future internet

MDPI

Article

Exploring the Dominance of the English Language on the Websites of EU Countries

Andreas Giannakoulopoulos [1,*], Minas Pergantis [1,*], Nikos Konstantinou [1], Aristeidis Lamprogeorgos [1], Laida Limniati [2] and Iraklis Varlamis [3]

1 Department of Audio and Visual Arts, Ionian University, 7 Tsirigoti Square, 49100 Corfu, Greece; nikoskon@ionio.gr (N.K.); a18labr@ionio.gr (A.L.)
2 Laboratory of Interactive Arts, Ionian University, 7 Tsirigoti Square, 49100 Corfu, Greece; laida.limniati@inarts.eu
3 Department of Informatics and Telematics, Harokopio University of Athens, 70 Eleftheriou Venizelou Str., 17676 Athens, Greece; varlamis@hua.gr
* Correspondence: agiannak@ionio.gr (A.G.); a19perg6@ionio.gr (M.P.)

Received: 5 March 2020; Accepted: 18 April 2020; Published: 22 April 2020

Abstract: The English language is the most dominant language in the Western world and its influence can be noticed in every aspect of human communication. It's increasing diffusion, especially since the turn of the century, is hard to measure with conventional means. The present research studies the use of language in websites of European Union (EU) member states, in order to collect data about the prevalence of the English language in the different countries and regions of the European Union. To achieve a realistic representation of today's landscape of the European Web, this study uses a vast population of websites and a representative sampling size and methodology. By analyzing and processing the findings from over 100,000 websites from every country in the EU, a solid foundation is set that is used to explore the dominance of the English language in the European World Wide Web in general. This is the first study that examines the presence of English content in the websites of all EU member countries and provides statistical evidence regarding the ratio of English content availability for each country. Conclusively, the results of the research demonstrate that the English language is available on more than one quarter of all websites of non-English speaking EU member states. Moreover, it is available in the vast majority of multilingual and bilingual websites, while at the same time being the only language that is available in a number of monolingual websites. In addition, it is shown preference over the national language in a significant number of cases. A moderate negative correlation is found between a member state's population and the availability of English in these countries' websites and the same holds true for a member state's Gross Domestic Product (GDP). Both these correlations indicate that smaller countries tend to provide more content in English in order to establish a stronger presence in the international environment. Taking into account the role of language in the expression of national identity, this study provides data and insights which may contribute to the discussion about the changes underway in the national identity of EU member states.

Keywords: web presence; geographical domains; internet statistics; European union; English language; national identity in EU; multilingual websites; globalization; world wide web

1. Introduction

In an effort to increase international communication, people and organizations within the EU area and Europe in general have been adopting the use of English as a common language for a long time. Although the use of the English language is considered an important factor for the internationalization of institutional and organizational websites [1,2], there is still, according to our knowledge, no study

that focuses on the EU countries and spans multiple disciplines from education to government and commerce. The facts that the EU member states are archetypal nation-states and language is the common denominator of national identity strengthen the need for a study such as the one at hand, which will try to provide data regarding the influence of a foreign language in member states of a multinational formation such as the EU. The purpose of this research is to study the diffusion of the English language, and measure its spread in countries of the EU, and more specifically in the ones that don't have English as one of their official languages.

In summary, the research goals of this work are: (i) to statistically examine and measure the existence of English content availability in a wide range of websites from each EU member state and (ii) to examine the proportion of websites in which English appears to have greater prevalence than each EU member state's official language or languages. In order to meet these goals, the study provides accurate and fully quantifiable data extracted by a very large and representative sample of EU websites. This information can become a valuable instrument in understanding not only the use of language to achieve better international reach, but also the changes underway in the national identity of EU member states.

This research makes use of the Internet and specifically the World Wide Web to develop a method to analyze the diffusion of English. The World Wide Web is a major driving force in creating a global community but is also a technological entity that can to some extent be measured. By studying what languages are available or hold prominence in a large number of websites from every country of the EU we can get an accurate metric of the diffusion of English.

For the purposes of this study we consider the use of a National Top-Level Domain (NTLD) as an intentional action from the website's owner to associate their website with a specific country. There are no technical requirements in procuring a Global Top-Level Domain (GTLD). The website hosting server does not need to be in a specific location and the language of the content presented in the website does not need to meet specific criteria. Hence, it is safe to assume that anyone who desired not to associate their website with a specific country or region would opt for the selection of a GTLD. On the other hand, purposefully selecting a NTLD can be seen as a clear indication of a relation between the website and the equivalent country. This notion is further reinforced as in the past use of the country's NTLD has been linked to cultural characteristics and national pride for example in Sweden [3] or India [4].

Following that reasoning, from a grand total of more than 5.9 million recorded websites that belong to European national top-level domains, a sampling pool of over 100 thousand websites was created. These websites were then automatically traversed, and their content analyzed in order to infer the language they use, as well as any additional languages they might be offering, through a Language Inference Algorithm (LIA) developed exclusively for the purposes of the research at hand. Further below the results of this process are presented and relevant conclusions are drawn.

2. Language on the Web

Half a century ago, Marshall McLuhan introduced and popularized the term "global village" as the effect of an interconnected world due to the massive consumption of media [5]. Internet as a medium of communication fulfilled this forecast by connecting people in distant locations and by intensifying relations. The Internet and the Web can be recognized as basic carriers or means of globalization when it comes to communication. Of course, we should take into account the digital divide [6], but in general, Internet usage is steadily growing with Internet penetration being at 58.8% of the world population at the end of 2019 [7].

Globalization does not have a specific definition [8] and it can take different meanings depending on the context someone is referring too. For instance, globalization can have financial aspects, cultural aspects, social aspects and more. Three decades ago, Anthony Giddens described globalization as the increase in worldwide social relations that connect remote localities in such a way that local events are formed by events taking place far away and vice versa [9]. We will use this definition because it is

relevant to the topic under discussion. Communication technologies are transforming localities not only at a national level but at a more intimate and personal one [9].

Although the aspects and consequences of globalization are disputed, what remains undisputable is the increase in interaction that the Internet (as part of globalization) has spurred [10]. One of the aspects of globalization feared by scholars was homogenization, in the sense that there will be a standardization of lifestyles at a worldwide level and that the oriental and traditional cultures will be "westernized" [11]. On the other hand, other scholars believe that homogenization will not happen, instead what will happen is hybridization [12] and of course, someone should not forget the very well-known by now term of "glocalization" which is a mix of global and local, meaning that globalization will adapt to the local context [13].

It is commonly accepted that language plays a crucial role in a culture, as it is an integral part of it, and in a way, it is a "symbolic representation of people" [14]. Gazzola [15] also points out the symbolic function of language mentioning that it is linked to people's sense of national identity. Crystal [16] characterizes language as one of the most immediate and universal symbols of people's identity. He argues that the need for identity preservation and intelligibility often lead people in different directions. The need for identity promotes the use of ethnic culture and language and the need for intelligibility pulls people towards learning an international language, with English being the first choice in most of the cases. Taking language as a part of culture in the context of globalization, it is worth investigating whether there is a homogenization, a hybridization, or glocalization of it in the Web context. For instance, we know by experience that there is a level of glocalization by accepting the use of terms such as "Greeklish", that means writing Greek words using Latin alphabet or the analogous case of "Franglais" which refers to the extended use of English words and expressions in modern French. There is also a level of hybridization when we adopt a foreign word and use it without translating it. On the other hand, homogenization is more difficult to observe. However, if cases where a language will be used as the main language of communication in the Web, then at this level we can talk about homogenization. Kim [17], in an article about the globalization of the English language, characterizes it as undeniably global and international. In fact, Kim compares English to previous world languages like Arabic, Latin, and Turkish, mentioning that in fact they were not global, but rather geographically and socially limited, while the English language has managed to transcend such boundaries and spread globally. The factors that have led to its dispersion are related not only to political and diplomatic factors, but also to the language structure itself. For instance, the English language has comparably simple grammar, it is not hesitant in adopting words from other languages, it has a great amount of literature, and it is written in an easy to learn way. The fact that English language is characterized as a global language is not a contemporary phenomenon. David Graddol [18] and David Crystal [16] referred to English language as global during the 1990s. When it comes to transferring the language expansion to the Internet, David Block [10] mentions that the English language was more prevalent than other languages even in the very beginning of the Web. According to Dor [19], this was attributed to the fact that, during the early days of the Internet, the majority of users were English speaking. Fishman [20] mentions that more than 80% of the content posted on the Internet is in the English language. Nonetheless, David Block [10] in his study found that, despite the initial estimations that the Internet will promote the English language above all as an international language, this is not exactly the case, since the web is a common space for other languages as well. Dor [19] argues that the spread of English language as the main language of communication between Internet users is regarded as the linguistic equivalent to financial globalization, but he believes that it is the consequence of other aspects of globalization, such as the financial and political. In addition, he also believes that the Internet is turning multilingual, but for reasons related to the economic aspects of globalization.

Although Edwards [21] describes language as a part of our identity and consequently a part of our national identity, he sees a link between nationalism and language communities in Europe, but in other continents such as Africa and Asia, the link seems to be replaced by religion. Regarding language prevalence in more official settings, it seems that every country can keep the official language

of the country as its working language. The same applies in cases where more countries that act as nation-states are included. For instance, Gazzola [15] mentions that each nation-state has its official and working language and multilingualism has been confirmed. Despite the fact that more than a decade passed since Gazzola's evaluation, still, each nation-state has its own official and working language, and by now there are 24 official languages in the European Union [22]. The European Union's website mentions that, "even after the withdrawal of the United Kingdom from the EU, English remains one of the official languages of Ireland and Malta". Furthermore, the EU tries to establish multilingualism by having as a goal that EU citizens will be able to communicate in two or more languages other than their mother tongue [22]. On the other hand, Wodak and Boukala mention that multilingualism is favored towards the languages of EU states and that proficiency in one of the national languages of its members is required in order to enforce this collective or (supra)national identity and exclude outsiders [23]. At the same time, Kuhn [24] mentions that there is an increase in the identity politics of the European Union and the citizens' collective identities are significant for European integration. This is attributed to various factors, most prominent among them being the increasing participation of the citizens in transnational transactions. Although it is not mentioned in her article, we can safely assume that the Web made those transactions easier while we can also argue that a common language among the EU citizens would help these international transactions.

Nowadays, English remains the most used language in the web, used by 25.2 % of Internet users worldwide, followed by Chinese that counts for the 19.3% of the Internet users [8]. In addition, in a study by Mongeon and Paul-Hus [25] using Web of Science and Elsevier's Scopus, it is suggested that journals written in the English language are overrepresented to the disadvantage of other languages. This is indicative of the fact that the authors are writing in the English language, even when it is not their mother tongue. A similar research regarding citations in Web of Science, Scopus, and Google Scholar by Martín-Martín [26] suggested that the majority of citations were published in English, with the percentage for unique citations ranging from 62% to 80%, depending on the field. These two studies imply the extended use of the English language, at least in the academic community, which in turn is one of the many communities that form the users of Web. The Mongeon and Paul-Hus [25] study also mentions that there was a considerable number of journals that had abstracts in more than one language, with the one of them being English. Although they did not use the data for the second language, the probabilities are that the second language was either their native tongue, or the language of the publisher or conference host. Except from the academic field, studies focused on specific parts of the world also indicate the use of English language when using the Internet, to the detriment of their native tongue or other languages. For example, Wei and Kolko [27] in their study regarding language and Internet diffusion patterns in Uzbekistan argue that users on average agreed that they have to use English too often while they are using the Internet.

Phillipson [28] notes that the majority of countries in continental Europe are promoting learning English as a foreign language while other languages fall short since English is the language used in a lot of conferences and publications of the EU, as well as one of the main working languages in the Union's institutions [29]. Noteworthy however is the fact that English was not included in the first official languages of the EU in 1958. These were Dutch, French, German, and Italian. English was added along with Danish in 1973, and Greek was added a few years later. The fact that the European Union wants to promote multilingualism and have its citizen be able to communicate fluently in at least three languages makes it more possible for website owners to include more than one language in their websites. Hillier [30] highlighted the importance of considering the cultural context when developing a multilingual website. He also argues that when someone chooses to create a multilingual website, each language version will have its own domain name, either at the country level domain or as a subdomain of the main domain, e.g., xxxx.com.gr, xxxxxx.com.fi etc. This implies that users/citizens correlate the domain with their language and consequently with their culture and national identity. Top-level domains (TLDs) are the letters (characters) forming the last part of a fully qualified domain name. There are two naming structures for TLDs. One is referred to as global (GTLDs), with the most

common endings being .com, .net, .org. The second naming structure is the national (NTLDs) and it is based on geographical criteria. It is usually composed endings with two letters such as .gr for Greece, .at for Austria, .fi for Finland etc. [31]. During the early days of the Web, the widespread use of the .com TLD over the rest led to the gradual adoption of GTLDs. By now, it is uncertain if in the vast cyberspace a user will choose a NTLD over GTLD due to a choice based on culture and national identity or due to a habit that has its roots in the early days of the Web. In general, although most works mentioned above are not strongly related to the present study, they make two very important points: i) the use of the English language is a means to achieve a broader international reach and ii) the choice of a national TLD over a global one indicates an intention to associate with the equivalent nationality. These two points of approach converge to create the necessity for research with the characteristics of this study, which in turn will provide data and quantified information to enrich the relevant theoretical discussions.

