Which granularity to bootstrap a multilingual method of document alignment: character N-grams or word N-grams?

Charlotte Lecluze, Loïs Rigouste, Emmanuel Giguet, Nadine Lucas

Which Granularity to Bootstrap a Multilingual Method of Document Alignment: Character N-grams or Word N-grams?

Charlotte Lecluze*, Loïs Rigouste, Emmanuel Giguet, Nadine Lucas

[a]*GREYC - CNRS UMR 6072 - Université de Caen Basse-Normandie, 14032 Caen Cedex*

Abstract

This article tackle multilingual automatic alignment. Alignment refers to the process by which segments that are translation of one another are automatically matched. Instead of comparing only pairs of languages at sentence level, as it is usually done to conform to human process in translation. The computer is used here for its capacity to infer semantic alignment from a collection of texts that are translations of the same content. The corpus contains press releases from Europa, the European Community website, available in up to 23 languages. The alignment process takes advantage of frequency similarity between different linguistic versions of a document by computing matching features for each repeated string in all versions. This is done to find reliable anchors in the process of linking versions. The question of the best granularity is raised to bring out some semantic equivalences, when comparing two linguistic versions, character N-grams or word N-grams. The alignment systems are traditionally based on word N-grams splitting. The observation of the morphological variety of languages, even inside a single linguistic family, quickly shows that the word granularity is inadequate to provide a widely multilingual system, i.e. a language independent system able to handle flexional languages as well as positional languages. Instead, when starting from a multilingual collection to focus on pairs of texts, we defend that character N-grams alignment is more efficient than word N-grams alignment.

Keywords: Corpus linguistic; Natural Language Processing (NLP); character N-grams based method; matching; alignment; multidocuments; multilinguism.

1. Introduction

International organizations issue information in different languages such as technical documentations, statutory documents, contractual documents, business information, press release... These various language versions of the same information have been carefully studied for several years in Natural Language Processing. They may be used as a basis to:

- learn how to translate new documents;
- populate multilingual terminology databases.

* Corresponding author. Tel.: +3-323-156-7398 ; fax: +3-323-156-7330.
E-mail address: charlotte.lecluze@unicaen.fr

In this study, we view the multidocument (Md) as a semantic unit. A multidocument gathers all the language versions of a single document.

It is worth noticing that when a reader closely observes multidocuments, some translation relations quickly appear to him, even without prior knowledge of the considered languages. These languages may even be very different genetically, in terms of written form dissimilarities and volume expansion.

Let us see for example those short extracts in French (fr), English (en), German (de) and Greek (el)[1] from one of our press releases.

fr Écouter, communiquer, agir au niveau local **-** Nouvelle approche de la Commission en matière de dialogue et de communication avec les citoyens européens

en Listen, Communicate, Go local – New Commission approach to dialogue and communication with European citizens

de Das neue Konzept der Kommission für die Kommunikation mit den europäischen Bürgern: Zuhören Kommuni zieren, Kontakte auf lokaler Ebene

el Ακούμε, επικοινωνούμε, προσαρμοζόμαστε τοπικά — η νέα προσέγγιση της Επιτροπής για τον διάλογο και την επικοινωνία με τους Ευρωπαίους πολίτες

Here, we quickly note the visual similarity of some words and punctuation marks (communiquer/communicate/ kommunikation approche/approach **-**/–) but also similarities of frequencies and positions of related words: . . . communiquer . . . communication . . . / . . . Communicate . . . communication . . . / . . . Kommunikation . . . Kommunizieren . . . / επικοινωνούμε . . . επικοινωνία . . .). There are therefore several clues, which, in combination, allow to infer semantic relationships, even considering an extreme case of a single sentence regardless of its context.

The increasing accessibility of those multidocuments opens the way to massive, weakly supervised reverse engineering operations on human-translated documents. Those practices indeed allow to extract linguistic informations and lexical resources for translators, lexicographers, linguists or terminologists.

Here we introduce the distinction that exists between matching and alignment (Kraïf, 2001), even if there is a continuum between them. Matching corresponds to equivalences listed in lexical resources. Alignment refers at translational equivalences observable in context. The first one is highly generalized while the other one is more contextual.

