A CORPUS-DRIVEN ANALYSIS OF N-GRAMS IN ENGLISH AND ITALIAN TERMS OF SERVICE

Patrizia GIAMPIERI
University of Camerino, Italy[1]

Gianluca MARRAS
Independent Researcher, Italy

Abstract: *This paper aims at exploring English-Italian equivalences in the N-grams generated from comparable corpora of online terms of service (ToS). To do so, 5-6-word N-grams composed of the most frequent English and Italian nouns, verbs, adjectives and adverbs are analyzed. The English-Italian nouns focused on are the following ones: "service(s)" and "servizi(o)", "agreement / contract" and "contratto", and "customer" and "cliente". The verbs are "be" and "essere", and "agree" and "accettare". The adjectives are "such" and "tale", and the adverbs are "expressly" and "espressamente". The analysis brings to the fore similarities across the N-grams. For example, the lemmas "terminate" and "recedere" can be considered as equivalents, as well as, to some extent, the verb phrases "agrees to be bound" and "prende atto ed accetta". In addition, false semantic equivalences appear, such as the N-grams composed of "such" and "tale", or those made up by "expressly" and "espressamente". The findings also report the highly repetitiveness of the language of the Italian corpus and, hence, the resulting high frequency of some Italian N-grams. Conversely, the English corpus features more varied lexical patterning with lower frequency rates per N-gram.*

Keywords: *N-grams; comparable corpora; legal linguistics; corpus linguistics; legal translation*

1 Introduction

N-grams, also referred to as lexical bundles or multi-words (Biber, Conrad), are sequences of *n* words or characters occurring in sequence in a text. When approaching a second language, users should become acquainted with corpus queries, one of which concerns the generation of N-grams (Bernardini, Ferraresi 212; Mikhailov 235). N-grams are considered useful for language teaching and learning because they highlight meaningful segments of linguistic data. Bernardini and Ferraresi (212), in fact, posit that N-grams help relevant corpus-sourced words or phrases come to the fore. In this way, they

[1] Although the authors have collaborated closely throughout the study, Patrizia Giampieri is personally responsible for the following sections: (1), (3), (4.1) and (5) and Gianluca Marras is personally responsible for the following sections: (2) (4.1), (4.2), (4.3) and (4.4). Section (6) is the result of a common effort.

allow language learners to grasp and familiarize themselves with units of language patterning.

1.1 N-grams in the legal language

In addition to the reasons mentioned above, N-grams are generated to understand the features and patterns of sector-based or technical language, such as the legal one (Katz, Bommarito, Seaman, Agichtein; Goźdź-Roszkowski *Discovering patterns and meanings* and *Corpus linguistics in legal discourse*; Torikai; Giampieri *Key N-grams*; Williams). According to Goźdź-Roszkowski (*Corpus linguistics in legal discourse* 1523), observing legal N-grams is useful because such contiguous word combinations go beyond the traditional corpus-sourced linguistic categorization: "[N-grams] elude the traditional predefined linguistic categories characteristic of the corpus-based approach".

Katz, Bommarito, Seaman and Agichtein carry out a diachronic analysis of judicial N-grams to trace the evolution of legal language in courts. Their investigation is particularly important because, as they state, with common law the relevance of legal rules is often associated with their recurrence over time, together with the frequency of use of the related terminology.

In another study, Goźdź-Roszkowski (*Discovering patterns and meanings* 60-66) explores the N-grams of a corpus of opinions of the Supreme Court of the United States of America from 2000 to 2007. The author focuses on the N-grams with the word "discovery", which is relevant in US trial practice and criminal proceedings. The author retrieves interesting collocations, such as "vacate the discovery", "dispute the discovery", "objections to the discovery", "broad discovery", "pretrial discovery", and so on.

Torikai compares the N-grams generated from the BNC (British National Corpus) with the ones produced by legal corpora. The findings highlight that the latter are limited in type but have high-frequency rates.

Giampieri (*Key N-grams*) explores and compares the key N-grams of a corpus of European directives on distance consumer contracts with a UK national legislation corpus on the same subject-matter. The author unveils five main features: lexical differences in the key N-grams owing to institutional or country-specific factors; use of *legalese*; adoption of Eurolect by EU institutions; divergences in the usage of some terms (despite referring to the same legal principles), and polysemous words (i.e., words with sector-dependent meanings).

