On the risk of cross-language plagiarism for less resourced languages such as Amazigh

Paolo Rosso

Natural Language Engineering Lab

ELiRF, Dept. SIC, Universidad Politécnica de Valencia, Spain

http://www.dsic.upv.es/grupos/nle/

prosso@dsic.upv.es

Abstract

The exact population of Amazigh speakers is hard to be said since most North African countries do not record language data. What is a fact is that Amazigh is a less resourced language with a very low degree of representation on the Web. In a society where information in multiple languages is available on the Web, cross-language plagiarism is occurring every day with increasing frequency, especially for less resourced languages. Potentially this could be the case of Amazigh. The lack of resources, such as Amazigh-Arabic and Amazigh-French, makes the detection of cross-language plagiarism a real challenge. This paper gives an overview of what plagiarism is and what are the available plagiarism detection tools, as well as the state-of-the-art plagiarism detection systems, focusing especially on the case where plagiarism occurs across languages. Special emphasis will be given to cross-language plagiarism in less resourced languages such as Amazigh.

1. Introduction

A relatively sparse population speaking a group of closely related and similar languages and dialects extends across the Atlas Mountains, the Sahara and the northern part of the Sahel in Morocco, Algeria, Niger, Mali, Tunisia, Libya, and the Siwa oasis area of Egypt[1]. There is a movement among speakers of the closely related languages to unite them into a single standard language: Amazigh. The exact population of Amazigh speakers is not easy to estimate, since most North African countries do not record language data. A survey included in the official Moroccan census of 2004 and published by several Moroccan newspapers[2] gave the following figures: 34% of people in rural regions spoke Amazigh and 21% in

[1] http://en.wikipedia.org/wiki/Amazigh_language

[2] http://www.bladi.net/marocain-berbere.html

urban zones did, the national average would be 28.4% or 8.52 millions. However, it is possible that the survey asked for the language "used in daily life" which would result of course in figures clearly lower than those of native speakers. Others estimate that the total number of speakers of Amazigh in the Maghreb appears to lie anywhere between 16 and 25 (30 millions if Sahel and the Siwa oasis are included) whose vast majority are concentrated in Morocco and Algeria.

In recent years, due to the large amount of text available on the WWW, plagiarism cases have increased. Moreover, in a society where information is available on the Web in multiple languages, cross-language plagiarism cases are also common, especially when the target language is a less resourced one (e.g. Amazigh) and the user is more likely to find the information s/he looks for in a more resourced language (e.g. English, French or Arabic). The rest of the paper is structured as follows. Section 2 defines what plagiarism is and what the different kinds of plagiarism are. The available plagiarism detection tools and the best state-of-the-art plagiarism detection systems participating at the first of plagiarism detection are also described. Section 3 is devoted to cross-language plagiarism and the first attempts to approach it. Special emphasis is given to the case where the target language is a less resourced one, such us Amazigh. Finally, in the last section some conclusions are drawn.

2. Plagiarism

Although often no distinction is made between text reuse and plagiarism and just the generic text reuse is employed, there is a narrow difference between the two. With text reuse we mean the activity whereby pre-existing written texts are used again to create a new text or version (Clough and Gaizauskas, 2009) but this does not mean that an infringement is intended: collaborative authoring (e.g. Wikipedia), news from press for newspapers (e.g. Reuters, Press Association, etc.), etc. In case the reuse of someone else’s prior ideas, processes, results, or words occurs without explicitly acknowledging the original author and source then we can talk about plagiarism (IEEE, 2008). It has to be said that often plagiarism could occur, for instance, in books from narrative and events that could resemble each other to plagiarism of ideas (that is not based on words dependency) and plagiarism of ideas is nowadays (practically) impossible to be detected automatically.

Surveys of the research done in automatic plagiarism detection can be read in (Clough, 2003) and (Maurer et al., 2006). Plagiarism detection can be divided into external plagiarism detection - when, given a suspicious fragment of a document, a set of potential source documents is available - and intrinsic plagiarism detection - when the lack of a set of potential source documents makes the detection of a suspicious fragment more difficult because based only on style changes.

2.1. Plagiarism Detection Tools

Many are the tools, some of them freely available, for plagiarism detection. All of them are external plagiarism detection tools, that is, their aim is to find the potential source fragment plagiarism has been committed from. Of course, this is possible only if the set of potential source documents is available. Moreover, they perform well only when a simple duplicate (copy-paste) or near-duplicate (use of synonyms) plagiarism of fragment occurs. Their performance decreases dramatically in case of paraphrasing (Barrón-Cedeño et al., 2010a) or translated plagiarism across languages (Potthast et al., 2011). Therefore, if from one hand due to the large amount of information available on the Web plagiarism has increased in recent years and this makes manual plagiarism detection infeasible (Weber, 2007; Kulathuramaiyer and Maurer, 2007), from the other hand texts can be easily found, manipulated – making usage of paraphrasing or translated plagiarism - and combined. Therefore, it is important to stress that automatic plagiarism detection has only to assists experts providing them linguistic evidence for the final decision.

