ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA

SCUOLA DI LINGUE E LETTERATURE, TRADUZIONE E INTERPRETAZIONE

SEDE DI FORLÌ

CORSO DI LAUREA IN

MEDIAZIONE LINGUISTICA INTERCULTURALE (Classe L-12)

ELABORATO FINALE

Web mining for translators: automatic construction of comparable, genre-driven corpora

CANDIDATO: RELATORE:

Simon Matthew Hoddinott Prof.ssa Silvia Bernardini

Anno Accademico 2015/2016

Primo Appello

Table of Contents

Abstract

The aim of this paper is to evaluate the efficacy of the application WebBootCaT to create specialised corpora automatically, investigating the translation of articles of association from Italian into English. The first section reflects on the relevant literature and proposes the utility of corpora for translators. The second section discusses the methodology employed, and the third section analyses the results obtained and comments on how language professionals could possibly exploit the application to its full. The fourth section provides a few concrete usage examples of the thus built corpora, to then conclude that WebBootCaT is a genuinely powerful tool that could be implemented by professional translators in order to save time and improve their translations in the long term.

1 Introduction

General language corpora have established themselves as an essential for most language professionals, but the impracticality of their construction means that they remain underexploited when dealing with specialised language. WebBootCaT, the application used throughout this paper, has revolutionised the corpus construction process, enabling users to create large, specialised corpora semi-automatically in a matter of hours. In this manner, translators can unleash the power of corpus linguistics in specific domains where previously only traditional sources were available.

The aim of this paper is to evaluate the effectiveness of WebBootCaT for automatic corpus construction and to examine how users can obtain the best results possible. In order to carry out my experiments, I examined how a translator might be expected to use automatically constructed corpora to translate the specialised language of articles of association; a definition of this domain is provided in 1.2. The first section reflects on the literature concerning automatic corpus construction and proposes the utility of corpora for translators, and subsequently illustrates the difference between the topic-driven and genre-driven approach in automatic corpus construction. The second section discusses the methodology used during the experiment and explains the rationale behind the methodology. The third section analyses the results obtained and comments on how language professionals can exploit the application to its full. The fourth section is aimed at providing a few concrete examples of how such corpora can be useful for translators. The paper then concludes that the WebBootCaT method is a powerful tool that could be implemented by professional translators in order to save time and improve their translations in the long term.

1.1 *Automatic construction of corpora*

The idea of constructing corpora automatically was first proposed by Baroni and Bernardini (2004) and consists in "bootstrapping" or piggybacking on Internet search engines in order to harvest the results of the queries. The procedure is relatively straightforward: a user inputs a number of seed terms which are sent to a search engine as queries; the identified hits are then automatically gathered, cleaned, de-duplicated and processed, resulting in a corpus. The corpus can then be enlarged by extracting the new seed terms from this first-pass corpus and using them in a new query; this final step can be repeated numerous times to create very large corpora in a short space of time.

1.1.1 *The WebBootCaT application*

The programme originally designed to perform this process is BootCaT,[1] which is currently available as a front-end version for desktop computers. Later, it was integrated into the Sketch Engine,[2] an online corpus software interface, where it assumed the name of WebBootCaT (Baroni et al., 2006). This paper makes use of WebBootCaT and not BootCaT solely because the current front-end version of BootCaT only supports html files, whilst WebBootCaT supports html, pdf, plain text and docx files, and the texts I investigated are typically published as pdf files. As we will see subsequently, the Sketch Engine also offers some powerful corpus query tools, such as its signature word sketches. Both BootCaT and WebBootCaT currently use the search engine Bing. BootCaT is entirely free of charge, requiring the user only to register for a Bing API, which is equally free until up to 5,000 queries per month. At the time of writing, the Sketch Engine provides a 30-day trial subscription, which allows users to store up to 1 million words on their account, after which they are required to pay for more storage.

1.2 *Definition of articles of association*

In order to interpret this paper correctly, it is necessary to have some awareness of what articles of association are. In sum, articles of association are a set of articles that form a document that governs how companies limited by shares are run. It sets forth the rights, duties, liabilities and powers of directors and shareholders and lays out the provisions regarding the proceedings at shareholders' meetings and the company's share capital (Cambridge Business English Dictionary).[3]

In the UK and Ireland, this document is referred to as a company's articles of association. In Italy it is referred to as *statuto sociale* or *statuto societario*. In North America they are called bylaws; in New Zealand and Australia the document is called a constitution and in South Africa it is called a memorandum of incorporation.

Whereas articles of association are seldom translated from English into Italian, very often Italian companies limited by shares have their articles of association translated into English so that foreign shareholders can understand the framework of rules that govern the company's meetings and shares.

[1] http://bootcat.sslmit.unibo.it/?section=home

[2] https://www.sketchengine.co.uk/

[3] http://dictionary.cambridge.org/dictionary/english/articles-of-association

The rationale behind examining articles of association and their translation is twofold. Firstly, it is a sector in which I have a certain degree of expertise, giving me a yardstick against which to measure the results produced by the corpora. Secondly, the language of articles of association is highly specialised and conventionalised, making an examination into its amenability to automatic corpus construction particularly interesting. Nevertheless, the choice to study articles of association is fundamentally arbitrary; this paper will try to show that it should be relatively easy to create corpora semi-automatically for any genre.

1.3 *Making the case for corpora*

Before the dawn of corpus linguistics, when dealing with specialised language, the translator could essentially rely on four resources, as put forward by Bowker and Pearson (2002, p. 14-18): dictionaries, printed texts, subject field experts and the translator's own intuition.

Ordinary dictionaries, be they bilingual or monolingual, typically do not contain relevant information about the genre. Specialised dictionaries, if they exist at all, will almost inevitably be expensive, out of date or hard to obtain (Baroni et al., 2006) and possibly unreliable or uninformative. "Printed texts" may be construed to mean any source of written information, such as encyclopedias and text books, or any sort of parallel text, in the sense of texts of the same genre written by mother-tongue experts in a non-translation context. Basing one's translatory decisions on such texts is arguably the best possible solution, but it is of course very impractical, considering the time it requires; even after having read thousands of pages, the information retrieved would represent only a fraction of the entire language subset. Consulting "subject field experts" presents analogous issues and would entail extortionate costs. The translator's own intuition is generally a reliable tool as far as general language and grammaticality is concerned, but when confronted with specialised language, it is often inaccurate (Bowker & Pearson, 2002; Reppen, 2010).

Corpus linguistics and corpus analysis tools can however remedy this predicament, allowing translators to sift through thousands of texts instantaneously and subsequently make their decisions on the basis of authentic examples of language that are considered to be almost entirely representative of the genre in question; in so doing, translators can verify their intuition (Bowker & Pearson, 2002) or "reassure" themselves (Varantola, 2003). In this respect, corpora aren't the be-all and end-all of translation, but they can certainly be considered a very powerful complementary resource. In comparison to dictionaries, corpora provide a veritable plethora of information: a word's collocational and colligational behaviour and its (relative) frequency; automatic keyword and terminology extraction;

information about phraseology and genre-specific conventions. Corpora also serve as a source of “distilled expert knowledge” (Bowker & Pearson, 2002), giving the translator valuable insights not only into purely linguistic aspects of the genre but also into relevant conceptual information.

Corpora present however one seemingly ineliminable shortfall: generally, the time required to compile a meaningfully representative corpus is disproportionate to the time at disposal for a given translation assignment. Whereas general langauge corpora, being readily available on the Internet, have established themselves as an essential element of the translator’s toolkit, specialised corpora seem to remain underexploited by translators. Apart from a few domains such as English for academic purposes, specialised corpora are simply not available and therefore need to be constructed by the translator ad hoc. However, this trade-off between the practicality and representativeness of corpora undestandably causes language professionals to shirk from constructing specialised corpora independently (Bernardini & Ferraresi, 2013; Reppen, 2010).

1.4 *Making the case for automatically constructed corpora*

With the above in mind, being able to construct corpora automatically would theoretically liberate corpus linguistics from its greatest Achilles heel: its poor practicality. Being automatic, translators no longer have to sacrifice their limited time gathering texts, thus pushing specialised corpora upwards along the cline of practicality. In turn, their automatic nature would allow the corpora to reach considerably larger sizes, pushing them up along the cline of representativeness. Since ideally users do not assess the texts before including them when creating corpora automatically, one could argue that this relative loss of control compromises the representativeness of the corpus population. However, both BootCaT and WebBootCaT provide means of discarding texts before adding them to the corpus; naturally each discarded text requires additional human intervention, but this feature potentially allows the corpus’ greater representativeness to remain unscathed.

In the light of this, if we were to ignore the issue of possible copyright infringements, the BootCaT method can create not only “ad hoc” or “disposable” corpora (Varantola, 2003) but also reliable corpora that could be used repeatedly, in the same vein as the WaCky web-crawled corpora (Bernardini, Baroni, & Evert, 2013).

1.5 *Genre and topic in corpus linguistics*

1.5.1 *Definition of genre and topic*

The notion of genre is highly contested among scholars; for the purposes of this paper, we shall only take into account the implications of genre that may be relevant to corpus construction. In sum, all texts can be said to belong to a particular genre, whereby different genres have different degrees of specificity with a varying set of lexical, rhetorical and structural conventions. Topic on the other hand may be interpreted as the general topic of discussion. In our case, the topic can be said to be company law, whereas the various underlying genres range from articles of association to proxy forms, management reports, notices of meeting, government legislation regarding company law, websites and handbooks containing information on how to set up a company, etc.

Much of the literature on genre analysis is based on Swales' definition of genre (1990, p. 58), but I believe Bhatia provides an interesting and more succinct insight:

> Genre essentially refers to language use in a *conventionalised* communicative setting in order to give expression to a specific set of *communicative goals* of a disciplinary or social institution, which give rise to stable structural forms by imposing constrains on the use of lexicogrammatical as we all discoursal resources. (Bhatia, 2004, pp. 23, my italics)

Here Bhatia underlines how extra-linguistic factors (the "communicative goals") necessarily shape the linguistic output of the author. If we were to analyse articles of association in terms of genre, we could conclude that their communicative goal is to inform shareholders how a company is run; articles of association are also a performative linguistic act, whereby the provisions set forth enter into force as soon as the articles of association are approved. By way of comparison, the communicative goal of a web-guide for company law is radically different and entails merely advising readers about setting up a company, void of any performative nature. Despite their distinct extra-linguistic characteristics, these two genres share an almost identical semantic field and therefore belong to the same topic.

1.5.2 *Genre-driven and topic-driven corpus construction*

1.5.2.1 The topic-driven approach

When introduced, the architects of BootCaT proposed a method that is now termed the "topic-driven" corpus construction pipeline, in that it is more suitable for retrieving texts belonging to a common topic as opposed to a common genre. This first-generation pipeline suggested that the user input a small list of unigram seed terms that are expected to be characteristic of the domain in question.[4] The authors then proposed to extract the most frequent keywords from the newly created first-pass corpus to then use them as new seeds, which, in light of their keyness, would be even more effective in comparison to the user-selected seeds. Although the results using this method were promising, even my personal first attempts at using unigram terms, be they keywords or user-defined, produced a large amount of noise (Baroni & Bernardini, 2004).

It seems as though keywords are only capable of reflecting the lexical aspects of a text; this incapability of reflecting extra-linguistic features means that keywords are unable to retrieve texts belonging to the same genre effectively, because they will always attract noise from different genres which however share a common topic.

1.5.2.2 The genre-driven approach

In a revision of the BootCaT method in (Bernardini & Ferraresi, 2013), the authors proposed a new "naive" method termed genre-driven corpus construction, which avoids unigram terms and opts for the most frequent n-grams of the corpus population as seed terms, "regardless of their being intuitively salient, syntactically complete, or *lexically rich*" (ibid. my italics). Despite doing away with lexical richness as well as keyness, seeing that n-grams are extracted locally, the results appear to be more auspicious.

The reason why n-grams are more effective is because they seem to be more capable of reflecting a genre's specific conventions, capturing the genre's phraseology and characteristic "turns of phrase". In this respect, Bernardini and Ferraresi (ibid.) support this theory by citing Biber and Conrad's use of "lexical bundles" to distinguish variations in register in conversation and academic prose (Biber & Conrad, 2011), also termed "lexical clusters" in other studies. Similarly, Greaves and Warren (2010) cite how Biber et al. (1999), Carter and McCarthy (2006) and Hyland (2008) have "all found that the analysis of n-grams in a register or genre affords an important means of differentiation."

