot. Complex elements such as compound words or phrases, which are widely presents in sentences, also pose critical problems. For example, alignment or extraction of glossaries, theoretically, consists of two tasks: identifying in each text and mapping the extracted terms in each language. However, these tasks cannot be completely independent, since expressions of a single graphic word in one language can be expressed by several graphic words in the other languages (see Tasble 2).

TABLE 2: DIFFERENCE BETWEEN AMAZIGH, ENGLISH AND FRENCH LANGUAGES

Language	Sentence	Words number
Amazigh	ⵍⵓⵣⵉⵔ ⵏ ⵓⴳⵍⵍⵉⴷ [luzir n ugllid]	3
English	The visit of the king	5
French	La visite du roi	4

4.3 Alignment Tools

Nowadays, we meet many automatic alignment tools:

- Alinèa[2:] developed by Olivier Kraif, director of educational department of Stendhal University of Grenoble. It is a program dedicated to the creation and editing bilingual word-and-sentence- aligned corpus.
- AlignFactory[3]: developed by the Canadian company Terminotix Inc. It allows sentence alignment.
- GIZA ++[4]: offers a couple of tools that can come in handy for sentence and word alignment of parallel corpora.
- WinAlign[5]: a software used for corpus sentence-alignment. It is a part of the SDL Trados Translation tools.
- OmegaT[6]: a free Computer Assisted Translation (CAT) tool. It offers a sentence-alignment tool.
- ClueAligner[7]: a web alignment tool designed for manual annotation and alignment of pairs of parallel corpus, representing both contiguous and non-contiguous multiword and phrasal expressions found in monolingual or bilingual parallel sentences.
- Mkalign[8]: created at the Sorbonne Nouvelle-Paris III University by Serge Fleury. It allows aligning parallel corpus on sentences and words.

However, the tool, which generates a better alignment at sentence's level and provides the necessary statistics on an aligned corpus, is "Alinèa". This tool allows significant timesaving on some stages, as well as a facility maintenance than the other tools [35].

V. CORPORA ENCODING STANDARDS

One of the many aspects to be taken when developing a new corpus is its encoding. Some corpora are made available in specific XML formats together with simple programs to process them. Some others are made available in formats like the Text Encoding Initiative (TEI) or the XML Corpus Encoding Standard (XCES)[9]. Unfortunately, these standards are not flexible enough for the tasks they are being used, and therefore each user expand and/or interpret the standard by their will. In this section, we will focus on three different formats that have been used by the research community to encode corpora, specifically parallel ones:

- The Text Encoding Initiative (TEI)[10] is an international organization founded, in 1987, to develop guidelines for encoding machine-readable texts in the humanities and social sciences. TEI includes a big variety of schemas

[2] Alinèa: http://olivier.kraif.u-grenoble3.fr/index.php?option=com content&task=view&id=27&Itemid=43 (Viewed 01/02/2017)

[3] AlignFactory: http://www.terminotix.com/index.asp?name=AlignFactory (Viewed 06/10/2017)

[4] Giza++: http://www.statmt.org/moses/giza/GIZA++.html (Viewed 01/10/2017)

[5] WinAlign: http://www.sdltrados.com/solutions/translation-alignment/ (Viewed 06/10/2017)

[6] OmegaT: http://omegat.org/fr/ (Viewed 06/10/2017)

[7] ClueAligner: http://uplug.sourceforge.net/doc/ica.html (Viewed 06/10/2017)

[8] Mkalign: http://mkalign.software.informer.com/2.0/ (Viewed 06/10/2017)

[9] http://www.xces.org/(Viewed 06/10/2017)

[10] www.tei-c.org/(Viewed 06/10/2017)

to encode texts, verses, transcription of speech, standard dictionaries, lists of places and names, tables, mathematical formulae, graphs, networks, trees and others.

- Translation Memory eXchange (TMX)[11] is an open XML standard, which ensures the correct exchange of data between various parallel corpora. TMX is designed to provide translations of sentences in different languages. It has been in existence since 1998. The exchange facilitates transfer of TM (Translation Memory) data among tools and translators with minimal or no loss of critical data.
- XML Localisation Interchange File Format (XLIFF) is an XML-based format created to standardize the way localizable data are passed between tools during a localization process. Its specification is aimed at the localization industry. It specifies elements and attributes to store content extracted from various original file formats and its corresponding translation.