Alignment methods, especially applied to corpora made of documents and theirs translations, allow to draw links between equivalent semantic zones. These textual zones may be of different kinds: paragraphs, sentences, words. . . (Harris, 1988). Several trends exist in the field of alignment. Therefore, the section 2 is devoted to an overview of the main methods to date and particularly clues and granularity on which they are based. In section 3, we focus on character N-grams as an interesting tool. Eventually, in the section 4, we describe an experiment consisting of matching character strings from two languages.

2. State of the art and issue

Alignment methods aims at putting into correspondence units in translation relationship at a certain level of granularity: word, clause. . . Different applications are considered: crosslingual information retrieval, statistic machine translation, dictionaries extraction. . .

Automatic alignment methods strech over a continuum from all statistical (Gale & Church, 1993) to lexical (Chen, 1993), through hybrid methods, combining various clues such as length, frequency or lexical properties (Langlais, 1997). Historically, research primarily focused on sentence alignment methods. Today, the state-of-the-art in sentence alignment is considered to have reached a sufficient quality to tackle other granularity. Furthermore, sentence alignment is closely linked to word alignment and more generally to subsentencial unit alignment. These considerations have led to investigate methods with smaller granularity than sentence: words (Gale & Church, 1991, Tiedemann & Nygard, 2004), chunks (Church, 1988, Abney, 1991, Lepage et al., 2007), clauses (Nakamura-Delloye, 2007),. . .

[1] We use here the language codes defined in the standard ISO 639-1

The classical approaches assume three mains assumptions:

- the order of sentences is identical or very close: **Quasi-synchronization**
- texts contain neither deletion nor addition: **Quasi-bijection**
- alignments (1:1) are largely predominant: **Quasi-univocal**

Moreover, beyond the fact that they mostly aim to align at the word level, they are often based on graphical invariants (Gale & Church, 1993, Simard et al., 1993) and above all, on parallelism assumption.

These characteristics are operational objectives that do not fit with the diversity found in "real" corpus. Indeed, corpora used by those methods are sentences aligned corpora as Hansard, JRC-ACQUIS Communautaire, JEIDA. . . or to the best paragraphs aligned corpora (Salem and Zimina 2006, Brixtel 2007). Those corpora do not have textual dimension.

On the contrary, our corpus aims to be full of linguistic diversity and marked of the rewriting work that is translation. Thereby, our first goal is to propose to exceed the parallelism assumption widely exploited by the state-of-the-art and which is a lock for the automatic processing of real translated documents. Indeed, those documents could contain flaps or deletions of text areas.

Below, we present our research on the best granularity: character N-grams or word N-grams, not to be define in an adhoc way, but in context, how a document has been translated. This experiment aims at establishing in context the diagnosis of parallelism (as appropriate synchronous, asynchronous with inversion and asynchronous with deletion), that is to say if the translation is overall litteral or not between two documents. In others words, we aim to find what is translated and what is not at the text granularity (Fig. 1) and to find order differences starting at the text level (Fig. 2), before going done the lower levels.

As we said, alignment systems, as most information retrieval systems, generally rely on a word segmentation step. But what is a word? The word is generally described as a segment of speech between two blank spaces and/or punctuation marks. But, the visual word, with regard to Western European languages, covers different realities from a semantic point of view because this division came later through printing. In addition, all writing systems do not mark the boundaries of the word by white spaces, as they do not exist, for example, in Chinese.

To illustrate this fact, consider translations of the phrase "les transports en commun" in four european languages with a significant disparity of the definition of word: French (fr), English (en), Hungarian (hu), Finnish (fi):

fr : les transports en commun ⇒ 4 words
en : public transport ⇒ 2 words: no article and no preposition
hu : a tömegközlekedés ⇒ 2 words: word agglutination and no preposition
fi : joukkoliikenteen ⇒ 1 word: word agglutination

