Comparative Legilinguistics
vol. 33/2018
DOI: http://dx.doi.org/10.14746/cl.2018.33.2

THE WEB AS CORPUS AND ONLINE CORPORA FOR LEGAL TRANSLATIONS

Patrizia GIAMPIERI, MSc
School of Law, University of Camerino
Via D'Accorso, 16, 62032 Camerino (MC)
patrizia.giampieri@unicam.it

Abstract: Legal language is hallmarked by a pedantic and user-*un*friendly jargon whose constructs are all but intuitive, not to mention the legal system specificity which makes it unique in every country. Second language (L2) learners or scholars, hence, may find it difficult to understand the language of the law; whereas translators may consider legal lexical phrases and patterns rather intricate to deal with. The literature claims that a practical way to deepen language knowledge can be found in the Web considered as corpus and in online corpora. This paper is aimed at exploring whether commercial search engines, Web concordancers and online specialised corpora can tackle the issues revolving around legal language. In particular, it will investigate whether Google advanced search and the Leeds Web concordancer can be used to meet the requirements of legal language learners, scholars and translators. Furthermore, it will address legal language queries (and results) in an online specialised corpus: the COCA. This paper will provide instances of the soundness of the above-mentioned online resources, especially when used jointly as cross-analysis tools. The shortcomings of one can, in fact, be compensated for by the other(s).

Key words: corpus linguistics, legal English, Web as corpus, online corpora, legal language, legal translations, technical translations, computational linguistics

SIEĆ JAKO KORPUS ORAZ KORPUSY ON-LINE NA POTRZEBY TŁUMACZENIA PRAWNICZEGO

Streszczenie: Język prawny i prawniczy cechuje się dokładnością i żargonowością a jego struktury nie są intuicyjne. Na to nakłada się określony system prawny, który sprawia, że język prawny i prawniczy jest w każdym kraju inny. Tak osoby uczące się drugiego języka (L2) jak i naukowcy mogą uznać język prawny za trudny do zrozumienia, tymczasem tłumacze mogą uważać, że jest on skomplikowany i zawiły, jak i jego przekład. Tymczasem literatura przedmiotu wskazuje, że remedium na te problemy może być sieć użytkowana jako korpus oraz korpusy on-line. Celem niniejszego artykułu jest weryfikacja tego, czy komercyjne przeglądarki internetowe, narzędzia konkordancji, korpusy specjalistyczne on-line mogą być przydatne w rozwiazywaniu problemów wynikających z natury języka prawnego i prawniczego. W szczególności badaniu poddaje się przeszukiwanie zaawansowane w przeglądarce Google i narzędzia konkordancji sieciowej Leeds i specjalistyczne korpusy on-line: COCA. W ten sposób wskazuje się sposób wykorzystania powyższych narzędzi sieciowych oraz ich działanie w sytuacji, gdy wykorzystywane są jednocześnie jako narzędzia do analizy krzyżowej.

Słowa kluczowe: językoznawstwo korpusowe, angielski język prawny i prawniczy, sieć jako korpus, język prawny i prawniczy, przekład prawniczy, tłumaczenie techniczne, językoznawstwo komputerowe

IL WEB COME CORPUS E CORPORA ONLINE PER LE TRADUZIONI GIURIDICHE

Riassunto: Il linguaggio giuridico è caratterizzato da un gergo pedante ed arcaico. Gli studiosi di una lingua straniera, i traduttori ed i professionisti che si approcciano al linguaggio giuridico in lingua straniera, devono tenere presente non solo le peculiarità tecnico-linguistiche, ma anche quelle legate al sistema giuridico di riferimento. Il presente articolo si pone l'obiettivo di mostrare come il Web, considerato come un corpus, può fornire risposte in ambito linguistico e giuridico. In particolare, analizzerà la sintassi di ricerca in Google, il Leeds ed il corpus online COCA. In tal modo si evidenzierà come, usati congiuntamente, questi strumenti possono fornire risposte attendibili in ambito giuridico.

Parole chiave: linguistica dei corpora; Inglese giuridico; il Web come corpus; corpora online; linguaggio giuridico; traduzioni giuridiche

1. The Specificity of the Legal Language

Legal jargon, also referred to as *legalese* (Tiersma 1999; Tiersma & Solan 2012: 22), is hallmarked by lexical peculiarities which make it very different from any other sector language (Tiersma 1999; Williams 2004, Williams 2011; Tiersma & Solan 2012). Amongst others, are nominalization, embeddings, subordinations, passive constructions, archaisms, influence from Law French and Law Latin (Laster 2001; Bhatia 2010; Tiersma & Solan 2012), anaphoric and cataphoric references (Abate 1998: 14-16), complex lexical phrases (Coulthard & Johnson 2010: 10) and ambiguity in the use of modal verbs (Williams 2005, Williams 2013) or in negations (Tiersma 1999; Coulthard & Johnson 2010: 10). All these features tend to make legal language very difficult to the layperson (Tiersma 1999; Tiersma & Solan 2012: 46; Giampieri 2016b) and very complex to the scholar or the legal translator (Giampieri 2016a). In addition to its lexical complexity, legal language is bounded to the legal system the country where it is used (Rotman 1995; De Groot & Van Laer 2008; Giampieri 2016a: 445). This means that second language (L2) scholars/learners and translators must be acquainted with the legal system of both the source and the target language, in order to fully understand the meaning of legal terms (Giampieri 2016a: 445-446). This may also entail that certain institutions, which are typical of a given legal system, may not be regulated in others. This is the case, for instance, of the Trust, which has no equivalent in the Italian legal system (Longinotti, 2009; Curzio 2014: 26). In addition, as with most of technical jargon, legal English is hallmarked by a wide array of fixed lexical bundles, also referred to as lexical phrases, or multi-words (O'Keeffe *et al.*2007: 63). Lexical bundles are "words which systematically co-occur with other words" (Biber and Conrad 1999: 181). Some examples in the legal sector are: *as laid down in*; *having regard to*; *hereinafter referred to as* and many others. Therefore, non-native speakers (NNS) are also confronted with the challenges of complex phrasal constructs, which