3. Methodology

In order to achieve the goals of this research, it was required that multiple websites from every member state of the EU were analyzed. This analysis provided us with information about the languages used in each website. The tools used in this process are described in detail below. All tools were developed using PHP and recorded their data in a MariaDB Server database. "PHP is a widely-used open source general-purpose scripting language" [32]. "MariaDB Server is one of the most popular database servers in the world [...] made by the original developers of MySQL" [33]. MariaDB was selected for its performance. PHP was selected because its popularity ensures there are plenty of tools for each task required by the study and because of the researchers' familiarity with the language.

The complete analysis process can be divided into several steps:

- Step 1: Collect a large number of websites to analyze.
- Step 2: Determine the sampling method and size.
- Step 3: Crawl websites and record relevant information.
- Step 4: Use that information to extract answers.

The process took place over the months of January 2020 and February 2020. The process of collecting websites provided more than 5.9 million websites and the crawling process included more than 100,000 websites. The large amount of data gathered and the representativeness of the sample which was automatically and without bias selected from a vast total population helped create a quite realistic estimate of how websites across the EU treat languages.

3.1. Collecting a Large Number of Websites to Analyze

The nature of the present research required as many websites as possible, so that both our total population and our sampling pool were as close a representation of reality as possible. For this purpose, we used information obtained from Common Crawl, a "repository of web crawl data that is universally accessible and analyzable" [34]. Among the data Common Crawl offers is an index of every available webpage for all member states of the EU amongst other countries. A process was developed in PHP: Hypertext Preprocessor (PHP) that used the CompounD indeX (CDX) server Application Program Interface (API) [35] to access Common Crawl's Uniform Resource Locator (URL) index [36] and created a MariaDB database with information about websites from every member state of the EU.

Although Common Crawl's index provides all available crawled pages, our process of data collecting only focused on recording the landing page of one website per domain. This way, we made sure that websites of different sizes got the same representation in our data. Whether a domain contained thousands of subpages or just a few, it was given a single record in our database. Later in the process, when the domain is crawled in order to determine the languages used, multiple subpages are processed. This helps us shift the focus from how many individual pages use a language to how many websites as a whole use a language, thus making the website the main entity of our research as opposed to treating every page as its own separate entity. This helps focus results around individual

real-life entities (people, businesses, organizations, groups, cities etc.) that are represented online instead of having such entities with a very large online presence in page count dominate over smaller but equally important ones.

On many occasions, a website is available both in normal HTTP and in the more secure HTTPS version. In addition to that, websites often make use of the www subdomain but are also available without it. In order to avoid domain duplicates our information gathering process made sure each domain was only accepted into the database once, while at the same time keeping track of what versions of the website were recorded in the Common Crawl Index (http, https, with www or without www).

For the purposes of this research, a decision was made to exclude subdomains from our website database. This was decided not only because subdomains are often subsections of one unified website, but also because subdomains are often used to provide different language versions of the same website (en.example.com vs de.example.com). Later in the process, when we crawl the websites ourselves to help detect their language, we extend our crawling to subdomains that are linked in a website's front page and consider them part of a single website entity.

In order to successfully distinguish websites that use their SLD (second level domain) as part of a second-level hierarchy for NTLDs as opposed to indicate the registrar we compiled all second level hierarchies used by the various member-states of the EU. Mozilla Foundation's Public Suffix List [37] along with a National Top-Level Domain for Europe provided by Global WHOIS Search [38] were used for the compilation of our list. The list is available on Table 1. All countries that only use NTLDs were omitted.

Table 1. List of TLDs and SLDs per country.

Country	TLD	SLDs
Austria	at	ac, co, gv, or, priv
Belgium	be	ac
Croatia	hr	iz, from, name, com
Cyprus	cy	ac, biz, com, ekloges, gov, ltd, name, net, org, parliament, press, pro, tm
Estonia	ee	edu, gov, riik, lib, med, com, pri, aip, org, fie
France	fr	asso, com, gouv, nom, prd, tm
Greece	gr	com, edu, net, org, gov
Hungary	hu	co, info, org, priv, sport, tm, 2000, agrar, bolt, casino, city, erotica, erotika, film, forum, games, hotel, ingatlan, jogasz, konyvelo, lakas, media, news, reklam, sex, shop, suli, szex, tozsde, utazas, video
Ireland	ie	gov
Italy	it	gov, edu
Latvia	lv	com, edu, gov, org, mil, id, net, asn, conf
Lithuania	lt	gov
Malta	mt	com, edu, net, org
Poland	pl	biz, com, info, net, org, waw
Portugal	pt	net, gov, org, edu, int, publ, com, nome
Romania	ro	arts, com, firm, info, nom, nt, org, rec, store, tm, www
Spain	es	com, nom, org, gob, edu
Sweden	se	a, ac, b, bd, brand, c, d, e, f, fh, fhsk, fhv, g, h, i, k, komforb, kommunalforbund, komvux, l, lanbib, m, n, naturbruksgymn, o, org, p, parti, pp, press, r, s, t, tm, u, w, x, y, z
United Kingdom	uk	ac, co, gov, ltd, me, net, nhs, org, plc, police, sch

Furthermore, the Common Crawl index provides a language annotation for every page, based on the detection result of Compact Language Detector 2, a library for the probabilistic detection of a written language [39]. Since our study mainly focuses on the linguistic aspect, we also kept a record of the language as detected by CLD2 and provided by the Common Crawl index. This is used in tandem

with our own crawling results to help determine the languages used by the different websites that were used in this research. A flowchart demonstrating the website collection process can be seen in Figure 1.

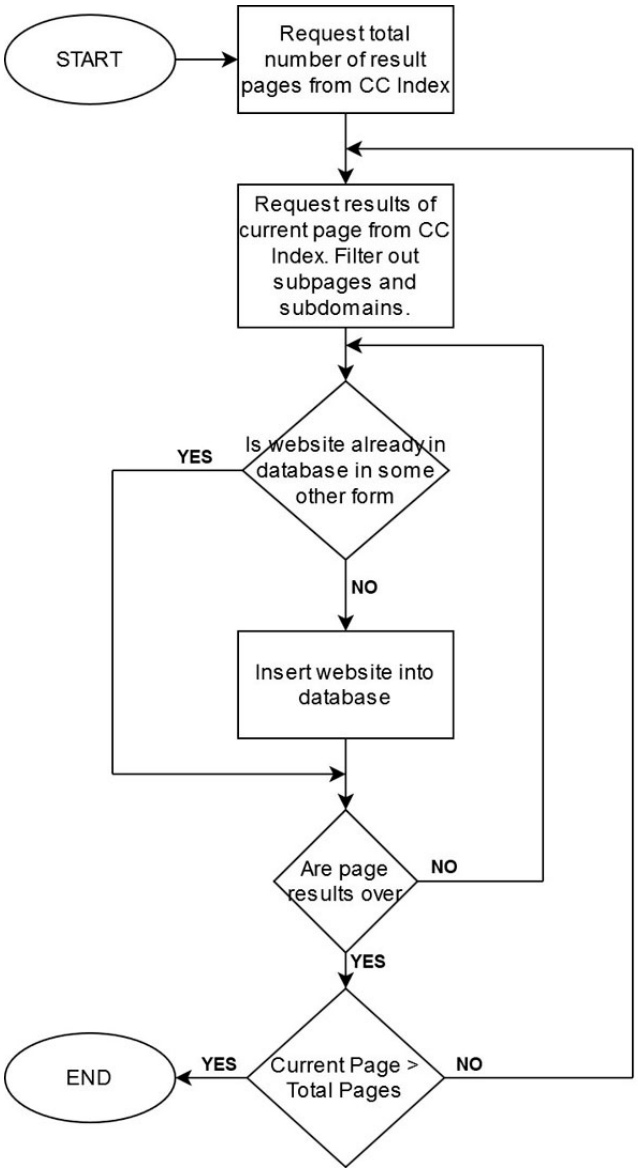


Figure 1. Flowchart of the website collection process.

3.2. Determining the Sampling Size

Time constrictions did not allow us to crawl every single website that was added in our website index which was more than 5.9 m websites. In order to make sure that our data were representative of reality we had to define an appropriate sampling size. Considering the number of websites available in our index, we set the confidence level high at 99%. In order to achieve a reasonable process timeframe and due to the rather large population size, we opted for an error margin of 2%. This led us to the sampling sizes that are available in Table 2.

Table 2. Number of available websites and appropriate sampling size per country.

Country	Websites	Sampling Size
Austria	172,815	4063
Belgium	173,471	4063
Bulgaria	13,445	3178
Croatia	21,603	3489
Cyprus	1126	887
Czechia	221,473	4084
Denmark	180,513	4067
Estonia	24,452	3556
Finland	98,232	3992
France	475,661	4125
Germany	1,627,339	4150
Greece	79,116	3953
Hungary	99,081	3993
Ireland	47,360	3825
Italy	458,727	4123
Latvia	18,779	3406
Lithuania	40,485	3773
Luxemburg	9304	2876
Malta	605	529
Netherlands	628,465	4133
Poland	334,824	4110
Portugal	45,026	3809
Romania	88,885	3975
Slovakia	75,469	3943
Slovenia	27,787	3619
Spain	234,791	4088
Sweden	221,795	4084
United Kingdom	480,802	4125
Total	5,901,431	102,018

The Cochran's standard sample size formula [40] that appears in Equation (1) was used in order to calculate the required sample for each country. Where N = population size, e = margin of error (percentage in decimal form), z = z-score which is the number of standard deviations away from the mean and can be found in specific tables, and p is sample proportion which is basically 0.5. The calculations were made using the help of SurveyMonkey's Sampling Size Calculator [41]. SurveyMonkey is a widely popular online survey platform [42].

$$SampleSize = \frac{\frac{z^2 \times p(1-p)}{e^2}}{1 + \frac{z^2 \times p(1-p)}{e^2 N}} \quad (1)$$

3.3. Crawling Websites and Recording Relevant Information

In the next step of our process, a number of websites equal to the sampling size were analyzed using a proprietary crawler which was developed using PHP and manipulated data in the MariaDB database developed during Step 1. The crawler first checked every available version of each website (http/https, with/without www). Priority was given to the https versions and the versions with www. The process followed the appropriate instructions available in the robots.txt file in order to determine whether crawling for that specific website was allowed or not. If it was, then the frontpage was scraped and analyzed. If not, the website was removed from the sampling pool and replaced. Additionally, if the page allowed crawling but returned an error page or a redirect it was also removed from the sampling pool and replaced.

In addition to that, the number of internal or subdomain links in each frontpage was counted and if that number was below 2 or above 250 the webpage was considered an extreme case and was removed from the sampling pool and replaced. The reasoning behind this is that, more often than not, low link pages are just placeholders and extremely high link pages are suspicious as they are often used to confuse search engines or are part of a back-link generation scheme. Additionally, they would increase the time required for the process to run without any clear benefit to the results.

First the lang attribute of the html tag was recorded if available. Then the crawler attempted to detect the actual written language of the page using the PHP language detection [43] library. The language detection function processed the HTML DOM using PHP's DOMDocument parsing library. With its help, it retained the text from each element of the frontpage while removing elements that don't traditionally contain written content such as <head>, <script>, <style>, <svg>, <img>, and <code>. The full text of the page was then run through the language detection algorithm with a setting of 9000 max Ngrams in order to achieve more confidence in the result. If the full text of a page did not exceed 150 characters, any language detection was considered unreliable due to the short string and as such the language detection function returned an empty result.

Afterwards, in order to detect whether another language is available on the website, the crawler scraped every internal or subdomain link present in the frontpage and used the same language detection function to determine the language of each website. The HTML tag's lang attribute of each subpage was also recorded. If a subpage's detected language or lang attribute was different to the frontpage's the crawler inserted this page into the database. These records would subsequently be used during the next step of the process to investigate the number of languages that a website supports.