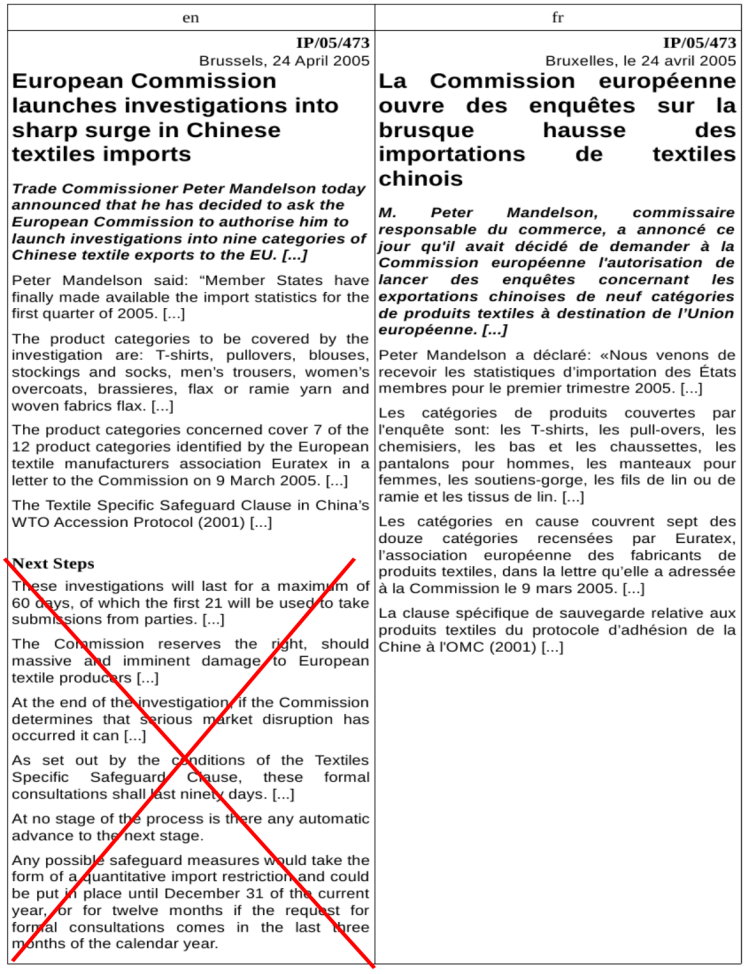
The answer to the question: "what is a word?" therefore seems to come from inner language properties. It depends on the morphological features of inflections and derivations of each language. Thus, the visual word is not universal enough to serve as a basis for an alignment and large-scale multilingual information retrieval system (Bender, 2009). This is even more true if the system is endogenous, i.e. resourceless.

According to our observations, a character N-grams split allows to identify interesting similarities across languages. The stake of our method is to show that character N-grams make languages more comparable revealing more repeated objects. The common factors highlighted by this method allow us to present a completely resourceless alignment method relies on text granularity.

3. Character N-grams

The idea of considering character n-grams rather than words has been used for author identification (Jardino, 2006), language identification (Cavnar et Trenkle, 1994; Dunning, 1994; Grefenstette, 1995; Sibun et Reynar, 1996) cited by Giguet (1998), speech analysis, text categorization (Damashek, 1995), numerical classification of multilingual documents (Biskri & Delisle, 2001) and information retrieval (Majumder, 2002, Mcnamee & Mayfield, 2004). However, to the best of our knowledge, there is only an attempt from Cromières (2006) to apply such a method to multilingual

alignment. Cromières particularly advises to use character level on asian languages, where the word is not easy to define, because, as we said, there is no special symbol to delimit word in many asian writing systems. As for western languages, Cromières has also applied his algorithm at the character level but only on a small corpus of bi-sentences from the Europarl, because of memory limitations.