Below the list of ten among the most well-known plagiarism detection tools (Vallés, 2010):

i. **Turnitin**[3] is not a free plagiarism detector tool. It has been developed by John Barrie (University of Berkeley) and it is used by more than 50 universities in the world[4].

ii. **WCopyFind**[5] is a tool which was developed in 2004 by Lou Bloomfield, University of Virginia. Plagiarism is detected on the basis of the comparison of word n-grams (sequence of n words). The size of n is decided by the users although for WCopyFind (Dreher, 2007) suggest using hexagrams.

iii. **Ferret**[6] is a tool to detect plagiarism that was developed in the University of Hertfordshire (Lyon at al., 2006). It is able to analyse documents in different formats (PDF, Word and RDF). It extracts trigrams obtaining a similarity measure on the basis of the common trigrams between two documents (Malcom and Lane, 2008).

[3] http://www.turnitin.com/
[4] Digital solutions for a new era in information. 2004. iparadigm: http://www.iparadigms.com
[5] http://plagiarism.phys.virginia.edu/
[6] http://homepages.feis.herts.ac.uk/_pdgroup/

iv. **CopyCatch**[7] is a tool designed by CFL Software. It is possible to calculate the similarity between two complete documents or some of its sentences. CopyCatch needs to have as input the document in order to investigate if some of its parts have been plagiarised. It succeeds in detecting the similarity also in case of simple paraphrasing: insertions, deletions or change in the order of the words. It works in different languages.

v. **iThenticate**[8] is a plagiarism detection service for preventing from Web-based plagiarism, content verification and intellectual property copyright. Given a document, it compares it against its large data base. A report is provided to the user in case a similarity is found with other(s) document(s).

vi. **Plagiarism Checker**[9] is a Web application which has been developed by the Department of Education of the University of Maryland. Its aim is to detect whether a text is suspicious to be copied. The suspicious text needs to be introduced and the application checks for similar texts using the API of Google. It is free and fast but, as most of these tools, it is quite unlikely to find the source text in case of paraphrasing or translated plagiarism.

vii. **Pl@giarism**[10] is a freely available tool that has been developed by the Law Faculty of the University of Maastricht in order to detect plagiarism cases in the essays of their students. Pl@giarism is a simple application for Windows which determines the similarity between two documents on the basis of the comparison of their trigrams. It returns a table with similarity percentages between the suspicious document and its similar documents.

viii. **DOC Cop**[11] is a freely available tool. It returns acceptable results especially if the comparison of the suspicious document is made against a smaller data base than the Web (Scaife, 2007). A report is sent by email and those fragments suspicious to be plagiarised are highlighted.

ix. **EVE2**[12] (Essay Verification Engine) is a tool developed by Canexus. EVE2 allows checking if students have plagiarised parts of their essay from the Web. It returns the links to the Web pages plagiarism is likely to have been committed from. Unfortunately it seems to be quite slow: Dreher (Dreher, 2007) carried out an

[7] http://csoftware.com/
[8] http://www.ithenticate.com/
[9] http://www.dustball.com/cs/plagiarism.checker/
[10] http://www.plagiarism.tk/
[11] http://www.doccop.com/
[12] http://www.canexus.com/

experiment in order to detect possible plagiarised texts in just 16 pages, containing 7,300 words, of a M.Sc. thesis and the tool took 20 minutes to process them.

x. **MyDropBox**[13] is an online service whose aim is to help the detection of plagiarism. The reports that the tool returns are quite well structured in order to highlight the links with the sources of the Web where plagiarism is likely to have been committed (Scaife, 2007).

2.2. External and Intrinsic Plagiarism Detection

As said previously, methods for automatic plagiarism detection can be divided in two main approaches: external plagiarism detection and intrinsic plagiarism detection.

External plagiarism detection can be considered as a task related to information retrieval. In fact, given a suspicious document *d* and a collection of potential source documents *D*, the task is to detect the plagiarised sections in *d* (if there are any), and their respective source sections in *D* (Potthast et al., 2009). Up to now, researchers have paid more attention to this approach (see, for instance, the previous section on plagiarism detection tools) because obtaining the source of a potential case of plagiarism provides better linguistic evidence to help the expets (e.g. forensic linguistics) to make their final decision on whether a fragment of text has been plagiarised or not. The problem is that it is not an easy task to find the potential source of plagiarism in case the set *D* of potential source documents is the Web itself. In fact, text plagiarism is observed at an unprecedented scale with the advent of the World Wide Web (the new term of cyber-plagiarism (Comas and Sureda, 2008) has been recently introduced to refer to the copy-paste syndrome) and this is the real scenario plagiarism detection systems should consider. In terms of number of comparisons, the size of the reference data set (e.g. the Web) could be a problem from a computational point of view. Therefore, it is important to reduce the number of exhaustive comparisons only to those between fragments that are more similar. In order to solve the problem of the size of the reference data set, in (Barrón-Cedeño and Rosso, 2009) the authors described a method based on the Kullback-Leibler distance (Kullback and Leibler, 1951) for reducing the search space (the Kullback-Leibler symmetric distance measures how close the probability distributions of the reference and suspicious documents are).