A decision was made to stop the crawling of the website after one level beyond the frontpage. In the vast majority of multilingual webpages, the frontpage contains a link to the different language versions of the website. This way, we reduced the running time of the crawler but did not cause heavy traffic or other issues to the website being crawled. Towards the same goal, a delay was added between subpage scrapes. This delay was set to 1 second, and in addition, with the running time of the crawler and the language detection algorithm for each subpage, there was enough time between successive scrapes so that the website server's performance wasn't negatively impacted. A flowchart demonstrating the crawling process can be seen in Figure 2.

3.4. Inferring Main and Other Languages

With the crawling process completed, large amounts of information were recorded in the database. The next step would be to use that information to reach a final conclusion about what the primary language of any website tested is, as well as what other languages are available in the website. In order to do that, an algorithm was developed that used the lang attribute of the HTML tag, the detected language of the php language detection library and the language provided by the Common Crawl Index which was detected by compact language detector 2. This language inference algorithm (LIA) was specifically developed for the purposes of this study, and as such, emphasis was given to the prevalence of the English language not only as a secondary supported language, but also as a primary language and its comparison with each country's official language or languages.

In order to make the comparison between the detected languages and lang attribute easier, an array was created that contained the multiple different notations of a country's official language that were encountered during the crawling process. In most cases, that included just the language's ISO 639-1 two letter code, that was used by the lang attribute and the PHP language detection library, and the language's ISO 639-3, that was used by CLD2. In some cases, more equivalent notations needed to be added in order to better infer each country's official language. A list of these notations can be found in Table 3.

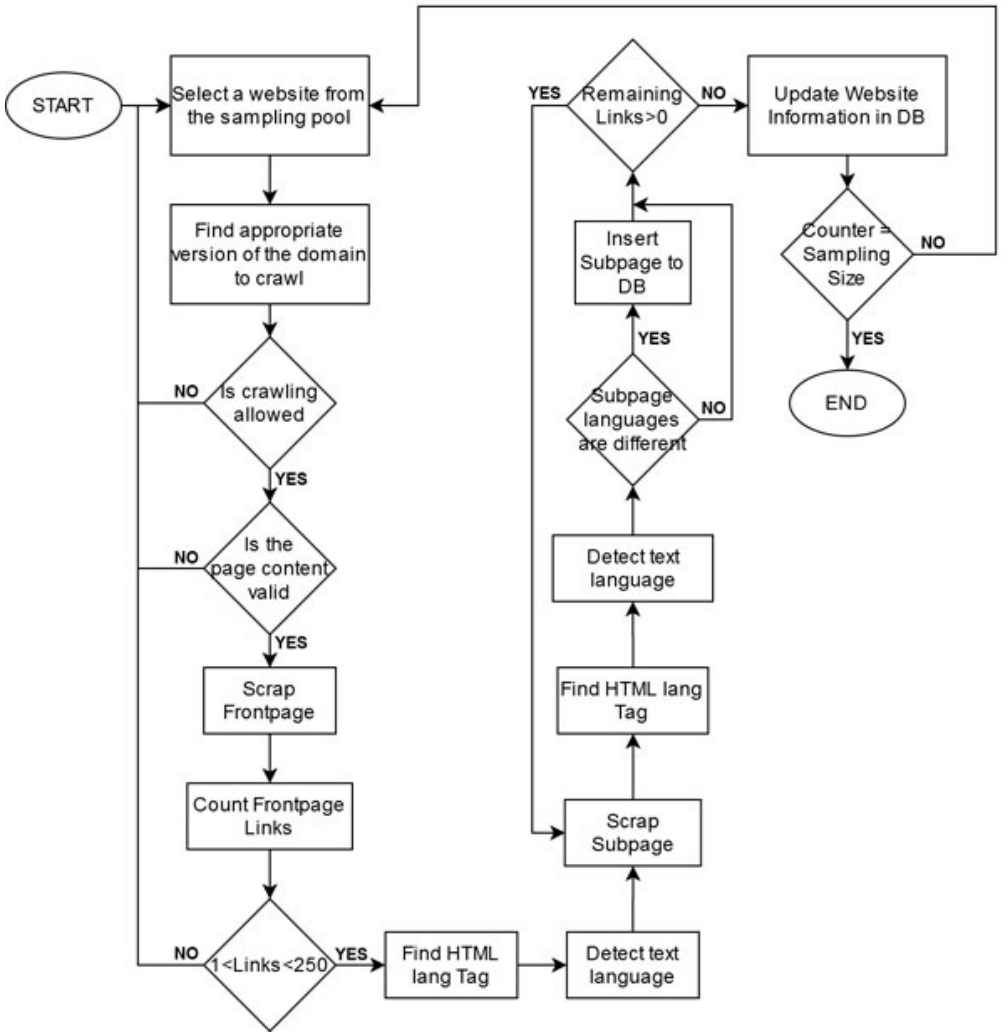


Figure 2. Flowchart of the crawling process.

In countries where English is an official language, if there is another official language, it was chosen as the primary language in order to facilitate the comparison between English and the other official language. This case includes Maltese for Malta and Gaelic for Ireland.

In countries with more than one non-English official language, the most prevalent was chosen based on how often it appeared in our already collected data. The most prevalent language was in every case identical in PHP's language detection library, CLD2, and the lang attribute of the HTML tag. This case includes Dutch for Belgium and French for Luxemburg.

In countries where the NTLD is different to the two-letter ISO 639-1 notation the NTLD was included in the equivalency table. This was decided because occasionally it is declared in the HTML tag's *lang* attribute even though it is not correct. This case included Austria, Belgium, Croatia, Cyprus, Czechia, Denmark, Estonia, Greece, and Sweden.

In the case of Croatia, the notation for both the Serbian and the Bosnian dialect of Serbo-Croatian was added.

With the notation equivalences in place, a priority rule system was implemented to infer each frontpages or subpage's language. The rules are presented in Table 4.

Table 3. Language notations used per countries.

Country	Primary Language Notation
Austria	de, deu, at
Belgium	nl, nld, be
Bulgaria	bg, bul
Croatia	hr, hrv, bs, sr
Cyprus	el, ell, gr, cy
Czechia	cs, ces, cz
Denmark	da, dan, dk
Estonia	et, est, ee
Finland	fi, fin
France	fr, fra
Germany	de, deu
Greece	el, ell, gr
Hungary	hu, hun
Ireland	ga, gle
Italy	it, ita
Latvia	lv, lav
Lithuania	lt, lit
Luxemburg	fr, fra
Malta	mt, mlt
Netherlands	nl, nld
Poland	pl, pol
Portugal	pt, por
Romania	ro, ron
Slovakia	sk, slk
Slovenia	sl, slv, si
Spain	es, spa
Sweden	sv, swe, se
United Kingdom	en, eng

Table 4. Rules used in LIA to infer the language of a webpage.

Rule Priority	Rule
Rule 1	If both the lang attribute of the HTML tag and the PHP detected language were the same, set the page's language as that value
Rule 2	If both the lang attribute of the HTML tag and the PHP detected language were equivalent according to the Notation Equivalence List, set the page's language as the ISO 639-1 value for the primary language.
Rule 3	If both the CLD2 detected language and either the lang attribute of the HTML tag or the PHP detected language were equivalent according to the Notation Equivalence List, set the page's language as the ISO 639-1 value for the primary language.
Rule 4	If both the CLD2 detected language and either the *lang* attribute of the HTML tag or the PHP detected language were English, set the page's language as the ISO 639-1 value for English.
Rule 5	If the lang attribute of the HTML tag is English but the PHP detected language is equivalent to the primary language, set the page's language as the ISO 639-1 value for the primary language.
Rule 6	If the lang attribute of the HTML tag is not empty and is a valid ISO 639-1 notation, set the page's language as that value.
Rule 7	If the PHP detected language is equivalent to the primary language, set the page's language as the ISO 639-1 value for the primary language.
Rule 8	If the PHP detected language is English, set the page's language as the ISO 639-1 value for English.
Rule 9	If the CLD2 detected language is equivalent to the primary language, set the page's language as the ISO 639-1 value for the primary language.
Rule 10	If the CLD2 detected language is equivalent to English, set the page's language as the ISO 639-1 value for English.
Rule 11	If the PHP detected language is not empty and the page is a Frontpage set the page's language as the detected value.

The rules are presented in order of priority. When a rule's conditions are met, the primary language is inferred, and no lower priority rule is checked. The priority of rules was chosen to focus on inferring primarily the primary language of each country and English. A bit of extra weight is being placed on the PHP detected language over the CLD2 because the selected string that was parsed by the algorithm was selected specifically for the purposes of this study.

The HTML tag lang attribute is highly valued when deliberately stated by the webpage's creator. An exception to this is embodied by rule 5. When the HTML tag lang attribute is set as English, but the detected language is the primary language of each country, the detected language is preferred (Rule 5). This was intentionally implemented because several popular CMS set the lang attribute to English by default without any regard to the page's actual content. In contrast, the lang attribute was used to infer the language of a page when the lang attribute was explicitly stated in the HTML tag as something other than English, as this clearly indicated the purpose of the website's developers. When the lang attribute is empty and the PHP detected language is not the country's primary language, a combination of the CLD2 detected language and PHP detected language is used to infer the page's language. In this process (rules 7–10), the PHP detected language is trusted more for the reasons mentioned above.

If the only information available to us after the crawling process is the PHP detected language, then it is only trusted in the case of a frontpage. This is meant to reduce the number of websites with no inferred language, but at the same time not to add a dubious detection as a secondary language when inferring the language of subpages.

If no information is available to us from either the lang attribute or the PHP detected language and the CLD2 detected language is neither the country's primary language nor English, the algorithm is unable to infer a website's language. This is very seldom the case and is usually true for image only websites or websites with non-html or client-side dynamically generated content. A flowchart demonstrating the language inferring process can be seen in Figure 3.

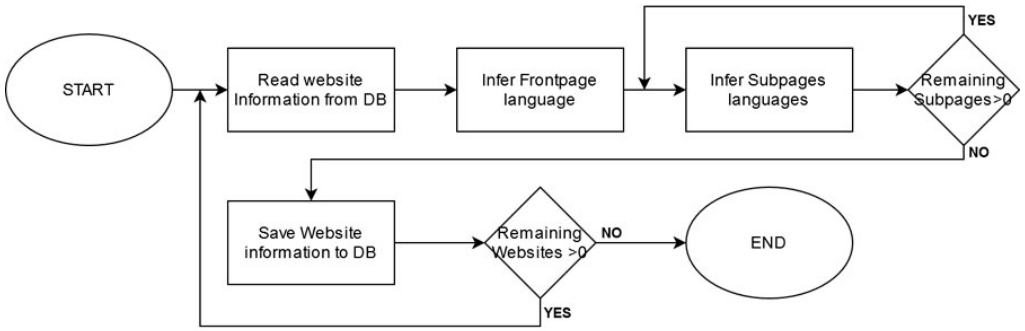


Figure 3. Flowchart of the language inferring process.

4. Results

4.1. Result Structure

After applying the LIA, a summary table was created for every EU member state. The full tables can be obtained at Appendix A. From these tables, the most important data for our research are variables inferred_lang, num_of_langs, and includes_en.

The variable inferred_lang has a value of the corresponding language code which was inferred as the primary language of the frontpage of each website and has been produced via the Language Inference Algorithm (LIA) which is described in Section 3.4. The variable num_of_langs has a numeric value which demonstrates how many languages were detected in the web site (1 for monolingual, 2 for bilingual, 3 and above for multilingual). The variable includes_en has a value of 1 if the English language was detected in the web site and 0 if the English language was not detected.

An analysis of the above collected detailed results is available for each country in Appendix B. The appearance of any language as each website's primary language is presented in a separate table for each country. The availability of the English language is demonstrated in a second separate table for each country. In each column, monolingual, bilingual, and multilingual websites are presented. The first row displays how many of them are not available in English, while the second row displays how many are available in English. The data in each table are then depicted in the form of two charts per country. A pie chart with the primary languages of each website and a bar chart with the

availability of the English language, both in total and in each different type of website (monolingual, bilingual, multilingual).

4.2. English Language Availability in Websites

By taking the average for the percentages of monolingual, bilingual, and multilingual websites for the main NTLD of each member state as detected by the LIA, we can get a unified picture of what's going on in the EU. Figure 4 displays a pie chart of that average for websites for non-English speaking member states and its equivalent for English speaking member states.

