A tagged and aligned corpus for the study of Proper Names in translation

Emeline Lecuit
LLL
Université François Rabelais
France
emeline.lecuit@univ-tours.fr

Denis Maurel
LI
Université François Rabelais
France
denis.maurel@univ-tours.fr

Duško Vitas
Faculty of mathematics
University of Belgrade
Serbia
vitas@matf.bg.ac.rs

Abstract

In this paper, we propose the creation of a tagged and aligned corpus for the study of a linguistic phenomenon, the translation of proper names. We try to modify the hypothesis according to which proper names cannot be translated and should therefore appear as borrowings in a target-language. To do so, we introduce a parallel multilingual corpus made of eleven versions in ten different languages of a novel. One of these versions, the French one, which appears to be the source-text, undergoes named entity extraction so as to localize more easily the phenomenon we try to study. We focus on the tools used for the creation of our corpus and present some results refuting the idea that proper names are not translatable.

1 Introduction

The idea according to which proper names cannot be translated seems to be unfortunately widely spread. This can lead to big translation mistakes. Nevertheless it can easily be explained by a long tradition of presenting proper names as belonging to a linguistic category often defined using very reductive criteria which seem to have a very long life ahead of them. We have a different opinion and believe that proper names can be translated and are translated more often than people seem to think. We therefore introduce a multilingual corpus which will help us defend our idea. This corpus is created using several NLP tools, including cascade transducers for the extraction of named entities and an alignment tool, for the alignment of the eleven versions of the same text composing our corpus.

This study is therefore a good example both of the creation of an annotated multilingual corpus and of its usability.

In Section 2 we present the text(s) composing our corpus and the problem we try to tackle, giving details about what a proper name can be. In Section 3 we describe the extraction and annotation of the proper names in the French text of our corpus, using the Named Entity extractor CasEN. In section 4, the different steps for the creation of our multilingual corpus, using the alignment tool XAlign, are presented. Section 5 contains some preliminary answers to our translation problem and a conclusion.

2 A corpus for the observation of a linguistic phenomenon

When talking about translating proper names, a French person could argue: "Je m'appelle Paul et mon nom ne change pas si je me rends à Londres"[1], which is correct. But Paul's plane is going to land in *London*, and not *Londres*.

A lot of people believe that proper names are never translated. This idea, though widely spread and defended by many (from Moore[2] to Kleiber, 1981) can be discussed.

Our hypothesis is that proper names are, just like any other linguistic unit, subject to translation processes of all sorts (from borrowing to adaptation through calque and literal translation, etc.) when transferred from a text in source-language to a text in target-language (as demonstrated by Agafonov *et al.*,

1 "I am called Paul and my name doesn't change if I go to London."

2 See Ballard, 2001.

Proceedings of Recent Advances in Natural Language Processing, pages 11–18,
Hissar, Bulgaria, 12-14 September 2011.

2006). To defend our hypothesis we use a parallel multilingual corpus, built with different versions (i.e. in different languages) of the same novel, *Le Tour du Monde en quatre-vingts jours* (*Around the World in eighty days*), written by the famous French author Jules Verne, in 1872. The choice of this novel amongst others was motivated by two main reasons. Firstly, there exist lots of translations of this novel. Indeed, Verne's novel was translated in many languages and is nowadays available on the Internet in almost all European languages. Secondly, there is an important number of proper names of all sorts in this novel. This may be due to the fact that the novel deals with the adventures of Phileas Fogg, a rich and enigmatic English character who, after a bet with his fellows from the Reform-Club, has to go around the world in less than 80 days and who therefore travels through many countries and also happens to meet a lot of people. The novel references proper names belonging to almost all the existing categories and sub-categories of proper names.