en	fr
IP/05/473 Brussels, 24 April 2005	**IP/05/473** Bruxelles, le 24 avril 2005
European Commission launches investigations into sharp surge in Chinese textiles imports	**La Commission européenne ouvre des enquêtes sur la brusque hausse des importations de textiles chinois**
Trade Commissioner Peter Mandelson today announced that he has decided to ask the European Commission to authorise him to launch investigations into nine categories of Chinese textile exports to the EU. [...]	***M. Peter Mandelson, commissaire responsable du commerce, a annoncé ce jour qu'il avait décidé de demander à la Commission européenne l'autorisation de lancer des enquêtes concernant les exportations chinoises de neuf catégories de produits textiles à destination de l'Union européenne. [...]***
Peter Mandelson said: "Member States have finally made available the import statistics for the first quarter of 2005. [...]	Peter Mandelson a déclaré: «Nous venons de recevoir les statistiques d'importation des États membres pour le premier trimestre 2005. [...]
The product categories to be covered by the investigation are: T-shirts, pullovers, blouses, stockings and socks, men's trousers, women's overcoats, brassieres, flax or ramie yarn and woven fabrics flax. [...]	Les catégories de produits couvertes par l'enquête sont: les T-shirts, les pull-overs, les chemisiers, les bas et les chaussettes, les pantalons pour hommes, les manteaux pour femmes, les soutiens-gorge, les fils de lin ou de ramie et les tissus de lin. [...]
The product categories concerned cover 7 of the 12 product categories identified by the European textile manufacturers association Euratex in a letter to the Commission on 9 March 2005. [...]	Les catégories en cause couvrent sept des douze catégories recensées par Euratex, l'association européenne des fabricants de produits textiles, dans la lettre qu'elle a adressée à la Commission le 9 mars 2005. [...]
The Textile Specific Safeguard Clause in China's WTO Accession Protocol (2001) [...]	La clause spécifique de sauvegarde relative aux produits textiles du protocole d'adhésion de la Chine à l'OMC (2001) [...]
Next Steps	
These investigations will last for a maximum of 60 days, of which the first 21 will be used to take submissions from parties. [...]	
The Commission reserves the right, should massive and imminent damage to European textile producers [...]	
At the end of the investigation, if the Commission determines that serious market disruption has occurred it can [...]	
As set out by the conditions of the Textiles Specific Safeguard Clause, these formal consultations shall last ninety days. [...]	
At no stage of the process is there any automatic advance to the next stage.	
Any possible safeguard measures would take the form of a quantitative import restriction and could be put in place until December 31 of the current year, or for twelve months if the request for formal consultations comes in the last three months of the calendar year.	

Fig. 1. Example of deletion: The end part of the English version was not translated in the French version.

Indeed, in most of the NLP applications mentioned above, the number N of characters considered (N-grams) is defined beforehand and constant. In general, the character N-grams considered are bigrams or trigrams (4-grams or 5-grams in the case of Mcnamee & Mayfield (2004). On the contrary, we consider character N-grams[2] of various size: we consider all repeated strings of our corpus (frequency equal or greater than 2). Note that we are only interested in the repeated strings of maximal length, i.e. for a given repeated string, all substrings of same frequency (meaning that they are only observed in the context of the larger string) are discarded. This methodological choice relies on two assumptions: an inner language assumption and a crosslanguage assumption.

[2] We use N in a generic way, its value is not predefined.

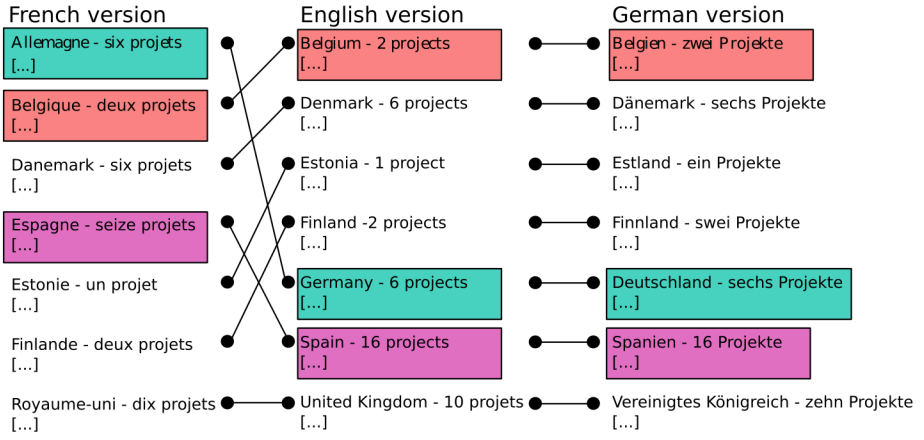


Fig. 2. Schematic view of the difference in order of the text areas between three versions of a press release listing alphabetically the European projects by country name.

3.1. Inner language assumption

For a given document in a language, there are more character N-grams repetitions than word n-gram repetitions, i.e. a character N-grams split highlights some common properties that a word n-gram split does not.