These three formats were designed for different objectives. The question is: which standard should we use to encode parallel corpora? A final decision on what encoding standard to use is highly depend on if the parallel corpus will be used as a translation memory for the machine translation system or not.

In the case of machine translation study, it is clear that TMX is the format to be chosen. Especially that in the case of an aligned multi-language corpus, TMX allows getting a separate XML file for each language and independent alignment files for each language pair. This way, the user can clearly choose what file to download and use.

VI. PROCESSING AND ALIGNMENT OF THE AMAZIGH-ENGLISH-FRENCH CORPUS

6.1 Processing

Based on the study we have accomplished about the construction of a corpus for an under resourced language, we found that despite the scarcity of parallel texts, and the difficulty of their construction, the use of a parallel corpora, help to obtain easily information about the characteristics of languages, which is very needed especially for a under-resourced one [27].

Building this kind of resources is a challenging task, especially when it deals with under-resourced language like Amazigh language. However, it is necessary to make the effort to building a corpus and enriching linguistic resources in favor of this language.

11 https://www.andovar.com/translation-memory-exchange-tmx/ (Viewed 06/10/2017)

Constructing an Amazigh corpus is not an easy task, especially that there is not so many electronic Amazigh text available. To undertake this study, we decided to collect, as a first step, bilingual texts as many as we can, as long as the texts are of good quality. Our sources are books from a variety of domains from IRCAM.

In the second step, the Amazigh, English and French texts were been normalized in a text-only form, by converting the various formats (for example rtf, doc, and pdf) to plain text files. Then, the corpus has been preprocessed, by replacing dates, numbers, and names by generic tags.

The total size of the corpus is about 265173 words divided into three main subgroups corresponding to the three languages: English (95362 words), Amazigh (73811) and French (96000). The difference between the numbers of words in the three languages is due, mainly to syntax and grammar diversity between languages. The corpus includes 10694 English sentences, 10032 Amazigh sentences and 11038 French sentences. Table 3 illustrates these statistics. The difference among the numbers of sentences is due to the reason that an Amazigh sentence can be translated by two or more sentences in the other languages.

TABLE 3: STATISTICS OF THE AMAZIGH-ENGLISH-FRENSH CORPUS

Languages	English	Amazigh	French
Number of words	95362	73811	96000
Number of sentences	10694	10032	11038

6.2 Alignment

After the preprocessing step, we proceed to sentence alignment, which is an essential requirement for the parallel corpus. To this end, we have used Alinèa tool.

The used aligning system is based on a statistical model for aligning sentences depending on the correlation between sentence length in parallel texts, and cognate pairs [35]. For each aligned pair, Alinéa provides a score that measures the association strength between units, calculated on the basis of similarities, relative positions in sentences, and co-occurrence statistics. This score is a relative value because it gives preference to some matches against others competing associations. The matching pairs are extracted using an iterative one-to-one matching algorithm: matches pairs of sentences based on their degree of overlap, where the overlap between a sentence pair is the total activation weight of terms common to both.

Actually, we have a parallel-aligned corpus Amazigh-English that contains 9747 couples of sentences, a parallel-aligned corpus Amazigh-French with 10091 couples of sentences, and an English-French one that contains 10135 couples of sentences.

6.3 Evaluation and analysis

Results obtained by Alinèa were satisfactory aligning more than 90% of the corpus, although the percentages differed from one language pair to another. We have evaluated our work by using manual evaluation. It is considered a slower method due to the important time spent in evaluating each couple of aligned sentences. The alignment evaluation is done by calculating the precision, recall and F-measure measurements. A perfect system would be the one which measurement of precision and a recall are equal to 1.

$$\text{Recall} = \frac{\text{Number of correctly aligned sentences}}{\text{Number of all sentences}} 100(\%)$$

$$\text{Precision} = \frac{\text{Number of correctly aligned sentences}}{\text{Number of aligned sentences}} 100(\%)$$

$$\text{F-measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} 100(\%)$$

We have calculated the F-measure, which combines recall and precision in a single efficiency measure [25], the following table show the results of the automatic alignment.