Proper names can refer to people (or group of people) real or fictitious (we call these proper names anthroponyms), to places (toponyms), to human productions (ergonyms), or to events (pragmonyms). Though the common idea of a proper name is a simple lexical unit (in the form of a family name, for example), proper names can be complex lexical units, composed of several proper names and/or adjectives, common names, etc. Consider the following examples: *Passepartout* and l'*Institution royale de la Grande-Bretagne* (the *Royal Institution of Great-Britain*), both taken from the novel, though very different in structure, are proper names.

In our corpus, we gather eleven versions of the novel: starting from the original French version. We also have two English versions (by two different translators, at two very different periods and oriented towards two very different audiences[3]), as well as one version in German, one in Spanish, one in Italian, one in Portuguese, one in Serbian (using a Roman alphabet), one in Bulgarian, one in Polish and one in Greek. This variety of languages allows us to observe the phenomenon on languages belonging to different families. Once the different versions of the text gathered, we need to isolate the units we want to study in the French version of our text and to align the different versions to facilitate the study.

3 Annotation of the proper names using CasEN

To have a clearer view of the items we want to study, it seems a good idea to isolate them using a named entity extractor. We decided to use the resource CasEN (Friburger and Maurel, 2004), which uses the tool CasSys, which is now available on the well-known platform Unitex (Paumier, 2006)[4]. The CasSys system applies a series of finite-state transducers to a text. Each transducer describes a local grammar for the recognition of some entities. The result is a text in which the objects to be studied are marked with indicative tags. The transducer cascade can only be applied to texts which have undergone a preprocessing (division of the text into sentences, tagging using dictionaries, etc.). Only after this first stage the series of transducers can be applied (one after the other, in a defined order) to the text and locate the different contexts that can indicate the presence of the object looked for. In our case, the objects looked for are all kinds of proper names. The transducers we use are extracted from a list of transducers created for the French Ester campaign.

The objects we need to extract are basically persons, organizations and places. Once localized, these objects receive the following tags:

pers (*person*)
  pers.hum (*human*), pers.anim (*animal*)
org (*organization*)
  org.pol (*political*), org.edu (*educational*), org.com (*commercial*), org.non-profit (*non commercial*), org.div (*media and recreation*), org.gsp (*administrative*)
loc (*location*)
  loc.geo (*geographical*), loc.admi (*administrative*), loc.line, loc.fac (*facilities*)
  loc.addr (*address*), loc.addr.post, loc.addr.tel, loc.addr.elec

[3] Comparing these two versions will show us if the phenomenon of translation of proper names can be affected when these factors vary.

[4] For more information, see http://tln.li.univ-tours.fr/Tln_CasEN.html. The transducers are available for download from this website.

prod (*product*)
prod.vehicule, prod.award, prod.art, prod.doc

Figure 1 below is an example of transducer.