would represent natural hindrances *per se* (Biber & Conrad 1999: 188). For these reasons, it is possible to infer that legal jargon is not L2 learner-friendly.

2. The Web as Corpus and Online Corpora: Literature Overview

Some scholars claim that "the corpus of the new millennium is the Web" (Kilgarriff 2001: 343), because "language is at the heart of the Internet" (Crystal 2006: 271). A corpus (plural: corpora) is a collection of texts of "naturally-occurring language" (Sinclair 1991: 171) in an electronic format, which is consulted in order to understand how language is used. For example, one of the advantages of using the Internet as corpus is the fact that it provides both qualitative and quantitative evidence of attested usage (Rosenbach 2007: 168). However, the Web itself cannot be considered as corpus in the traditional sense of the word, because it is a "sprawling, gargantuan, inexhaustible entity" (Gatto 2014: 2), whose data are ever-changing, overwhelmed by duplicates and too dynamic to be fully relied on. To this highly-debated question, however, some scholars reply by arguing that the constantly flowing water of a river shares the same fate, which, however, does not prevent it from being tested (Kilgarriff 2001: 343). Therefore, if scholars wish to query terms on the Internet, they would need to use a commercial search engine such as Google and some common sense. It is argued, in fact, that most of the Internet users look for terms lazily and naively; consequently, they tend to misuse the Web as a linguistic resource (Battelle 2005: 23-25; Gatto 2008: 53; Gatto 2014: 79). Therefore, a cautious approach should always be adopted when submitting queries and interpreting results. As a matter of fact, "webidence", or "Web as linguistic evidence" (Fletcher 2007: 36 also quoted in Gatto 2008: 58 and Gatto 2014: 87); i.e., high matches (or hits) simplistically and mistakenly considered as evidence of attested usage, is very likely to lead inexperienced users astray. Itis claimed that "Googleology is bad science" (Kilgarriff 2007: 1), because the number and type of matches are not consistent over time. Furthermore, commercial search engines are not designed

for linguistic purposes (Gatto 2008; Gatto 2014: 75), as they normally find "contents, not linguistic forms" (Ferraresi 2009: 2). What is also criticised about the Web as corpus and the use of Google to explore it linguistically, is the fact that Google is a "poor concordancer" (Sharoff 2006: 64). A concordancer is a programme which retrieves and displays data from a given corpus for further analysis (Gatto 2014: 18). A concordancer shows concordance lines, which are instances of sentences containing the term(s) in question, displayed and ordered in a manner suitable for readers (Gatto 2014: 9). It goes without saying that Google cannot provide concordance lines in a such a way to carry out systematic and organised linguistic analyses. Furthermore, Google shows neither collocations nor colligations, which are important linguistic aspects. Collocations concern "patterns of usage" (Gatto 2014: 29-30) and refer to the likelihood of co-occurrence of lexical items (Lehecka 2015). Colligations, instead, regard the co-occurrence of syntactic categories, or better the "occurrence of a grammatical class or structural pattern with another one, or with a word or phrase" (Sinclair 2003: 173). Therefore, "what collocation is on a lexical level of analysis, colligation is on a syntactic level" (Römer 2005: 13). As can be seen, the linguistic richness of a text can be multifaceted; consequently, specific tools of analysis are mandatory. In this respect, by using Google advanced search, queries can be quite precise. For example, the Boolean operators OR, AND, NOT (Gatto 2008: 55) allow to include or exclude terms from the search. By searching exact phrases within inverted commas (e.g. *"contract termination"*), it is possible to narrow the search down to specific words in a given sequence. Furthermore, it is possible to instruct Google to search only within a given domain by using the command *site:*, or to exclude other domains, by using the command *site:-*. As can be seen, Google can be "a versatile tool for various forms of empirical language research" (Bergh 2005: 34).