TABLE 4: RESULTS OF THE AUTOMATIC ALIGNMENT

Measure	Precision	Rappel	F-measure
Amazigh-English	95%	94%	94%
Amazigh-French	95%	93%	93%
English-French	99%	98%	98%

Evaluating the alignments in the three language pairs shows that the automatic alignment achieved the best results in case of English-French with a percentage of 98% followed by the Amazigh-English pair 94% then by the Amazigh-French 93%.

As shown in Table 4, we found that the measurements are satisfactory and the alignment of sentences is of very good quality. The results obtained in this stage show that the best results were obtained among the most similar language pair; English-French. However, we analyzed all errors of sentences' alignment in order to find reasons behind the errors in the case of the Amazigh-English pair and the Amazigh-French one.

While analyzing the alinement results, firstly, we found that 2.1% of the errors is related to the unbalanced number of sentences. In the English text, we found 10694 sentences, while in the corresponding Amazigh text there are only 10032 sentences, whereas the French one contains 11038 sentences.

Secondly, we remarked that the main reason behind errors is the grammatical differences between the languages. 2.7% of errors depends on the fact that English has two or more words in expressions that constitute only one word in Amazigh (see Table 2). The same thing for the Amazigh-French pair because more words are needed in French than in Amazigh to say the same thing.

Finally, other examples of the wrongly aligned material (1.2%) cannot be explained by grammatical differences between languages, rather than to be due to segmentation errors, which caused two or more sentences to be regarded as a single unit (see Table 5). Majority of these sentences were adjusted manually, so that the resulting corpus contained a set of sentences and their counterparts in 1-to-1 relation (see Table 6).

TABLE 5: MISALIGNMENT DUE TO INCORRECT SEGMENTATIONS

English	**Amazigh**	**French**
[s42 s43] "I will go and steal some bread again." Her brother said to her :	[s42] "ⵔⴰⴷ ⴼⵜⵓⵖ ⴰⴷ ⴷ ⴰⵡⵉⵖ ⴷⴰⵖ ⴰⵖⵔⵓⵎ." ⵉⵏⵏⴰ ⴰⵙ ⴳⵯⵎⴰⵙ : ["rad ftur ad d awigh dagh aghrum." inna as g^{w}mas:]	[s42 s43] "Je vais aller voler du pain à nouveau."Son frère lui a dit:

TABLE 6: A MANUALLY CREATED 1-TO-1 ALIGNMENT

English	Amazigh	French
[s42] "I will go and steal some bread again."	[s42] "ⵔⴰⴷ ⴼⵜⵓⵖ ⴰⴷ ⴷ ⴰⵡⵉⵖ ⴷⴰⵖ ⴰⵖⵔⵓⵎ." ["rad ftur ad d awigh dagh aghrum."]	[s42] "Je vais aller voler du pain à nouveau."
[s43] Her brother said to her :	[s43] ⵉⵏⵏⴰ ⴰⵙ ⴳⵯⵎⴰⵙ : [Inna as g^{w}mas:]	[s43] Son frère lui a dit:

For the corpus encoding, we chose to encode our corpus on the TMX standard. A TMX file is a header with some meta-information and a body with a sequence of translation units (tu). A translation unit is a sequence of translation unit variants (tuv) with a segment (seg). Fig. 2 presents a simple example of Aligned parallel texts (Amazigh-English) in TMX format.

VII. CONCLUSION AND FURTHER DEVELOPMENT

In this paper, we have presented an overview of multi-lingual corpus in general, specifically the parallel and comparable ones. We have made a comparative study between the two types of corpus. We have introduced the alignment concept: its approaches, main tools and techniques. We gave a brief insight of the three major schemas available to encode parallel corpora. Then, we have shown a workflow of building a multi-lingual parallel corpus (Amazigh-English-French). This workflow is based on a sentence-aligned parallel corpus using “Alinèa”.