Williams (189-193) uses Google Books N-gram Viewer to investigate the usage of the expressions "plain language" and "plain English" since 1800.

1.2 Terms of service

The literature has dedicated a great deal of research in investigating the language of (online) terms of sale/service (Daiza; Martínez, Mollica, Liu, Podrug, Gibson; Williams; Giampieri *Corpus-based translation*). Williams (140) posits that online terms of sale/service "have become an increasingly pervasive presence in our daily lives, especially for internet users". The B2C field (i.e., the one that involves businesses and consumers) has particularly raised scholarly interest (Martínez, Mollica, Liu, Podrug, Gibson; Williams). Some of the reasons are due to the fact that consumers are the vulnerable party to such contracts (Moresco; Williams) and, at the same time, terms of sale/service are easily retrievable from the Internet (Williams 139; Giampieri *Corpus-based translation* 38).

2. Research questions

In light of the above, this paper wishes to explore and compare N-grams in English and Italian online terms of service in order to find equivalents, if any. To this aim, English and Italian comparable corpora of terms of web hosting services (Giampieri *Corpus-based translation*) are consulted via the Sketch Engine platform. The research questions that this paper aims to address are the following ones: 1) "Is it possible to find English-Italian equivalents in the analysis of 5-6-word N-grams in comparable corpora of online terms of service?", and 2) "Which language patterns mostly emerge from the analysis?".

3. Methodologies

N-grams are generated from comparable corpora of online terms of service (ToS) (Giampieri *Corpus-based translation*), that are uploaded to the Sketch Engine interface. Although the corpus building process is described in Giampieri (*Corpus-based translation* 40-46), it is reported in this section for reasons of completeness.

The texts composing the English corpus are online terms of web hosting services collected in 2021. The English terms and conditions of service are drawn up according to the laws of England and Wales. The documents are anonymized: companies' names are replaced by "xx" and their addresses, websites, VAT and phone numbers are removed. The English corpus is composed of 168,199 words (7,924 word types). The same procedure is followed to build the Italian corpus, which comprises Italian terms of web hosting services written in Italian according to the laws of Italy. It is composed of 133,681 words (6,998 word types). For more information on the creation of the English and Italian ToS corpora, see Giampieri (*Corpus-based translation* 40-46).

The corpora are uploaded to the Sketch Engine platform, where N-grams are easily generated. For the purposes of this paper, the N-grams composed of the most frequent nouns, verbs, adjectives and adverbs in each ToS corpus are addressed, compared and analyzed. In this way, possible equivalent N-grams may come to the fore. The most frequent English nouns and their Italian equivalents in the ToS corpora are "service(s)/*servizi(o)*", "customer/*cliente*" and "agreement/*contratto*". The most recurrent verbs are "be/*essere*" and "agree/*accettare*"; the most frequent adjectives and adverbs are "such/*tale*", and "expressly/*espressamente*".

4. Analysis

This section focuses on the N-grams generated from English and Italian corpora of online terms of web hosting services (Giampieri *Corpus-based translation*). In particular, the 5-grams and 6-grams containing the frequent words above-mentioned are dealt with. In order to make the searches all-inclusive, such words are queried case-insensitively. The analyses follow a comparative approach and present N-grams from the English and Italian ToS corpora. Data are reported in tables, showing the N-grams and their frequency per corpus in percentage values. Sample phrases are also extracted from the English and Italian corpora.

4.1 N-grams composed of frequent nouns

As mentioned, the most frequent nouns in the English ToS corpus are "service(s)", "customer", and "agreement/contract".

In the Italian ToS corpus, the most recurrent nouns are *servizi(o)*, *cliente*, and *contratto*, which are the equivalents of the English terms. Table 1 below shows the English N-grams containing "service" and "services".

For reasons of space, the data concerning the related equivalents, i.e., *servizi* and *servizio* are reported in different tables.