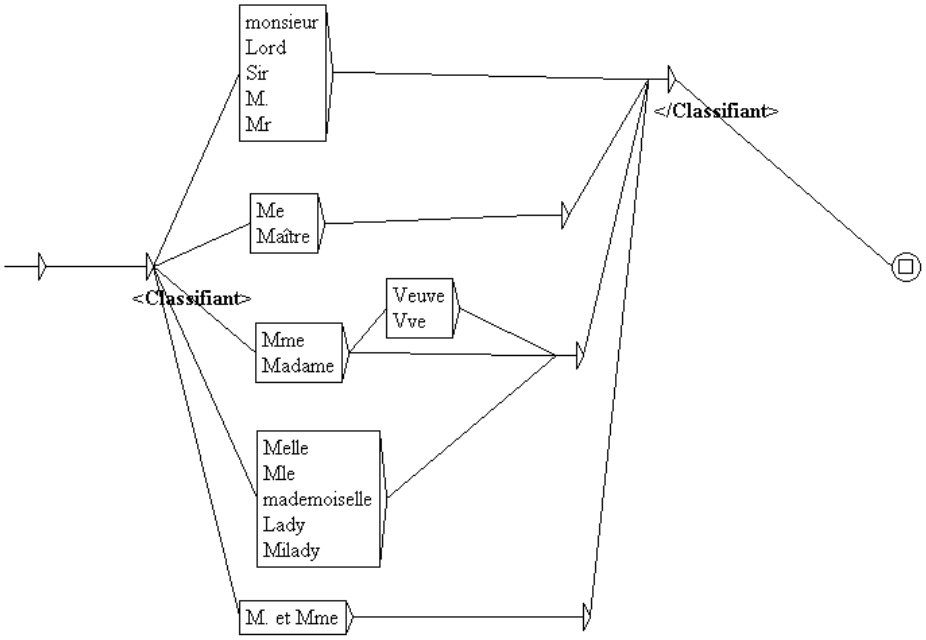

Figure 1: A transducer describing titles introducing person proper names

Let us illustrate the annotation of proper names in our corpus. When we apply the selected transducers (as explained above) to our French text; the input text:

> En l'année 1872, la maison portant le numéro 7 de Saville-row, Burlington Gardens - maison dans laquelle Sheridan mourut en 1814 - , était habitée par Phileas Fogg, esq., l'un des membres les plus singuliers et les plus remarqués du Reform-Club de Londres, bien qu'il semblât prendre à tâche de ne rien faire qui pût attirer l'attention.

becomes the tagged text:

> En l'année 1872, la maison portant le numéro 7 de **<ENT type="loc.line">**Saville-row **</ENT>**, **<ENT type="loc.line">** Burlington Gardens**</ENT>** -- maison dans laquelle **<ENT type="pers.hum">** Sheridan**</ENT>** mourut en 1814 --, était habitée par **<ENT type="pers.hum">**Phileas Fogg, esq. **</ENT>**, l'un des membres les plus singuliers et les plus remarqués du **<ENT type="org.div">**Reform-Club de Londres **</ENT>**, bien qu'il semblât prendre à tâche de ne rien faire qui pût attirer l'attention.

where each recognized proper name receives a tag indicating its category and sub-category.

After applying the cascade, a checking was carried out and some corrections were manually made (some tags were expanded or reduced, i.e. the brackets were moved to adjust to the entity, and others were deleted, added, or modified).

This first phase of our work provides us with a tagged text, in which all the proper names can be easily located using simple requests.

Our French version of the text comprises 3415 proper names (519 different). These proper names represent 8.6% of all the characters in the text and 8% of all the words in the text[5].

The following stage consists in aligning all the different versions of our text with the French one and all together.

[5] According to Coates-Stephens (1993), this figure can reach 10% in newspaper articles, which shows the importance of these units in texts and explains our involvement in this subject.

4 Alignement of the texts using XAlign

XAlign is a text aligner developed by the LORIA (2006) and available on the Unitex Platform. It combines the performances of an alignment tool to those of a well-know corpus processing system. One of the advantages offered by XAlign is the possibility to reuse an alignment already existing. This NLP tool allows the treatment of two texts at a time, which means that to obtain our multilingual corpus, we first have to align the texts two by two. In fact, we align the French text with all the other versions individually.

Prior to the alignment each translation is transformed into a TEI format and marked at a sentence, paragraph and division level with respectively <s>, <p> and tags. Id attributes are also added to the texts. All these markers will function as explicit anchor points which will help the alignment of the texts. Other potential anchor points, such as proper names, for example will also help the alignment. The alignment will extract the complete optimum path (following a pre-defined set of transitions, 1:1 equivalence, 1:2 equivalence, 2:1 equivalence, etc.). This alignment is represented as a double window, with one version of the text (in one language) on the left side and the other version of the text (in another language) on the right side. Between these two versions, red lines link the translation equivalents. The alignment is therefore visual and easy to consult (see Paumier and Dimitriu, 2008). Below is an example of alignment.