Example:

We seek repeated word N-grams in a French document sample:
fr: Donner aux collectivités les moyens de développer les transports en commun. La Commission européenne a adopté aujourd'hui une proposition révisée d'un règlement qui contribuera au développement de services publics de transport en commun.
⇒ 3 word N-grams are repeated. Here 1-gram (one word) or 2-grams (two words, like "en commun")

We seek repeated character N-grams (here, at least 3 characters, including spaces) in the same sample:
fr: Donner aux collectivités les moyens de développer les transports en commun. La Commission européenne a adopté aujourd'hui une proposition révisée d'un règlement qui contribuera au développement de services publics de transport en commun.
⇒ 5 character N-grams are repeated.

We seek repeated word N-grams in a document sample in Finnish:
fi: Paikallisviranomaisille tarjotaan keinot joukkoliikenteen kehittämiseen. Euroopan komissio hyväksyi tänään tarkistetun ehdotuksen asetukseksi jolla edistetään julkisten joukkoliikennepalvelujen kehittämistä.
⇒ 0 repeated word N-gram.

We seek repeated character N-grams in a document sample in Finnish (here, at least 3 characters, including spaces) in the same sample:
fi: Paikallisviranomaisille tarjotaan keinot joukkoliikenteen kehittämiseen. Euroopan komissio hyväksyi tänään tarkistetun ehdotuksen asetukseksi, jolla edistetään julkisten joukkoliikennepal velujen kehittämistä.
⇒ 6 character N-grams are repeated.

Beside the basic observation that there are more repeated character N-grams than repeated word N-grams, it is interesting to analyse their nature. It seems that a character N-grams split highlights more signifiers. Here, even if we had implemented a singular/plural processing for the word N-grams, we would have found the repetition 'transport'/ 'transports'. But we would have missed other interesting elements such as the derivation 'développer'/ 'développement'. To handle those cases, it is common practice to use more or less complex linguistic resources. However, this approach is

time-consuming in terms of construction, maintenance and extension of the system to new languages. The character N-grams split is cheeper in this respect.

3.2. Crosslanguage assumption

Among the repeated character N-grams, extracted thanks to monolingual common properties, those which have a similarity of distribution (frequency and positions) are semantic equivalents.

Character N-grams extraction highlights further interesting equivalences. Therefore, the repeated character N-grams extraction is not just a computer convenience, it is also an adequate tool for large scale comparison of multilingual corpora. Again, extracting the same semantic equivalences with exogenous approaches would require excellent and comprehensive resources for all languages considered, with a serious cost drawback. Then, the example below shows us both the differences and the similarities between the signifiers of a signified in several languages.

Table 1. List of graphical words meaning "transport" in a texts sample in French, Spanish and Greek, and their (frequency).

Language	Graphical words meaning "transport" and (their frequency)
French (fr)	transports (3), transport (3)
Spanish (es)	transporte (5), transportes (1)
Greek (el)	μεταφορών (3), μεταφορέας (1), μεταφορές (1), μεταφορέα (1)

Here, as the frequencies in the Table 1 show, the longer is the text sample, the striker is the word frequency gap problem. If we now pay attention to the character string repetitions (Table 2), we see that in each language there is a common substring among all the semantic equivalences of "transport".

Table 2. Character N-grams (of at least 3 characters) that are common to words meaning "transport" in the same texts sample in French, Spanish and Greek and their respective frequency.

Language	Repeated character N-grams meaning "transport"	Frequencies
French (fr)	transport- (3+3)	6
Spanish (es)	transporte- (5+1)	6
Greek (el)	μεταφορ- (3+1+1+1)	6

Starting from those substrings to establish matchings, and then alignments, appears to be a promising way to get a language comparing tool working on large volumes of data especially without having to maintain expensive resources. The frequency gaps between equivalent (or partially equivalent) words are softened. In the experience detailed below, we show the extraction of some equivalences. Our method is an associative approach. It first aims at extracting translation equivalences, i.e. matchings; before infering some links, i.e. alignments (contextual correspondences). This extraction will not be exhaustive on the document.