Using n-grams as seed terms therefore allows us to refine the granularity of the queries in WebBootCaT, adjusting the metaphorical net with which we trawl on the Web in order to ignore irrelevant genres. This phenomenon can be explained by relating it to the notion of genre. Identifying and respectively retrieving a text according to its topic is relatively easy, because topic accounts solely for the lexical peculiarities of a text. It is so

[4] It was also suggested that users could extract the most frequent keywords from a relevant Wikipedia article and use those as seed terms.

easy that a search engine can perform the process very well. Conversely, identifying and retrieving a text according to its genre is much more complicated, because as described in 1.5.1, genre is also characterised by extra-linguistic factors. Search engines are capable of processing a restricted range of extra-textual information such as a text's publication date, the host server's location, file format, length and language, but they are not (yet) capable of recognising or processing implicit extra-linguistic features such as a text's purpose, addresser or addressee, which are all fundamental elements of genre analysis. Therefore, when using a computational approach, in order to trick the search engine into finding the right given genre, we must adapt our purely linguistic queries (we are after all dealing with words) so that they reflect the extra-linguistic aspects that characterise the text's genre (cf. Bernardini & Ferraresi, 2013). As stated above, n-grams often contain the linguistic expression of these extra-linguistic features.

2 Methodology

2.1 *Methodology outline*

In order to assess the effectiveness of WebBootCaT, I tried to simulate as realistically as possible how a translator may be expected to use it, that is, seeking a large, "quick-and-dirty" corpus, employing minimum effort. On WebBootCaT the user can determine three main parameters, all of which will have varying impact on recall and precision, namely: number of seeds, tuple length, type of seeds. My analysis consisted in adjusting these three parameters to see what particular combination might be most effective for the present genre. As well as the abovementioned parameters users can also input whitelist and blacklist words, which have effect not upon querying but only when processing the URLs; initial trials established that the effect of whitelists and blacklists is quite hit-and-miss, so I left them out of my experiment.

2.2 *Constructing the manual corpora for seed selection*

The first step of my analysis was to manually construct a small ad hoc corpus and then extract the keywords, key terms and n-grams to be used as seeds. To do this I downloaded 15 articles of association in English and Italian from the Internet. For the Italian articles of association, I referred to the Wikipedia article "Lista delle maggiori aziende italiane per fatturato", selecting 15 companies with no particular preference. I repeated the process to gather the English articles of association, referring to the following Wikipedia articles: "FTSE 100 Index", "List of largest public copmanies in Canada by profit", "Fortune 500", "S&P/ASX 20" and "NZX 50 Index"; I attempted to gather an equal number of articles of association from each country (the above articles correspond to the UK, Canada, USA, Australia and New Zealand) in order to include a greater variety of language.

After downloading the articles of association, I used Anthony Laurence's AntFileConverter[5] to convert the pdf files into plain text files. I then uploaded them on the Sketch Engine as a zip file and compiled the corpus. The first observation to be made is that the English manual corpus contained 331,188 words, whereas the Italian manual corpus contained only 105,253. One could say that this is not a good start in terms of corpus comparability, but this is simply due to the fact that Italian articles of association are shorter than their English-language counterparts and necessarily differ slightly in terms of content.

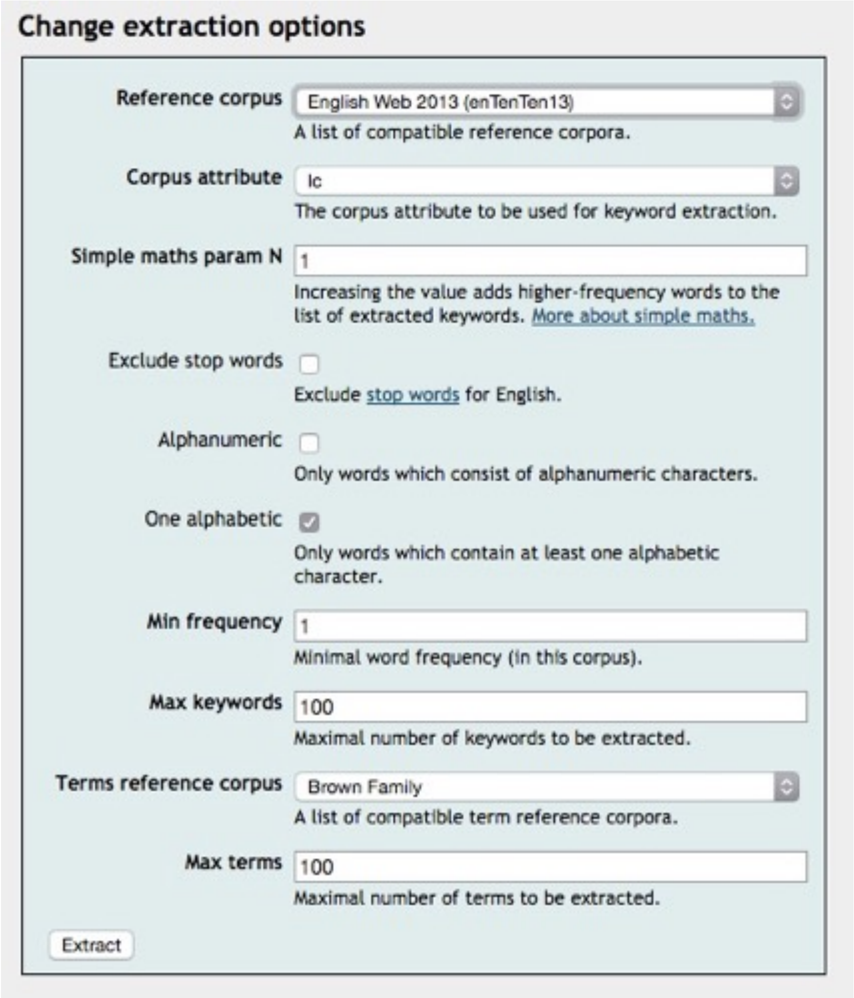
The next step was to extract the seed terms. As can be observed in the tables containing the keywords in Appendices A and B, I did not clean the texts before importing them into WebBootCaT. The 30th keyword in the Italian manual corpus is "blea", evidently owing to frequent linebreaks in the word "assemblea". The 32nd is "stratori", the broken form of "amministratori". The 18th keyword in the English manual corpus is "lon15010141", probably the name of a file used in a header or footer. As regards proper nouns, (which wouldn't have been removed even if I had cleaned the files) the 29th keyword in the Italian manual corpus is "Bancoposta", followed by the 43rd keyword "Pirelli". Whenever I used keywords in my queries, I simply skipped these aberrant items, making the 31st item become the new 30th and so forth.

2.2.1 *Keyword and key term extraction*

Under the "manage corpus" menu item on the Sketch Engine, one can have access to all the features of an ordinary corpus analysis tool. The menu item "Keywords/terms" under "Search corpus" allows users to extract keywords and multiword key terms from a given corpus, with the option to choose from a variety of pre-loaded reference corpora. To extract the keywords, I used the standard corpus "English Web 2013 (enTenTen13)" and all the other standard parameters. See Figure 1 for an example of the extraction options. On this page the Sketch Engine gives users the opportunity to select keywords or key terms with checkboxes and use them as seeds in a WebBootCaT run, intended to speed up the entire web-mining process; see Figure 2 for an example. Interestingly, this feature is not available on the n-gram extraction page.

[5] http://www.laurenceanthony.net/software/antfileconverter/

To extract the key terms, I likewise used all the standard parameters. I attempted to use the enTenTen13 corpus as a reference corpus, but its size (almost 20 billion words) caused the process to be very slow, to the extent that I aborted the attempt and reverted to the standard Brown Family corpus. The reference corpus used to extract keywords from the Italian manual corpus was the itTenTen 2010 corpus; to extract key terms, a sample of the same corpus was used as reference corpus. See Appendices A and B for the tables containing keyword and key term lists for each corpus.



Change extraction options

Reference corpus: English Web 2013 (enTenTen13)
A list of compatible reference corpora.

Corpus attribute: lc
The corpus attribute to be used for keyword extraction.

Simple maths param N: 1
Increasing the value adds higher-frequency words to the list of extracted keywords. More about simple maths.

Exclude stop words □
Exclude stop words for English.

Alphanumeric □
Only words which consist of alphanumeric characters.

One alphabetic ☑
Only words which contain at least one alphabetic character.

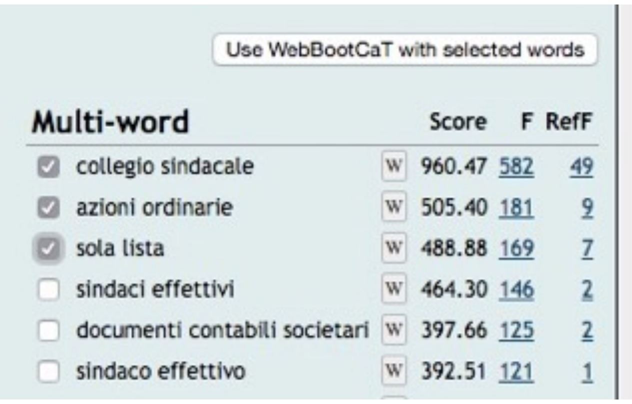
Min frequency: 1
Minimal word frequency (in this corpus).

Max keywords: 100
Maximal number of keywords to be extracted.

Terms reference corpus: Brown Family
A list of compatible term reference corpora.

Max terms: 100
Maximal number of terms to be extracted.

Figure 1. *"Change extraction options" pane on the keyword and multiword term extraction page*

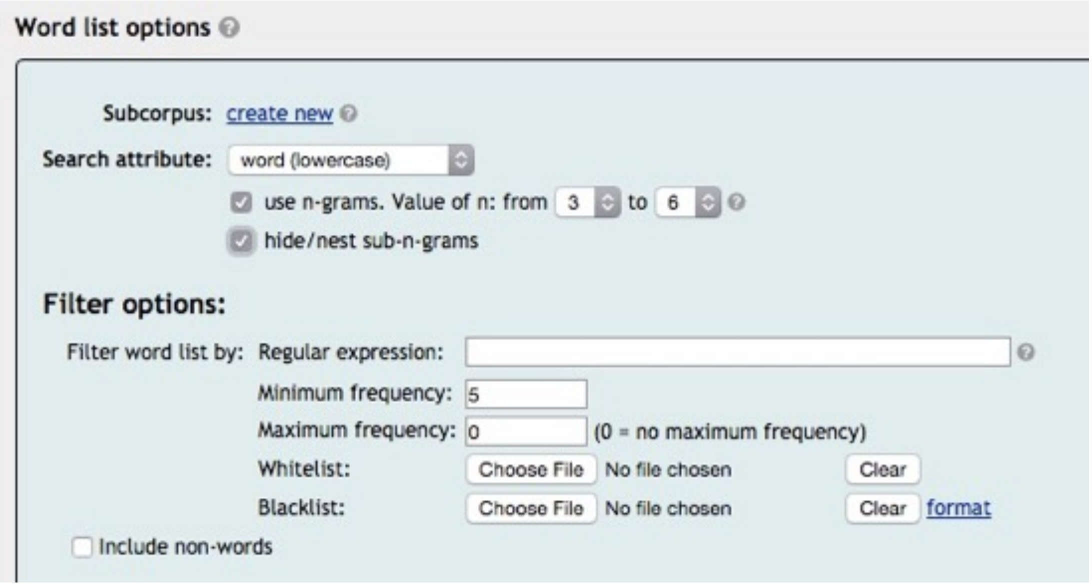


Multi-word		Score	F	RefF
☑ collegio sindacale	W	960.47	582	49
☑ azioni ordinarie	W	505.40	181	9
☑ sola lista	W	488.88	169	7
□ sindaci effettivi	W	464.30	146	2
□ documenti contabili societari	W	397.66	125	2
□ sindaco effettivo	W	392.51	121	1

Figure 2. Example of checkboxes and hyperlink to WebBootCaT

2.2.2 *N-gram extraction*

Under the menu item "Word List" users can create word lists and extract n-grams or keywords. Figure 3 shows the word list and n-gram creation pane with the settings I used to create mine: I chose to treat all data as lowercase (not the standard setting), searching for n-grams from 3 to 6 words in length, activating the option to nest sub-n-grams; an example of nested sub-n-grams is shown in Figure 4. The rationale for choosing nested n-grams was because otherwise the most frequent n-grams would have represented smaller chunks of the same lexical cluster, whereby these smaller chunks would essentially all pivot on the same element or "turn of phrase" in a given text; i.e. the n-grams "the chairman of", "chairman of the", "of the meeting", "the chairman of the" and "chairman of the meeting" are all contained in the five-gram "the chairman of the meeting". Moreover, using nested n-grams makes sure that any given lexical cluster is only represented once, allowing the query as a whole to pivot on a greater variety of clusters and thereby theoretically retrieving fewer duplicates. See Appendix C for the lists of n-grams of each corpus.