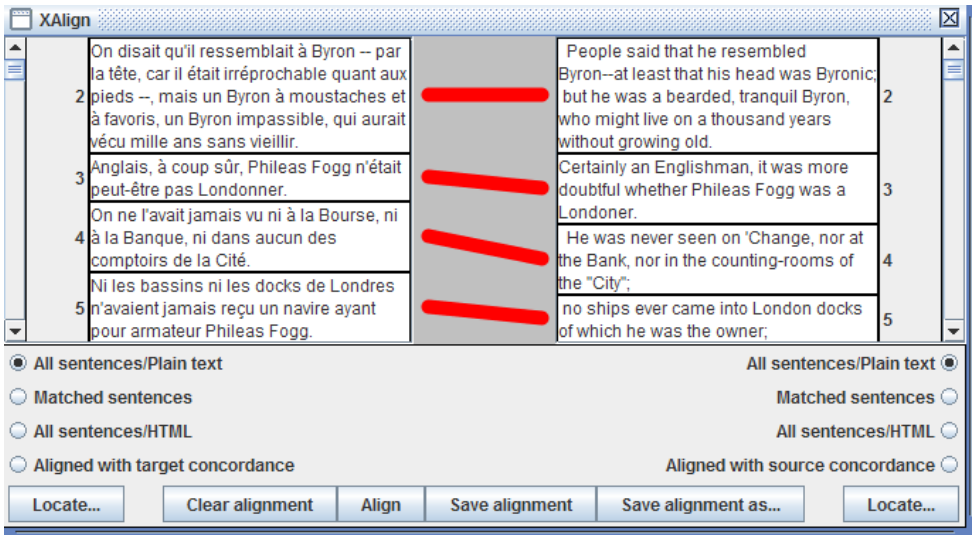

Figure 2 : Extract of an alignment using XAlign

This alignment will be saved as an alignment file in the XAlign directory in Unitex. The alignment file, in XML format, lists all the "linkings" and "alignments" between the two texts. The linkings correspond to links between two (or more) segments of one of the two versions, when the alignments are of type 1:2 or 2:1, for example, meaning that two segments in one version correspond to one segment in the other version or vice versa. The linking indicated in the alignment file (see Figure 3) means that the two segments will be considered as a whole.

```
<link targets="\\Private\Unitex2.0\French\Corpus\vern-fr-01-37-fixed.xml#d2p8s6 \\Private\Unitex2.0\French\Corpus\vern-fr-01-37-fixed.xml#d2p8s7" type="linking" xml:id="l1" />
```

Figure 3 : XAlign Alignment file (linking)

The alignments indicate, using the id codes applied during the preprocessing, equivalent segments in the first and second texts (see Figure 4).

```
<link
targets="\\Private\Unitex2.0\French\Corpus\ve
rn-fr-01-37-fixed.xml#d6p20s1 \\Private
\Unitex2.0\English\Corpus\vern-En-01-37-
fixed.xml#d6p20s1" type="alignment" />
```

Figure 4: XAlign alignment file (alignment)

One of the advantages offered by XAlign is that, because it is hosted by Unitex, it is quite easy to do requests on the texts, thanks to the option "XAlign Locate Pattern". Another advantage is that the alignment can easily be modified/corrected.

Now that we have created all our bitexts (alignments of the French text with the other versions individually), proofread and corrected them when needed, we can gather all the bitexts in one big multitext (alignment of all the texts). This is easily done manually. We obtain a big table, allowing us to visualize all the equivalent segments of our texts in the different languages. The table in Annexe is an extract of our Multilanguage corpus (It shows the first sentence of the text aligned in the different versions, with the French tagged version on the left side).

This tagged and aligned corpus allows us to carry on our study of proper names in translation.