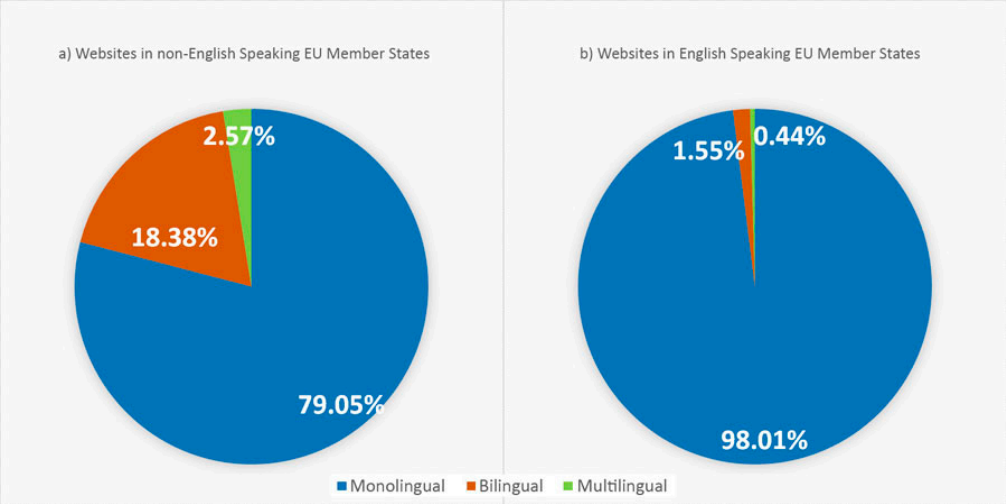


Figure 4. Percentage of Monolingual, Bilingual and Multilingual Websites (**a**) Non-English Speaking, (**b**) English Speaking.

Having established an overview of the number of languages available, the next step is focusing on which languages are preferred. Using the results presented in Appendix B, it is calculated that an estimate of 25.64% of websites in non-English speaking EU countries are available in English as seen in Figure 5. This includes monolingual websites that only offer the English language and bilingual or multilingual websites that offer English among other choices. In English speaking countries, this percentage reaches 99.1%.

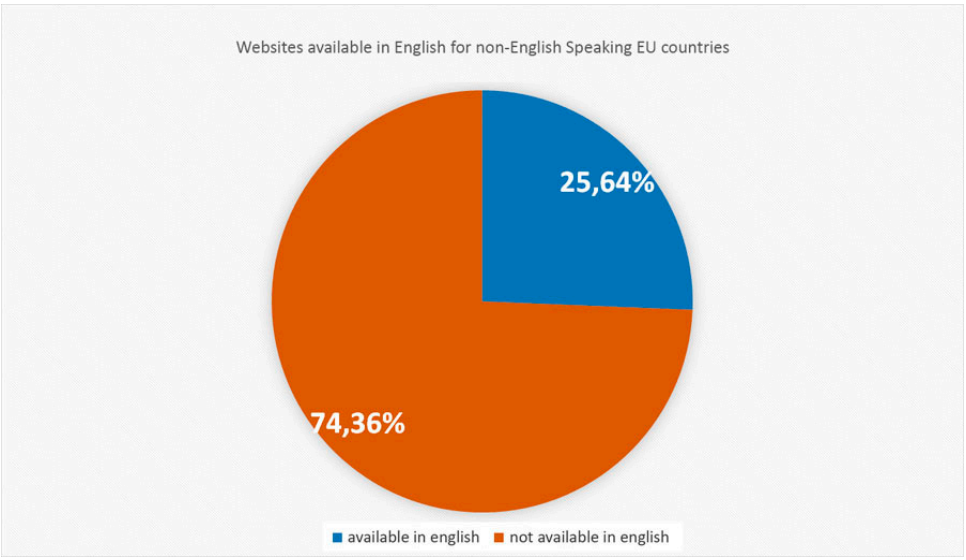


Figure 5. Percentage of websites Available in English for non-English Speaking EU countries.

Having the English language available on a website can be achieved by developing a website only in English, by adding English on top of each country's established official language as an option, or by providing it alongside a group of other languages. In Figure 6, we can compare the availability of the English Language in monolingual, bilingual, or multilingual sites.

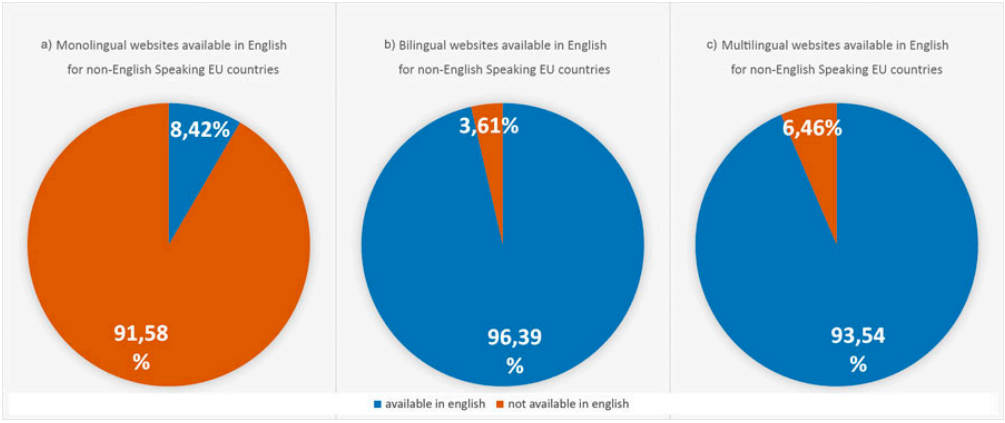


Figure 6. Percentages of Websites Available in English for non-English Speaking EU member states by number of available languages (**a**) monolingual, (**b**) bilingual, (**c**) multilingual.

While looking at averages can provide a good summary of the language landscape in the European part of the World Wide Web, it is also important to notice the trends in each individual country. In Figure 7, the availability of the English language can be seen for non-English speaking countries in a comprehensive chart.

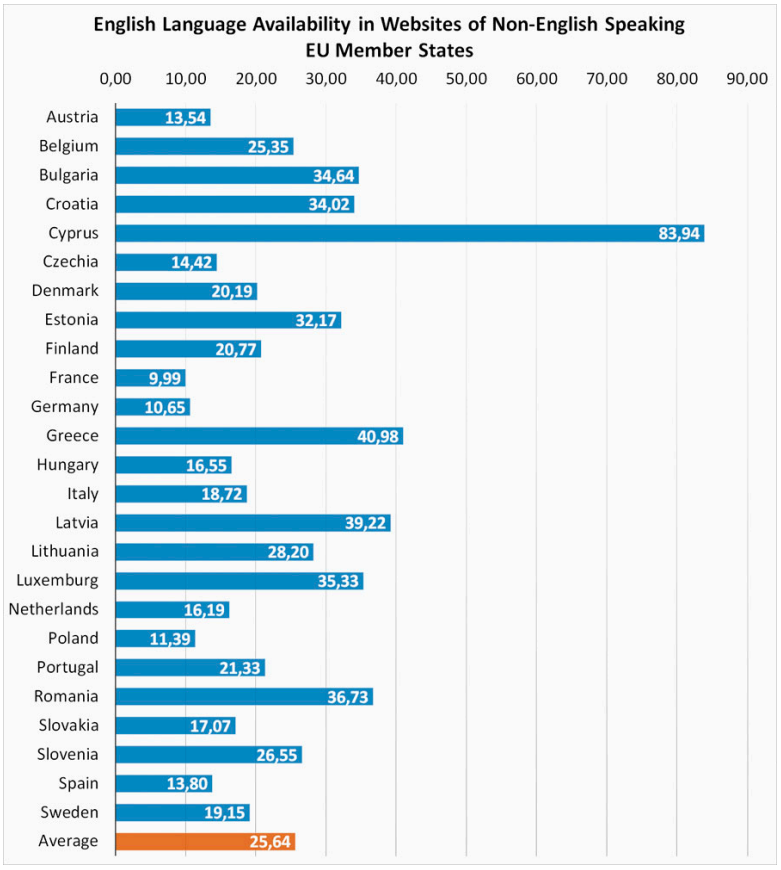


Figure 7. English Language Availability in Websites of Non-English Speaking EU Member States.

In an effort to paint a clearer picture regarding English language availability in the different member states, the results of Figure 7 have been integrated into a map of the EU area as seen in Figure 8.

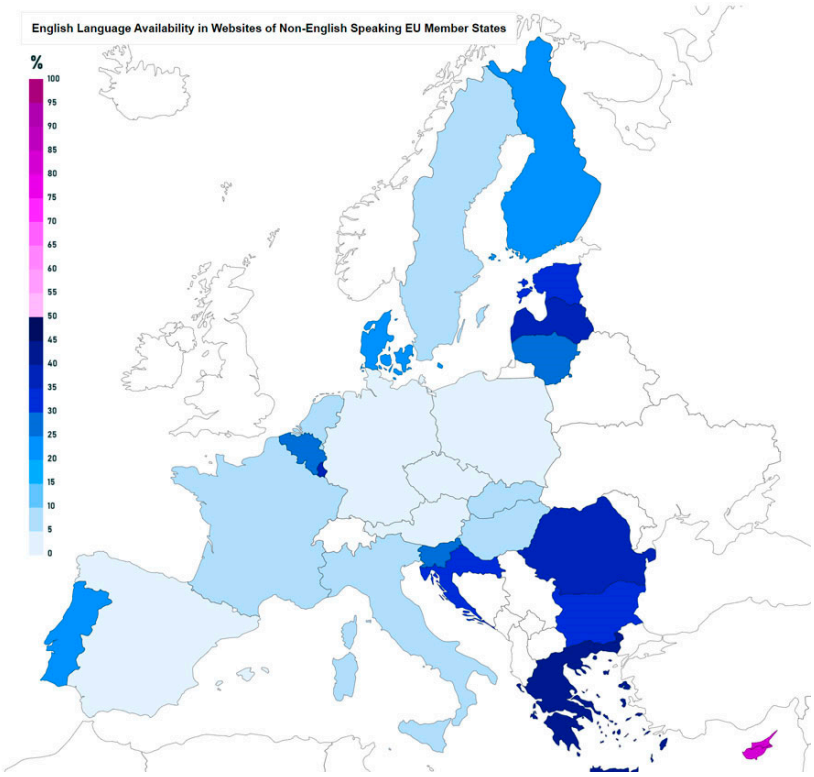


Figure 8. Map of English Language Availability in Websites of Non-English Speaking EU Member States.

4.3. English as a Primary Language in Websites

The first language a user encounters on a website's landing page can be safely considered the primary language of that website. As seen in Figure 9, the English language appears as the primary language of 9.92% of websites on average in non-English speaking EU member states.

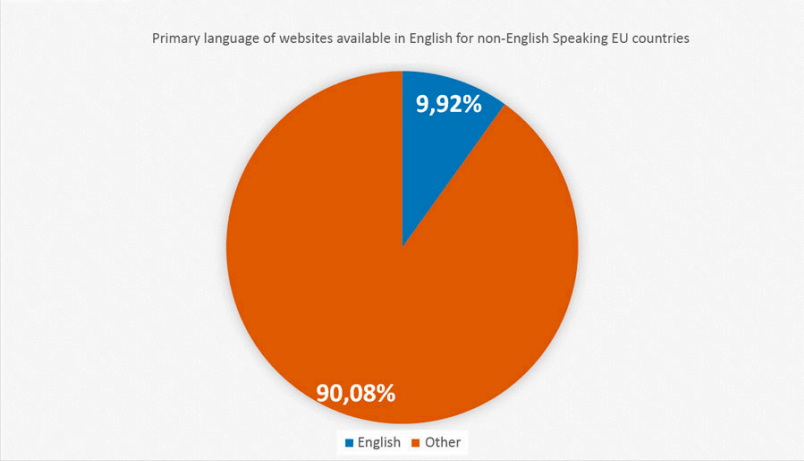


Figure 9. Percentage of Websites with English as Primary Language in non-English Speaking EU Member States.

Moving away from the average and trying to establish an overview of the situation in each individual country, Figure 10 displays the percentage of websites that have English as their primary language in the non-English speaking countries of the EU.