5-6-grams with "service"	**Freq. %**	**5-6-grams with "services"**	**Freq.%**
Terms and conditions of service	0.06	The provision of the services	0.62
The provision of the service	0.06	With the provision of the services	0.20
Service interruption in the availability	0.05	Your use of the services	0.16
Product or service sold by the	0.05	In respect of the services	0.15
The supply of the service	0.05	In relation to the services	0.13

Table 1. N-grams with "service/services" in the English ToS corpus

As can be noticed in Table 1, the N-grams containing "service" show lower frequencies than the N-grams with "services". This is to be expected, as providers generally offer multiple services. However, a common recurrent multi-word is "the provision of the service(s)". Some relevant terms are the nouns "supply" and "provision" which precede "of the service" (left column of Table 1). More precisely, "supply" and "provision" can be considered as synonymous, as the following corpus-driven phrases highlight: "a contract shall be created between XX and the Customer for the supply of the Service", and "make payment of any sum due for the provision of the Services". Interesting prepositional phrases are "in respect of the services" and "in relation to the services" (right column of Table 1) in sentences such as "the charges paid by you in respect of the Services" and "the Fees paid by the Customer in relation to the Services".

Table 2 below exhibits the N-grams containing *servizi(o)* and their translations.

5-6-grams with *servizio*	**Translation**	**Freq. %**	**5-6-grams with *servizi***	**Translation**	**Freq. %**
Non ha usufruito del servizio	Did not use the service	0.23	*Per il mancato utilizzo dei servizi*	For not using the services	0.17
Abbia acquistato un servizio con spazio	Have purchased a service with space	0.23	*Il mancato utilizzo dei servizi*	Not using the services	0.17
Acquistato un servizio con spazio	Purchased a service with space	0.23	*Dei servizi offerti da xx*	Of the services offered by xx	0.15
Un servizio con spazio web	A service with web space	0.23	*Per il tramite dei servizi*	By means of the services	0.13
Acquistato un servizio con spazio web	Purchased a service with web space	0.23	*I servizi offerti da xx*	The services offered by xx	00.13.00

Table 2. N-grams with *servizio/servizi* in the Italian ToS corpus

As observable, the most frequent 5-6-word N-grams containing *servizio* are highly repetitive. Some sample phrases are the following ones: *usufruito del servizio* (translation: "used the service"), and *acquistato un servizio* (translation: "purchased a service"). The N-grams with *servizi* also focus on the use (or non-use) of the services, as in *i servizi offerti* (translation:

"the services offered") or *il mancato utilizzo dei servizi* (translation: "not using the services").

It is apparent that the Italian N-grams are more repetitive than the English ones. For example, *acquistato un servizio* is featured in three N-grams (column to the left, Table 2), whereas *mancato utilizzo dei servizi* and *i servizi offerti da* are noticed in two N-grams each (column to the right, Table 2). Conversely, the English N-grams are lexically richer (see Table 1). Therefore, it can be argued that the English corpus offers a wider variety of lexical patterns.

Equivalent N-grams can be found in the phrases "the provision of the services" and "the supply of the service" (see Table 1), corresponding to *i servizi offerti da xx* (translation: "the services offered by xx") (see Table 2).

Table 3 below shows data on the N-grams with "customer" and *cliente* in the two ToS corpora.

5-6-grams with "customer"	**Freq. %**	**5-6-grams with *cliente***	**Translations**	**Freq. %**
If the customer fails to	0.20	*Il cliente prende atto ed*	The customer acknowledges and	2.40
Be liable to the customer	0.20	*Cliente prende atto ed accetta*	Customer acknowledges and accepts	2.40
Not be liable to the customer	0.15	*Il cliente prende atto ed accetta*	The customer acknowledges and accepts	2.39
The customer in respect of	0.14	*Cliente prende atto ed accetta che*	Customer acknowledges and accepts that	1.83
Responsibility of the customer to	0.13	*Il cliente si impegna a*	Customer undertakes to	0.81
To the customer in respect	0.11	*Caso in cui il cliente*	Case in which customer	0.51

Table 3. N-grams with "customer" and *cliente* in the English and Italian ToS corpora

It is evident that the Italian ToS corpus is quite repetitive and produces very high frequencies of the lexical phrase *il cliente prende atto ed accetta che* (translation: "the customer acknowledges and accepts that"). A synonymous expression which emerges from Table 3 is *il cliente si impegna a* (translation: "the customer undertakes to").