Word list options

Subcorpus: create new

Search attribute: word (lowercase)

☑ use n-grams. Value of n: from 3 to 6

☑ hide/nest sub-n-grams

Filter options:

Filter word list by: Regular expression:

Minimum frequency: 5

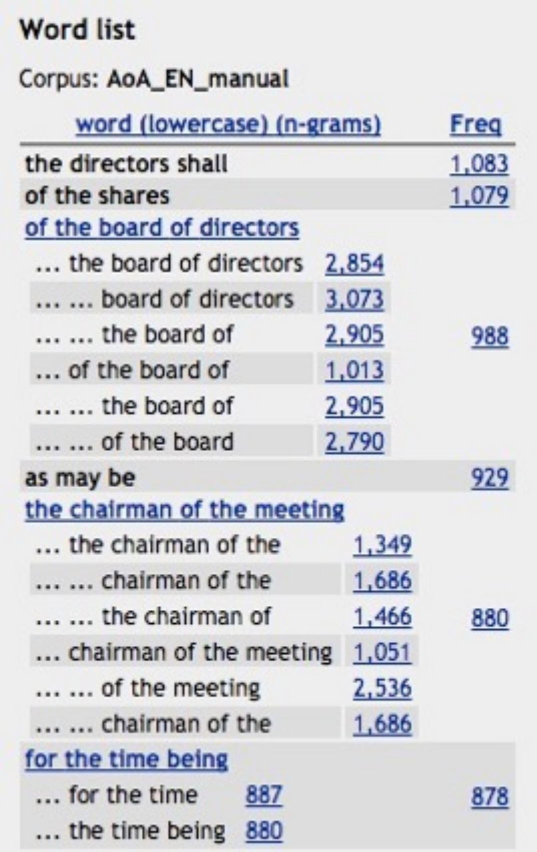
Maximum frequency: 0 (0 = no maximum frequency)

Whitelist: Choose File No file chosen Clear

Blacklist: Choose File No file chosen Clear format

□ Include non-words

Figure 3. Word list creation pane



Word list

Corpus: **AoA_EN_manual**

word (lowercase) (n-grams)		Freq
the directors shall		1,083
of the shares		1,079
of the board of directors		988
… the board of directors	2,854	
… … board of directors	3,073	
… … the board of	2,905	
… of the board of	1,013	
… … the board of	2,905	
… … of the board	2,790	
as may be		929
the chairman of the meeting		880
… the chairman of the	1,349	
… … chairman of the	1,686	
… … the chairman of	1,466	
… chairman of the meeting	1,051	
… … of the meeting	2,536	
… … chairman of the	1,686	
for the time being		878
… for the time	887	
… the time being	880	

Figure 4. N-gram list pane showing n-grams with sub-n-grams unhidden

2.3 *Determining the parameters*

2.3.1 *Number of seeds*

As far as the number of seeds is concerned, unlike Bernardini and Ferraresi (2013) and Dalan (2013), I decided to use a small seed set composed of 15 seeds. Bernardini and Ferraresi used

43 and 45 keywords respectively for their topic-driven corpora and 41 and 46 trigrams respectively for their genre-driven corpora. Dalan used 28 keywords and 28 trigrams in both her corpora for both languages. Both of these experiments used 10 tuples with a tuple length of 3, which meant that it was possible for a maximum of 30 individual seeds to be represented in the query, whereby "query" is to be understood as the whole series of tuples sent to Bing.

Given that the tuples are generated randomly, when searching with 10 tuples and tuple length set at 3, in a set of 30 seeds, it is almost certain that one seed will not be represented, and very likely that a handful of seeds will likewise be left out of the query or repeated. In a similar fashion, when searching with 45 seeds, at the very least 15 seeds will not be taken into consideration for each query. If we were to take my English-language manual corpus into consideration, the top three keywords have a keyness score of 515, 444 and 378, whilst the 43rd, 44th and 45th keywords have a keyness score of 80, 79 and 77; this means that the last three keywords are averagely 6 times less "key" than the first three keywords. With that in mind, the exclusion of even one of the top three keywords could have a serious impact on the effectiveness of the query as a whole. This means that if a query deriving from 45 seeds is unsuccessful, I could simply regenerate the tuples and consequently create a relatively successful query; this naturally has considerably negative implications as far as the reproducibility of the experiment is concerned.

As corroboration of my decision to use a restricted seed set, the literature on the Sketch Engine website also states that 20 seeds is "recommended", whilst 8 is "too low" and 40 is "useless (Creating and Compiling a Corpus Using the Interface)". Similarly, on the WebBootCaT start page, underneath the text input field, the user is prompted to input "3 to 20" seeds. On another page, the authors of the Sketch Engine respond to the question of a user who sought to create a 10-million-word corpus, suggesting that the user input 20-60 seeds and pointing out that the user can "repeat the process with the same seeds multiple times [because] *there is only a very small probability the same seed tuples will be chosen*." *(*Questions and Answers on Using WebBootCaT*,* my italics)

The authors of the Sketch Engine go on to recommend that this user split his/her seeds into sets of 10, presumably so that he/she can be sure to have exploited those seeds fully, repeating if necessary, allowing him/her to pass onto a new seed set which should harvest new URLs, avoiding possible duplicates. Moreover, the authors suggested that the user input 20-60 seeds precisely in as much as he/she was aiming to build a very large corpus; whether one performs the queries in sets of 10 or not, without enough seeds, the user would probably start retrieving more duplicates than fresh URLs. In addition, WebBootCaT imposes a retrieval limit of 50 URLs for each tuple, making more seeds a necessity after exhausting a given number of them. I, on the other hand, was aiming for a smaller, six or possibly seven-figure specialised corpus, for which three or four moderately successful BootCaT runs suffice.

As pointed out above, using a smaller seed set tends to produce more duplicates within a query. This can be rather exasperating, as one may go through the hassle of checking whether the URLs contain relevant information or not, only to discover that many of the URLs are identical. This happens because WebBootCaT does not automatically remove duplicates immediately upon querying; instead it performs this task *after* the user has ticked the checkboxes and confirms their selection. Theoretically, with 10 tuples retrieving 10 URLs each, it is possible to harvest 100 texts; but in practice, even when the user decides to include all the checked URLs in the corpus, the actual number of URLs depends on how varied the seeds are and on how many duplicates were present in the URL selection pane. Naturally, it would be a great improvement if WebBootCaT removed duplicates before presenting them to the user on the relevant pane.

As stated above, I used a seed set of 15 for all queries; however, during my analysis I noticed that some queries were returning a large number of duplicates. In order to confirm my hypothesis that using more seeds would lower precision to a greater extent than it would increase recall, I carried out a query with 30 seeds using the most successful query that I had found until that moment: the top 15 n-grams with tuple length at six (*EN_n-gram_6* and *IT_n-gram_6*). The results (*EN_n-gram_6_30* and *IT_n-gram_6_30*) are shown at the end of the chart on Figures 7 and 8, and indeed it was less effective than the same trial using only 15 seeds.

2.3.2 *Tuple length*

Tuple length is also an interesting parameter, and considering how easy it is to change it, potentially a very useful one. In an online tutorial designed for the BootCaT front-end programme, users are advised to use three seeds per tuple if they want to build a specialised corpus and two seeds if they wish to build a general-language corpus and are using general-language words (BootCaT front-end tutorial - Part 2). In contrast, on the Sketch Engine website, it is suggested that a seed length of three or four is optimal, specifying that a tuple length of four may produce fewer but more accurate results (Creating and Compiling a

Corpus Using the Interface). It may be interesting to note that the maximum tuple length on WebBootCaT is seven.

After a few preliminary trials, I had already established that varying the tuple length could have a strong impact on precision, and that in my case, a longer tuple length could be rather effective. To investigate the matter further, I chose to experiment with a tuple length of three and six.

2.3.3 *Type of seeds*

I also wanted to investigate which type of seeds would be most effective. To do this, I used the 15 most frequent keywords, key terms and n-grams in each corpus, carrying out one query with a tuple length of three and another set of queries with tuple length at six. Previous studies into the efficacy of the BootCaT process, such as Bernardini and Ferraresi (2013) or Dalan (2013), only used keywords and n-grams, thus the use of key terms constitutes a novelty. I also carried out a query using a mixture of the five most frequent keywords, key terms and n-grams. Over the course of these investigations, I kept track of particularly effective tuples and decided to use them in user-defined queries, which in the results appear under the label "custom". See Table 1 below for the seed sets used in these two user-defined queries; the seed sets for the other queries correspond to the first 15 words in the respective lists provided in the appendix. Notice that in the Italian custom seed set I experimented with repeating seeds, namely "*presente statuto*" "*arbitratore*", "*adunanze*" and "*rieleggibili*".[6] I chose to do this because I observed that they were very effective in other queries; the seed "*presente statuto"* was particularly effective, because it can almost exclusively be contained in articles of association.

Table 1. Tuples for "custom" queries. Multiword seeds are enclosed in quotation marks

EN_custom_3 & EN_custom_6	IT_custom_3 & IT_custom_6
adjourned	"presente statuto"
adjournment	ineleggibilità
certificated	rieleggibili
forfeiture	"collegio sindacale"
stockholders	arbitratore
uncertificated	rieleggibili
"electronic form"	"azioni ordinarie"
"ordinary resolution"	"presente statuto"
"record date"	adunanze
"registered address"	"presente statuto"

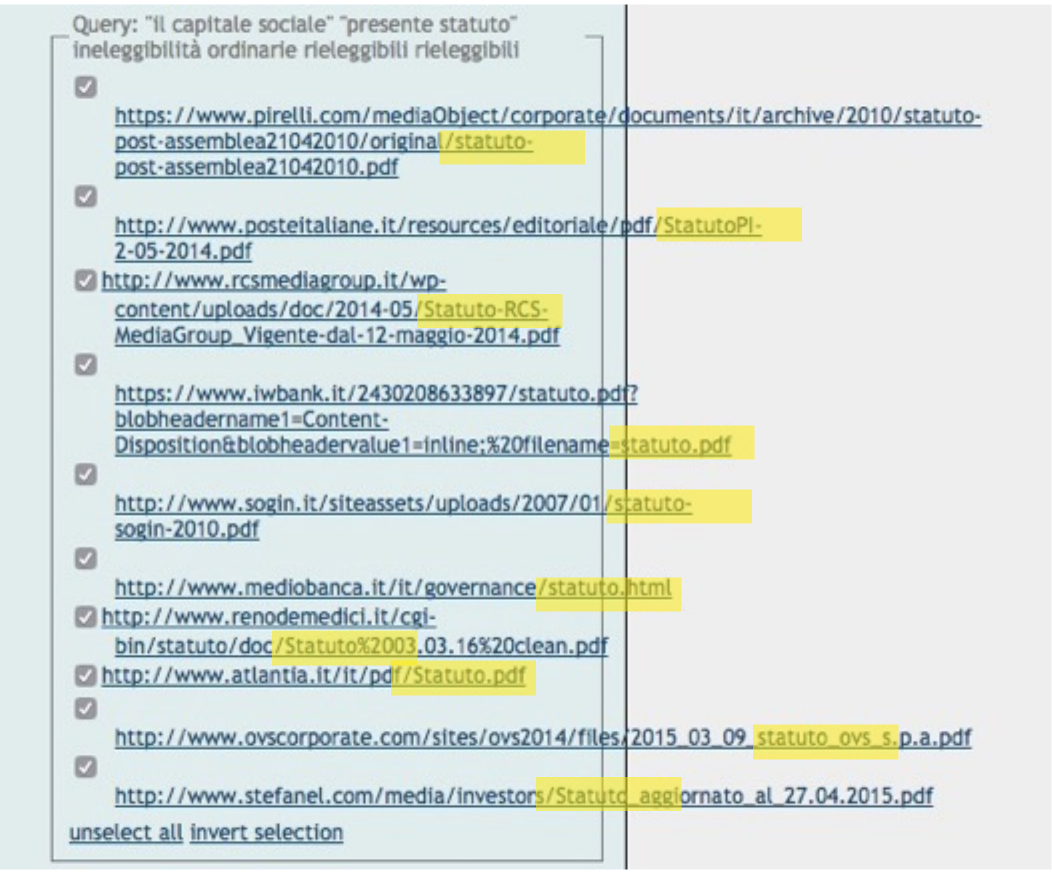
[6] English translation: "these articles (of association)", "arbitrator", "meetings", "re-electable".