Most of state-of-the-art plagiarism detection systems base their approach on the comparison of word n-grams of the fragments of the suspicious document *d* and those of the documents of the reference data set *D* (Kasprzak et al., 2009) also taking into account vocabulary expansion, for instance with Wordnet[14] (Kang et al.,

[13] http://www.mydropbox.com/
[14] http://wordnet.princeton.edu/

2006). The comparison could also be made on the basis of character n-grams (Grozea et al., 2009) where character n-grams of the suspicious documents are matched against the character n-grams of the source document (see Figure 1). A dot means that the character n-gram exists in both documents. A diagonal provides linguistic evidence of a possible plagiarism case (e.g. left corner of the graph). A diagonal together with a cluster of dots gives less evidence but a certain similarity between the two fragments of the suspicious document and the source still occurs and deserves to be manually further investigated by the forensic linguistic expert who has to make the global decision whether it is a plagiarism case or not. A similar plot approach was also employed by (Basile et al., 2009) but instead of plotting character n-grams, after a pre-process in which each word was substituted by its length (e.g. length = 6), n-grams of numbers were plotted (e.g. substituted by its length = 11 2 3 6). Once more, a dot means that the number n-gram exists in both documents, and a diagonal provides linguistic evidence of a possible plagiarism case (i.e., a sequence of words of the same length is found both in the suspicious document and in the source one).

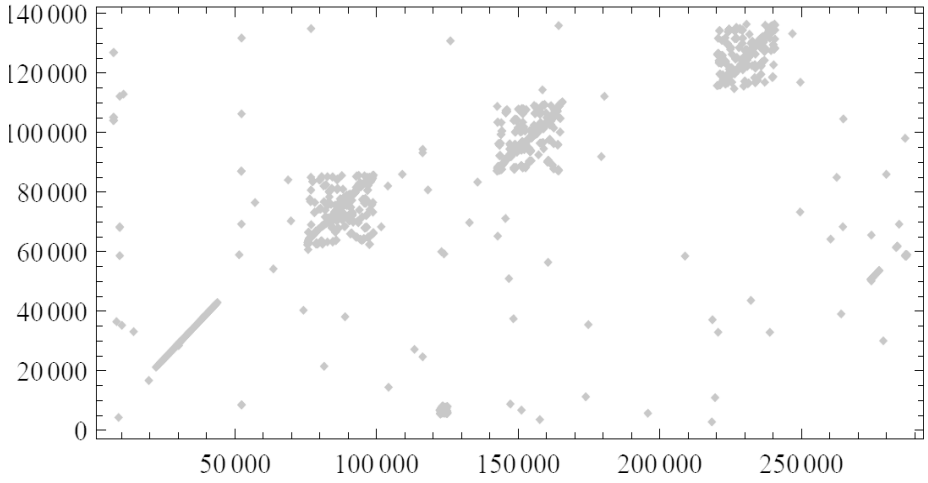
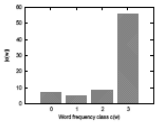
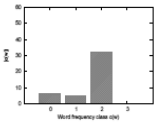
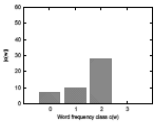


Figure 1. ENCOPLOT: visual approach for external plagiarism detection (Grozea et al., 2009)

In case of lack of the reference set of potential source documents *D*, the detection of plagiarised fragments has to rely only on changes in the writing style in the document. A person could be often able to manually identify potential cases of plagiarism by detecting text inconsistencies (unexpected irregularities through a document such as changes of style, vocabulary, or complexity are triggers of suspicion) or by resembling previously consulted material. Nevertheless, the large amount of potential source texts available nowadays makes infeasible this manual

plagiarism detection based on writing style change. In order to assist experts, automatic intrinsic plagiarism detection methods have been developed aiming to detect whether the document *d* contains text fragments written by a different author.