We have created parallel corpus for pairs of languages with a relatively different typology Amazigh, English and French. We have attempted to bridge between three linguistic theories commonly used for their description. Nevertheless, we notice that the alignment of the pairs Amazigh-English and Amazigh-French texts was difficult, due to their syntactic structure differences.

The value of a parallel corpus grows with its size and with the number of languages for which translations exist. In the future, we plan to collect more data for parallel corpus to improve the quality and give more confidence for the results induced from it. The corpus is still under development, and we hope that we may enlarge it further with material that can be made freely available to the public.


```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE tmx SYSTEM "tmx14.dtd">
<tmx version="1.4">
<body>
<tu>
<tuv xml:lang="EN">
<seg>A story about two children and an
ogress.</seg>
</tuv>
<tuv xml:lang="ZGH">
<seg>[illegible]</seg>
</tuv>
</tu>
<tu>
<tuv xml:lang="EN">
<seg>There once lived a man and a woman.</seg>
</tuv>
<tuv xml:lang="ZGH">
<seg>[illegible]
[illegible]</seg>
</tuv>
</tu>
</body>
```

Fig. 2. Exemple of the aligned parallel texts in TMX format

REFERENCES

[1]. F. Ataa Allah, Traitement automatique des langues peu dotées : Cas de la langue amazighe, Thesis for the fulfillment of the degree of Researcher qualified to supervise research. IRCAM, Rabat, Morocco 2015.

[2]. N. S. Dash, Corpus Linguistics: An Introduction, (Pearson Education India, 2008).

[3]. M. Ameur et al., Initiation à la langue amazighe, (Rabat: IRCAM, 2004).

[4]. P. Andries, Unicode 5.0 en pratique : Codage des caractères et internationalisation des logiciels et des documents, (France : Dunod, 2004).

[5]. M. Amrouch et al., Handwritten Amazighe Character Recognition Based On Hidden Markov Models, International Journal on Graphics, Vision and Image Processing, 10(5), 2010, 11–18.

[6]. K. El Gajoui, F. Ataa Allah and M. Oumsiss, Recognition of Amazigh Language Transcribed into Latin based on Polygonal Approximation, International Journal of Circuits, Systems and Signal Processing, vol. 10, 2016, 297-305.

[7]. A. Abenaou, F. Ataa Allah and B. Nsiri, Vers un système de reconnaissance automatique de la parole amazighe basé sur les transformations orthogonales paramétrables, Asinag, n°9, 2014, 133-145.

[8]. H Satori and F. El Haoussi, Investigation Amazigh speech recognition using CMU tools, International Journal of Speech Technology, 17(3) , 2014, 235-243.

[9]. F. Ataa Allah and J. Frain, Amazigh Converter based on WordprocessingML, Proc. 6th edition of Language & Technology Conference, Poznań, Pologne, 7-9 december 2013.

[10]. N. Yakoubi, J. Frain and F. Ataa Allah, Convertisseur Numérique : Tifinaghe – Braille, Proc. of the 7th edition de la conférence internationale sur la Technologie de l’Information et de Communication, IRCAM, Rabat, Morocco, 28-29 November 2016.

[11]. S. Boulaknadel and F. Ataa Allah, Online Amazigh Concordancer, Proc. of the International Symposium on Image Video Communications and Mobile Networks, Rabat, Morocco, September 30th – October 2nd, 2010.

[12]. S. Boulaknadel and F. Ataa Allah, Building a Standard Amazigh Corpus, Advances in Intelligent Systems and Computing: Proc. of the International Conference on Intelligent Human Computer Interaction, Prague, Czech Republic, August 29-31, 2011, vol. 179/2013: 91-98. Springer Berlin Heidelberg, ISBN: 978-3-642-31602-9.

[13]. M. Outahajala et al., Tagging Amazigh with AnCoraPipe, HLT & NLP Workshop, Malta, 17 may 2010.