“share certificate”	adunanze
“such person”	arbitratore
quorum	“presente statuto”
“electronic transmission”	“il capitale sociale”
“such meeting”	rieleggibili

2.4 *Assessing precision*

In order to assess the precision of their queries, Bernardini and Ferraresi (2013, p. 312) submitted a sample of 10 randomly selected URLs to a group of approximately 30 people composed of translation trainers and translation students. This method is highly practical, in that it asks real translators what value they give to the results in terms of relevance. Conversely, in my case the only condition that a given text needed to satisfy in order to be relevant was for it to be a set of articles of association. As such, there were no degrees of relevance; i.e. either a text is a set of articles of association or it is something else. Recognising if a text met this criterion was straightforward enough for me to be able to do it reliably by myself, because all articles of association have an explicit title and a rigid, distinct form.

Moreover, instead of taking samples, as Bernardini and Ferraresi did, I decided to evaluate the relevance of every URL. Instead of reading the contents of the webpage every time, the URL name itself often gave me very strong clues so as to be almost certain that it contained articles of association. This method is obviously prone to human error, but in a realistic situation, a translator would probably also take advantage of this shortcut. Figure 5 shows one of the panes in which WebBootCaT shows the URLs retrieved by each query. Notice how these URLs give reasonably fool-proof clues about the content of the webpage. In this case I have 10 URLs, all of which at some point contain the word “*statuto*” and other insightful words such as “corporate”, “investors”, “governance” or “*statuto vigente*” and “*statuto aggiornato*”. I have highlighted the word “*statuto*” for each URL in yellow. After numerous checks, I came to the conclusion that only an extremely deceitful webmaster would name a document *“statuto”* without it actually containing articles of association.



Query: "il capitale sociale" "presente statuto" ineleggibilità ordinarie rieleggibili rieleggibili

☑ https://www.pirelli.com/mediaObject/corporate/documents/it/archive/2010/statuto-post-assemblea21042010/original/statuto-post-assemblea21042010.pdf

☑ http://www.posteitaliane.it/resources/editoriale/pdf/StatutoPI-2-05-2014.pdf

☑ http://www.rcsmediagroup.it/wp-content/uploads/doc/2014-05/Statuto-RCS-MediaGroup_Vigente-dal-12-maggio-2014.pdf

☑ https://www.iwbank.it/2430208633897/statuto.pdf?blobheadername1=Content-Disposition&blobheadervalue1=inline;%20filename=statuto.pdf

☑ http://www.sogin.it/siteassets/uploads/2007/01/statuto-sogin-2010.pdf

☑ http://www.mediobanca.it/it/governance/statuto.html

☑ http://www.renodemedici.it/cgi-bin/statuto/doc/Statuto%2003.03.16%20clean.pdf

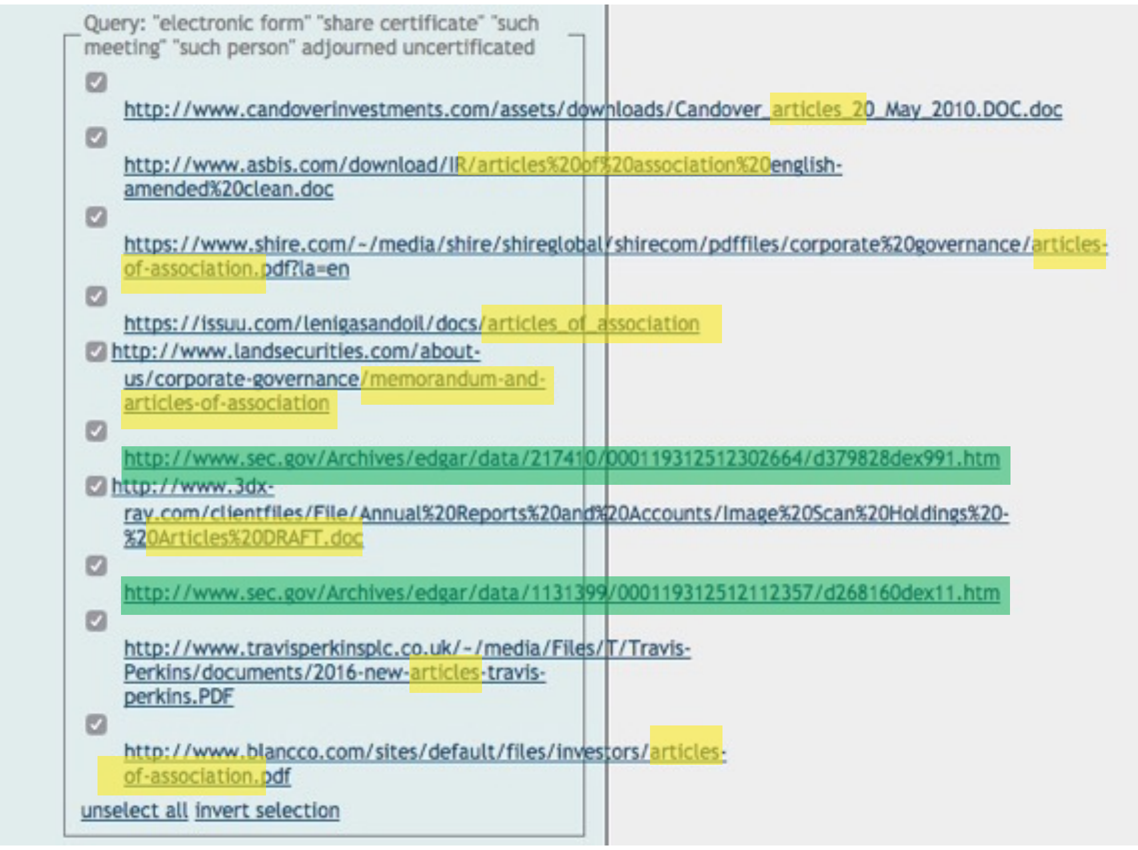
☑ http://www.atlantia.it/it/pdf/Statuto.pdf

☑ http://www.ovscorporate.com/sites/ovs2014/files/2015_03_09_statuto_ovs_s.p.a.pdf

☑ http://www.stefanel.com/media/investors/Statuto_aggiornato_al_27.04.2015.pdf

Figure 5. Example of a WebBootCaT manual URL selection pane

The names of the URLs for the English queries were generally less insightful, in that very often they originated from online national archives and therefore gave no clue as to the content of the page. An example is provided in Figure 6, with the unhelpful URLs highlighted in green. In order to verify the relevance of the URL, it was necessary to visit the webpage; in general, it was possible to understand the content of the whole page simply by viewing the first section, but this limitation slowed the process greatly, as I could no longer make an act of blind faith as I did with the Italian queries. Naturally, the possibility of using the name of the URL to judge its relevance probably varies from genre to genre.



Query: "electronic form" "share certificate" "such meeting" "such person" adjourned uncertificated

☑ http://www.candoverinvestments.com/assets/downloads/Candover_articles_20_May_2010.DOC.doc

☑ http://www.asbis.com/download/IR/articles%20of%20association%20english-amended%20clean.doc

☑ https://www.shire.com/~/media/shire/shireglobal/shirecom/pdffiles/corporate%20governance/articles-of-association.pdf?la=en

☑ https://issuu.com/lenigasandoil/docs/articles_of_association

☑ http://www.landsecurities.com/about-us/corporate-governance/memorandum-and-articles-of-association

☑ http://www.sec.gov/Archives/edgar/data/217410/000119312512302664/d379828dex991.htm

☑ http://www.3dx-ray.com/clientfiles/File/Annual%20Reports%20and%20Accounts/Image%20Scan%20Holdings%20-%20Articles%20DRAFT.doc

☑ http://www.sec.gov/Archives/edgar/data/1131399/000119312512112357/d268160dex11.htm

☑ http://www.travisperkinsplc.co.uk/~/media/Files/T/Travis-Perkins/documents/2016-new-articles-travis-perkins.PDF

☑ http://www.blancco.com/sites/default/files/investors/articles-of-association.pdf

Figure 6. Example of manual URL selection pane with typical results for an English-language query

3 Results and discussion

This section will describe the results of my queries, shown in Figures 7 and 8 below. The names of the queries are to be interpreted in the following way: LANGUAGE_type-of-seeds_tuple-length, whereby “term” in the chart stands for “key term”. I have reproduced the queries with six-seed tuples in black and queries with three-seed tuples in grey. The last bar in both charts represents the query attempted with 30 seeds. The y-axis shows the number of relevant URLs retrieved per query.

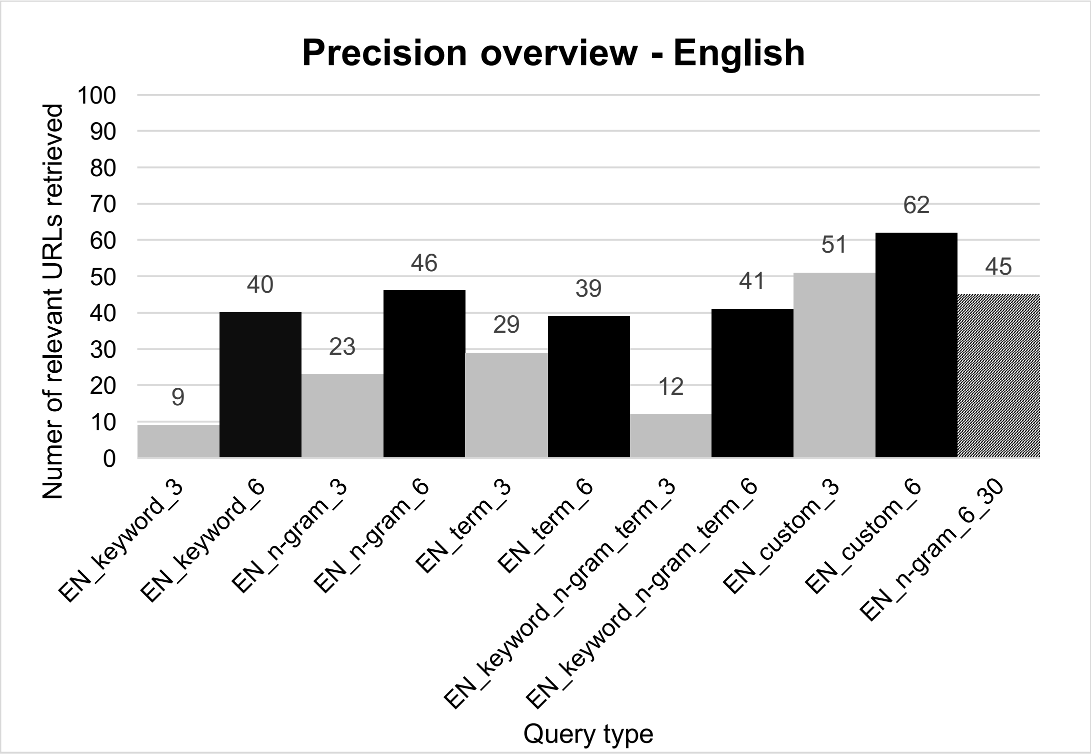


Figure 7. Bar chart illustrating precision for English-language queries. Three-seed tuples are depicted in grey, six-seed tuples in black

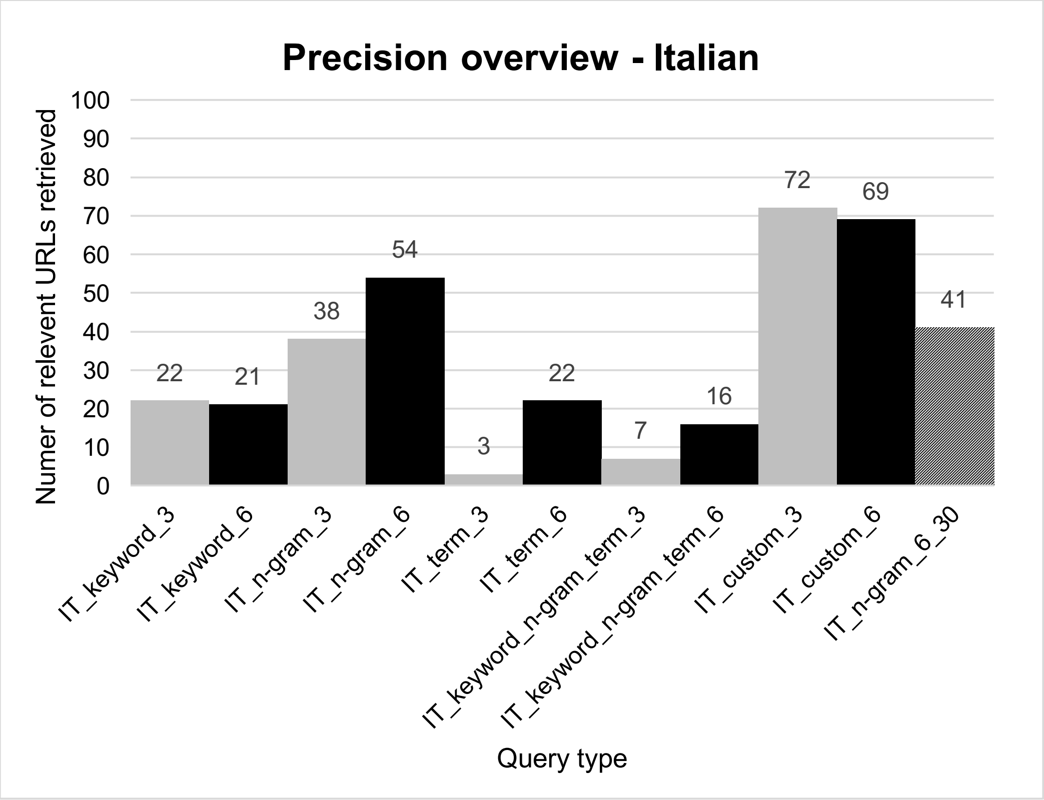


Figure 8. Bar chart illustrating precision for Italian queries

3.1 *Observations*

3.1.1 *Recall, duplicates and number of seeds*

I have named the charts "precision overview" but in reality precision and recall could be considered two sides of the same coin in the case of WebBootCaT. In my experience, the greater a query's recall (number of distinct URLs), the lesser its precision (number of relevant URLs) and vice versa. For example, the query *IT_term_6* was actually very effective in retrieving relevant texts, but the vast majority of them were duplicates, decreasing the overall number of URLs substantially. The same applies to *IT_custom_6*.