4. Character N-grams and matching

The above assumptions lead us to implement two different processes: a monolingual one and a multilingual one. The first allows extracting some common properties of segments of speech. In line with the work of Cromières (2006), we perform a search for N-grams in the context of repeated characters, i.e. *populations*. Populations are derived from an array of suffixes. They are obtained by computing patterns non-gapped character strings as described by Ukkonen (Ukkonen, 2009)[3]. These strings have the following characteristics:

- **repeated**: strings occur twice or more;

[3] The code for computing these strings is provided in *Python* at http://code.google.com/p/py-rstr-max/

- **maximal**: strings cannot be expanded to the left nor to the right without lowering the frequency.

These computational objects are not always directly interpretable by humans; they are over-generated. The second and multilingual process involves discarding these over-generated strings using frequency and position properties; it matches some semantic equivalences.

Thus, we compute the matching between strings of different languages, taking into account the similarity of distributions in all bi-documents in the collection. An example of distribution for two N-grams of characters is given in Table 3.

Table 3. Example of distributions of two characters N-grams in Greek and French. White space is represented by the character "_".

language	N-gram	frequency corpus	frequency by multidocument			
			doc_0	doc_1	[...]	doc_n
el	'_αερολμέν'	(23)	4	2	[...]	3
fr	'aéroports'	(21)	4	2	[...]	2

To limit the combinatorial explosion caused by an exhaustive comparison of all maximum repeated strings, we simply compare the frequency close strings. We use a normalized distance L1. This is done to both characters N-grams (s_1 et s_2) in two different languages, the ratio of the sum of frequency differences per document and the sum of the frequency of the two N-grams in the collection bi-documents in these languages are:

$$distance(s_1, s_2) = \frac{\sum_{doc} |effectif(s_1, doc) - effectif(s_2, doc)|}{effectif\_corpus(s_1) + effectif\_corpus(s_2)} \quad (1)$$

This distance calculation can produce characters N-grams populations matching with a distance within $[0, 1]$. A distance of 0 means that two N-grams have identical distributions in the corpus. This distance allows the computation of highly generalized correspondences in a collection of multidocuments. It makes the processing insensitive to differences in discourse order or local suppression of text areas between versions. We give some examples of matches and calculated in Table 4.

5. Results

From an extract of our system outputs, on the corpus presented below, we qualify the equivalences highlighted by a character N-grams matching. Then we insist on the semantic equivalences that cannot be found with a word n-gram extraction or using dictionaries.

This experiment has been conducted on a morphologically different language pair in order to show its wide applicability. For this article, we chose a set of 8 multidocuments in two languages among the European language family. We focus in this paper on the fr-el pair, a language pair from two linguistic groups (latin vs hellenic languages). This pair allows us to show that the method is alphabet-independent. In order to find some repetitions and then quickly establish matchings, we used a corpus of thematically close documents[4].

The alignment algorithm uses frequency and position properties in the collection. In Table 4, we show some examples of output of the system on our corpus. Each matching is presented on two lines of the table, each corresponding to one language and listing the following information: language, string, frequency and positions of occurrences in the collection as (Md number: offset). The offsets or text positions are here normalized by the document size, they can be read as percentages. 0 means first character of the text and 99 the end of the text. Thus the first occurrence of '2005' in greek is in the first percentage (1%) of the greek version of the first multidocument (1:).

The examples of matching character strings presented in Table 4 show that the matchings revealed by our method can go from less than a word to more full phrases. They can also concern some visual invariants such as date.

[4] In others experiences, we have used wider (around 40-50 Multidocuments) and non thematic collections of multidocuments.