Conversely, the English ToS corpus is more varied and it features different language patterns. For example, there are phrases regarding the possible non-performance or non-observance of the terms (introduced by "fails to"), and instances concerning the responsibility of the parties ("(not) liable to" and "responsibility of"). In Table 3, however, no English-Italian equivalent N-grams are noticed.

The last frequent nouns under investigation are "agreement/contract" and *contratto*. Table 4 and Table 5 display the frequencies of the N-grams containing "agreement" and "contract" from the English ToS corpus, and *contratto* from the Italian one. Equivalent N-grams are written in bold. The reason why two words are addressed in English (namely, "agreement" and "contract") is due to the fact that the word "agreement" is often considered as a full synonym of "contract" in the English legal language and in legal practice[2]. On the contrary, *accordo* ("agreement") has a wider scope than *contratto* ("contract") in Italian and is not perceived as its synonym[3]. The following corpus-sourced sample phrase clarifies this aspect: *il Cliente riconosce di aver dato espressamente il suo accordo per l'esecuzione del servizio* (translation: "the customer recognizes/acknowledges to have expressly given his/her agreement/consent for the performance of the service"). In this case, *accordo* is a synonym of *consenso* ("consent"). For this reason, *accordo* may have a broader meaning than *contratto* and, hence, it is not tackled in this analysis. In addition, it is not as frequent as *contratto* in the Italian ToS corpus: it shows only 73 hits (1.06%), whereas *contratto* appears 932 times (13.55%). In the English corpus, the word "agreement" occurs 1,047 times (13.21%) whereas "contract" 534 (6.74%). Therefore, also as regards frequencies, "agreement" and "contract" can be considered as equivalents of *contratto*. Table 4 and Table 5 report the N-grams with "agreement", "contract", and *contratto*. Equivalences are marked in bold.

5-6-grams with "agreement"	**Freq. %**	**5-6-grams with "contract"**	**Freq.%**
The terms of this agreement	0.25	In connection with the contract	0.10
In connection with the agreement	0.23	Set out in the contract	0.09

[2] The Black's Law Dictionary (8th Edition) (Garner) confirms that the two terms are often used interchangeably. For instance, it quotes "[t]he writing that sets forth such an agreement <a contract> is valid" (Garner 970).

[3] Art. 1321 of the he Italian Civil Code establishes that a *contratto* is *l'accordo di due o più parti per costituire, regolare o estinguere tra loro un rapporto giuridico patrimoniale* (back-translation: "the agreement between two or more parties to constitute, regulate or extinguish a legal relationship among them").

5-6-grams with "agreement"	**Freq. %**	**5-6-grams with "contract"**	**Freq.%**
In connection with this agreement	0.20	Or in connection with the contract	0.08
Or in connection with this agreement	0.18	Its obligations under the contract	0.06
The terms of the agreement	0.18	**Or termination of the contract**	0.06
Its obligations under this agreement	0.16	Sub-contract or deal in any	0.05

Table 4. N-grams with "agreement", "contract" in the English ToS corpus

5-6-grams with *contratto*	**Translation**	**Freq.%**
Presenti condizioni generali di contratto	**These general contract conditions**	0.38
Di recedere dal presente contratto	**To withdraw from / terminate this contract**	0.31
Facoltà di recedere dal contratto	Power/faculty to withdraw from the contract	0.29
La facoltà di recedere dal contratto	The power/faculty to withdraw from the contract	0.22
Contratto in qualsiasi momento e senza	Contract at any time without	0.16
Contratto in qualsiasi momento e	Contract at any time and	0.16

Table 5. N-grams with *contratto* in the Italian ToS corpus

It can be easily understood that "the terms of this agreement" (Table 4, first N-gram to the left) is an equivalent of the Italian *presenti condizioni generali di contratto* (translation: "these/the present general contract conditions", Table 5). The two N-grams also produce similar frequencies (20 in the English ToS corpus, 0.25%, against 26 in the Italian, 0.38%).

Other N-grams from the English corpus mostly focus on the lexical phrase "in connection with the/this agreement/contract". In Italian, conversely, the legal multi-word *facoltà di recedere dal contratto* (translation: "power to withdraw from the contract" or "faculty to terminate the contract") is particularly frequent.