When one considers that ten tuples at tuple length six means that the query will contain 60 seeds created from the 15 original seeds, it is quite predictable that a great deal of the URLs will be duplicates. Increasing the number of original seeds however would mean having to use seeds that ranked lower, whose overall effectiveness will probably be lesser than that of the former seed set. In any case, as illustrated at an earlier point, it could be more useful for a user to split up his/her seeds into smaller groups, in order to know that he/she has depleted the seed set entirely, so as to pass onto a new seed set without worrying about underexploiting effective seeds.

As predicted, the experiment using a seed set of 30 seeds was less effective in comparison with the same query carried out with 15 seeds, although interestingly *EN_n-gram_6_30* retrieved only one relevant URL less in comparison with *EN_n-gram_6*, where the standard 15 seeds were used.

After experimenting with user-defined seed sets, I carried out another interesting experiment by using only the following seed: *"presente statuto"*. Tuple length was set at 1, and the number of URLs to the maximum of 50. Naturally, only one tuple was generated, but 46 of the 50 URLs were relevant. This is most likely due to the fact that no other genre could possibly contain the deixis expressed in *"[il] presente statuto"*. However, only 13 out of 15 texts in the Italian manual corpus had this exact term, which would undermine the URLs' representativeness. Nevertheless, the possibility of harvesting 46 URLs in less than five minutes is extremely useful. Having said that, the limit of 50 URLs per tuple gives little room to exploit these custom/user-defined seed sets fully. I could similarly comment the effectiveness of *IT_custom_3* or *IT_custom_6* with their 72 and 69 relevant texts resepectively, but given that creating such a custom seed set requires quite a lot of thought (and a stroke of luck), it was more a proof-of-concept trial than a realistic query.

3.1.2 *Tuple length*

The first trend to notice is that six-seed tuples are almost always more effective than three-seed tuples, with the exception of *IT_keyword_6* and *IT_custom_6*. The three-seed tuples attracted a large amount of noise, whereas the six-seed tuples were evidently able to filter out the noise from the signal. I conjecture that three-seed tuples were ineffective because of the very large overlap between the genre of articles of association and other similar genres; perhaps with other genres a smaller tuple length would be just as effective.

3.1.3 *Type of seed*

Another interesting observation is the fact that, as predicted, of all the automatically created seed sets, the n-grams were the most effective, apart from *EN_n-gram_3*, which retrieved 23 in comparison with the 29 relevant URLs retrieved by *EN_term_3*.

I had hypothesised that a hybrid query using the five most frequent keywords, key terms and n-grams would be very effective, leveraging on the high scores obtained in their relative lists; however, in the Italian queries, the result was very poor, but *EN_keyword_n-gram_term_6* was actually quite successful, returning 41 relevant texts.

3.1.3.1 N-grams

The fact that n-grams are more effective is still quite surprising. I expected that the key terms would have been the most effective of the automatically extracted seeds, because they supposedly combine the keyness of keywords with the length and genre-specificity of n-grams. Taking a look at the key terms in Tables 4 and 5 however, they are exclusively nouns and adjectives; this means that they reflect merely semantic aspects of the genre, which as explained, are shared by texts of the same topic. This does not mean however that all the key terms were not effective; for example, the key term "*presente statuto*" was incredibly effective and the key term "*capitale sociale*" is coincidentally also contained in the n-gram *"il capitale sociale"*.

If we take into consideration the n-gram "the chairman of the meeting" again, perhaps I will be able to give a concrete example of how n-grams can be considered the linguistic expression of the genre's extra-linguistic features. Grammatically speaking, there are two ways expressing the concept of possession in English, so we could say either "the chairman of the meeting" or "the meeting's chairman". Any astute speaker of English can already perceive that using the Saxon genitive here is rather infelicitous, but this is grammatically possible and acceptable in informal speech, and as stated above, the translator's intuition is often deceitful, let alone if the translator is translating into an acquired language.

Searching the two variants on Google.co.uk with quotation marks affords some interesting observations. "The chairman of the meeting" returns 21,800,000 hits and although on the first page no articles of association are in sight, they are all authoritative texts, mainly consisting in rules and procedures of shareholders' meetings. "The meeting's chairman" returns 13,200 hits, the first of which is "A Guide to Parish Meetings and Parish Polls – Dorchester Town Council", followed by "Agenda – Hospital Broadcasting Association", "Parish Polls – South Norfolk Council" and a Google Books result originating from a history book. These were the only four results from an English-speaking country, the rest (six) were a series of business-related webpages with domains in Jordan, Germany, Italy, Angola and Spain. The German text had the name of "*Procedural information* for the Annual General Meeting" (my italics), which reeks of "translationese".

Referring back to Swales' definition of genre,[7] one can see how the "[recognition] by the expert members of the parent discourse community" is vital in order to distinguish one genre from another. A few parishes, a hospital radio station, a history book author and some non-native speakers of English can hardly be considered authoritative figures in the genre of articles of association. One can therefore conclude that the Saxon genitive does not belong to the set of conventions for expressing the relationship of possession between "chairman" and "meeting" in articles of association. And this is precisely why the n-gram "the chairman of the meeting" was effective in finding relevant URLs: because it encapsulated this particular style convention. The two key words "chairman" and "meeting" alone, even though realistically they were not ranked highly enough as keywords for me to have used them, would not have had this distinctive function and could have retrieved irrelevant or unauthoritative texts with the wording "the meeting's chairman" as well as texts containing the wording "the chairman of the meeting".

Admittedly, the 13,200 hits of "the chairman's meeting" in comparison with the 21,800,000 of "the chairman of the meeting" on Google.co.uk would probably render this particular n-gram only partially relevant in terms of its genre-specificity, but is only one example; there are also other n-grams that could harbour genre conventions within their

[7] A genre comprises a class of communicative events, the members of which share some set of communicative purpose. These purposes are recognized by the expert members of the parent discourse community, and thereby constitute the rationale for the genre. This rationale shapes the schematic structure of the discourse and influences and constrains choice of content and style. (Swales, 1990: p. 58)

linguistic form. For example, the 4th Italian n-gram "*nel caso in cui*" could be said to reflect the genre convention according to which it is preferable to use this locution as opposed to the simple "*se*". Other n-grams include "not less than", which presumably is used more often that the simple "at least". The n-gram "for the purpose of" is also very peculiar, in that I would have instinctively opted for something simpler such as "in order to".

3.2 *Query effectiveness*

The highest number of relevant URLs retrieved by an automatically created seed set was 54 with *IT_n-gram_6*. One could argue that 54 out of 100 is a meagre result, but in reality, the ability to harvest 54 texts in one fell swoop is unprecedented, especially because WebBootCaT does the rest automatically. Locating, downloading and converting 15 texts for the manual corpus took around 45 minutes; with WebBootCaT, if one is lucky with the parameters, it is possible to harvest hundreds of texts within less than an hour. Perhaps the genre I investigated was particularly pernicious considering its large overlap with other genres; in comparison to the maximum precision of 54% achieved in my results, Bernardini and Ferraresi's (2013) experiments, which examined a different genre, proved that an automatically produced query using n-grams could reach up to an average of 70% precision.

3.3 *Manually selecting URLs on WebBootCaT*

Considering the question of query effectiveness described above, one could could conclude that if the BootCaT method requires so much human intervention to manually select the URLs on the relevant pane, perhaps it is still too time-consuming to be considered a viable tool for the translator. Indeed, considering the results of Bernardini and Ferraresi (2013) and Dalan (2013) and in light of my personal findings, perhaps it is too early to speak of fully automatic corpus construction.

Moreover, the possibility of judging a URL's relevance just from the name could change radically from genre to genre; perhaps the fact that articles of association are almost always published as pdf files ensures that they are recognisable.

However, in reality, if a translator were strapped for time and needed to build a set of comparable corpora for an assignment, he/she could use the parameters which he/she predicts would be appropriate, and then simply use the corpus with a pinch of salt. For example, if we hypothesise that a given corpus population contains 55% relevant texts and 45% irrelevant texts, perhaps when using word lists or keyword and key term lists, the sought-after candidate translation may be ranked lower than it otherwise would be or concordances might show anomalous results. Moreover, I would suggest that translators greatly prefer spotting out a translation amongst a set of authentic examples as opposed to inventing a translation from scratch or by using dictionaries, translation memories or parallel corpora. Continuing with this conjecture, the translator might spot an interesting candidate translation, and then he/she could click on the word(s) to view to original source file. In this manner, he/she could confirm whether the candidate translation originates from a relevant or irrelevant text. Furthermore, if a translator identifies an irrelevant text, within seconds he/she can click on the text and remove it from the corpus, gradually improving the corpus' representativeness.

3.4 *Web mining as an unbiased sampling method*

In this section, I suggest that web mining provides an objective method of harvesting texts, doing away with the biases that humans will necessarily have when selecting texts manually for corpus construction. This bias often undermines the representativeness of the corpora, as was the case with my manual corpora. For both corpora, I chose texts originating from large companies, ignoring smaller companies entirely. As far as the English manual corpus is concerned, I tried to take a sample of texts from a variety of countries, but naturally my attempt was fundamentally biased and flawed. I did not consider South Africa or countries like India, Singapore or Hong Kong where English is an official language and is widely used in business contexts. When searching with WebBootCaT however, the queries act as an unbiased sampler, harvesting texts simply according to their relevance and thereby allowing unexpected texts to be found; during my investigations I even came across URLs originating from the Cayman Islands and Jamaica. In light of this, I believe that web mining allows us to create a more balanced corpus population, which increases representativeness and thereby allows translators to draw more authoritative conclusions about the language under investigation.

4 Using the corpora

In order to put my corpora to the test, I built an English-language corpus totalling 3,709,337 words and an Italian-language corpus totalling 955,262 words. To do this, I performed four WebBootCaT runs, adding these to the original 15 texts from the manual corpora. Obviously I had a great advantage in knowing which seeds and tuples were effective, but I believe that four runs are sufficient, and in total it took no more than an hour to build both corpora. Considering that renowned general-language corpora such as the BNC have 100 million words, one must acknowledge that the possibility to create 7-figure corpora for specific domains in a matter of hours is quite revolutionary. As mentioned before, in light of the difference in size, my corpora could be seen as poorly comparable, but in reality coincidentally the English corpus contains 173 texts and the Italian corpus contains 171 texts, which should guarantee that a similar number of linguistic features occur in each corpus.

Instead of focusing on lexical features, I decided to dedicate this section to complex linguistic phenomena where traditional sources are pushed to their boundaries and where corpora can give the translator a genuine cutting edge.