For these reasons, Web concordancers have been developed, which explore the Web linguistically and consider it as corpus. One of these, is the Leeds (Wilson *et al.*2010). The Leeds has the advantages of providing instances of language use from the Web in a form which is suitable for linguistic analysis (Gatto 2008: 80). For instance, it generates viewer-friendly concordance lines showing the searched term in a bold character. Furthermore it is provided with POS (part of speech) annotation. Annotation

is "adding interpretative linguistic information to a corpus" (Leech 2005: 25). In practice, POS tagging indicates the word class of each word. This makes search easier but most of all, it helps find collocations and colligations. Another important feature, is the search for lemmas. A lemma (or headword) is "a set of lexical forms having the same stem and belonging to the same major word class, differing only in inflection and / or spelling" (Francis and Kučera 1982: 1). For example, *terminate* is the headword of *terminating, terminated* and *terminates*. Nonetheless, given that the Leeds is grounded in the Web, it shares the same shortcomings (Gatto 2008:99; Gatto 2014: 107); namely, the volatility of the data retrieved. Furthermore, in the Leeds, it is not possible to narrow the search down to specific domains.

In view of the argumentation provided, it could be claimed that the Web might be too vast and disorganised to provide scholars/learners or translators with the right legal terminology and translation equivalents. For this reason, in order to either corroborate or confute this claim, an online specialised corpus will be addressed: the COCA (Corpus of Contemporary American English) (Davies 2008; Davies 2010), in particular its Academic Law Political Science (Acad-LawPolSci) sub-corpus (8,600,386 words).

In light of the above, this paper is aimed at exploring whether the Web and Web corpora can be reliable tools to help scholars unfold the many layers of the language of the law. In order to do so, the Web will firstly be considered as corpus and investigated by means of Google commercial search engine and a Web concordancer. Afterwards, the COCA will be queried in order to verify whether it can provide useful insights into legal language and help scholars/learners and translators deal with its complexity.

3. Analysis of The Web as Corpus and Web Concordancers

As stated above, the complexity of legal language constructs cannot go unnoticed to L2 learners, as meticulous scholars and translators likely to find the specific jargon rather difficult. For this reason, it could be argued that commercial search engines and Web

concordancers are not suitable for legal linguistic research because they tend to be too vast and unspecific. This paper will explore whether this claim is actual or not. As a matter of fact, the "vexed question" (Tognini-Bonelli 2001: 57) of the representativeness of a corpus is of paramount importance when carrying out linguistic analyses and searching for terms (Sinclair 2005). Although it is self-evident that the whole Web is representative *per se*, it cannot be denied that, as claimed above, it is by far too vast and disorganised to allow clear-cut linguistic analyses. Therefore, if on the one hand commercial search engines might be rich in any kind of unmethodically organised legal terms, on the other hand, legal corpora might be scarce in highly specialised terminology. For instance, one might argue that it is difficult to explore the differences between *rent* and *hire* or *tribunal* and *court* in any given legal corpus, especially if not large. In this respect, however, the literature claims that highly specialised corpora are generally small but, nevertheless, accurate (Aston 1999; Granger 2013: 11).

This paper will hence explore to what extent the Web can be a reliable source of legal terminology and, at the same time, whether online corpora can be consulted for highly-specialised term search. In practice, it will try to find the right balance between managing overwhelming data and finding highly technical terms.

3.1. Google

The literature abounds in guidelines and suggestions on how to write queries in commercial search engines (Baroni & Bernardini 2006; Gatto 2008; Zanettin 2012; Gatto 2014). For example, in Google it is advisable to use the advanced search or at least to narrow the search down by using inverted commas in order to look for exact phrases. It would also be sensible to search only in reliable domains (for instance: .gov.uk or .gov) and eschew non-native websites. Very insightful is also the wildcard character (*), which allows to search for unspecified words in a given sentence or phrase. Finally, the Boolean operators (AND, OR, NOT) could be used effectively.

The following pages will show how to make legal queries fruitful by using Google.

Google_Example 1

It is argued that collocations are difficult to learn by NNS, because "their inherent fuzziness makes them difficult objects for language teaching" (Sinclair *et al.*2004: xxiv). Therefore, it could be interesting to investigate the verbs which collocate with *agreement* and *contract*. In order to do so, the search strings would be *" * a contract"* and *"to * an agreement"*. Interestingly enough, in the first case the following results would be retrieved: *to award a contract, to enter into a contract, to execute a contract, to draw up a contract, to end a contract, to make a contract*; whereas in the second: *to have an agreement, to reach an agreement, to find an agreement, to come to an agreement, to execute an agreement*. At this point, the distinction between *contract* and *agreement* could also be made clear by writing, for example: *define: contract site:.businessdictionary.com* and *define: agreement site:.businessdictionary.com*.

Google_Example 2

As suggested by the literature, also colligations are worthwhile exploring (Sinclair 2003; Römer 2005; Gatto 2014: 29-31). For instance, law scholars/learners might be intrigued by the syntactic categories which precede and follow the words *virtue*, or *derogation* which form recurrent, formulaic legal lexical phrases. A good way to discover such colligations would be by writing the following strings: *"* virtue *"* and *"* derogation *"*. However, in order to make the research more adherent to the legal sector, the query should be restrained to legal domains, such as .justice.gov.uk, which corresponds to the British justice domain. Therefore, the search strings could be *"* virtue *" site:.justice.gov.uk* and *"* derogation *" site:.justice.gov.uk*. In the first case, the lexical phrase *by virtue of* prevails; whereas in the second, lexical and non-lexical phrases appear: *a derogation under, a(ny) derogation from, new derogation for, operational derogation that, designed derogation order, unless derogation has been agreed*. It is self-evident that in this case, a thorough cross-analysis with other linguistic tools (such as dictionaries, Web concordancers or specialised corpora) would be called for, in order to find an unequivocal match, if any.