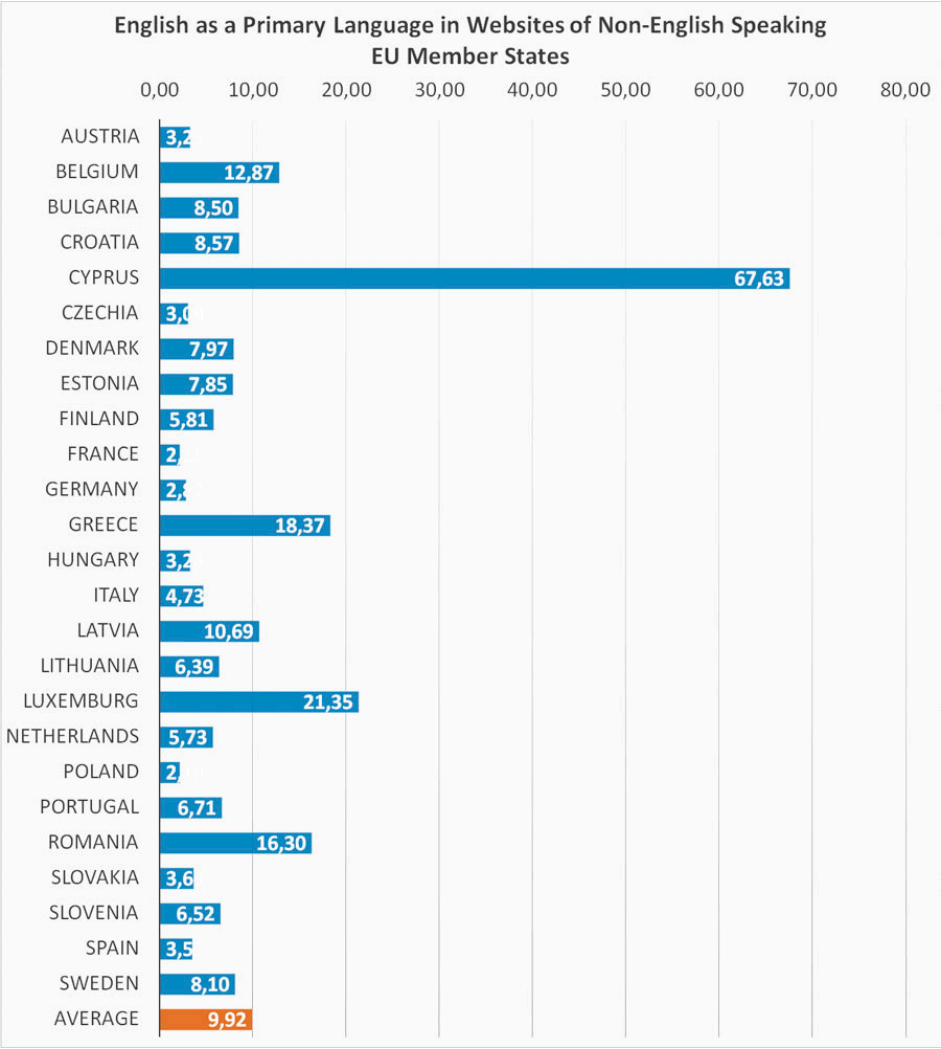


Figure 10. English as a Primary Language in Websites of Non-English Speaking EU Member States.

In an effort to paint a clearer picture regarding English as a primary language in websites in the different member states, the results of Figure 10 have been integrated into a map of the EU area as seen in Figure 11.

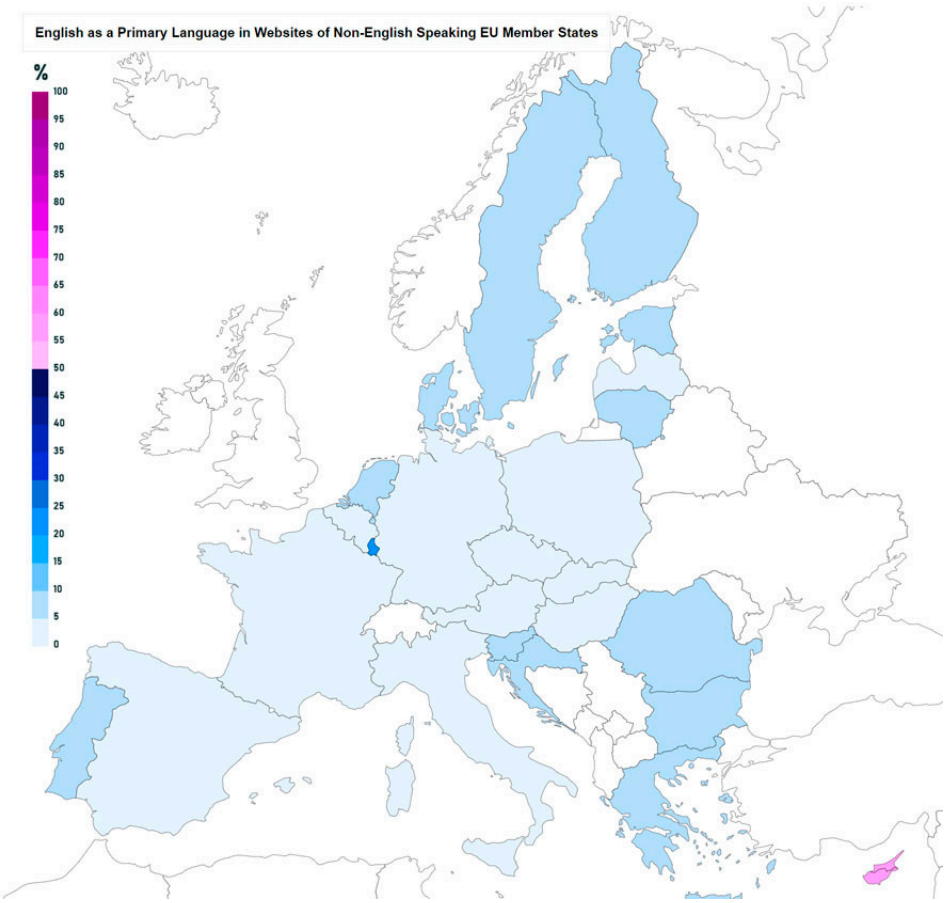


Figure 11. English as a Primary Language in Websites of Non-English Speaking EU Member States.

5. Discussion

5.1. English Language Availability in Websites

One of the major cornerstones of international communication is a common language. In an effort to increase their respective reach, websites within the European Union often make themselves available in more than one language. Browsing through the results of the previous section, it is made abundantly clear that websites of non-English speaking member states are keener to provide their users with alternative languages.

It is made clear by Figure 4 of the results section that there is a discrepancy between English and non-English speaking countries. The percentage of monolingual sites drops from 98.01% in English speaking countries to 79.05% in non-English speaking countries. The major force behind that discrepancy is the effort to make non-English websites available in English as we will see further down.

Additionally, there is a notable difference between bilingual and multilingual websites. Making a website available in more than one language requires a lot of work, so having three or more languages is a costly endeavor. If the main purpose is increasing a website's reach and that can be accomplished by using a dominant language, this makes the bilingual option much more attractive to website owners.

If we take into account the relative percentage of websites available in English in non-English speaking member states, as presented in Figure 5, coupled with the fact that English speaking countries not only have a very low number of bilingual or multilingual websites, but also have a very high percentage of websites available in English, we come to the conclusion that there is a consensus that the English language is considered enough to cover the need of EU websites for international reach.

Having the English language available on a website can be achieved by adding it on top of each country's established official language and that can increase costs. In some cases, especially when international or pan-European reach is the main objective of the website, the official language is abandoned, and the website is presented only in English. Additionally, when the investment for providing multiple languages is made, occasionally, other languages get priority over English.

As seen in Figure 6a, the choice to forgo the official language of a country in order to accommodate English is not very popular, although it is still significant. On the other hand, when multiple languages are supported, it is very rare for English to be omitted as seen in Figure 6b,c. This observation reinforces our earlier assumption that the availability of multiple languages is driven primarily by the need to include English in order to increase reach. The most popular choice that accommodates both the use of English for greater reach and the relatively low cost of adding a singular secondary language makes the model of a bilingual website that supports English the most popular model. This comes in line with the findings of Mongeon and Paul-Hus [25] which also demonstrated the popularity of bilingual content with one of the two languages being English (in their case only in relevance to abstracts of scientific publications). On average 17.08% of all websites are bilingual websites that support the English language. This percentage comes really close to the total of bilingual websites and represents the largest part of all websites that are available in English.

Studying both Figures 7 and 8, we can see that the lowest percentages of English availability come from the largest EU member states (Germany, France, and Poland). Websites of smaller countries seem much more eager to provide their content in the English language. More than 35% of websites in Latvia, Belgium, Romania, Greece, Luxemburg, and Cyprus have English as their primary language.

Reasons for the variation in percentages may include local culture, economy (for example focus on tourism or exports), and whether a country has more than one official language among others. However, a trend seems to be emerging in that smaller countries tend to put greater effort into making their websites available in English. In order to further investigate this trend, we proceeded to study the interrelation between both a country's GDP and a country's total population in relation to the availability of its websites in English.

To analyze the interrelation between the population (population) variable and availability in English, Pearson's correlation coefficient (Pearson's r) has been applied [44]. The results are shown in Table 5 where there appears to be a negative correlation between population and availability in English (−0.462). This leads us to reject the null hypothesis (there is no correlation between population and availability in English). The significance level is 0.01, confirming the statistically significant negative moderate correlation.

Table 5. Pearson correlation coefficient between (population—availability in English).

Variable	Type	Population	Availability in English
population	Pearson Correlation	1	−0.462 *
	Sig. (1-tailed)		0.010
	N	25	25
availability in English	Pearson Correlation	−0.462 *	1
	Sig. (1-tailed)	0.010	
	N	25	25

*. Correlation is significant at the 0.05 level (1-tailed).

In order to investigate if there is a correlation between the availability in English and the economic situation for each member state of the EU, we decided to use the gross domestic product (GDP) indicator. Information about each country's GDP was gathered from Eurostat [45]. Pearson's correlation coefficient (Pearson's r) has been applied in order to see if there is a correlation between GDP and availability in English. As we can see in Table 6, the Pearson correlation is −0.446 at 0.013 one-sided significance, which shows a negative correlation between the two variables.

Table 6. Pearson correlation coefficient between the Availability in English and GDP.

Variables	Type	Availability in English	GDP
availability in English	Pearson Correlation	1	−0.446 *
	Sig. (1-tailed)		0.013
	N	25	25
GDP	Pearson Correlation	−0.446 *	1
	Sig. (1-tailed)	0.013	
	N	25	25

*. Correlation is significant at the 0.01 level (2-tailed).

The moderate negative correlation observed between both population and GDP in relation to the availability of the English language in non-English speaking countries of the EU indicates that smaller and less affluent countries offer higher English language availability in their websites. From a cultural perspective, the largest countries often carry a heavier cultural impact which manifests itself in the form of national pride, while smaller countries might be keener to facilitate communication beyond their own borders. In addition to the cultural perspective, the economy is also a major driving force for internationalization. Countries with a smaller GDP offer a smaller marketplace, which would create the need for businesses to try and increase their reach beyond the country's limitations. As a result, business websites would put extra effort in providing their content in English, something that might not be necessary for a business in a larger country and hence a wider marketplace. In other words, the room to grow as an economic entity is larger in a bigger country which makes the incentive to grow internationally less impactful.

5.2. English Language as a Primary Language in Websites

As mentioned in the results section we consider the first language that a user encounters on a website's landing page to be the primary language of that website. In Figure 9, we saw that the English language appears as the primary language in 9.92% of all websites on average in non-English speaking EU member-states, making this another metric of the prevalence of the English language throughout the European part of the World Wide Web.

Despite this percentage being relatively low, it is still a significant percentage. On average, almost one out of ten websites on the TLD of non-English speaking EU member states do not use its own official language as the language of choice for its homepage, but the English language.

Furthermore, studying that percentage on individual member states through Figures 10 and 11 we can determine that the lowest percentages of English availability come from the largest EU member states (Germany, France and Poland), while more than 10% of websites in Latvia, Belgium, Romania, Greece, Luxemburg, and Cyprus have English as their primary language. Comparing these results with their equivalents regarding English language availability makes it safe to say that the same values that govern the availability of English also govern, to some extent, whether a website has English as its primary language or not.

In an effort to prove this, Pearson's correlation coefficient (Pearson's r) has been applied between English as primary language and availability in English in order to examine the null hypothesis that there is no linear relationship between these two variables. The results are shown in Table 7, where there appears to be a strong correlation [44] between them, which leads us to reject the null hypothesis. The significance level is 0.00, confirming a statistically significant correlation.