By analyzing the word "contract" in the English ToS corpus, the phrase "or termination of the contract" comes to the fore. It corresponds to the Italian

di recedere dal presente contratto (translation: "to withdraw from / terminate this contract"), although with fewer frequencies (5 in the English ToS corpus, 0.06%, against 21 in the Italian, 0.31%).

4.2 N-grams composed of frequent verbs

The most frequent verbs in the English and Italian ToS corpora are "be" and "agree" (in English), and *essere* and *accettare* (in Italian), which also seem full equivalents. Table 6 lists the frequencies of the verbs "be" and *essere* in the infinitive forms. Equivalent N-grams are marked in bold.

5-6-grams with "be"	**Freq. %**	**5-6-grams with *essere***	**Translations**	**Freq. %**
Not be liable for any	0.53	*Non potere essere ritenere responsabile*	**Cannot be held liable**	0.42
Will not be liable for	0.35	*Il cliente essere tenere ad*	Customer be held to (the)	0.39
But be not limit to	0.34	*Il cliente essere tenere a*	Customer be held to	0.38
We will not be liable	0.32	*Potere essere ritenere responsabile per*	May be held liable for	0.35
Shall not be liable for	0.32	*Non potere essere ritenere responsabile per*	**May/cannot be held liable for**	0.32
Shall not be liable for any	0.27	*Xx non potere essere ritenere*	Xx may not be held	0.29
It be your responsibility to	0.27	*Xx non potere essere ritenere responsabile*	**Xx may not be held liable**	0.28
Will not be liable for any	0.27	*Essere essi di proprietà di*	Be of property of	0.22

Table 6. N-grams with "be" and *essere* in the English and Italian ToS corpora (verbs are generated in infinitive forms)

It is very interesting that both the English and the Italian N-grams use the lexical phrase "be / held liable for". Frequencies are also similar in both languages (e.g., 0.53% in English and 0.42% in Italian). Other similarities across the N-grams do not come to the surface.

Table 7 below exhibits the frequencies and sample phrases of the frequent verbs "agree" and *accettare* (all verbs are in the infinitive forms).

5-6-grams with "agree"	**Freq. %**	**5-6-grams with *accettare***	**Translations**	**Freq. %**
Agree to be bind by	0.20	*Prendere atto ed accettare che*	**Acknowledge and agree that**	2.12
You agree that you will	0.20	*Cliente prendere atto ed accettare*	**Customer acknowledge and agree**	2.01
Agree to indemnify and hold	0.11	*Il cliente prendere atto ed accettare*	**The customer acknowledge and agree**	1.99
You agree that we may	0.10	*Cliente prendere atto ed accettare che*	**Customer acknowledge and agree that**	1.83
Unless otherwise agree in writing	0.09	*Il cliente prendere atto e accettare*	**The customer acknowledge and agree**	0.42
The customer agree to indemnify	0.09	*Cliente prendere atto e accettare*	**Customer acknowledge and agree**	0.42
The customer agree that the	0.08	*Prendere atto e accettare che*	**To acknowledge and agree that**	0.41

Table 7. N-grams with "agree" and *accettare* in the English and Italian ToS corpora (verbs are in the infinitive form)

From Table 7 it is evident that the Italian N-gram *prendere atto ed accettare che* shows far higher frequencies than the English one and, hence, it is particularly repetitive.

The most frequent English lexical phrase "agree and be bound by" is slightly similar to the Italian *prendere atto ed accettare che* (translation: "acknowledge and agree that"). The two N-grams, in fact, entail an understanding between the parties. Although the Italian phrase does not clearly refer to a binding relationship, this may be implied in the verb *accettare*.

As a whole, the two N-grams can be considered as partly equivalent. More varied lexical phrases emerge from the English ToS corpus, such as "agree to indemnify and hold", and "unless otherwise agreed in writing".

4.3 N-grams composed of frequent adjectives

This section addresses the most frequent adjectives in the English and Italian ToS corpora. The equivalent adjectives considered in this section are "such"

and *tale* as they apparently serve similar purposes and have the same meaning. Table 8 below reports data and frequencies in this respect.