5 Results and conclusion

We have created a tagged and aligned corpus for the study of a linguistic phenomenon. Tagged corpora and aligned corpora exist. What makes our corpus interesting is the high number of languages represented and the nature of the text used. Indeed, most multilingual corpora are made of versions of law texts (see for example multilingual corpora of the European Union law texts). Vaxelaire (2006) explains that choice of non-literary texts for the study of proper names because in literary texts "tous les types de noms propres peuvent être modifiés […]ou changés par des noms qui ne peuvent être considérés comme des équivalents que dans ce contexte précis[..].", which can be translated as follows : " all the types of proper names can be modified […] or changed into names which cannot be considered as equivalents except on this special occasion".We propose to study a novel. We will therefore study proper names translated by their equivalents but will also meet the case when a proper name is translated with names which cannot be considered as equivalents of translation. The novel we chose is a bit dated but this makes it available and free of use. Moreover, our corpus is extendable. Indeed, there are lots of other versions of the text not considered here which could easily be added to our corpus. We have already mentioned that our corpus is ideal for the study of proper names, since there are many of them in the text and of very various types, as can be observed in the table below (see Figure 5).

hypertypes	total number of occurrences	number of different occurrences
anthroponyms	2079	162
toponyms	1142	320
ergonyms	186	31
pragmonyms	8	6

Figure 5 : The proper names in the original version

Our study is still in progress. We only present here figures concerning 10% of the proper names of each of the types presented above, i.e. 16 anthroponyms, 32 toponyms, 3 ergonyms and 1 pragmonym. These samples are made of the most used items in each category. These 52 proper names represent 2029 occurrences in the French version, i.e. about 60% of all the proper names in the text.

Figure 6 is a table containing the results for these occurrences.

Target language	Borrowing	assimilation	Partial or total calque	Absence of translation	Other processes
English (1st version)	69,1%	11,3%	2,2%	12,1%	5,4%
English (2nd version)	74,2%	13,2%	2,0%	6,6%	4,1%
German	79,7%	10,6%	3,7%	5,2%	0,6%
Polish	31,1%	53,4%	4,5%	10,7%	0,2%
Serbian (Latin alphabet)	4,9%	89,1%	4,5%	0,3%	1,3%
Bulgarian	0,0%	90,6%	6,3%	2,5%	0,7%
Greek	0,0%	86,8%	4,6%	3,5%	5,1%
Italian	72,0%	21,5%	2,6%	3,0%	1,0%
Portuguese	73,7%	16,0%	5,9%	4,1%	0,2%
Spanish	51,6%	15,5%	25,1%	7,4%	0,4%

Figure 6: Translation processes (results)

What we can conclude from the study of these linguistic units is that the wide-spread hypothesis according to which proper names cannot be translated can be discussed.

Indeed, it appears that according to their type (fictitious or real proper names), according to their category (anthroponyms, toponyms, ergonyms, etc.), according to their use (as simple references, or as metaphors for example), according to their construction (simple or complex units), according to the target-language (sometimes implying different morphologic behaviors, sometimes using different alphabets, etc.), proper names can undergo a variety of translation processes. These phenomena are easily observable thanks to our corpus. Indeed, we can use the French tagged part of our corpus to identify a segment of the text containing a proper name. Then, on the same line of our table, we can visualize the translations of this proper name in the various different languages and analyse them.

If most proper names are simple borrowings from the source-text, as can be seen in Figure 6, many are subject to various assimilation (graphic and/or phonetic), as illustrated in the following example (for complete details about translation processes, see Vinay and Darbelnet, 2004).

FRA	ENG2	SPA	ITA
{ENT type'"pers.hum"}Mrs. **Aouda**{/ENT}, ne voulant pas être vue, se rejeta en arrière.	Not wishing to be seen, Mrs **Aouda** jumped back.	Mistress **Aouida**, no queriendo ser vista, se echó para atrás.	Mrs **Auda**, non volendo esser visita, si ritrasse indietro.