Table 7. Pearson's correlation coefficient (English as primary language—availability in English).

Variables	Type	English as Primary Language	Availability in English
English as primary language	Pearson Correlation	1	0.921 **
	Sig. (1-tailed)		0.000
	N	25	25
availability in English	Pearson Correlation	0.921 **	1
	Sig. (1-tailed)	0.000	
	N	25	25

**. Correlation is significant at the 0.01 level (2-tailed).

English being the first language a user encounters on a website's landing page (i.e., the primary language) in any website that intentionally exists in a national TLD is not something that one would expect. Despite this fact, the percentage of such websites is consistent in most EU countries and although small, it is still significant. The strong correlation between the availability of the English language and English language as a primary language in a website indicates that the variable of English as a primary language follows similar trends. This means that it can also be used as an objective metric in order to study the prevalence of the English language throughout the World Wide Web. Given all the above, it would come as no surprise if this particular metric increases in the course of time.

5.3. National and Regional Factors

As mentioned above, cultural and geographical factors, besides a country's population, may play a role in the influence of the English language in its websites. Having established a strong correlation between website availability in English and English as a primary language for a website, it is safe to assume that whatever conclusions we draw from differences in specific member states will play a strong role in both these metrics.

Looking at individual countries, we note that Cyprus appears to be an outlier, having 83.94% of websites available in English and 67.6% using it as a primary language. This occurs despite the fact that English is not an official language in Cyprus. This can be attributed mostly to cultural factors. Cyprus was a British colony up until 1960, so both the cultural and the linguistic influence of the British Empire are strong. Additionally, international banking and tourism are both prevalent in Cyprus, creating an environment that encourages the use of English as a primary language in websites. This, combined with the small population of Cyprus and its position in the periphery of the EU, create the conditions which make the English language more popular than even the island nations' official languages.

Taking a country's geographical location in relation to the EU more into consideration, we notice the highest percentages of English availability and prevalence are noticed in peripheral countries. Countries of the Balkan peninsula, countries of the Baltic Sea and Scandinavia, as well as Portugal all have higher percentages than the EU average both in English availability and in English as a primary language. This might indicate an effort from peripheral countries to achieve greater integration with both European culture and the common marketplace. Most peripheral countries tend to also be later additions to the EU roster, which reinforces their need to put more effort into internationalization and EU integration.

On a different note, Belgium and Luxemburg also seem to display a high influence of the English language. Both these countries have more than one non-English official language. This creates a multilingual culture and somewhat diminishes the sense of national pride that might be otherwise connected with language. Additionally, the cost of integrating a secondary language into a website is lower if the website already has support for more than one language. The multilanguage functionality is there and all that remains is the addition of the translated content.

In addition to observing each country individually, some of the factors that influence the availability or status of the English language in the websites of non-English speaking EU member states might be

attributed to the wider region. Cultural relations often arise from common history or the geo-political status of a wider region, all of which may play a part in the adoption of English in the World Wide Web. In order to better understand regional factors, we separated the 25 non-English speaking countries of the EU to four different regions based on how they are defined by the European Vocabularies [46] of the Publications Office of the European Union, which is an "interinstitutional office whose task is to publish the publications of the institutions of the European Union" [47]. Table 8 shows how the countries were divided.

Table 8. Non-English speaking Countries divided to four European Regions as defined by EuroVoc.

Central and Eastern Europe	Northern Europe	Southern Europe	Western Europe
Bulgaria, Croatia, Czechia, Hungary, Poland, Romania, Slovakia, Slovenia	Denmark, Estonia, Finland, Latvia, Lithuania, Sweden	Cyprus, Greece, Italy, Portugal, Spain	Austria, Belgium, France, Germany, Luxemburg, Netherlands

The average percentage of websites that have English available and the average percentage of Websites that have English as their primary language in each different region are shown in Figures 12 and 13, respectively.

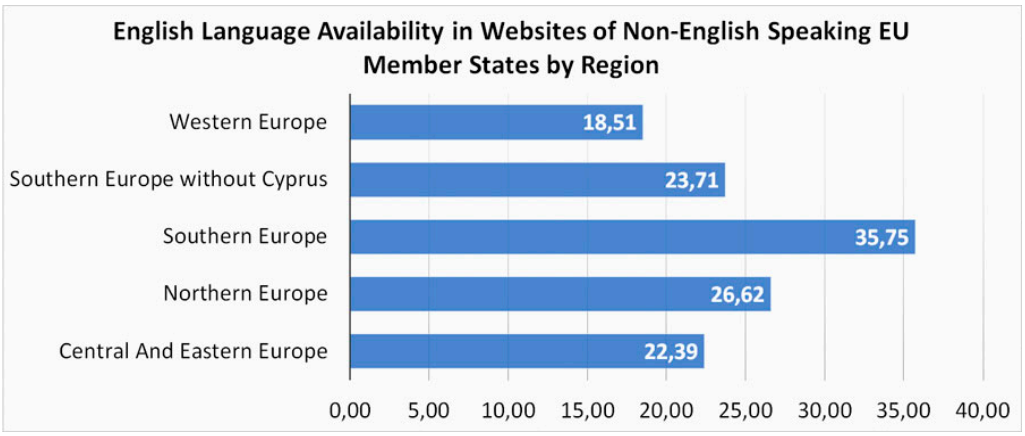


Figure 12. English Language Availability in Websites of Non-English speaking EU Member states by region.

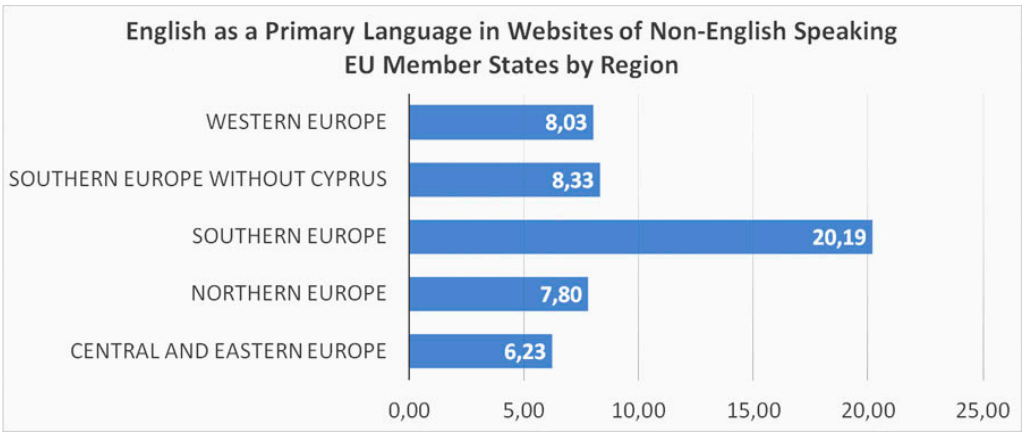


Figure 13. English as a Primary Language in Websites of Non-English speaking EU Member States by Region.

Southern Europe appears to be noticeably keener to include English either as a primary language or as just an available language in most cases. This is in large part due to the influence of Cyprus, which is an outlier. Calculating Southern Europe without Cyprus will lead to 23.31% instead of 35.75% on availability and 8.33% instead of 20.19% on primary language. This brings Southern Europe in tow with the other regions. Western Europe seems to be somewhat more reluctant to provide English as an available language, but seems to be keener to have it as a primary language.

6. Conclusions

The present study has demonstrated explicitly that the English language represents the number one choice for international communication in the European Union through the World Wide Web. More than one quarter of websites belonging to NTLDs of non-English speaking countries, on average, offer their content in the English language. On the other hand, websites of English speaking countries very rarely offer the option for any other language besides English. This clearly indicates that, for both English speaking countries and non-English speaking ones, the use of the English language is viewed as the most efficient way to attain international reach. On top of that, a significant number (almost 10%) of websites from non-English speaking countries prioritize English over their own official language or languages. This further reinforces the status of English as the most prevalent language for international communication in the wider European Union.

When studying the reasons that lead to a greater or lesser availability of the English language in websites of non-English speaking EU member states, a statistically significant moderate negative correlation was discovered between both the population and GDP of a member state and the availability of English in its websites. This signifies that, in general, larger countries population-wise and more affluent countries put less effort into providing their website content in English than smaller countries. This correlation reinforces the notion that language is a hard barrier in achieving greater reach. Countries with a smaller number of people speaking their official language need to compensate by putting more effort into providing their content in English, in order to attain the reach that an equivalent website in a larger non-English speaking country or in an English speaking country would have.

Besides population and GDP, some characteristics that were identified to influence the prevalence of the English language were: i) a country's position in the EU periphery (in the Balkans, the Baltic states, Scandinavia, and Portugal), ii) an already existing multilingual culture due to more than one official language (in Belgium and Luxemburg), and iii) a close historical connection with British culture (in the case of Cyprus).

The fact that, to the best of our knowledge, there are no directly related works, makes it difficult to compare the results of this study to previous research on the field, Yet, it is clear that the quantified information presented in the Results section, as well as the hermeneutical insights elaborated in the Discussion section provide numerical evidence of i) the theoretical approaches regarding national identities in the era of globalization, ii) the discussion about the use of English language as a means of international communication, and iii) the political analysis of the role of national identities in the multinational environment of the EU. Thus, the study is in common perspective with loosely related works along the above-mentioned axes. The main contribution of this study lies in the provision of quantified information regarding the usage of the English language in the websites of EU, obtained from a large-scale data research all over the web of EU member states. In addition, this research paves the way for the exploration of the current status of European integration in quite a different way to the traditional approaches. Instead of opinion polls and theoretical analysis, this study indicates that more accurate results may be achieved by the use of unintentionally and freely provided data on the web.

Pushing this research further, it would be interesting to examine the language situation of the .eu top-level domain which was launched in 2005. Although not explicitly an NTLD, it acts as a representation for the EU. Its popularity in different member states or the popularity of different languages in websites using that particular TLD might lead to some interesting conclusions. Additionally, applying the same or a similar methodology, the prevalence of other languages in different

parts of the world can be studied. For example, the use of Chinese in the Far East and Indonesia or a comparison between English and Spanish in Latin America can be explored (continent-wide reach versus local/international reach). The World Wide Web and the Internet in general provide a vast amount of data that can be used to study the diffusion of language world-wide through clearly defined metrics and can help reach conclusions about language which also hold true in the offline world.

Author Contributions: Conceptualization, A.G.; Methodology, A.G. and M.P.; Software, M.P.; Validation: Iraklis Varlamis; Formal analysis, M.P. and N.K.; Investigation, M.P., A.L., L.L. and I.V.; Resources, L.L.; Data curation: M.P. and N.K.; Writing—original draft, M.P., N.K. and L.L.; Writing—review & editing, M.P. and I.V.; Visualization, A.L.; Supervision, A.G. and I.V.; Project administration, A.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A Data Files

Giannakoulopoulos, A; Pergantis, M; Konstantinou, N; Lamprogeorgos. A; Limniati, L; Varlamis, I. Data set of the article: Exploring the Dominance of the English Language on the Websites of EU Countries, (Version 1) (Data set). Zenodo. Available online: https://doi.org/10.5281/zenodo.3698008 (accessed on 05 March 2020).

Appendix B Detailed Result Tables and Charts for Each Country

B.1. Austria

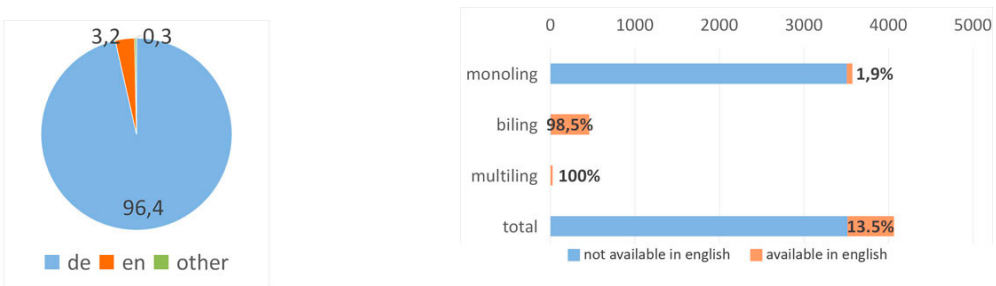
Below you may find the results for the websites in the TLD .at (Austria) (Table A1, Table A2, and Figure A1).

Table A1. Primary languages (at).

Languages	Total	Percent
de	3918	96.4%
en	132	3.2%
other	13	0.3%

Table A2. English language availability (at).