Google_Example 3

It is argued that NNS might be puzzled about noun pre or post-modification (Gatto 2008: 61-64; Gatto 2014: 96). For example, it might be wondered whether the chunk *employment contract* is more common than *contract of employment*. If one wishes to follow Gatto's advice (2014: 98) and search only in Google Books, for example, the following string could be typed: *"employment contract" Google Books* and *"contract of employment" Google Books*. Then, it would be possible to decide on the basis of the number of matches. In the first case, more than 100,000 matches would be retrieved; whereas in the second only 28,100. It goes without saying that the first bundle is more common. At this point, however, it would be interesting to verify whether the results are corroborated by other English-speaking domains. In order to do so, the following strings could be written: *"employment contract" site:.ie* and *"contract of employment" site:.ie*. In the first case, approximately 27,000 matches would be retrieved; whereas in the second 40,000. This is a case where results lead to discrepancies. Therefore, further linguistic investigations should be called for.

Google_Example 4

This example will address a translation issue. In particular, translation candidates of the Italian *foro competente* will be searched by relying on Google. It is self-evident that the words *"foro competente" English* could be typed to find a translation equivalent. However, if one wishes to be accurate, reliable alternatives should be found. First of all, the word *competente* could be looked up in any online Italian-English dictionary and the word *competent* would be found. Then, the domain *tribunalsdecisions.service.gov.uk* could be chosen in view of its (supposed) targeted content. Hence, the search string is as follows: *"competent *" site:.tribunalsdecisions.service.gov.uk*. Unfortunately, the search would not provide clear-cut results: *competent authority, competent representative, competent doctors, competent under national Law, competent solicitor, competent manner* and *competent court*. The latter could be a possible translation candidate, but its occurrences are too low to be taken for granted (i.e., only 1). Therefore, other search must be undertaken. It would be advisable to exploit the Italian fixed collocation *legge applicabile e foro competente* and opt

for a calque (Longinotti 2009: 29; Scarpa 2014: 233) of *legge applicabile*, which is *applicable law*. At this point, the search string could be the following: *"* and applicable law"* or *"applicable law and *"*. The results are striking, as most of the phrases retrieved are *jurisdiction and applicable law*, which can be considered a perfect translation candidate of *legge applicabile e foro competente*. Hence, *foro competente* means *jurisdiction*.

These examples proved that, to some extent, commercial search engines can help find not only legal terms, but also collocations, colligations and translation candidates. It goes without saying that many are the shortcomings. First of all, as claimed by the literature, the volatility of the information retrieved (Gatto 2008; Gatto 2014: 191), which heavily relies upon the existence or non-existence of (private or public) Websites. Secondly, the fact that translation candidates, collocations and colligations are not easy to find: one must formulate the query correctly, otherwise overwhelming and unreliable information would be retrieved. Thirdly, commercial search engines provide neither a word frequency list, nor recurrent collocations. Lastly, it is not possible to formulate a query which would help find, for example, the adjectives or verbs which precede or follow a noun. This is what POS (part of speech) tagging would perform, but it is self-evident that the whole Web cannot be furnished with annotation.

In view of these shortcomings, it is now interesting to verify whether Web concordancers such as the Leeds can address them. The next pages will deal with examples which will not only overcome issues, but will also raise some questions.

3.2. The Leeds

The Leeds (Wilson *et al* .2010) is a Web concordancer which uses annotation (or POS tagging). In practice, apart from the standard term search, in the Leeds it is possible to investigate which syntactic categories follow a specific verb, or a noun, etc. POS tagging obviously entails knowing the tag (or abbreviation) which corresponds to each part of speech. A list of the tags is provided in the Leeds Website; therefore, tagging is straightforward. The interface also

arranges the searched terms in concordance lines and shows the *urls* which generated them. It is possible to obtain collocates, whose span (or desired position) can be selected (e.g. within 2 words before and after the searched term). Furthermore, in order to find words or terms between two, it is possible to write two dots between the words in question; whereas lemmatization is instructed by using the symbol %.

Leeds_Example 1

As in Google_Example 1, it could be interesting to verify which verbs and determiners precede *contract*. In order to do so, the search string is as follows: *[pos="VV.*"] [pos="DT"] contract*. Some of the results are the following: *finalise a contract, locating the contract, view the contract, accepted the contract, approve this contract, argue that contract, awarded a contract, breached the contract, end the contract, enforced the contract,* etc. As can be seen, some terms are similar to the ones found in Google_Example 1 above. This, however, comes as no surprise, given the fact that the Leeds is a Web concordancer; i.e., it is rooted in the Web.

Leeds_Example 2

It is argued that a contract cannot be *terminated* by *default* or *breach* (Giampieri 2016a), where *default* and *breach* are consequences of the non-payment by a party. In such a case, in fact, a contract is *cancelled*, not *terminated* (UCC 1972; Giampieri 2016a). It would be interesting to investigate whether the lemma (or headword) *terminate* collocates with *default* and/or *breach* in the Leeds. The search string could be written as follows: *[lemma="terminate"] .. default* or *terminate% .. default*. The results are interesting, as only two concordance lines are retrieved, which, however, are unrelated to legal matters. Table 1 here below shows the concordance lines obtained.