Table 4. Examples of character N-grams matchings

el ’ 2005’ (11): 1:1% 2:1% 2:10% 3:1% 3:87% 4:1% 5:3% 6:1% 7:1% 7:92% 8:2%
fr ’ 2005’ (11): 1:1% 2:1% 2:9% 3:1% 3:88% 4:1% 5:2% 6:1% 7:1% 7:92% 8:2%
el ’ της παλαιστινιακής οικονομίας’ (3): 3:36% 3:63% 3:69%
fr ’ l’économie palestinienne’ (3): 3:36% 3:62% 3:69%
fr ’La Commission européenne ’ (6): 2:2% 2:4% 3:1% 6:3% 6:4% 8:6%
el ’H Ευρωπαϊκή Επιτροπή ’ (6): 2:2% 2:4% 3:1% 6:3% 6:5% 8:6%
fr ’ palestinienne’ (7): 3:34% 3:36% 3:51% 3:53% 3:56% 3:62% 3:70%
el ’αλαιστινιακή’ (7): 3:33% 3:36% 3:52% 3:54% 3:57% 3:63% 3:69%
el ’ αερολιμέν’ (23): 1:4% 1:10% 1:14% 1:18% 1:22% [. . .]1:92% 3:82% 3:99%
fr ’aéroports’ (21): 1:4% 1:10% 1:14% 1:18% 1:22% [. . .] 1:92% 3:82% 3:98%
fr ’ marché ’ (9): 1:34% 3:67% 6:7% 6:14% 6:30% 6:58% 6:63% 7:7% 7:35%
el ’ αγορά’ (9): 1:33% 3:68% 6:6% 6:14% 6:31% 6:60% 6:66% 7:6% 7:34%

Comparing positions of character N-grams is therefore an interesting method to find visual invariants, avoiding false cognates that a method based on string comparison would propose. We now examine the n-grams in context, to show which common factors repeated character N-grams revealed and that a word split would have missed.

In French, “aéroports” matched with “αερολιμέν-” [aérolimén] which illustrates well the link between the languages made by the process of character N-grams. In context, it corresponds to two different words in Md3: “αερολιμένες” and “αερολιμένας”.

It is also the case of the signified “marché” in french matched with its greek equivalent “αγορά” [agora] yet present in two different words just in Md6, “αγορά” and “αγοράς”; or even of the signified “palestinienne” beginning alternatively by an uppercase letter in the greek version of the Md3 which has also been matched with its greek equivalent “-αλαιστινιακή” [alestiniaki].

However, we must raise three limitations to the segmentation/alignment characters N-grams. These are solved through the implementation of a specific and/or adapted computer processing.

First, lexical or polylexical words where one or more letters change, in the case of diphthongization such as in the verb “Contar” in Spanish, at the first persons of the present: “Cuento”, “cuentas”, “cuenta” (i.e. skip-grams for Mcnamee and Mayfield (Mcnamee & Mayfield, 2004) or MFS, Maximum Frequent Sequences with the possibility of having a gap between the words of the sequence) Secondly, the risk to make semanticaly unmotivated links, between strings not related to the level of the word like “transmission” and “transparency” or “transport” and “transparency”, for instance.

Finally, the overgeneration of “uninteresting” repeated strings in the purpose of lexical resources construction by a alignement method. To assume that each character N-grams of a language can be aligned with any N-gram in another language we can find many associations but requires to lay down rules to cover this very large search space. We solved this problem by comparing the positions of N-grams with similar frequencies.

Conclusion

Despite their undeniable interest, character N-grams are not much used in alignment and creation of language resources. We think this is partly due to a reluctance to work on a material that does not make sense to humans. It is not natural for a human to validate an equivalence such as: fr “aéroports”/el “αερολιμέν-” even if it is correct in this case (to compare with fr “aéroport”/ el “αερολιμένας” in a dictionary). However, a postprocessing step to get words or phrases from character N-grams equivalence is not out of reach. This allows to fulfill evaluation needs and to use the resulting equivalences for tasks performed by humans. We believe that this postprocessing step to “return to words” is a small price to pay comparing to the added-value offered by the detection of new equivalences. In addition, the N-gram matchings can be used such as bootstraps for other automated phases. We apply this principle on alignment in an experiment of areas alignment. This experiment aims at establishing in context the diagnosis of parallelism (as

appropriate synchronous, asynchronous with inversion and asynchronous with deletion). It can reveal if the translation is overall litteral or not between two documents.

In a multidocument or in a set of multidocuments, some character N-grams are consistently translated into other character N-grams in other languages. We showed here that the extraction of these equivalences is not only feasible, but also inexpensive in terms of resources. Furthermore, these equivalences prove useful to draw links between documents, thereby providing a first reliable step to full multidocument alignment, even between morphologically different languages. Character N-grams split makes languages much more comparable.