5-6-grams with "such"	**Freq. %**	**5-6-grams with *tale***	**Translations**	**Freq. %**
In such a way as	0.08	*Le parti su tale oggetto*	The parties on that/such subject-matter	0.10
The possibility of such damages	0.08	*Tra le parti su tale oggetto*	Between the parties on that/such subject-matter	0.10
From the use of such	0.06	*Nel caso in cui entro tale*	In the event within such	0.10
Of the possibility of such	0.06	*Tra le parti su tale*	Between the parties on such	0.10
Or such other address as	0.06	*Conclusi tra le parti su tale*	Concluded between the parties on that/such	0.10
The use of such elements	0.06	*Caso in cui entro tale*	Case in which such	0.10
From the use of such elements	0.06	*Caso in cui entro tale periodo*	Case in which within such period	0.10
Of the possibility of such damages	0.06	*In cui entro tale periodo*	In which within such period	0.10

Table 8. N-grams with "such" and *tale* in the English and Italian ToS corpora

Table 8 shows that the English ToS corpus generates at least three different language samples with "such": "in such a way as"; "the possibility of such damages" and "from the use of such elements".

The Italian ToS corpus, by contrast, is slightly more repetitive. The most frequent 5-6-grams with the word *tale* are less varied. As a matter of fact, the phrases *conclusi tra le parti su tale oggetto* (translation: "concluded between the parties on that/such subject-matter") and *nel caso in cui entro tale periodo* (translation: "in the event/in the case in which within such period") are particularly repetitive.

An assumed equivalence between English and Italian N-grams may be found in "of the possibility of such" and *nel caso in cui entro tale* (5 hits the first one, 0.06%, and 7 the second, 0.10%). However, if analyzed in context, the two prepositional phrases differ quite substantially. The former, in fact, relates to damages (e.g., "we have been advised of the possibility of such damages"), whereas the latter refers to a timely (re)action by the customer (e.g., *nel caso in cui entro tale periodo il cliente non sollevi alcuna eccezione*

in merito, translation: "in the event the customer will not raise any claims / will not communicate anything to the contrary within such period of time"). Therefore, despite having the same meaning, "such" and *tale* seem to be used in different situations or contexts, and Table 8 does not feature any equivalent N-gram.

4.4 N-grams composed of frequent adverbs

The adverbs "expressly" and its Italian equivalent, *espressamente*, are now analyzed as they are the most frequent in their respective ToS corpus. Table 9 reports data in this regard.

5-6-grams with "expressly"	**Freq. %**	**5-6-grams with *espressamente***	**Translations**	**Freq. %**
Expressly stated in this agreement	0.06	*Accetta espressamente le seguenti clausole*	Expressly accepts the following clauses	0.09
That are not expressly stated	0.06	*Ed espressamente approvata per iscritto da*	And expressly approved in writing from/since	0.07
Expressly stated in these terms	0.06	*Ed accetta espressamente le seguenti*	And expressly accepts the following	0.07
Expressly referred to in them	0.06	*Specificatamente ed espressamente approvata per*	Specifically and expressly approved for/to	0.07
That are not expressly stated in	0.06	*Se non specificatamente ed espressamente*	If not specifically and expressly	0.07
Are not expressly stated in	0.06	*Approva e ed accetta espressamente*	Approves and accepts expressly	0.07
As expressly set out in	0.05	*Ed accetta espressamente le seguenti*	And expressly accepts the following	0.07
Any document expressly referred to	0.05	*Ed accetta espressamente le seguenti clausole*	And expressly accepts the following clauses	0.07

Table 9. N-grams with "expressly" and *espressamente* in the English and Italian ToS corpora

It is remarkable that although "expressly" and *espressamente* are one the literal translation of the other, there are actually no equivalent N-grams with these two adverbs, at least from a lexical (more than a syntactical)

perspective. The English ToS corpus, in fact, mostly produces N-grams with declaratory verbs containing "expressly", such as "state", "refer to" and "set out". Conversely, the Italian N-grams mainly highlight acceptance or approval (as in *accetta espressamente*, translation: "expressly accepts"). Therefore, there are neither similarities nor equivalences among the N-grams of Table 9.