The features considered by these models are, among others, word length average, sentences length average, stop-words average, as well as readability and vocabulary richness (Meyer zu Eißen and Stein, 2006). The readability of a text could be measured, for instance, on the basis of the complex words used (complex words are those with three or more syllables) employing indexes such as Gunning fog or Flesch (DuBay, 2008). Figure 2 shows how linguistic evidence for plagiarism could be provided on the basis of the above measures for intrinsic plagiarism detection. In the example, two text fragments (last two columns) are compared with the all document (column named as Global). Linguistic evidence is provided with respect to the use of more complex words in the first text fragment (complexity measure of 17 vs. approx. 14). Once more, the automatic approach has the aim to simply assist the forensic linguistic expert who has to be the one making the decision. Finally, like for the external plagiarism detection, there are methods

Measure	Global	■	■
tokens	135	63	72
types	78	44	46
W. avg. freq. class			
avg. sentence length	19.28	21.00	18.0
avg. word length	4.93	5.38	4.54
Complexity measures	16.72	17.07	13.82

that apply character n-gram profiles to characterise an author's style and search for irregularities in the document *d* (Stamatatos, 2009).

Figure 2. Measures for intrinsic plagiarism detection

2.3. Plagiarism Detection Competition

The development of plagiarism detection models is not new although the large amount of information available on the Web plagiarism has increased in recent years. One of the first approaches we have track of goes back to the 1970s (Ottenstein, 1976). However, after more than 30 years, no standard evaluation framework (i.e., standard text collections with documented cases of plagiarism and

evaluation measures) existed in order to compare the performance of the different plagiarism detection methods. In fact, researchers often used small and private (80% of cases (Potthast et al, 2010a)) collections of documents that cannot be freely provided to other researchers for ethical reasons. Moreover, they estimated the quality of the models by considering different evaluation measures. Therefore, with the aim of providing a standard evaluation framework on automatic plagiarism detection, together with the Webis research group of Weimar University[15] and the universities of the Aegean[16] and of Bar-Ilan[17], the first International Competition on Plagiarism detection[18] was organised. In 2011 its third edition, sponsored by Yahoo! Research Barcelona[19], will be organised again as one of the benchmarking activities of CLEF evaluation campaign[20].

In the first edition (Stein et al., 2009) two tasks were organised: external plagiarism detection and intrinsic plagiarism detection. The best approach for external plagiarism detection was the ENCOPLOT of (Grozea et al., 2009) and for intrinsic plagiarism detection the one of (Stamatatos, 2009). Both approaches were based on the comparison of character n-grams. The teams who participated with two of the software tools previously described (WCopyFind and Ferret) did not obtain a good performance (Potthast et al., 2009).

In the second edition no distinction between external and intrinsic plagiarism detection was made. The best approach was the one of (Kasprzak and Brandejs, 2010) that was based on word n-grams. In the first edition (Potthast et al., 2009), 10 teams participated in the external plagiarism detection task and only 4 teams in the intrinsic plagiarism detection one. In the second edition (Potthast et al., 2010a), although no distinction was made and only one plagiarism detection task was organised, many of the 18 teams that participated had their overall performance penalised because they did not solve properly (or they did not solve at all) the intrinsic plagiarism cases (30% of total plagiarism cases (Potthast et al., 2010b)). The above shows that less attention has been paid from the research community to the intrinsic plagiarism detection both because more difficult also in terms of giving linguistic evidence without a source document where the plagiarism has been committed from.

[15] http://www.uni-weimar.de/cms/medien/webis/home.html
[16] http://www.icsd.aegean.gr/lecturers/stamatatos/
[17] http://u.cs.biu.ac.il/~koppel/
[18] http://pan.webis.de/
[19] http://labs.yahoo.com/Yahoo_Labs_Barcelona
[20] http://clef2011.org/index.php?page=pages/labs.html

The results of the competition, as well as the description of the evaluation measured and the data set (8.4 Giga Bytes, 162,000 plagiarism cases, between training and test samples) are available at: http://pan.webis.de .

3. Cross-language Plagiarism

In a society where information is available on the Web in multiple languages, cross-language plagiarism occurs every day with increasing frequency. This behaviour was simulated in the data set of the competition where 14% of plagiarism cases were translated plagiarisms from Spanish or German into English (Potthast et al., 2010b).

3.1 Cross-language Plagiarism Detection

Cross-language plagiarism detection deals with the automatic identification and extraction of plagiarism in a multilingual setting. In this setting, a suspicious document is given, and the task is to retrieve the source documents of the suspicious fragments from a large, multilingual document collection.Up to the present time, cross-language plagiarism detection has not been approached sufficiently due to its intrinsic complexity. Whereas some commercial tools are able to perform plagiarism analyses on different languages, detecting cases of translated plagiarism is still in its infancy. In the first edition of the competition no team tried to detect the cross-language plagiarism cases (Potthast et al., 2009). In the second edition, some teams approached the problem on a monolingual basis translating the source documents in Spanish or German into English (Potthast et al., 2010a). No matter the large size of the data set (8.4 GB, 162,000 plagiarism cases) this is still a close scenario but in the open (and more realistic) scenario of the Web, it would be not feasible from a computational time point of view translating all the documents into the target language plagiarism needs to be investigated (e.g. Amazigh).