Figure 7 : Borrowings with graphic and/or phonetic assimilation

Our corpus also highlights the transcription processes (also accounted for in the "assimilation" column of Figure 6), which are not surprising in Bulgarian and Greek, both languages using a non Latin alphabet, but more striking in our Serbian (using a Latin alphabet) version. In this version, *Passepartout*, the name of the hero's manservant becomes *Paspartu*, for instance. Partial or total calques mentioned in Figure 6 (see Figure 8 for an example), mainly concern proper names which Jonasson (1994) described as "mixed" and "descriptive-based"[6] proper names, i.e. composed of "pure" proper names and/or other lexical elements, such as adjectives, common names, etc. The absences of translation, especially in the first English version and in

[6] "Mixtes" or " à base descriptives" in the original version.

the Polish version, mainly concern anthroponyms, which are replaced either by pronouns or defined descriptions.

The "other processes" are various: transpositions, free translations, to name just a few. The examples below (Figure 9, Figure 10) illustrate some of these translation techniques.

FRA	ENG	GREK	POR	POL	SPA
Vous allez à {ENT type"'loc.admi"}New York{/ENT}?	Are you going to New York?'	Πηγαίνετε στη Νέα Υόρκη;	Vai para Nova York?	- Jedzie pan do Nowego Jorku?	¿Vais a Nueva York?

Figure 8: Partial Calques from the English *New York* (except for the French, borrowing)

FRA	ENG	POL
Je suis un agent de la {ENT type"'org.com"}Compagnie péninsulaire{/ENT}. Litterally, the Peninsular Company	I work for P. and O.' For Peninsular and Oriental	- Jestem agentem Towarzystwa Morskiego Indii Wschodnich. Literally, the Society Maritime of Oriental India

Figure 9: Free translation (also called adaptation)

FRA	ENG	BUL	GER	POL	SRP
Canal de Suez	Suez Canal	Суецкия канал	Suez-Canal	Kanał Sueski (Sueski is an adjective)	Suecki kanal

Figure 10: Transposition (same semantic content but different syntactic structure)

All these examples, which are just a few of all the examples localized thanks to our corpus, seem to prove that it is wrong to promote the systematic use of borrowings when translating proper names.

References

Agafonov, Claire, Grass, Thierry, Maurel, Denis, Rossi-Gensane, Nathalie. and Agata Savary 2006. " La traduction multilingue des noms propres dans PROLEX ". *Méta*, vol.51, n°4, p.622-636.

Ballard, Michel 2001. *Le Nom propre en traduction*. Paris : Ophrys.

Coates-Stephens, Stephen 1993. " The analysis and acquisition of proper names for the understanding of free text ", in Computers and the Humanities, vol.26, p.441-456.

Friburger, Nathalie and Denis Maurel 2004. "Finite-state transducer cascades to extract named entities in texts", *Theoretical Computer Science*, vol.313, p.94-104.

Jonasson, Kerstin 1994. *Le nom propre: constructions et interpretations*. Louvain-la-Neuve : Duculot.

leiber, Georges 1981. *Problèmes de référence : descriptions définies et noms propres*. Paris : Klincksieck.

LORIA 2006. XAlign (Alignement multilingue). http://led.loria.fr/outils/ALIGN/align.html

Paumier S.ébastie,n 2006. *Unitex 2.0 User Manual*, http://igm.univ-mlv.fr/~unitex~/

Paumier Sébastien and Dumitriu Dana-Marina 2008. "Editable text alignments and powerful linguistic queries", in *Proceedings of the 27th Conference on Lexis and Grammar*, L'Aquila.

Vinay Jean-Paul and Darbelnet Jean 2004. "A Methodology for Translation", in *The Translation Studies Reader*. Venuti L. (eds) : Routledge.

Vitas, Duško, Koeva, Svetla, Krstev, Cvetana and Ivan Obradovic. 2008. “*Tour du monde* through the dictionaries”, in *Proceedings of the 27th Conference on Lexis and Grammar*, L’Aquila.