References

Abney, S. P. (1991). Parsing by chunks. In S. A. Robert Berwick, & e. Carol Tenny (Eds.), *Principle-Based Parsing* (pp. 257–278). Dordrecht: Kluwer Academic Publishers.

Bender, E. M. (2009). Linguistically naïve! = language independent: Why NLP needs linguistic typology. In *ILCL'09, Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?, European chapter of the Association for Computational Linguistics (EACL).* (pp. 26–32). Athènes, Grèce.

Biskri, I., & Delisle, S. (2001). Les n-grams de caractères pour l'extraction de connaissances dans des bases de données textuelles multilingues. In *Actes de la 8ème conférence annuelle sur le Traitement Automatique des Langues Naturelles, 2-5 juillet* (pp. 93–102). Tours, France.

Chen, S. F. (1993). Aligning sentences in bilingual corpora using lexical information. In *Proceedings of the 31st annual meeting on Association for Computational Linguistics* (pp. 9–16). Columbus, Ohio: Association for Computational Linguistics.

Church, K. W. (1988). A stochastic parts program and noun phrase parser for unrestricted text. In *Proceedings of the second conference on Applied natural language processing* (pp. 136–143). Austin, Texas: Association for Computational Linguistics.

Cromières, F. (2006). Sub-sentential alignment using substring co-occurrence counts. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL* (pp. 13–18). Australia.

Damashek, M. (1995). Gauging similarity with n-Grams: Language-Independent categorization of text. *Science*, *267*, 843–848.

Gale, W. A., & Church, K. W. (1991). Identifying word correspondence in parallel texts. In *Proceedings of the workshop on Speech and Natural Language* (pp. 152–157). Pacific Grove, California: Association for Computational Linguistics.

Gale, W. A., & Church, K. W. (1993). A program for aligning sentences in bilingual corpora. *Comput. Linguist.*, *19*, 75–102.

Giguet, E. (1998). *Méthode pour l'analyse automatique de structures formelles sur documents multilingues*. Ph.D. thesis Université de Caen/Basse-Normandie Caen.

Harris, B. (1988). Bi-text, a new concept in translation theory. *Language Monthly (UK)*, *54*.

Jardino, M. (2006). Identification des auteurs de textes courts avec des n-grammes de caractères. In *Actes des 8es Journées internationales d'analyse statistique des Données Textuelles* (pp. 543–549). Besançon volume 1.

Kraïf, O. (2001). *Constitution et exploitation de bi-textes pour l'aide à la traduction*. Ph.D. thesis Université de Nice Sophia- Antipolis.

Langlais, P. (1997). Alignement de corpus bilingues : intérêts, algorithmes et évaluations. *Bulletin de Linguistique Appliquée et Générale*, *numéro Hors Série*, 245–254.

Lepage, Y., Migeot, J., & Guillerm, E. (2007). A measure of the number of true analogies between chunks in japanese. In Z. Vetulani, & H. Uszkoreit (Eds.), *LTC* (pp. 154–164). Springer volume 5603 of *Lecture Notes in Computer Science*.

Majumder, P. (2002). N-gram : a language independent approach to IR and NLP. In *Proceedings of the International Conference on Universal Knowledge and Language, November 25-29*. Goa, India.

Mcnamee, P., & Mayfield, J. (2004). Character N-Gram tokenization for european language text retrieval. *Information Retrieval*, *7*, 73–97.

Nakamura-Delloye, Y. (2007). Méthodes d'alignement des propositions : un défi aux traductions croisées. In *Actes de la 14ème conférence annuelle sur le Traitement Automatique des Langues Naturelles, June 12-15* (pp. 223–232). Toulouse, France.

Simard, M., Foster, G. F., & Isabelle, P. (1993). Using cognates to align sentences in bilingual corpora. In *Proceedings of the 1993 conference of the Centre for Advanced Studies on Collaborative research: distributed computing - Volume 2* (pp. 1071–1082). Canada.

Tiedemann, J., & Nygard, L. (2004). The OPUS corpus - parallel and free. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'2004)* Parallel corpora (pp. 1183–1186). Lisbon, Portugal.

Ukkonen, E. (2009). Maximal and minimal representations of gapped and non-gapped motifs of a string. *Theorie in Computer Science*, *410*, 4341–4349.