	Monoling	Biling	Multiling	Total
not available in English	3506	7	0	3513
available in English	68	454	28	550
total	3574	461	28	4063
English percentage	1.9%	98.5%	100.0%	13.5%



(a) (b)

Figure A1. Visualization of the results for domain at. (**a**) Primary languages pie chart and (**b**) English language availability bar chart.

B.2. Belgium

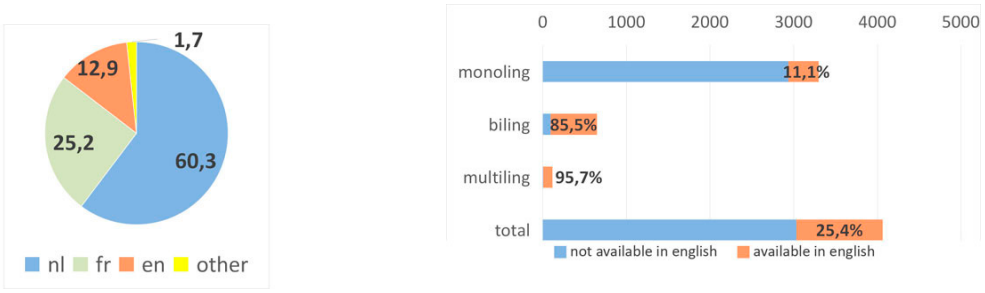
Below, you may find the results for the websites in the TLD .be (Belgium) (Table A3, Table A4, and Figure A2).

Table A3. Primary languages (be).

Languages	Total	Percent
nl	2449	60.3%
fr	1022	25.2%
en	523	12.9%
other	69	1.7%

Table A4. English language availability (be).

	Monoling	Biling	Multiling	Total
not available in English	2934	94	5	3033
available in English	365	553	112	1030
total	3299	647	117	4063
English percentage	11.1%	85.5%	95.7%	25.4%



(a) (b)

Figure A2. Visualization of the results for domain be. (**a**) Primary languages pie chart and (**b**) English language availability bar chart.

B.3. Bulgaria

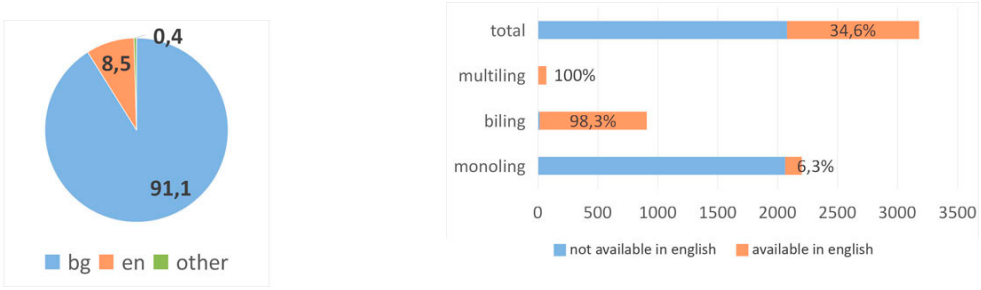
Below, you may find the results for the websites in the TLD .bg (Bulgaria) (Table A5, Table A6, and Figure A3).

Table A5. Primary languages (bg).

Languages	Total	Percent
bg	2894	91.1%
en	270	8.5%
other	14	0.4%

Table A6. English language availability (bg).

	Monoling	Biling	Multiling	Total
not available in English	2062	15	0	2077
available in English	139	891	71	1101
total	2201	906	71	3178
English percentage	6.3%	98.3%	100.0%	34.6%



(a) (b)

Figure A3. Visualization of the results for domain bg. (**a**) Primary languages pie chart and (**b**) English language availability bar chart.

B.4. Croatia

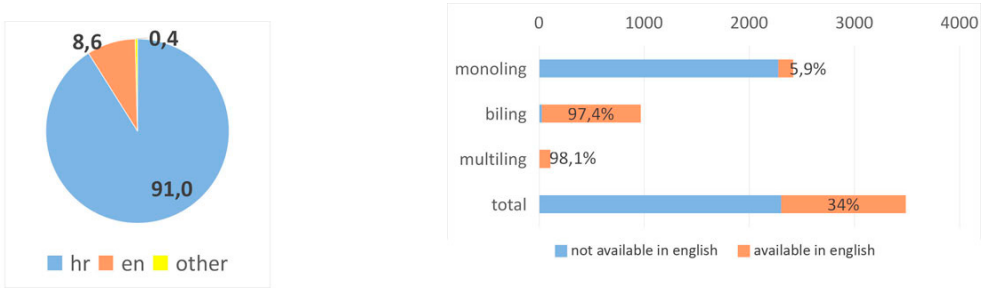
Below, you may find the results for the websites in the TLD .hr (Croatia) (Table A7, Table A8 and Figure A4).

Table A7. Primary languages (hr).

Languages	Total	Percent
hr	3176	91.0%
en	299	8.6%
other	14	0.4%

Table A8. English language availability (hr).

	Monoling	Biling	Multiling	Total
not available in English	2275	25	2	2302
available in English	143	940	104	1187
total	2418	965	106	3489
English percentage	5.9%	97.4%	98.1%	34.0%



(a) (b)

Figure A4. Visualization of the results for domain hr. (**a**) Primary languages pie chart and (**b**) English language availability bar chart.

B.5. Cyprus

Below you may find the results for the websites in the TLD .cy (Cyprus) (Table A9, Table A10, and Figure A5).

Table A9. Primary languages (cy).

Languages	Total	Percent
el	265	32.0%
en	560	67.6%
other	3	0.4%

Table A10. English language availability (cy).

	Monoling	Biling	Multiling	Total
not available in English	133	0	0	133
available in English	421	255	19	695
total	554	255	19	828
English percentage	76.0%	100.0%	100.0%	83.9%

(a) (b)

Figure A5. Visualization of the results for domain cy. (**a**) Primary languages pie chart and (**b**) English language availability bar chart.

B.6. Czechia

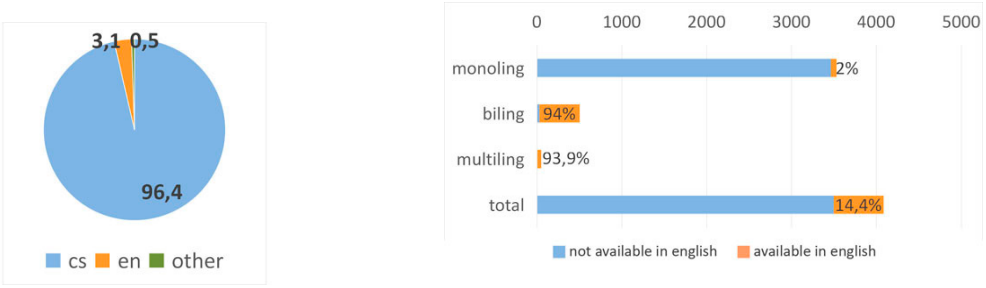
Below, you may find the results for the websites in the TLD .cz (Chechia) (Table A11, Table A12, and Figure A6).

Table A11. Primary languages (cz).

Languages	Total	Percent
cs	3937	96.4%
en	126	3.1%
other	21	0.5%

Table A12. English language availability (cz).

	Monoling	Biling	Multiling	Total
not available in English	3462	30	3	3495
available in English	69	474	46	589
total	3531	504	49	4084
English percentage	2.0%	94.0%	93.9%	14.4%



(a) (b)

Figure A6. Visualization of the results for domain cz. (**a**) Primary languages pie chart and (**b**) English language availability bar chart.

B.7. Denmark

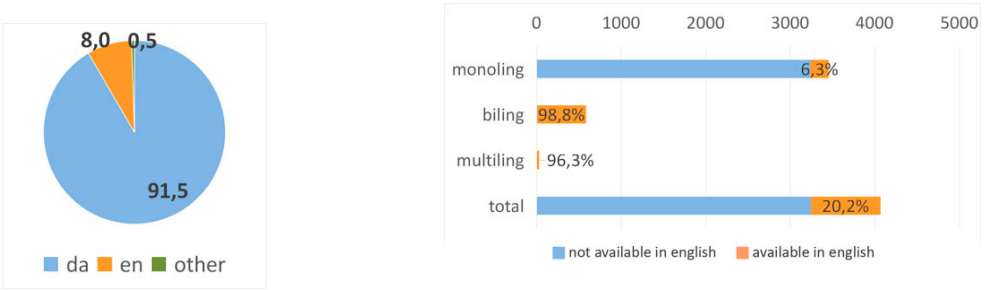
Below, you may find the results for the websites in the TLD .dk (Denmark) (Table A13, Table A14, and Figure A7).

Table A13. Primary languages (dk).

Languages	Total	Percent
da	3723	91.5%
en	324	8.0%
other	20	0.5%

Table A14. English language availability (dk).

	Monoling	Biling	Multiling	Total
not available in English	3238	7	1	3246
available in English	218	577	26	821
total	3456	584	27	4067
English percentage	6.3%	98.8%	96.3%	20.2%



(a) (b)

Figure A7. Visualization of the results for domain dk. (**a**) Primary languages pie chart and (**b**) English language availability bar chart.

B.8. Estonia

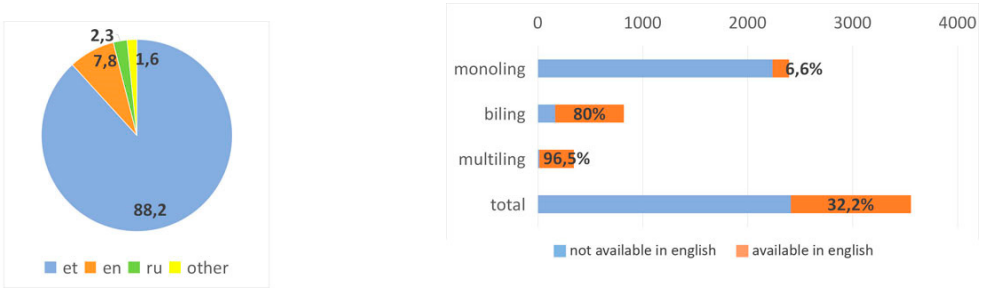
Below, you may find the results for the websites in the TLD .ee (Estonia) (Table A15, Table A16, and Figure A8).

Table A15. Primary languages (ee).

Languages	Total	Percent
et	3136	88.2%
en	279	7.8%
ru	83	2.3%
other	58	1.6%

Table A16. English language availability (ee).

	Monoling	Biling	Multiling	Total
not available in English	2236	164	12	2412
available in English	158	656	330	1144
total	2394	820	342	3556
English percentage	6.6%	80.0%	96.5%	32.2%



(a) (b)

Figure A8. Visualization of the results for domain ee. (**a**) Primary languages pie chart and (**b**) English language availability bar chart.

B.9. Finland

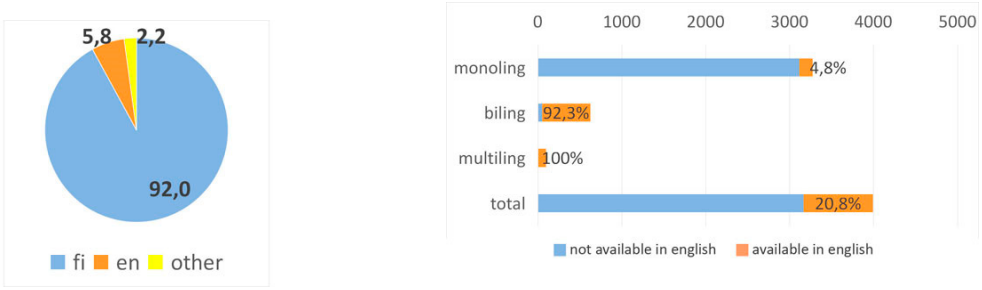
Below, you may find the results for the websites in the TLD .fi (Finland) (Table A17, Table A18, and Figure A9).

Table A17. Primary languages (fi).

Languages	Total	Percent
fi	3674	92.0%
en	232	5.8%
other	86	2.2%

Table A18. English language availability (fi).

	Monoling	Biling	Multiling	Total
not available in English	3115	48	0	3163
available in English	157	578	94	829
total	3272	626	94	3992
English percentage	4.8%	92.3%	100.0%	20.8%



(a) (b)

Figure A9. Visualization of the results for domain fi. (**a**) Primary languages pie chart and (**b**) English language availability bar chart.