FRA-NP	BUL	ENG1	GER	GREK	POL	POR	SPA	ENG2	ITA
En l'année 1872, la maison portant le numéro 7 de {ENT type"'loc.line"}Saville-row{/ENT}, {ENT type"'loc.line"}Burlington Gardens{/ENT} -- maison dans laquelle {ENT type"'pers.hum"}Sheridan{/ENT} mourut en 1814 --, était habitée par {ENT type"'pers.hum"}Phileas Fogg, esq.{/ENT}, l'un des membres les plus singuliers et les plus remarqués du {ENT type"'org"}Reform-Club de Londres{/ENT}, bien qu'il semblât prendre à tâche de ne rien faire qui pût attirer l'attention.	През 1872 година в къщата на "Савил роу" № 7, Бърлингтън Гардънс – същата, в която през 1814 година почина Шеридан, – сега живееше Филиас Фог. ой беше един от най-странните и видни членове на Реформаторския лондонски клуб, въпреки че сякаш се стараеше да не привлича с нищо вниманието на другите.	Mr. Phileas Fogg lived, in 1872, at No. 7, Saville Row, Burlington Gardens, the house in which Sheridan died in 1814.He was one of the most noticeable members of the Reform Club, though he seemed always to avoid attracting attention;	Im Jahre 1872 wohnte in dem Hause Nummer 7, Saville-Row, Burlington Gardens, - worin Sheridan im Jahre 1814 starb, -- Phileas Fogg, Sq., eines der ausgezeichnetsten und hervorragendsten Mitglieder des Reformclubs zu London, der jedoch dem Anschein nach beflissen war nichts zu thun, was Aufsehen erregen konnte.	Το έτος 1872, στο σπίτι με τον αριθμό 7 της Σαβύλ-ρόου, στην περιοχή Μπέρλινγκτον Γκάρντενς -σπίτι στο οποίο πέθανε ο Σέρι-νταν το 1814- κατοικούσε ο αξιότιμος κύριος Φύλεας Φογκ. Ο κύριος αυτός ήταν ένα από τα πλέον ιδιόρρυθμα και επίλεκτα μέλη της Λέσχης της Μεταρρύθμισης του Λονδίνου, παρ' ότι φαινόταν πως προσπαθούσε να μην κάνει τίποτα που θα μπορούσε να κινήσει την προσοχή των άλλων.	W roku 1872 dom opatrzony numerem siódmym przy Saville Row w Burlington Gardens - dom, w którym w 1814 zmarł Sheridan * - zamieszkiwał szanowny pan Fileas Fogg, jeden z najwybitniejszych i najbardziej oryginalnych członków klubu "Reforma" w Londynie, choć ze swej strony zdawał się czynić wszystko, by nie zwracać na siebie uwagi.	Em 1872, a casa de número 7 da Saville Row, Burlington Gardens - casa em que Sheridan morrera em 1814 - era habitada por Phileas Fogg, esquire, um dos membros mais singulares e destacados do Reform Club de Londres, apesar de todo seu esforço em evitar, segundo parecia, chamar a atenção sobre si.	En el año 1872, la casa número 7 de Saville-Row, Burlington Gardens -- donde murió Sheridan en 1814- estaba habitada por Phileas Fogg, quien a pesar de que parecia haber tomado el partido de no hacer nada que pudiese llamar la atención, era uno de los miembros más notables y singulares del ReformClub de Londres.	IN the year 1872, No. 7 Savile Row, Burlington Gardens - the house where Sheridan died in 1814 - was occupied by Phileas Fogg, Esq. This gentleman was one of the most remarkable, and indeed most remarked upon, members of the Reform Club, although he seemed to go out of his way to do nothing that might attract any attention.	Nell'anno 1872 la casa numero sette di Saville Row, Bullington Gardens (la casa in cui morì Sheridan nel 1816), era abitata da Phileas Fogg, Esq., che sebbene sembrasse aver giurato di non far nulla che potesse dar nell'occhio, era uno dei più eccentrici e più noti soci del Reform Club di Londra.

Annexe : Extract of the multitext corpus