Table 1: Concordance lines of the search *[lemma="terminate"] .. default*.

) certificate cannot be found) the session will also be [*sic*.]	**terminated.**	The default is never. 6.4 slapd. c onf Backend Directives [*sic*.]

on the number of backtracks allowed before a search is	**terminated (**	default: 125). The limit prevents some legitimate,

By following the same search syntax, it is possible to investigate whether the lemma *terminate* collocates with breach. In such a case, no concordance lines would be retrieved. Hence, literature findings (UCC 1972; Giampieri 2016a) are underpinned.

Leeds_Example 3

The nouns and verbs *rent* and *hire* are considered synonyms by many bilingual dictionaries and translated *affitto* or *locazione* indistinctly. In order to better grasp their differences, it would be useful to search for their collocates. The query should be formulated in order to search for nouns which collocate with *rent* and *hire* within a span of 4 words. Table 2 highlights how to formulate the query.

Table 2: Search for collocations of *hire*

Search query	hire
Context	4 words on the left 4 words on the right
POS tag of the collocate	NN.* POS tags

Note: The tag *NN.** means "any noun".

The same can be repeated for *rent*. Table 3 reports some of the collocations of *hire* and *rent*.

Table 3: Noun collocates of *rent* and *hire*.

Collocates of *rent*	Collocates of *hire*
rent ~~ car	hire ~~ company
rent ~~ property	hire ~~ employee
rent ~~ apartment	hire ~~ someone

rent ~~ month	hire ~~ staff
rent ~~ payment	hire ~~ people
rent ~~ house	hire ~~ car
rent ~~ disclaimer	hire ~~ lawyer
rent ~~ tenant	hire ~~ employer
rent ~~ landlord	hire ~~ consultant

From Table 3 above, it is possible to infer that *rent* collocates with immovable goods (*property*, *apartment*, *house*). In particular, the last two words (*tenant* and *landlord*) describe the people involved in house letting. *Hire*, instead, collocates with people and in particular with the world of work (*employee*, *people*, *employer*, *staff*, *lawyer*, *consultant*). It would be possible to guess that *hire* refers both to people who work for a company on a stable basis (*staff*, *employee*) and people who work independently on a case-by-case basis (*lawyer*, *consultant*). Finally, both *hire* and *rent* collocate with movable goods (*car*).

Leeds_Example 4

The online English-Italian Collins dictionary translates both *tribunal* and *court* as *tribunale*. In order to explore the differences between these two terms, it might be useful to search for collocates. In order to make the research as broad as possible, collocations should be searched up to 4 words before and after the term in question. Unfortunately, function words (or grammatical words, such as determiners: *the, a, an, this, his, her*..) cannot be excluded. Table 4 here below shows how to formulate the query.

Table 4: Search for collocates of *tribunal*.

Search query	tribunal
Context	4 words on the left 4 words on the right

After excluding the function words, the collocations of *tribunal* are the following: *military, crime, war, international,*

employment, competent, independent, Hussein, industrial, Hague; whereas the words which collocate with *court* are: *appeal, district, federal, order, case, ruling, rule, decision, trial, state, judge, supreme*. Therefore, it can be inferred that *tribunal* is a term used for specific purposes (*military, international, crime, war, employment*); whereas *court* is the term commonly used to describe the place where justice is governed. Furthermore, it is apparent that *court* is used in North American (*federal, state*).

In light of these examples, it can be claimed that the Leeds is a useful tool to explore language patterns. The POS tagging, for example, is particularly insightful. Nonetheless, the Leeds is grounded in the Web and it is not based on a legal corpus. Furthermore, the fact that domains cannot be selected makes search quite random and unspecific.

In light of the above, it can be stated that the Leeds is an effective language aid, especially if used in conjunction with other tools, such as Google search and dictionaries. Nonetheless, it might not completely fulfil the eagerness for learning of legal scholars as it does not always address the legal language specificity. Furthermore, the Leeds is not provided with a site-restriction function, which makes its results quite unspecific. For these reasons, legal scholars may find online specialised corpora more useful

3.3 Analysis of an Online Specialised Corpus: the COCA

The COCA (Corpus of Contemporary American English; henceforth COCA) is a corpus organised in many sub-corpora. The law section relies on an Academic Law and Political Science sub-corpus (8.6 Mln words approximately). It is provided with POS tagging; hence, queries and results can be extremely precise. In the COCA it is possible to obtain concordance lines, collocates and KWiC (key words in context, Sinclair 2003: 176; Bergh 2005; Wilson *et al.*2010; Zanettin 2012; Gatto 2014). The POS function can be applied both to the search term and to its collocate(s), which makes the search particularly versatile and the results very accurate. Furthermore, the position of the collocate(s) can be chosen. Finally, the wildcard

character "*" can be used to search for lemmas. Many others are its features and the literature abounds in examples and guidelines on how to exploit its full potential (Davies 2008; Davies 2010). For reasons of space, however, the following pages will focus on some of its main features.