Few are the cross-language plagiarism detection approaches that have been investigated so far. Probably the two methods with a certain impact are CL-ASA (cross-language alignment-based similarity analysis) and CL-ESA (cross-language explicit semantic analysis). CL-ASA (Barrón-Cedeño et al., 2008; Pinto et al., 2009) is based on the IBM-M1 statistical machine translation model (Brown et al. 1993) and needs a parallel data set to be trained[21]. It estimates the likelihood of two text fragments of being valid translations of each other. CL-ESA is another interesting method for cross-language plagiarism detection (Potthast et al., 2008). CL-ESA intends to estimate, at semantic level, how similar two texts written in

[21] The JRC-Acquis data set was used : http://wt.jrc.it/lt/Acquis/

different languages are. This estimation is carried out on the basis of a comparable data set, such as Wikipedia (Figure 3). The CL-ASA and CL-ESA models have been compared in (Potthast et al., 2011) with the cross-language character n-gram model (CL-CNG). Despite its simplicity, CL-CNG results to be a good choice to compare text fragments across languages if they are syntactically related.

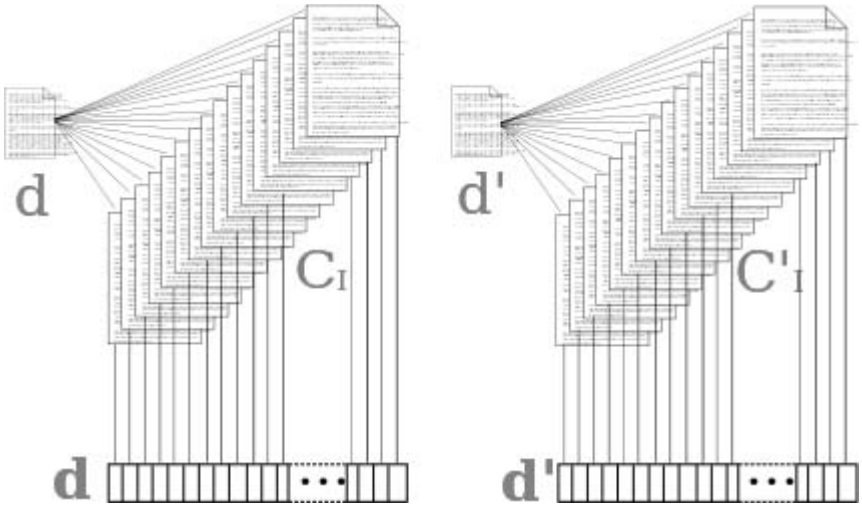


Figure 3. Cross-language explicit semantic analysis (Potthast et al., 2008).

Similarity between documents *d* and *d'* is computed on the basis of the vector space model with indexes the subset of Wikipedia common articles in both languages

3.2 Cross-language Plagiarism Detection in Less Resourced Languages

A less resourced language is that with a low degree of representation on the Web (Alegria et al., 2009). This makes not always possible to employ previous approaches such as CL-ASA and CL-ESA. CL-CNG results to be a good choice but only if the two languages are syntactically related.

If few attempts have been made to solve the problem of cross-language plagiarism detection, even less work has been done to tackle this problem for less resourced languages. One of the few works is the one of (Barrón-Cedeño et al., 2010b) on plagiarism detection across distant language pairs where the authors investigated the case of Basque, a language where, due to the lack of resources, cross-language plagiarism is often committed from texts in Spanish and English. Basque has no known relatives in the language family; however it shares some of its vocabulary

with Spanish. Therefore, the CL-CNG method based on character n-grams was investigated. CL-CNG was compared with CL-ASA and a method that approached the problem from a monolingual perspective calculating the similarity after employing a machine translation pre-process (Figure 4). The translation and monolingual similarity analysis (T+MA) performed better than the other models. As previously said, approaching the problem of cross-language plagiarism detection from a monolingual point of view after translating all the documents into the target language, would not be computationally possible in a realistic scenario such as the Web.

3.3 The Difficulty of Detecting Cross-language Plagiarism in Amazigh

Maghreb states such Morocco and Algeria have created institutions such as the Institute Royal de la Culture Amazighe (IRCAM[22]) and the Haut Commissariat à l'Amazighité (HCA[23]) in order to promote the Amazigh language. In Morocco, Amazigh has been introduced in mass media (an Amazigh television channel was launched in 2010) and in the educational system (Amazigh is taught in various Moroccan primary schools). Moreover, IRCAM during just 8 years since its creation has published more than 150 books related to the Amazigh language and culture, a number which exceeds the whole amount of Amazigh publications in the 20th century. No matter these efforts, from a computational linguistic point of view Amazigh is still a less resourced language. In fact few are the annotated large data set (e.g. (Outahajala et al., 2011)).