5. Discussion

Similarities and discrepancies emerge from the analysis of the N-grams composed of the most frequent nouns, verbs, adjectives and adverbs in the ToS corpora.

In particular, the Italian language is highly repetitive and, for this reason, tends to be characterized by higher frequencies. Examples are phrases such as *il cliente prende atto ed accetta che* (translation: "the customer acknowledges and accepts that"), and *facoltà di recedere dal contratto* (translation: "power/faculty to withdraw from the contract"). These Italian expressions are recurrent when generating N-grams with frequent nouns. The same occurs to the verbs "agree" and *accettare*, as many Italian N-grams are composed of the phrase *il cliente prende atto ed accetta che*.

By contrast, the language of the English ToS corpus is more varied and richer in language patterns. Examples are the following phrases with frequent nouns and verbs: "the provision of the service(s)"; "terms and conditions of service"; "fails to", and "(not) liable to". With regard to the verb "agree", the English ToS corpus shows a certain lexical variety, as evidenced in the phrases "agree to be bound"; "agree to indemnify", and "unless otherwise agreed in writing". Therefore, from the analyses above, it can be stated that the N-grams of the Italian ToS corpus are lexically less varied than those of the English one.

As far as frequent adjectives are concerned, no equivalent N-grams with similar meaning(s) and/or usages in context are found in the ToS corpora. As a matter of fact, the common adjectives "such" and *tale* do not generate any particularly relevant or similar N-grams. Furthermore, N-grams with "such" are lexically more varied than those with *tale*.

With regard to the adverbs "expressly" and *espressamente*, it is interesting to highlight that the language of the English corpus uses "expressly" in declaratory forms, such as with the verbs "state", "refer to" and "set out". On the contrary, the language of the Italian corpus tends to use *espressamente* to reinforce acceptance or approval. Therefore, no equivalents are noticed.

It would be sensible to argue that longer lists of N-grams may yield more results and, perhaps, show more equivalences. However, in several instances the repetitiveness of certain Italian patterns and their high frequencies speak by themselves, as in the case of the phrase *il cliente prende*

atto ed accetta, whose preponderance does not leave any room for misinterpretation.

In light of these findings, it can be stated that the language of the Italian corpus is highly repetitive and features lexical phrases and multi-words mainly revolving around a reduced number of patterns. The language of the English corpus, by contrast, uses words and terms in a more varied way and, hence, it is lexically richer.

6. Conclusion

The linguistic analysis carried out in this chapter was aimed at exploring and shedding light on the N-grams hallmarking English and Italian ToS corpora of general terms and conditions of web hosting services. To do so, 5-6-grams were focused on; more precisely, the N-grams containing the most frequent words (nouns, verbs, adjectives and adverbs) in both corpora.

The analysis of the N-grams showed the highly repetitive character of the Italian language in the ToS corpus. The lexical phrase *il cliente prende atto ed accetta che* (translation: "the customer acknowledges and accepts that") came to the fore in the majority of the cases. Conversely, the language of the English corpus was more varied. It showed equally distributed frequencies on a larger number of terms. Consequently, the English ToS corpus produced more varied lexical patterns.

The research questions that this paper aimed to answer were the following ones: 1) "Is it possible to find English-Italian equivalents in the analysis of 5-6-word N-grams in comparable corpora of online terms of service?", and 2) "Which language patterns mostly emerge from the analysis?". The answer to the first question is partly affirmative. The fully equivalent N-grams noticed during the analysis were the following ones: "the terms of this agreement" and *presenti condizioni generali di contratto* ("these general contract conditions"); "termination of the contract" and *recedere dal presente contratto* ("to withdraw from this contract" or "to terminate this contract"); "shall/will not be liable for" and *non potere essere ritenuto responsabile per* ("cannot / may not be held liable for"). A partial equivalent N-gram was as follows: "agree to be bound by" and *prendere atto ed accettare che* ("acknowledge and agree that").

The analysis of terms in contexts helped uncovering false equivalences, such as the ones regarding "such" and *tale*, or "expressly" and *espressamente*. Unfortunately, the presence of highly repetitive patterns in the Italian language did not allow other lexically relevant equivalences to emerge. At the same time, given the highly recurrence of the same Italian patterning, generating longer N-gram lists may have not yielded more valuable results. The same patterns would have probably surfaced.