COCA_Example 1

It would be useful to understand the differences between *liable for* and *liable to*, which seem to be similar. A good way to proceed, is by generating and analysing concordance lines. Therefore, from the menu tab we select *List* and type *liable for* in the field. By selecting the Acad-LawPolSci sub-section and clicking on *Find matching string*, 178 concordance lines would be retrieved, such as *was held liable for tort damages; liable for alleged flaws in communicating information; hold manufacturers liable for the external risks*. When searching for *liable to* by following the same methodology, 57 concordance lines would be retrieved, such as *liable to trigger procedural defects; liable to be a long process; liable to forget important points*. From the concordance lines obtained, it is possible to infer that *liable for* means *legally responsible for*; whereas *liable to* means *likely to*. Hence, the first one is more frequent in legal texts, which is also underpinned by the higher matches.

COCA_Example 2

Leeds_Example 2 proved that the lemma *terminate* does not collocate with *default*. It would be sensible to verify this in the Acad-LawPolSci section of the COCA. To this aim, we choose *Collocates* from the menu tab, write *terminat** in the *Word/phrase* field and *default* in the *Collocates* field. No concordance lines are generated. However, in Leeds_Example 2 it was claimed that *breach* is a synonym of *default*. It would make sense to write *breach* instead of *default* in the *Collocates* field. Strangely enough, no hits are found in the Acad-LawPolSci sub corpus, but one is retrieved from a Magazine section: *claiming wrongful termination, breach of contract*. One might argue, however, that this is not a reliable source of legal terminology.

COCA_Example 3

Google_Example 3 revealed that with the words *employment* and *contract*, pre-modification (*employment contract*) prevailed over

post-modification (*contract of employment*). It would be interesting to verify whether the COCA corroborates Google findings. In order to do so, we use the *Collocates* function and write *employment* in the *Word/phrase* field and *contract* in the *Collocates* field. With the view to narrowing the search down, the word span is restricted to 2 words before and 1 after the term in question. 14 concordance lines with *employment contract* are retrieved and only 3 with *contract of employment*. Hence, Google Books findings are corroborated. This, however, does not imply that *employment contract* is *per se* the most used phrase. A corpus, in fact, "can only tell us what is or is not present in the corpus" (Bennet 2010: 3).

COCA_Example 4

Leeds_Example 3 highlighted the differences between *rent* and *hire* by showing their collocates. It could be useful to explore them in the Acad-LawPolSci section of the COCA. By typing *rent* in the *Word/phrase* field and hitting the button, the system automatically types an asterisk in the *Collocates* field. Some of the collocates retrieved are: *seeking, pay, risk-free, tenants, extraction, less, market, landlord, reflects, charge, space, fully, land, office, apartment, costs, room, two-bedrooms*. Some of the collocates of *hire* are, instead: *you, lawyer, firms, attorney, refuse, fire, workers, want, employers, him, applicant, temporary*. Hence, the COCA results corroborate Leeds findings; i.e., that *rent* collocates with immovable goods (and, again, with the two main parties of a tenancy agreement; i.e., *landlord* and *tenant*); whereas *hire* collocates with people and professionals. There is no mention of movable goods, instead.

As could be seen, the COCA is a reliable legal language tool, which provides useful information on both general and highly specialised legal matters. Furthermore, it can be used in conjunction with other linguistic resources (such as dictionaries and Web concordancers) in order to corroborate legal language pattern

4. Conclusions

Legal language is hallmarked by complex constructs which makes it very different from any other technical language (Tiersma 1999; Tiersma & Solan 2012; Williams 2004, 2011). The path to deepen

the knowledge of legal English is, hence, treacherous and L2 scholars and translators are called on painstaking activities in order to learn the peculiarities and the formulaic, fixed terms of the language of the law. Nonetheless, the Web and online corpora could be helpful, although some precautions should be taken in order avoid naïve Internet search or unfruitful page consultations (Battelle 2005: 23-25; Gatto 2008: 53). The literature provides instances on how the Web, Web concordancers and online corpora can be valid alternatives (and supplements) to dictionary search in language learning and translation (Kilgarriff 2001; Baroni & Bernardini 2006; Zanettin 2012; Gatto 2014). Hence, this paper was aimed at exploring how the Web could be used as corpus for legal purposes. In addition, it highlighted how specialised corpora could be a reliable resource to help dissipate linguistic doubts. In particular, it investigated how cross analyses and targeted search could help eager law scholars and translators overcome language hindrances. To this aim, Google structured queries were firstly tackled and it was underpinned how, by narrowing search down and restricting domains or searched terms, it became a useful language tool. Nonetheless, Google reliability could not always be taken for granted (see Google_Example 2 and 3). Therefore, other online tools needed considering. A Web concordancer such as the Leeds (Wilson *et al.*2010), for instance, proved to be satisfactory, albeit sharing the volatility which is typical of commercial search engines. The Leeds was a practical tool provided with POS tagging and a Collocation search function, although it did not have any commands to exclude function words or to search for terms in specific domains. Nonetheless, if used together with other online tools, it proved to be fruitful, despite being based on the Web. In order to address these shortcomings, an online specialised corpus was also tackled: the COCA and its Acad-LawPolSci sub-corpus (Davies 2008; Davies 2010). The COCA provided an answer to every query, even the most specialised and intricate ones (e.g. COCA_Example 4: *rent* vs. *hire*). Therefore, it can be considered an effective language tool, in particular if used in conjunction with other resources, such as dictionaries, Google queries and Web concordancers.