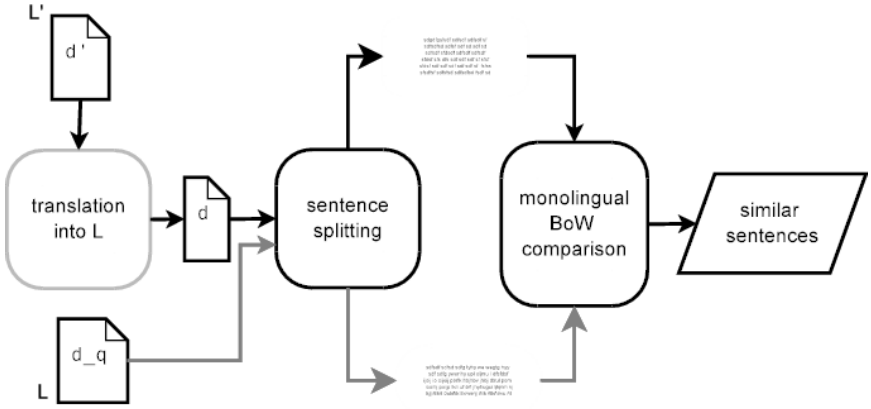


Figure 4. The translation and monolingual similarity analysis model (Barrón-Cedeño et al., 2010b)

[22] http://www.ircam.ma/
[23] http://hcamazighite.org/

The low degree of representation of Amazigh on the Web potentially could be the cause of translated plagiarism from languages such as English, Arabic or French (Figure 5) where the information could be found more easily. Amazigh is not syntactically related to English, French or Arabic and this makes not feasible using the CL-CNG model to detect cases of cross-language plagiarism cases. The lack of large parallel (in Amazigh and Arabic, French or English) and comparable data sets (e.g. Wikipedia) makes a real challenge the use of the CL-ASA and the CL-ESA models previously described. Up to the present time, IRCAM developed three parallel lexicons containing words in Amazigh and their equivalent in French[24] (Ameur et al., 2006), in French and Arabic about media[25] (Ameur et al., 2009), and in French-Arabic-English about Amazigh grammar[26] (Boumalk and Naït-Zerrad, 2009). However they are small and not parallel data sets of equivalent sentences.

Last, with respect to the possibility of employing the translation and monolingual similarity analysis (T+MA) model an automatic machine translator (French-Arabic-English into Amazigh) is needed. The possibility of having to deal with data sets in Amazigh written in both Latin or Tifinaghe characters (Figure 5) is also a further problem, although it seems that recently texts written in Tifinaghe Unicode are increasingly used.

iga usinag agldan n tussna tamaziGt yat tmrslt takadimit nna tsskr tgldit tamGrabit ittindazn i uHTTu d tnkra n tussna tamaziGt. yusa d usinag g RbaT,tamaZunt n lmGrib, ittusnta ass n 17/10/2001 s unbaD n ugllid muHmmad 6 s uDahiR agldan yusin azwil 01-01-299. iga usinag tamrslt Gur tlla tsimant taZRfant d tsmGurt.	ⵉⵣⴰ ⵙⵙⵉⵏⴰⵣ ⴰⵣⵎⴰⵏⴰⵍ ⵏ ⵜⵙⵙⵏⴰ ⵜⴰⵎⴰⵙⵉⵢⵜ ⵖⴰⵜ ⵜⵎⴰⵙⵙⵉⵏⵜ ⵜⴰⵣⴰⵎⵓⵔⵜ ⵏⴰⵏ ⵜⵙⵙⵏⴰ ⵜⴰⵎⴰⵙⵉⵢⵜ. ⵢⵓⵙⴰ ⴷ ⵓⵙⵉⵏⴰⵣ ⵖ ⵙⵍⴻⴻⵙ ⵏ ⵜⵏⵣⵔⴰ ⵏ ⵜⵙⵙⵏⴰ ⵜⴰⵎⴰⵙⵉⵢⵜ. ⵖⵙⵙ ⵏ ⵙⵙⵉⵏⴰⵣ ⵠ ⵕⵙⴻⵙ,ⵜⴰⵎⴰⵙⵙⵏⵜ ⵏ ⵍⵎⵢⵙⵜ, ⵉⵜⵜⵙⵙⵏⵜⴰ ⴰⵙⵙ ⵏ 17/10/2001 ⵙ ⵙⵏⵙⴰⴻ ⵏ ⵙⵠⵍⵍⵉⵏ ⵎⴰⵢⵓ 6 ⵙ ⵙⴻⴰⵙⵣ ⴰⵣⵎⴰⵏⴰⵍ ⵖⵙⵙⵉⵏ ⴰⵙⵍⵉⵎ 01-01-299. ⵉⵣⴰ ⵙⵙⵉⵏⴰⵣ ⵜⴰⵎⴰⵙⵙⵏⵜ ⵢⵙⵙ ⵜⵍⵍⴰ ⵜⵙⵉⵎⵙⵏⵜ ⵜⴰⵙⵕⵓⵜⵏⵜ ⵏ ⵜⵙⵉⵎⵙⵏⵜ.