[14]. N. Miftah, F. Ataa Allah and I. Taghbalout, Sentence-Aligned Parallel Corpus Amazigh-English, Proc. of the International Conference on Information and Communication Systems, 2017, Irbid, Jordanie.
[15]. F. Ataa Allah, Finite-State Transducer for Amazigh Verbal Morphology, Literary & Linguistic Computing, Oxford University Press, 2014.
[16]. F. Nejme et al., AmAMorph: Finite State Morphological Analyzer for Amazighe, Journal of Computing and Information Technology, 24(1), 2016.
[17]. S. Boulaknadel et al. Amazighe Named Entity Recognition using a rule based approach, Proc. of AICCSA 2014, 478-484.
[18]. I. Taghbalout, F. Ataa Allah and M. El Maraki, Towards UNL based machine translation for Moroccan Amazigh language, International Journal of Computational Science and Engineering, 2018.
[19]. A. Le Serrec, Analyse comparative de l'équivalence terminologique en corpus parallèle et en corpus comparable : application au domaine du changement climatique, Ph.D. Thesis, University de Montréal, Montréal, Canada. 2008.
[20]. P. Koehn, Europarl: A Parallel Corpus for Statistical Machine Translation, Proc. of the 10th Machine Translation Summit MT Summit X, 2005, Phuket Islan, Thailand.
[21]. R. Steinberger, B. Pouliquen, A. Widiger, C. Ignat, T. Erjavec, D. Tufi and D. Varga, The JRC-Acquis: A Multilingual Aligned Parallel Corpus with 20+ Languages, Proc. of the 5th International Conference on Language Resources and Evaluation, 2006.
[22]. R. Harastani, Alignement lexical en corpus comparables: le cas des composés savants et des adjectives relationnels, Ph.D. Thesis , University of Nantes, Nantes, France, 2014.
[23]. L. Bowker and J. Pearson, Working with Specialized Language: A practical guide to using corpora, (London & New York: Routledge, 2002).
[24]. A. Hazem and M. Emmanuel, Extraction de lexiques bilingues à partir de corpus comparables par combinaison de représentations contextuelles, TALN-Récital, Les Sables d'Olonne, 2013.
[25]. T. Do, Extraction de corpus parallèle pour la traduction automatique depuis et vers une langue peu dotée, Ph.D. Thesis University de Grenoble, Saint-Martin-d'Hères, France. 2012.
[26]. B. Li and E. Gaussier, Improving Corpus Comparability for Bilingual Lexicon Extraction from Comparable Corpora, Proc. of the 23th International Conference on Computational Linguistics, 2010.
[27]. N. Miftah et al., Corpus multilingues pour les langues peu dotées, Proc. of the 7th TICAM Conference, Rabat, Morocco, 2016.
[28]. G. Rappazzo, Evaluation de la fonction de recherche dans les outils d'exploitation d'un corpus parallèle, Ph.D. Thesis, University de Genève, Suisse 2014.
[29]. D. Bouamor, Constitution de ressources linguistiques multilingues à partir de corpus de texts parallèles et comparables, Ph.D. Thesis, University paris sud, Paris, France. 2014
[30]. I. Dagan, K. Church, and W. Gale, Robust bilingual word alignment for machine aided translation, natural language processing using very large corpora, Speech and Language Technology, vol 11, 1999, 209-224.
[31]. J. Véronis, Parallel text processing: alignment and use of translation corpora, Computational Linguistics, 27(4), 2001.
[32]. G. Lunter, A. J. Drummond, I. Miklos, and J. Hein, Statistical alignment: Recent progress, new applications, and challenges, in Rasmus Nielsen (Ed.), Statistical Methods in Molecular Evolution, (New York: Springer, 2005).
[33]. M. Simard, G. Foster and P. Isabelle, Using Cognates to Align Sentences in Bilingual Corpora, Proc. of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation, 1992, Montreal, Canada 67-81.
[34]. M. Kay and M. Roscheisen, Text - Translation Alignment, Computational Linguistics, 1993.
[35]. O. Kraif, Alignement multilingue pour l'étude contrastive: outils et applications, in M. Hédiard (Ed.), Linguistica dei corpora, Strumenti e applicazioni, Edizioni dell'Università degli Studi di Cassino, 2008, 83-99.

Fadoua ATAA ALLAH "The First Parallel Multi-lingual Corpus of Amazigh "International Journal of Engineering Research and Applications (IJERA) , vol. 8, no.6, 2018, pp.05-12