4.1 *Translating "fermo restando"*

Let us hypothesise that a translator has come across the expressions "fermo restando" or "fermo…" such as in "**fermo restando** quanto previsto nel precedente Art. 10" or "**fermo** il disposto dell'art. 2344 del Codice Civile". The dictionary Zingarelli 2016 defines "fermo restando" as "restando valido, inteso, stabilito che…",[8] the De Mauro defines it as "restando valido, essendo stabilito che…".[9] These definitions are helpful, but the concept is still somewhat unclear and these dictionaries provide no usage examples. Before attempting a translation, we can simply look in our Italian language corpus to try to spot patterns and identify conceptual knowledge. One easy way to do this is to create a concordance and sort the results by the text to the right of the node, seeing that in our case the term has a cataphoric function. Here is one example taken from the corpus that may be able to elucidate the concept further:

> Il diritto di recesso è disciplinato dalla legge, **fermo restando** che non hanno diritto di recedere gli azionisti che non hanno concorso all'approvazione delle deliberazioni riguardanti la proroga del termine della Società […]

One could translate this sentence loosely as: "the right of withdrawal shall be governed by applicable regulations, but any shareholder who has not voted on resolutions regarding the extension of the duration of the Company shall not have the right to withdraw." One could also express this relationship as: provision *x* does not change provision *y* in any way. When taken apart and analysed, it seems rather straightforward to translate this concept, but many traditional sources do not lead us to an appropriate translation.

Taking a look at the bilingual dictionary Il Ragazzini (2015), under the usage notes for the lemma "fermo" we can find the proposed translation of "it being understood that" for

[8] lo Zingarelli 2016 Vocabolario della lingua italiana
[9] Il Nuovo De Mauro, def. 3, (retrieved: 25/06/16) http://dizionario.internazionale.it/parola/fermo

"fermo restando che".[10] Fernando Picchi's dictionary Economics & Business (1986)[11] has no relevant entry, nor does Francesco De Franchis' Law Dictionary (1996);[12] note that these dictionaries are relatively old and that the only reason I had access to them was because my institutional library has copies of them. IATE states "provided that" as a translation.[13] WordReference provides the translation "it being understood that"[14] as well as an incomprehensible and contradictory thread composed of 48 entries that leaves the reader more confused than at the beginning of their search, suggesting translations among "it being understood" (without the conjunction that), "notwithstanding", "without prejudice to", "provided that", "sticking to what expressed and contemplated by" [sic], "further to what" [sic]. One user even admits, "I've been translating Italian to English for almost 15 years now, and EVERY time I get stuck on this expression."[15] The forums on ProZ are somewhat more insightful, one suggesting "subject to" and "provided that",[16] another suggesting "without prejudice to", "considering that" and "leaving untouched",[17] and another suggesting "without prejudice to", "it being understood that" and "not withstanding" [sic].[18] Linguee produces similarly mixed results.

One must acknowledge that in order to translate this seemingly innocuous term, I have consulted approximately 10 traditional sources, and in doing so have spent more than 30 minutes. Even after this research, I have no way of identifying which translations are reliable or if any of the suggested translations are reliable at all. I could attempt to read English-language articles of association to identify a translation, but as stated in 1.3, this would probably take weeks. To use one of these translations would amount to a linguistic stab in the dark; and of course, as the user on WordReference underlines perfectly, even after this

[10] il Ragazzini 2015 dizionario italiano-inglese inglese-italiano (2015); G. Ragazzini; Zanichelli

[11] Economics & Business, Dizionario enciclopedico economico e commerciale inglese-italiano italiano-inglese; F. Picchi; Zanichelli

[12] Dizionario giuridico - Law dictionary (1996); F. De Franchis; Giuffrè

[13] IATE (retrieved 25/06/16) http://iate.europa.eu/SearchByQuery.do

[14] Wordreference.com (retrieved 25/06/16) http://www.wordreference.com/enit/it%20being%20understood%20that

[15] Wordreference.com (retrieved 25/06/16) http://forum.wordreference.com/threads/fermo-restando-che.1838552/

[16] ProZ.com (retrieved 25/06/16) http://www.proz.com/kudoz/italian_to_english/law_contracts/2739055-fermo_restando_quanto_precede.html

[17] ProZ.com (retrieved 25/06/16) http://ita.proz.com/kudoz/italian_to_english/bus_financial/73346-fermo_restando.html

[18] ProZ.com (retrieved 25/06/16) http://www.proz.com/kudoz/italian_to_english/law_contracts/2999389-fermo_restando.html

investment of time, the translator has still not identified a suitable translation, and every time the translator is confronted with the same term, he/she will be in the same position.

Using our corpus, on the other hand, allows us to make conclusions founded upon real examples. As stated in 1.3, we could use our corpus to verify our intuition or alternatively to verify the translations that I gleaned from traditional sources. If we perform a simple search for the translation proposed by Il Ragazzini and WordReference ("it being understood that"), no results are returned, even when searching the form "being understood". Indeed, the half-baked progressive form and the dummy subject sounds very unidiomatic and inelegant to the native ear, and searches on Google.co.uk return mainly non-native texts or native texts belonging to an entirely different genre and a distinctly lower register. Searching the other translations gives us confirmation that they are genuinely used, but still we're left with a handful of possible translations and only one gap to fill.

Instead of verifying our intuition or translations provided from other sources, we could try to identify an equivalent from within the corpus itself. When I sorted the concordance of "fermo restando" in Italian, I noticed that one pattern was "fermo restando quanto previsto nel precedente articolo…". In order to discover the unknown translation, we can start from a certainty, such as the word "precedente", which I know is translated as "foregoing". If I hadn't known this, perhaps after searching for "preceding" (the more immediate translation), I would have noticed that there were too few results and I would have used a bilingual dictionary to identify other translations of "precedente" until finding a translation with a satisfactory number of results. This is one of the reasons why in 1.3 I stated that corpora are a complementary instrument, to be used in combination with other sources.

I created a concordance of "foregoing" and sorted the results to the left, seeing that our unknown term should necessarily be located a few words before the node. The pattern was very easy to identify: the strongest collocation was by far "without prejudice to". The translation "notwithstanding" was also relatively frequent, but the translations "provided that" and "subject to" were almost entirely absent. Thanks to our corpus, the translator can quickly identify the most common translation and use it with much greater confidence than in the case of traditional sources.

4.2 *Translating "regolarmente costituita"*

Another case that lends itself to interesting analysis is that presented by the term "regolarmente costituita", for example in "l'Assemblea Ordinaria si reputa **regolarmente costituita** con la presenza di almeno i due terzi più uno dei soci". Monolingual dictionaries do not cover this very specific use of the verb “costituire”; similarly, traditional bilingual dictionaries and IATE provide no information. Bab.la provides an inadequate translation and EUR-Lex provides the translation “duly established”, which is a possible translation but not a suitable one in this context, because it refers to a company established in accordance with applicable law, not to a company meeting that satisfies certain requirements in order to be considered valid. However, the WordReference[19] and ProZ[20] forums as well as Linguee, along with a deluge of red herrings, at some point provide what I had previously identified as a suitable translation. Needless to say, the translator would require extensive knowledge of the field in order to fish out a suitable translation among these red herrings.

For example, one user on ProZ suggested the translation “quorate”, and indeed the Oxford English Dictionary defines quorate as “a meeting attended by a quorum and so having valid proceedings”.[21] As such, “quorate” would seemingly be a perfect translation, and many translators might be attracted by this apparent exact equivalent. A quick search in our English corpus however shows that only 39 results were found. Incidentally, on the results pane I discovered the very frequent pattern “duly convened **and** quorate” or “at a duly convened, quorate meeting”, which apart from providing us with another candidate translation (duly convened), also shows us that “quorate” must be a sort of sub-condition of meetings that possess the quality of being “duly convened”. My assumption would be that “quorate” could refer to the number of people present and “duly convened” might require that certain figures are present, such as the chairman, a notary public or members of the board of statutory auditors. Again, the great advantage of corpus linguistics is that I do not have to be an expert of the field to make such assumptions, because my corpus is relatively representative and I can infer knowledge by pinpointing a single linguistic phenomenon simultaneously in a large quantity of texts.

Let us hypothesise that I did not notice the candidate translation “duly convened” when I searched for “quorate”. Again, instead of verifying candidate translations, I could decide to start searching from within my native corpus population. We can take an absolute certainty, “meeting” as the translation of “assemblea”, and create a concordance. At this point we can create a list of candidate collocations by using the relevant tool on WebBootCaT and

[19] WordReference.com (retrieved 26/06/16) http://forum.wordreference.com/threads/lassemblea-si-reputa-regolarmente-costituita.1563120/

[20] ProZ.com retrieved (26/06/16) http://www.proz.com/kudoz/italian_to_english/other/912409-sono_validamente_costitutite_dichiarata_validamente_costituita.html

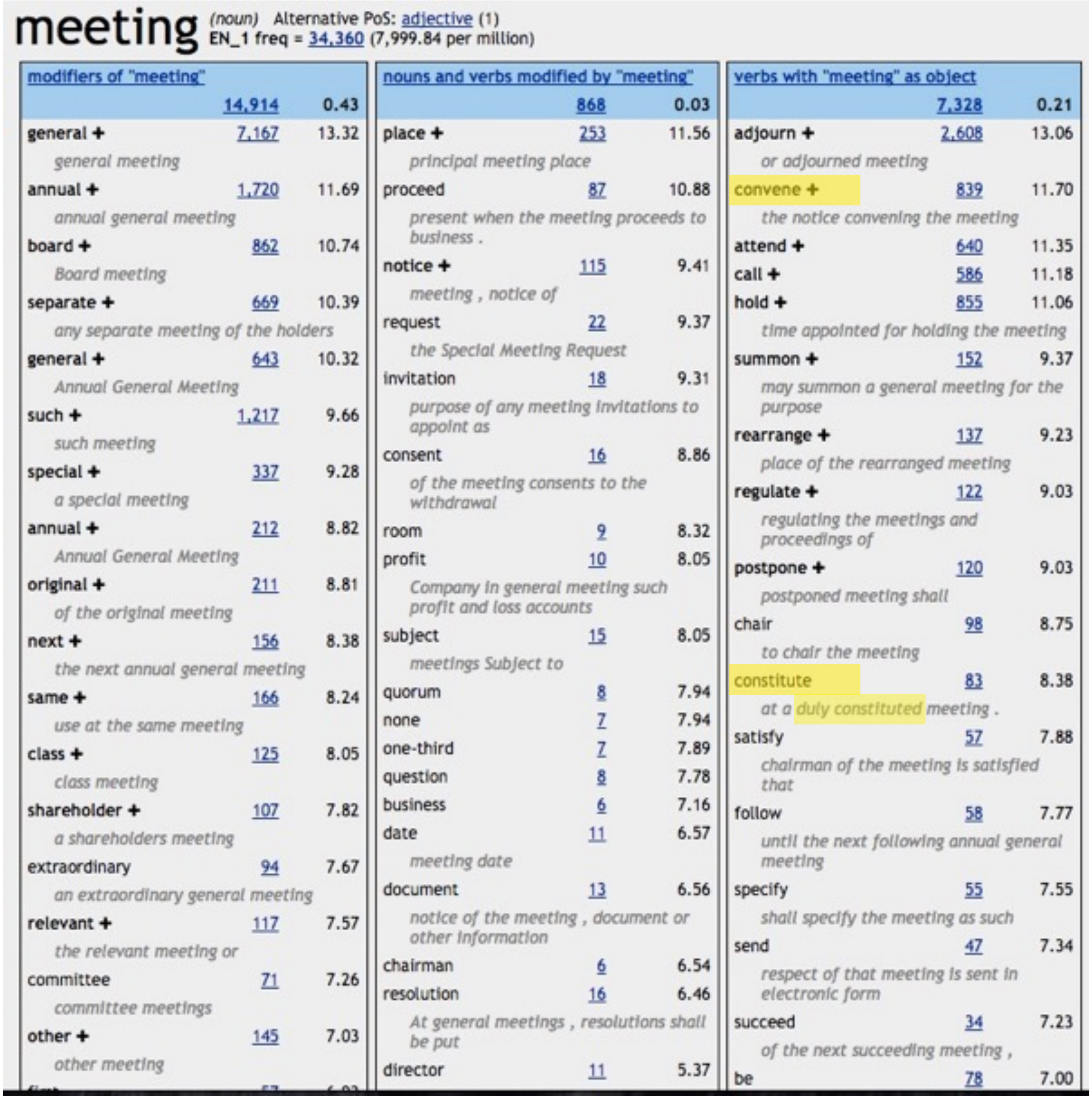
[21] Oxford English Dictionary (retrieved 26/06/16) http://www.oxforddictionaries.com/definition/english/quorate

on the first page we can spot "convened"; subsequently from the candidate "convene" we could create another concordance and further collocations. Alternatively, we could search for "meeting" using the Sketch Engine's signature word sketch, which presents the user with a word profile with collocations categorised according to their grammatical function; see Figure 9. Clicking on the plus symbol allows the user to access a multiword word sketch that the Sketch Engine identifies automatically; see Figure 10.