Figure 5. French-Amazigh cross-language plagiarism: Latin (left) and Tifinaghe scripts (right).

[24] http://www.ircam.ma/fr/index.php?soc=publi&pg=5&rd=64
[25] http://www.ircam.ma/fr/index.php?soc=publi&pg=2&rd=109
[26] http://www.ircam.ma/fr/index.php?soc=publi&pg=2&rd=118

Source: http://fr.wikipedia.org/wiki/Institut_royal_de_la_culture_amazighe

4. Conclusions

Although the problem of plagiarism is well-known, not always people know what the available tools for its detection and their limitations are. Moreover, in case of less resourced languages such as Amazigh, plagiarism from other languages is more likely to occur. Automatic cross-language plagiarism detection is still in its infancy. Therefore, the detection of translated plagiarism is not possible using the available tools. This paper gives an overview of plagiarism detection and, in particular, cross-language plagiarism detection: a problem that will have to be addressed with special emphasis in the future because every time occurring more often especially for less resourced languages.

Acknowledgements

Most of what was described in this paper is the result of the joint work done together with Alberto Barrón-Cedeño in the framework of his Ph.D. and partially also with Enrique Vallés in his M.Sc. This research work was funded by the MICINN research project TEXT-ENTERPRISE 2.0 TIN2009-13391-C04-03 (Plan I+D+i). Last but not least, I have to thank also my Ph.D. students Mohamed Outahajala and Lahsen Abouenour for helping me with Amazigh/e .:-) Tanmmirt bzzaf!

References

Alegria, Iñaki, Mikel L. Forcada, and Kepa Sarasola, editors. (2009). Proc. of the SEPLN 2009 Workshop on Information Retrieval and Information Extraction for Less Resourced Languages, Donostia, Basque Country. University of the Basque Country.

Ameur M., Bouhjar A., Boumalk A., Elazrak N., Abdellaoui R. (2009). Vocabulaire des médias Français-Amazighe-Anglais-Arabe. Publications de l'IRCAM.

Ameur M., Bouhjar A., Elmedlaoui M., Iazzi E. (2006). Vocabulaire de la langue Amazighe. Publications de l'IRCAM.

Barrón-Cedeño A., Rosso P., Pinto D., Juan A. (2008). On Cross-lingual Plagiarism Analysis using a statistical model. In: Proc. 2nd Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse, PAN-2008, Patras, Greece, July 21-24.

Barrón-Cedeño A., Rosso P. (2009) On the relevance of search space reduction in automatic plagiarism detection. In: Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN), num. 43, pp. 141-149

Barrón-Cedeño A., Vila M., Rosso P. (2010a) Detección automática de plagio: de la copia exacta a la paráfrasis. In: Panorama actual de la lingüística forense en el ámbito legal y policial: teoría y práctica. Jornadas (In)formativas de Lingüística Forense, Madrid, 21-22 October. Ed.: Euphonia Ediciones SL.

Barrón-Cedeño A., Rosso P., Agirre E., Labaka G. (2010b) Plagiarism detection across distant language pairs. In: Proc. of the 23rd International Conference on Computational Linguistics, COLING-2010, Beijing, China, August 23-27

Basile C., Benedetto D., Caglioti E., Cristadoro G., Degli Esposti M.. (2009). A plagiarism detection procedure in three steps: selection, matches and "squares". In Stein et al. (2009), pages 1–9. http://ceur-ws.org/Vol-502, pp. 19-23.

Boumalk A., Naït-Zerrad, K. (2009). Amawal n tjrrumt -Vocabulaire grammatical. Publications de l'IRCAM.

Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., & Mercer R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. Computational Linguistics, 19(2), 263–311.

Clough P. (2003). Old and new challenges in automatic plagiarism detection. National UK Plagiarism Advisory Service. http://ir.shef.ac.uk/cloughie/papers/pasplagiarism.pdf

Clough P., Gaizauskas R. (2009). Corpora and Text Re-Use. In Lüdeling, A., Kytö, M., and McEnery, T., editors, Handbook of Corpus Linguistics, Handbooks of Linguistics and Communication Science, pages 1249—1271. Mouton de Gruyter.

Comas R., Sureda J., editors (2008). Academic cyberplagiarism, volume 10 of Digithum. Universitat Oberta de Catalunya.

Dreher H. (2007). Automatic conceptual analysis for plagiarism detection. Journal of Issues in Informing Science and Information Technology 4, pages 601-614.

DuBay W.H. (2004). The principles of readability.Impact Information,

http://www.impact-information.com/impactinfo/readability02.pdf

Grozea C., Gehl C., Popescu M. (2009). ENCOPLOT: Pairwise sequences matching in linear applied to plagiarism detection. In Stein et al. (2009), pages 1–9. http://ceur-ws.org/Vol-502, pp. 10-18

IEEE. (2008) A plagiarismFAQ. http://www.ieee.org/web/publications/rights/plagiarism_FAQ.htm, 2008. [Online; accessed 3-March-2010].