The answer to the second question ("Which language patterns mostly emerge from the analysis?") revolves around the fact that the English language showed varied patterns, where N-grams were lexically and syntactically rich constructions. Examples in this regard were the phrases "the provision of the services" and "the supply of the service"; "will be liable for" and "shall be liable for", or "in respect of the services" and "in relation to the services". In Italian, the expression *il cliente prende atto ed accetta che* (translation: "the customer acknowledges and accepts that") prevailed in many different N-grams.

The lexical and syntactical richness of the English corpus can be exploited in legal language training in order to help students familiarize with legal collocations and colligations. As for the language of the Italian corpus, conversely, the repetitiveness of many N-grams could prevent users from grasping the peculiarities of legal language and/or of the language of terms of service. Nonetheless, it may teach them some of the other characteristics of legal discourse, such as its redundancies, verbosity, and tautological nature.

As mentioned, the limits of this paper lie in the reduced number of N-grams generated per word. However, it could be posited that longer lists may not have produced more insightful results.

Further research could focus on the first 20 or 30 most frequent N-grams in each corpus and verify whether the Italian language features more varied patterns. Alternatively, scholars may focus on N-grams composed of other nouns, verbs, adjectives or adverbs in the same or in other legal documents or domains.

Works Cited:

Bernardini, Silvia, Adriano Ferraresi. Corpus linguistics. *The Routledge Handbook of Translation and Methodology*. Ed. Federico Zanettin, Christopher Rundle. New York: Routledge, 2022. 207-222.

Biber, Douglas, Susan Conrad. Lexical bundles in conversation and academic prose. In *Out of Corpora: Studies in Honour of Stig Johansso*. Ed. Hilde Hasselgård, Signe Oksefjell. Amsterdam: Rodopi, 1999. 181-190.

Daiza, Heather. "Wrap contracts: How they can work better for businesses and consumers." *California Western Law Review* 54 (1) (2018): 201-239.

Garner, Bryan A. *Black's Law Dictionary. Eighth edition*. Eagan: Thomson West, 2004.

Giampieri, Patrizia. "Key N-Grams in EU Directives and in the UK National Legislation on Consumer Contracts." *International Journal for the Semiotics of Law* 37 (2023): 59–75.

Giampieri, Patrizia. *Corpus-based translation of private legal documents*. Amsterdam: John Benjamins, 2024.

Goźdź-Roszkowski, Stanisław. "Discovering patterns and meanings: Corpus perspectives on phraseology in legal discourse." *Roczniki Humanistyczne* 8 (2012): 47–68.

Goźdź-Roszkowski, Stanisław. "Corpus linguistics in legal discourse." *International Journal for the Semiotics of Law* 34 (2021): 1515–1540.

Katz, Daniel Martin, Michael J. Bommarito, Julie, Seaman, Eugene Agichtein. "Legal n-grams? A simple approach to track the evolution of legal language." *Proceedings of JURIX 2011: The 24th international conference on legal knowledge and information systems*, Vienna, 2011.

Martínez, Eric, Francis Mollica, Yufei Liu, Anita Podrug, Edward Gibson. "What did I sign? A study of the impenetrability of legalese in contracts." *Proceedings of the Annual Meeting of the Cognitive Science Society* 43 (2021): 140-146. https://escholarship.org/uc/item/5k09w2td.

Mikhailov, Mikhail. "Text corpora, professional translators and translator training." *The Interpreter and Translator Trainer* 16:2 (2022): 224-246.

Moresco, Matteo G. M. "Condizioni generali di contratto e tutela della concorrenza." *PhD Thesis*. Università degli Studi di Milano, 2019.

Torikai, Shinichiro. "Multi-Word Sequences in Legal Discourse." *Language, Culture, and Communication* 9 (2017): 113-147.

Williams, Christopher. *The impact of plain language on legal English in the United Kingdom*. London / New York: Routledge, 2023.

Other resources

Codice Civile (Italian Civil Code): https://www.gazzettaufficiale.it/sommario/codici/codiceCivile.