In comparison to looking for translations using traditional sources, finding candidate translations with corpora seems like child's play; the second collocate under the category "verbs with 'meeting' as object" shows us the verb "convene", which any good translator should identify as a candidate translation. Further down the same list, we see "constitute" and the example "a duly constituted meeting". Clicking on the frequency to the right of the word links the user directly to a concordance list, which we can sort in order to identify patterns and ascertain whether the usage corresponds to the usage of "regolarmente costituita". All this takes a matter of minutes, and we can then double-check our hypotheses by searching "duly convened" and "duly constituted" in the corpus. The former returns 159 instances and the latter 121, so we can safely conclude that, for all intents and purposes, both of these translations are equally common. A ProZ forum suggested the translations "validly constituted" and "legally constituted",[22] but a quick search in our corpus shows that there was only one case of the former and no cases of the latter.

As (Tognini-Bonelli, 2001) and (Greaves & Warren, 2010) underline, all words occur in connection with other words, and are characterised not only by the meaning that we associate them with, but also simply by the words with which they commonly occur, that is, by the word's collocational profile or co-selection. Very few monolingual dictionaries take this into account, and I do not believe there is a single bilingual dictionary that has been able to answer to this need. And this is precisely where the power of corpus linguistics makes all the difference, because no dictionary or other source would allow you to search for "convene" or "constitute" and find the adverb "duly", because this collocation is far too uncommon in general language.

[22] ProZ.com (retrieved 26/06/16) http://www.proz.com/kudoz/italian_to_english/other/912409-sono_validamente_costitutite_dichiarata_validamente_costituita.html



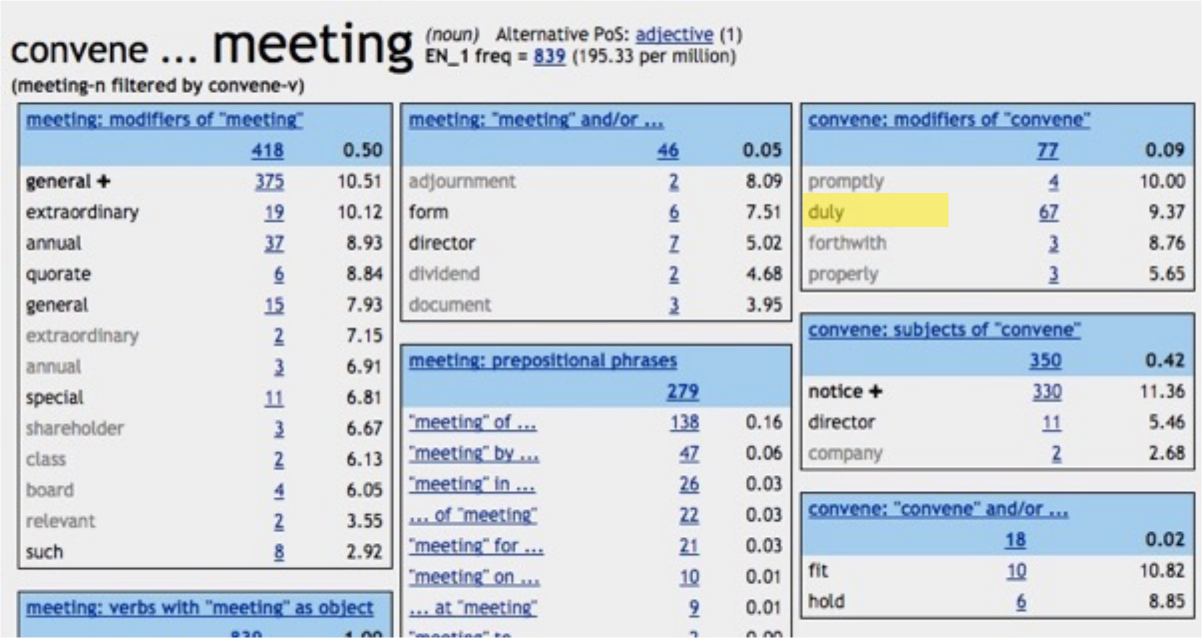
meeting *(noun)* Alternative PoS: adjective (1)
EN_1 freq = 34,360 (7,999.84 per million)

modifiers of "meeting"	14,914	0.43
general +	7,167	13.32
general meeting		
annual +	1,720	11.69
annual general meeting		
board +	862	10.74
Board meeting		
separate +	669	10.39
any separate meeting of the holders		
general +	643	10.32
Annual General Meeting		
such +	1,217	9.66
such meeting		
special +	337	9.28
a special meeting		
annual +	212	8.82
Annual General Meeting		
original +	211	8.81
of the original meeting		
next +	156	8.38
the next annual general meeting		
same +	166	8.24
use at the same meeting		
class +	125	8.05
class meeting		
shareholder +	107	7.82
a shareholders meeting		
extraordinary	94	7.67
an extraordinary general meeting		
relevant +	117	7.57
the relevant meeting or		
committee	71	7.26
committee meetings		
other +	145	7.03
other meeting		

nouns and verbs modified by "meeting"	868	0.03
place +	253	11.56
principal meeting place		
proceed	87	10.88
present when the meeting proceeds to business .		
notice +	115	9.41
meeting , notice of		
request	22	9.37
the Special Meeting Request		
invitation	18	9.31
purpose of any meeting invitations to appoint as		
consent	16	8.86
of the meeting consents to the withdrawal		
room	9	8.32
profit	10	8.05
Company in general meeting such profit and loss accounts		
subject	15	8.05
meetings Subject to		
quorum	8	7.94
none	7	7.94
one-third	7	7.89
question	8	7.78
business	6	7.16
date	11	6.57
meeting date		
document	13	6.56
notice of the meeting , document or other information		
chairman	6	6.54
resolution	16	6.46
At general meetings , resolutions shall be put		
director	11	5.37

verbs with "meeting" as object	7,328	0.21
adjourn +	2,608	13.06
or adjourned meeting		
convene +	839	11.70
the notice convening the meeting		
attend +	640	11.35
call +	586	11.18
hold +	855	11.06
time appointed for holding the meeting		
summon +	152	9.37
may summon a general meeting for the purpose		
rearrange +	137	9.23
place of the rearranged meeting		
regulate +	122	9.03
regulating the meetings and proceedings of		
postpone +	120	9.03
postponed meeting shall		
chair	98	8.75
to chair the meeting		
constitute	83	8.38
at a duly constituted meeting .		
satisfy	57	7.88
chairman of the meeting is satisfied that		
follow	58	7.77
until the next following annual general meeting		
specify	55	7.55
shall specify the meeting as such		
send	47	7.34
respect of that meeting is sent in electronic form		
succeed	34	7.23
of the next succeeding meeting ,		
be	78	7.00

Figure 9. Partial view of a word sketch for the node "meeting" with candidate translations highlighted.



convene ... meeting (noun) Alternative PoS: adjective (1)
EN_1 freq = 839 (195.33 per million)
(meeting-n filtered by convene-v)

meeting: modifiers of "meeting"	418	0.50
general +	375	10.51
extraordinary	19	10.12
annual	37	8.93
quorate	6	8.84
general	15	7.93
extraordinary	2	7.15
annual	3	6.91
special	11	6.81
shareholder	3	6.67
class	2	6.13
board	4	6.05
relevant	2	3.55
such	8	2.92

meeting: verbs with "meeting" as object

meeting: "meeting" and/or ...	46	0.05
adjournment	2	8.09
form	6	7.51
director	7	5.02
dividend	2	4.68
document	3	3.95

meeting: prepositional phrases	279	
"meeting" of ...	138	0.16
"meeting" by ...	47	0.06
"meeting" in ...	26	0.03
... of "meeting"	22	0.03
"meeting" for ...	21	0.03
"meeting" on ...	10	0.01
... at "meeting"	9	0.01

convene: modifiers of "convene"	77	0.09
promptly	4	10.00
duly	67	9.37
forthwith	3	8.76
properly	3	5.65

convene: subjects of "convene"	350	0.42
notice +	330	11.36
director	11	5.46
company	2	2.68

convene: "convene" and/or ...	18	0.02
fit	10	10.82
hold	6	8.85

Figure 10. Partial view of multiword word sketch: "meeting" filtered by "convene"; candidate translations highlighted.

4.3 *Translating "anche non soci"*

Let us now examine another difficult term for Italian to English translators: "anche non socio" (and similar expressions) as in "Il presidente dell'Assemblea nominerà un segretario, **anche non socio**, e qualora necessario anche uno o più scrutatori, **anche non soci**". I could content myself with something like "[…] a secretary, who does not have to be a shareholder" or "[…] who must not necessarily be a shareholder", but the radical jump in register is jarring. Again, I performed a search using traditional sources for "anche non soci" and only Linguee provided some information[23]; all the other bilingual sources (il Ragazzini, il Sansoni Online, WordReference, ProZ, bab.la, Glosbe, IATE, EUR-Lex) were of no help. This is not entirely surprising, because this is a marginal use of the word "anche". A very common translation among the Linguee results was "including non-shareholders", which is very unidiomatic; a search for "including non-shareholders" in my corpus retrieved no results, and the term "non-shareholder" retrieved only two results, which allows us to conclude that the term "non-shareholder" is probably a non-standard neologism. Another translation on Linguee was "who may or may not be shareholders", which as regards style and register is acceptable but fails to convey the sense of exceptionality that "anche" does; the formulation "*anche* non socio" leads the reader to believe that usually the secretary is a shareholder, but *even* people who are not shareholders can *also* become secretaries.

[23] Linguee.com (retrieved 26/06/16) http://www.linguee.com/english-italian/search?source=auto&query=anche+non+soci

We can perform the same process as in the previous cases, attempting to identify a translation from within the corpora. In the Italian concordance for “anche non socio”, I noticed that this expression co-occurred quite often with appointment of scrutineers (scrutatore) at general assemblies. I created a concordance for “scrutineer” and even without sorting the results I was able to see that the preferred form for expressing this concept was “who need not be members”, as in “the Chairman may appoint scrutineers, who need not be members.” Not only does this finding suggest that we should opt for “member” as opposed to “shareholder”, but it also gives us the turn of phrase “who need not be”; this inversion is practically absent in daily speech and even a native translator would have had to have been an expert in the field in order to have used it. Just to confirm that “need not” is genuinely common in this genre, I performed a search which returned 1,014 results, allowing me to conclude that it is used quite extensively.

5 Conclusion

We can conclude that the WebBootCaT method is a very powerful tool for translators working with specialised language. Not only is it time-saving, but the corpora produced are reliable and, moreover, the Sketch Engine is relatively user-friendly in comparison with other corpus analysis tools. Obviously, the WebBootCaT method is not suitable for every translation assignment: it will always require a considerable investment of time before the translation process, on average around 2-3 hours for a corpus containing approximately 150 texts such as the ones I made. In order to quicken the process, one could even start off with a manual corpus of only 5 texts, performing the first BootCaT runs with a little more caution, considering the weak representativeness of such a small manual corpus.

The optimal settings for the three parameters that users can adjust probably change from genre to genre and according to the desired size of the corpus. When building a small 6-figure corpus, I suggest one can probably count on a seed set of 10-20 seeds; when aiming for a larger corpus, one will necessarily start requiring more seeds in order to avoid duplicates. As far as the tuple length is concerned, I conjecture that with highly conventionalised genres a tuple length of at least 5 is advisable, whereas with less specific genres a shorter tuple length may be enough, allowing the user to create a greater number of individual tuples from the same seed set. As far as the type of seed is concerned, I believe it is safe to say that n-grams are generally more effective than any other automatically produced seed.

If the translation assignment is a one-off, then perhaps this investment is not so profitable, but if the translator is interested in the field and intends to specialise in the general topic (e.g. legal translation, medical translation etc.), then the investment is certainly worthwhile. Instead of misusing one's time by trying to find amateurish translations on the Internet, a smart translator might choose to sacrifice some of their time in advance and reap the benefits during the translation assignment and during all future similar assignments. Not only does the translation process have the potential to be quicker, but also of much higher quality. Translators could then store their corpora and build up a library of corpora for the specific genres that they work with; of course these corpora could be enlarged or fine-tuned at any time.