Kang N., Gelbukh A., Han S. (2006). PPChecker: Plagiarism pattern checker in document copy detection.In Proc. of the Text, Speech and Dialogue, 10th Int. Conf. TSD-2006. LNAI(4188), pp. 661–667, Springer-Verlag.

Kasprzak J., Brandejs M., Křipač M. (2009). Finding plagiarism by evaluating document similarities. In Stein et al. (2009), pages 1–9. http://ceur-ws.org/Vol-502, pp. 24-28

KasprzakJ., Brandejs M. (2010). Improving the reliability of the plagiarism detection system - Lab report for PAN at CLEF 2010. . In: Braschler M., Harman D., and Pianta E.(Eds.), Notebook Papers of CLEF 2010 LABs and Workshops, CLEF-2010, Padua, Italy, September 22-23

Kulathuramaiyer N., Maurer H. (2007). Coping With the Copy-Paste-Syndrome. In E-Learn 2007, pages 1072—1079, Quebec, CA.

Kullback S., Leibler R. (1951). On Information and sufficiency. Annals of Mathematical Statistics, 22(1):79–86.

Lyon C., Barrett R., Malcolm J. (2006). Plagiarism is easy, but also easy to detect. Plagiary: Cross-Disciplinary Studies in Plagiarism, Fabrication, and Falsification, 1.

Malcolm J., Lane P.C.R. (2008). Efficient search for plagiarism on the web. Proc. of the Int. Conf. on Technology, Communication and Education, pp. 206-211.

Maurer H., Kappe F., Zaka B. (2006). Plagiarism - A Survey. Journal of Universal Computer Science, 12(8):1050–1084

Meyer zu Eißen S., Stein B. (2006). Intrinsic Plagiarism Detection. In Advances in Information Retrieval, Proc. of the 28th European Conf. on IR Research, ECIR 2006, LNCS(3936):565–569, Springer-Verlag.

Ottenstein K.J. (1976). An Algorithmic approach to the detection and prevention of plagiarism. ACM SIGCSE Bulletin, 8(4):30–41.

Outahajala M., Zenkouar L., Rosso P. (2011) Building an annotated corpus for Amazighe. 4ème atelier international sur l'amazighe et les TIC. Rabat.

Pinto D., Civera J., Barrón-Cedeño A., Juan A., Rosso P. (2009). A statistical approach to crosslingual natural language tasks. In: Journal of Algorithms, vol. 64, num. 1, pp. 51-60. DOI: 10.1016/j.jalgor.2009.02.005

Potthast M., Stein B., Eiselt A., Barrón-Cedeño A., Rosso P. Overview of the 1st International Competition on Plagiarism Detection. In Stein et al. (2009), pp. 1–9. URL http://ceur-ws.org/Vol-502.

Potthast, M., Stein, B., & Anderka, M. (2008). A Wikipedia-based multilingual retrieval model. In Advances in Information Retrieval, 30th European Conf. on IR research, ECIR 2008, Glasgow , LNCS(4956), pp. 522–530, Springer-Verlag.

Potthast M., Barrón-Cedeño A., Eiselt A., Stein B., Rosso P. (2010a). Overview of the 2nd International Competition on Plagiarism Detection. In: Braschler M., Harman D., and Pianta E.(Eds.), Notebook Papers of CLEF 2010 LABs and Workshops, CLEF-2010, Padua, Italy, September 22-23

Potthast M., Barrón-Cedeño A., Stein B., Rosso P. (2010b). An Evaluation Framework for Plagiarism Detection. In: Proc. of the 23rd International Conference on Computational Linguistics, COLing-2010, Beijing, China, August 23-27

Potthast M., Barrón-Cedeño A., Stein B., Rosso P. (2011). Cross-Language Plagiarism Detection. In: Languages Resources and Evaluation. Special Issue on

Plagiarism and Authorship Analysis, vol. 45, num. 1. DOI: 10.1007/s10579-009-9114-z

Scaife B. (2007). Evaluation of plagiarism detection software. Technical report, IT Consultancy,

Stamatatos E. (2009). Intrinsic plagiarism detection using character n-gram profiles. In Stein et al. (2009), pages 1–9. http://ceur-ws.org/Vol-502, pp. 38-46.

Stein B., Rosso P., Stamatatos E., Koppel M., Agirre E., Eds. SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 09), San Sebastian, Spain, 2009. CEUR-WS.org. http://ceur-ws.org/Vol-502.

Vallés E. (2010). Empresa 2.0: Detección de plagio y análisis de opiniones. M.Sc. thesis. Universidad Politécnica de Valencia.

Weber S. (2007). Das Google-Copy-Paste-Syndrom.Wie Netzplagiate Ausbildung und Wissen gefahrden.Telepolis.