In light of these findings, this paper claims that in their linguistic search, legal English learners/scholars and translators can be supported by the Web as corpus and online specialised corpora. However, this can take place as long as they are cautious

and forbearing enough to use an array of online resources and do not rely only on one linguistic tool.

References

Abate, Salvatore Claudio. 1998. *Il documento legale anglosassone*. [The Anglo-Saxon legal document]. Milan: HOEPLI.

Aston, Guy. 1999. "Corpus use and learning to translate", *Textus*, 12 (1999): 289-314. http://www.sslmit.unibo.it/~guy/textus.htm (accessed October 13 2018).

Baroni, Marco, and Silvia Bernardini. 2006. *Wacky! Working Papers on the Web as Corpus*. Bologna: Gedit. http://wackybook .sslmit.unibo.it/wackycontents.html (accessed October 13 2018)

Battelle, John. 2005. *The Search: How Google and Its Rivals Rewrote the Rules of Business and Transformed Our Culture*. Boston: Nicholas Brealey Publishing.

Bennet, Gena R. 2010. *Using corpora in the langauge learning classroom: corpus linguistics for teachers*. Michigan: University of Michigan Press ELT.

Bergh, Gunnar. 2005. "Min(d)ing English language data on the Web. What can Google tell us?", *ICAME Journal*, 29: 25-46.

Bhatia, K.L. 2010. *Textbook on legal language and legal writing*. New Delhi: Universal law publishing Co. Pvt. Ltd.

Biber, Douglas, and Susan Conrad. 1999. "Lexical bundles conversation and academic prose". In *Out of Corpora: Studies in Honour of Stig. Johansson*, ed. H. Hasselgard and S. Oksefjell, 181-189. Amsterdam: Rodopi.

Carter, Ronald A., and Michael McCarthy, J. 2006. *Cambridge Grammar of English*, Cambridge: Cambridge University Press.

Cobb, Tom. 2004. "The Compleat Lexical Tutor, v.4", *TESL-EJ*, 8.3, http://www.teslej.org/wordpress/issues/volume8/ej31/ej31m2/ ?wscr (accessed October 13 2018)

Coulthard, Malcolm, and Alison Johnson. 2010. *The Routledge Handbook of Forensic Linguistics*. Abingdon: Routledge.

Curzio, Stefano. 2014. *Tutela del patrimonio e Trust*. [Assets Custody and Trust]. Santarcangelo di Romagna: Maggioli Editore.

Crystal, David. 2006. *Language and the Internet. Second Edition*. Cambridge: Cambridge University Press.

Davies, Mark. 2008. "The Corpus of Contemporary American English (COCA): 520 million words, 1990-present", *Davies Mark*, 2008, http://corpus.byu.edu/coca/ (accessed October 13, 2018).

Davies, Mark. 2010. "The Corpus of Contemporary American English as the First Reliable Monitor Corpus of English", *Literary and Linguistic Computing*, 25, 4: 447-64.

De Groot, Gerard-René, and Conrad Van Laer, J. P. 2008. *The Quality of Legal Dictionaries: an assessment*. Maastricht: University of Maastricht.

Ferraresi, Adriano. 2009. "Google and beyond: web-as-corpus methodologies for translators", *Revista Tradumàtica*, 7, http://webs2002.uab.es/tradumatica/revista/num7/articles/04/04art.htm (accessed October 13, 2018).

Fletcher, William H. 2007. "Concordancing the Web. Promise and problems, tools and techniques". In *Corpus linguistics and the Web*, ed. M. Hundt, N. Nesselhauf and C. Biewer, 25-46. Amsterdam: Rodopi.

Francis, Nelson W., and Henry Kučera. 1982. *Frequency analysis of English usage: Lexicon and Grammar*. Boston: Houghton Mifflin.

Gatto, Maristella. 2008. *From body to Web: an introduction to Web as corpus*. Roma, Laterza.

Gatto, Maristella. 2014. *Web as corpus: Theory and Practice*. London: Bloomsbury.

Giampieri, Patrizia. 2016a. "A Critical Compartive Analysis of Online Tools for Legal Translations", *The Italian Law Journal*, 2.2: 445-461.

Giampieri, Patrizia. 2016b. "Is the European legal English legalese-free?" *Italian Journal of Public Law* 8.2: 424-440.

Granger, Sylviane. 2013. *Learner English on Computer*. Abingdon: Routledge.

Greaves, Chris. 2009. *Concgram 1.0: a phraseological search engine. CD-ROM*. Amsterdam: John Benjamins.

Kilgarriff, Adam. 2001. "Web as corpus". In *Proceedings of the Corpus Linguistics 2001 Conference*, ed. P. Rayson, A.

Wilson, T. McEnery, A. Hardie and S. Khoja, 342-344. Lancaster University.

Kilgarriff Adam. 2007."Googleology is bad science", *Association for computational linguistics*, 33.1: 147-151.

Laster, Kathy. 2001. *Law as Culture*. Sydney, The Federation Press.

Leech, Geoffrey. 2005. "Adding Linguistic Annotation". In *Developing Linguistic Corpora: a Guide to Good Practice*, ed. M. Wynne, 25-39. Oxford: Oxbrow Books.