6 References

Baroni, M., & Bernardini, S. (2004). BootCaT: Bootstrapping corpora and terms from the web. *Proceedings of LREC 2004* (pp. 1313-1316). Lisbon: ELDA.

Baroni, M., Kilgarriff, A., Pomikálek, J., & Rychlý, P. (2006). WebBootCaT: a web tool for instant corpora. *Proceedings of EuraLex*, (pp. 123-132).

Bernardini, S., & Ferraresi, A. (2013). Old Needs, New Solutions: Comparable Corpora for Language Professionals. In S. Sharoff, R. Rapp, P. Zweigenbaum, & P. Fung, *Building and Using Comparable Corpora*. Springer Berlin Heidelberg.

Bernardini, S., Baroni, M., & Evert, S. (2013). *A WaCky Introduction*. Retrieved May 07, 2016, from http://wackybook.sslmit.unibo.it/pdfs/bernardini.pdf

Bhatia, V. (2004). *Worlds of Written Discourse*. London: Continuum.

Biber, D., & Conrad, S. (2011). Lexical bundles in conversation and academic prose. In A. Kruger, K. Wallmach, & J. Munday (Eds.), *Corpus-Based Translation Studies* (pp. 211-236). London: Continuum.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finnegan, E. (1999). *The Longman Grammar of Spoken and Written English*. Harlow: Pearson Education.

BootCaT front-end tutorial - Part 2. (n.d.). Retrieved May 14, 2016, from docs.sslmit.unibo.it: http://docs.sslmit.unibo.it/doku.php?id=bootcat:tutorials:basic_2#tuple_generation

Bowker, L., & Pearson, J. (2002). *Working with specialized language: a practical guide to using corpora*. London; New York: Routledge.

Carter, R. A., & McCarthy, M. J. (2006). *Cambridge Grammar of Spoken English*. Cambridge: Cambridge University Press.

Chatrand, M., Millar, C., & Wiltshire, E. (1997). *English for Contract and Company Law*. London: Sweet & Maxwell.

Creating and Compiling a Corpus Using the Interface. (n.d.). Retrieved May 13, 2016, from Sketch Engine: https://www.sketchengine.co.uk/creating-and-compiling-a-corpus-using-the-interface/

Dalan, E. (2013). Costruzione automatica di corpora orientati al genere e fraseologia: Il caso delle guide web in inglese degli Atenei europei. MA thesis; University of Bologna, SSLMIT Forlì: (unpublished).

Ferri, V. (2014). Estrazione terminologica automatica: sistemi a confronto. MA thesis; University of Bologna, SSLMIT Forlì: (unpublished).

Fortune 500. (2016, May 07). Retrieved from Forbes: http://fortune.com/fortune500/

FTSE 100 Index. (2016, May 10). Retrieved from Wikipedia: https://en.wikipedia.org/wiki/FTSE_100_Index

Greaves, C., & Warren, M. (2010). What can a corpus tell us about multi-word units? In *The Routledge Handbook of Corpus Linguistics* (pp. 212-226). Abdingon: Routledge.

Hyland, K. (2008). As Can Be Seen: Lexical Bundles and Disciplinary Variation. *English for Specific Purposes* (27(1)), 4-21.

Kilgarriff, A. (2013). *Term finding and more in SkE*. Retrieved May 07, 2016, from https://www.sketchengine.co.uk/xdocumentation/raw-attachment/wiki/AK/Papers/TermfindingAndMoreInSkE.docx?format=raw

Kilgarriff, A., PVS, A., & Pomikálek, J. (2011). Electronic Lexicography in the 21st Century: New Applications for New Users. *eLex*, (pp. 122-128).

List of largest public copmanies in Canada by profit. (2016, May 07). Retrieved from Wikipedia: https://en.wikipedia.org/wiki/List_of_largest_public_companies_in_Canada_by_profit

Lista delle maggiori aziende italiane per fatturato. (n.d.). Retrieved May 10, 2016, from Wikipedia: https://it.wikipedia.org/wiki/Lista_delle_maggiori_aziende_italiane_per_fatturato

NZX 50 Index. (2016, May 07). Retrieved from Wikipedia: https://en.wikipedia.org/wiki/NZX_50_Index

Questions and Answers on Using WebBootCaT. (2016, May 13). Retrieved from Sketch Engine: https://www.sketchengine.co.uk/questions-and-answers-on-using-webbootcat/

Reppen, R. (2010). Building a corpus: what are the key considerations? In A. O'Keeffe, & M. McCarthy, *The Routledge Handbook of Corpus Linguistics* (pp. 31-37). Abdingon: Routledge.

S&P/ASX 20. (2016, May 07). Retrieved from Wikipedia: https://en.wikipedia.org/wiki/S%26P/ASX_20

Swales, J. (1990). *Genre analysis*. Cambridge: Cambridge University Press.

Tognini-Bonelli, E. (2001). *Corpus Linguistics at Work*. Amsterdam; Philadelphia: J. Benjamins.

Varantola, K. (2003). Translators and disposable corpora. In F. Zanettin, S. Bernardini, & D. Stewart, *Corpora in Translator Education*. Manchester: St. Jerome Publishing.

Zanettin, F. (2012). *Translation-driven corpora*. Oxon; New York: St Jerome Publishing.

Zanettin, F., Bernardini, S., & Stewart, D. (2003). *Corpora in Translator Education*. Manchester, UK; Northampton MA: St. Jerome Publishing.

Appendix A - Keyword tables

Table 2. Most frequent keywords in English manual corpus

Rank	Word	Score	Freq	Ref. Freq
1	uncertificated	515.44	223	228
2	stockholder	443.50	260	8370
3	quorum	377.50	329	23487
4	certificated	324.42	175	5909
5	adjourned	313.66	243	18373
6	depositary	270.98	142	5108
7	stockholders	239.68	210	23768
8	shareholder	239.44	571	103663
9	proxy	229.62	732	146203
10	forfeiture	226.05	181	19778
11	directors	224.47	2660	604964
12	moneys	207.98	138	12521
13	dividend	202.40	535	117369
14	shares	199.23	3014	778562
15	adjournment	179.93	106	8597
16	debentures	160.39	87	6140
17	appointor	160.17	69	228
18	lon15010141	159.45	68	0
19	authorise	155.83	93	9024
20	forfeited	140.93	140	30044
21	transferee	136.39	73	5783
22	pursuant	133.27	352	117319
23	dividends	132.78	344	114650
24	payable	125.42	367	132421
25	transacted	123.63	68	6585
26	holder	120.19	685	279295
27	holders	116.87	553	228060
28	capitalisation	116.46	64	6571
29	discretions	113.09	50	887
30	duly	110.56	181	64178

Table 3. Most frequent keywords in Italian manual corpus

Rank	Word	Score	Freq	Ref. Freq
1	ineleggibilità	219.06	59	3644
2	onorabilità	178.55	42	2798
3	supplenti	164.88	66	6910
4	quozienti	162.42	25	775
5	determinandone	146.32	26	1369
6	societari	138.45	45	5039
7	trasferente	136.45	17	48
8	rieleggibili	134.77	22	1011
9	maggioranze	133.97	50	6240
10	supplente	133.16	62	8541
11	convertibili	132.03	24	1473
12	deliberazioni	130.27	139	23517
13	azionisti	129.96	116	19173
14	ordinarie	127.19	116	19659
15	convocazione	121.98	216	41044
16	cadauna	117.30	19	983
17	quoziente	115.15	43	6249
18	deliberare	111.63	57	9666
19	convocata	106.31	95	19205
20	assemblea	104.53	902	211826
21	nominative	101.57	16	876
22	validamente	99.40	35	5722
23	prelazione	99.39	35	5723
24	effettivi	95.68	83	18556
25	amministratori	95.23	318	80112
26	controllate	94.75	96	22186
27	adunanze	92.88	23	3123
28	liste	91.73	299	78125
29	bancoposta	90.33	19	2195
30	blea	86.56	11	123

Appendix B - Key term tables

Table 4. Most frequent key terms in English manual corpus

Rank	Word	Score	Freq	Ref. Freq
1	general meeting	477.11	457	10
2	record date	275.95	118	0
3	alternate director	268.96	115	0
4	ordinary resolution	266.63	114	0
5	such person	237.89	152	4
6	such meeting	210.45	101	1
7	electronic form	175.76	75	0
8	special resolution	175.76	75	0
9	uncertificated form	171.10	73	0
10	relevant system	168.77	72	0
11	absolute discretion	150.13	64	0
12	share certificate	145.47	62	0
13	registered address	138.48	59	0
14	such shareholder	133.82	57	0
15	electronic transmission	124.50	53	0
16	special meeting	119.35	114	10
17	such notice	116.75	62	2
18	preference share	115.17	49	0
19	eligible shareholder	115.17	49	0
20	such manner	98.99	63	4
21	nominal amount	98.86	42	0
22	electronic communication	96.53	41	0
23	share capital	91.79	73	7
24	shareholder nominee	89.54	38	0
25	other company	88.69	47	2
26	such share	87.21	37	0
27	nominal value	86.83	46	2
28	restricted share	84.88	36	0
29	copy form	83.89	40	1
30	hard copy form	83.89	40	1

Table 5. Most frequent key terms in Italian manual corpus

Rank	Word	Score	Freq	Ref. Freq
1	collegio sindacale	857.94	212	49
2	azioni ordinarie	581.85	85	9
3	sindaci effettivi	475.65	61	2
4	sola lista	461.11	65	7
5	sindaco supplente	357.10	44	0
6	sindaco effettivo	326.29	41	1
7	documenti contabili societari	304.45	39	2
8	ordine progressivo	300.39	40	4
9	valore nominale	284.70	48	18
10	capitale sociale	284.38	132	135
11	documenti contabili	277.68	44	14
12	assemblea delibera	276.17	34	0
13	presente statuto	258.69	39	11
14	dirigente preposto	257.95	35	5
15	revisione legale	249.98	32	2
16	codice civile	224.94	68	71
17	sede sociale	220.26	41	25
18	numero progressivo	219.01	33	11
19	normativa vigente	214.61	92	121
20	casi previsti	206.54	28	5
21	sindaci supplenti	199.34	25	1
22	assemblea ordinaria	199.13	28	7
23	assemblea straordinaria	195.82	30	12
24	modalità previste	188.94	38	31
25	esercizio sociale	188.09	25	4
26	medesima lista	172.16	22	2
27	sedi secondarie	168.75	22	3
28	oggetto sociale	165.42	27	16
29	applicabili disposizioni	162.86	20	0
30	numero minimo	160.50	42	55

Appendix C - N-gram tables

Table 6. Most frequent n-grams in English manual corpus. Italicised items denote nested n-grams

Rank	Word	Freq
1	the company may	268
2	*the chairman of the meeting*	170
3	of the shares	157
4	as a director	148
5	is to be	135
6	of the corporation	134
7	the directors shall	122
8	the company is	115
9	*for the time being*	114
10	as may be	112
11	of any such	108
12	*for the purpose of*	107
13	the board shall	106
14	meetings of the	105
15	*the date of the*	103
16	whether or not	102
17	the company has	101
18	not less than	101
19	*of the company and*	101
20	of the relevant	99
21	*subject to the provisions of*	97
22	the companies act	95
23	more than one	95
24	with the company	93
25	*shall be entitled to*	93
26	*in person or by proxy*	92
27	*in the case of a*	90
28	an alternate director	88
29	the meeting and	87
30	with respect to	86

Table 7. Most frequent n-grams in Italian manual corpus. Italicised items denote nested n-grams

Rank	Word	Freq
1	del codice civile	118
2	ai sensi dell	83
3	del capitale sociale	68
4	*nel caso in cui*	58
5	di cui al	53
6	*presidente del consiglio di amministrazione*	50
7	*al consiglio di amministrazione*	49
8	*il maggior numero di voti*	49
9	*il consiglio di amministrazione può*	49
10	*nell' avviso di convocazione*	46
11	*di equilibrio tra i generi*	42
12	*attività di direzione e coordinamento*	42
13	*ottenuto il maggior numero di*	41
14	la società può	40
15	il capitale sociale	40
16	*assemblea straordinaria dei soci del*	40
17	*redazione dei documenti contabili societari*	39
18	*preposto alla redazione dei documenti*	39
19	*materia di equilibrio tra i*	38
20	*in materia di equilibrio tra*	38
21	*alla redazione dei documenti contabili*	38
22	*dirigente preposto alla redazione dei*	36
23	per la nomina	34
24	di cui all	34
25	del presente statuto	34
26	*di cui all' articolo*	34
27	*ai sensi di legge*	34
28	*in possesso dei requisiti di*	34
29	sindaci effettivi e	33
30	in ogni caso	33