Lehecka, Tomas. 2015. "Collocation and colligation". In *Handbook of Pragmatics Online*, ed. J.-O. Östman and J. Verschueren, 1-20. Amsterdam: John Benjamins.

Longinotti, Daniela. 2009. "Problemi specifici della traduzione giuridica: traduzioni di sentenze dal Tedesco e dall'Inglese" [Specific Problems of Legal Translation: Translation of Judgement from German and English]. *Quaderni di Palazzo Serra*. 17: 1-38. http://www.disclic.unige.it/pub/17/longinotti.pdf (accessed October 13, 2018)

O'Keeffe, Anne, Michael McCarthy, and Ronald Carter. 2007. *From Corpus to Classroom: Language Use and Language Teaching*. Cambridge: Cambridge University Press.

O'Keeffe, Anne, and Michael McCarthy. 2010. *The Routledge Handbook of Corpus Linguistics*. London: Routledge.

Renouf, Antoinette, and John Sinclair. 1991. "Collocational Frameworks in English". In *English Corpus Linguistics: Studies in Honour of Jan Svartvik*, ed. K. Aijmer and B. Altenberg, 128-144. London: Longman.

Römer, Ute. 2005. *Progressives, Patterns, Pedagogy: A Corpus-driven Approach to English progressive forms, functions, contexts and didactics*. Amsterdam: John Benjamins.

Rosenbach, Anette. 2007. "Exploring constructions on the Web: a case study". In *Corpus Lingusitics and the Web*, ed. M. Hundt, N. Nesselhauf and C. Biewer, 167-190. Amsterdam: Rodopi.

Rotman, Edgardo. 1995. "The Inherent Problems of Legal Translation: Theoretical Aspects", *Indiana International and Comparative Law Review*, 6.1: 187-196.

Scarpa, Federica. 2014. "L'influsso dell'inglese sulle lingue speciali dell'italiano" [The influence of English on Italian special languages], *Rivista internazionale di tecnica della traduzione*. Trieste: Edizioni Università di Trieste, 16.14: 225-243.

Sharoff, Serge. 2006. "Creating general-purpose corpora using automated search engine queries". In *Wacky! Working Papers on the Web as Corpus*, ed. M. Baroni and S. Bernardini, 63-98. Bologna: Gedit.

Sinclair, John. 1991. *Corpus Concordance Collocation*. Oxford: Oxford University Press.

Sinclair, John. 2003. *Reading Concordance: an introduction*. London: Longman.

Sinclair, John. 2005. "Corpus and Texts: basic principles". In *Developing Linguistic Corpora: a Guide to Good Practice*, ed. M. Wynne, 5-24. Oxford: Oxbrow Books.

Sinclair, John, Susan Jones, and Robert Daley. 2004. *English Collocation Studies: The OSTI Report*. London: Continuum.

Tiersma, Peter M. 1999. *Legal Language*. University of Chicago Press.

Tiersma, Peter M. and Lawrence M. Solan. 2012. *The Oxford Handbook of Language and Law*. New York: Oxford University Press.

Tognini-Bonelli, Elena. 2001. *Corpus Linguistics at Work*. Amsterdam: John Benjamins.

Uniform Commercial Code (UCC) Act 174 of 1962 § 440.2106. *Contract, agreement, contract for sale, sale, present sale; definitions of certain terms* http://legislature.mi.gov/doc.aspx?mcl-Act-174-of-1962 , web (accessed October 13, 2018)

Williams Christopher. 2004. "Legal English and Plain Language: an introduction", *ESP Across Culture*, 1: 111-124.

Williams, Christopher. 2005. "Vagueness in legal texts: is there a future for shall?". In *Vagueness in Normative Texts*, ed. N. Gotti, V. Bhatia, J. Engberg and D. Heller, 201-224. Bern: Peter Lang.

Williams, Christopher. 2011. "Legal English and Plain language: an update", *ESP Across Cultures*. 8: 139-151.

Williams, Christopher. 2013. "Changes in the verb phrase in legislative language in English". In *The Verb Phrase in English: Investigating Recent Language Change with Corpora*, ed. B. Aarts, J. Close, G. Leech and S. Wallis, 353-371. Cambridge: Cambridge University Press.

Wilson, James, Anthony Hartley, Serge Sharoff, and Paul Stephenson. 2010. "Advanced corpus solutions for humanities researchers". In *Proceedings of the 24th Pacific Asia*

Conference on Language, Information and Computation, ed. Ryo Otoguro, Kiyoshi Ishikawa, Hiroshi Umemoto, Kei Yoshimoto and Yasunari Harada, 769-778, http://anthology.aclweb.org/Y/Y10/Y10-1089.pdf (accessed October 13, 2018)

Zanettin Federico. 2012. *Translation-Driven Corpora: Corpus Resources for Descriptive and Applied Translation Studies*. London: Routledge.

Online Resources

COCA – Corpus of Contemporary American English: http://corpus.byu.edu/bnc

Collins Dictionary: https://www.collinsdictionary.com

Google Advanced Search: https://www.google.com/advanced_search

Leeds: http://corpus.leeds.ac.uk/internet.html