B.10. France

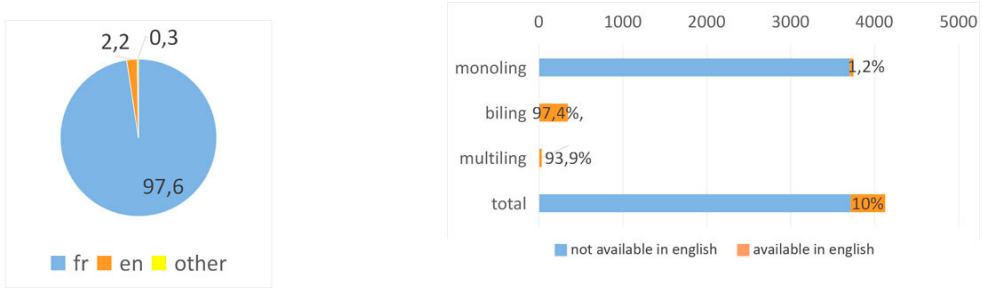
Below, you may find the results for the websites in the TLD .fr (France) (Table A19, Table A20, and Figure A10).

Table A19. Primary languages (fr).

Languages	Total	Percent
fr	4025	97.6%
en	89	2.2%
other	11	0.3%

Table A20. English language availability (fr).

	Monoling	Biling	Multiling	Total
not available in English	3702	9	2	3713
available in English	46	335	31	412
total	3748	344	33	4125
English percentage	1.2%	97.4%	93.9%	10.0%



(a) (b)

Figure A10. Visualization of the results for domain fr. (**a**) Primary languages pie chart and (**b**) English language availability bar chart.

B.11. Germany

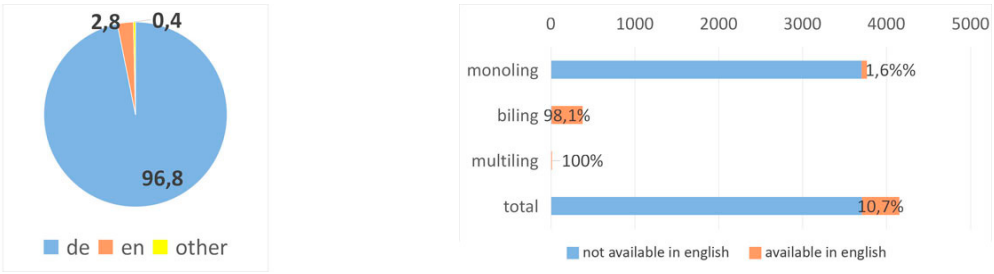
Below, you may find the results for the websites in the TLD .de (Germany) (Table A21, Table A22, and Figure A11).

Table A21. Primary languages (de).

Languages	Total	Percent
de	4018	96.8%
en	117	2.8%
other	15	0.4%

Table A22. English language availability (de).

	Monoling	Biling	Multiling	Total
not available in English	3701	7	0	3708
available in English	62	370	10	442
total	3763	377	10	4150
English percentage	1.6%	98.1%	100%	10.7%



(a) (b)

Figure A11. Visualization of the results for domain de. (**a**) Primary languages pie chart and (**b**) English language availability bar chart.

B.12. Greece

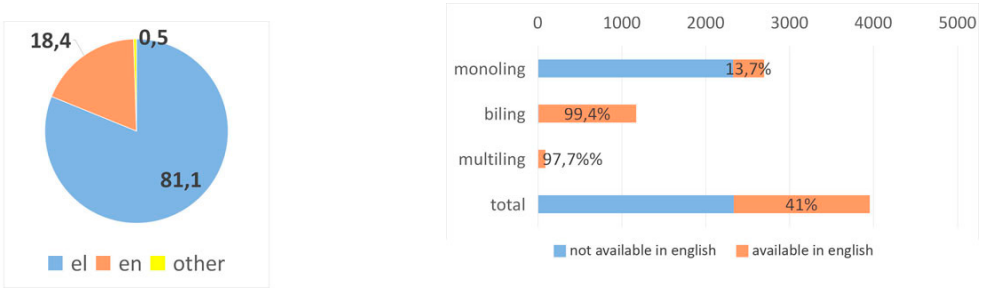
Below you may find the results for the websites in the TLD .gr (Greece) (Table A23, Table A24, and Figure A12).

Table A23. Primary languages (gr).

Languages	Total	Percent
el	3207	81.1%
en	726	18.4%
other	20	0.5%

Table A24. English language availability (gr).

	Monoling	Biling	Multiling	Total
not available in English	2324	7	2	2333
available in English	369	1165	86	1620
total	2693	1172	88	3953
English percentage	13.7%	99.4%	97.7%	41.0%



(a) (b)

Figure A12. Visualization of the results for domain gr. (**a**) Primary languages pie chart and (**b**) English language availability bar chart.

B.13. Hungary

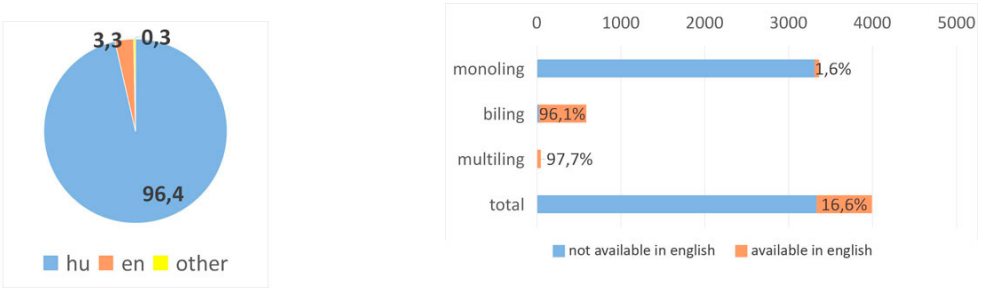
Below, you may find the results for the websites in the TLD .hu (Hungary) (Table A25, Table A26, and Figure A13).

Table A25. Primary languages (hu).

Languages	Total	Percent
hu	3849	96.4%
en	131	3.3%
other	13	0.3%

Table A26. English language availability (hu).

	Monoling	Biling	Multiling	Total
not available in English	3308	23	1	3332
available in English	55	564	42	661
total	3363	587	43	3993
English percentage	1.6%	96.1%	97.7%	16.6%



(a) (b)

Figure A13. Visualization of the results for domain hu. (**a**) Primary languages pie chart and (**b**) English language availability bar chart.

B.14. Ireland

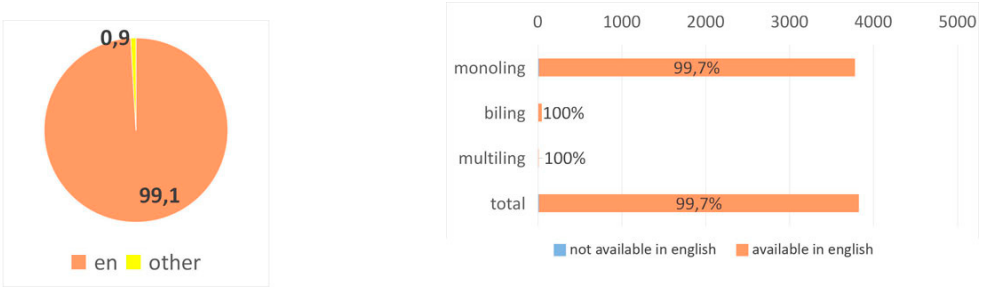
Below, you may find the results for the websites in the TLD .ie (Ireland) (Table A27, Table A28, and Figure A14).

Table A27. Primary languages (ie).

Languages	Total	Percent
en	3790	99.1%
other	35	0.9%

Table A28. English language availability (ie).

	Monoling	Biling	Multiling	Total
not available in English	13	0	0	13
available in English	3763	47	2	3812
total	3776	47	2	3825
English percentage	99.7%	100.0%	100.0%	99.7%



(a) (b)

Figure A14. Visualization of the results for domain ie. (**a**) Primary languages pie chart and (**b**) English language availability bar chart.

B.15. Italy

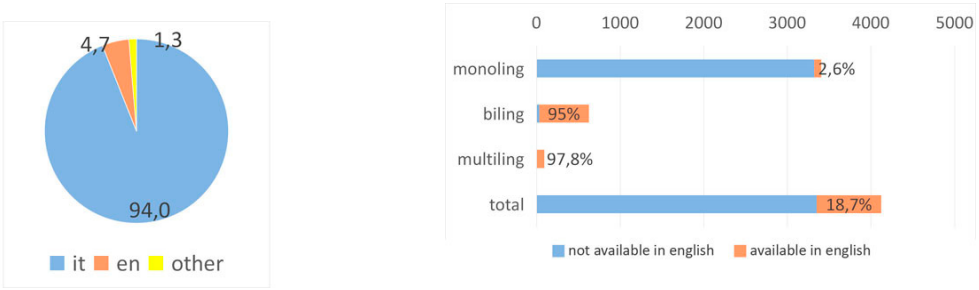
Below, you may find the results for the websites in the TLD .it (Italy) (Table A29, Table A30, and Figure A15).

Table A29. Primary languages (it).

Languages	Total	Percent
nl	3874	94%
en	195	4.7%
other	54	1.3%

Table A30. English language availability (it).

	Monoling	Biling	Multiling	Total
not available in English	3318	31	2	3351
available in English	88	595	89	772
total	3406	626	91	4123
English percentage	2.6%	95%	97.8%	18.7%



(a) (b)

Figure A15. Visualization of the results for domain it. (**a**) Primary languages pie chart and (**b**) English language availability bar chart.

B.16. Latvia

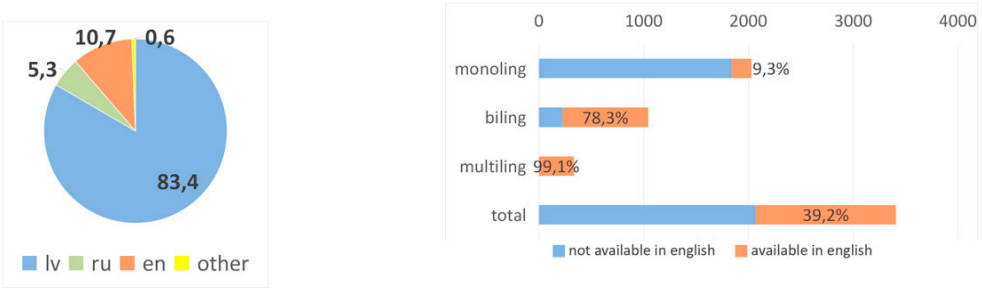
Below, you may find the results for the websites in the TLD .lv (Latvia) (Table A31, Table A32, and Figure A16).

Table A31. Primary languages (lv).

Languages	Total	Percent
lv	2839	83.4%
ru	181	5.3%
en	364	10.7%
other	22	0.6%

Table A32. English language availability (lv).

	Monoling	Biling	Multiling	Total
not available in English	1840	227	3	2070
available in English	188	817	331	1336
total	2028	1044	334	3406
English percentage	9.3%	78.3%	99.1%	39.2%



(a) (b)

Figure A16. Visualization of the results for domain lv. (**a**) Primary languages pie chart and (**b**) English language availability bar chart.

B.17. Lithuania

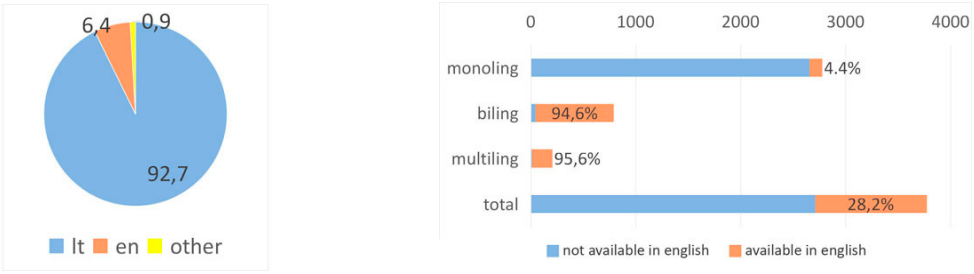
Below, you may find the results for the websites in the TLD. lt (Lithuania) (Table A33, Table A34, and Figure A17).

Table A33. Primary languages (lt).

Languages	Total	Percent
lt	3497	92.7%
en	241	6.4%
other	35	0.9%

Table A34. English language availability (lt).

	Monoling	Biling	Multiling	Total
not available in English	2657	43	9	2709
available in English	121	747	196	1064
total	2778	790	205	3773
English percentage	4.4%	94.6%	95.6%	28.2%



(a) (b)

Figure A17. Visualization of the results for domain lt. (**a**) Primary languages pie chart and (**b**) English language availability bar chart.

B.18. Luxemburg

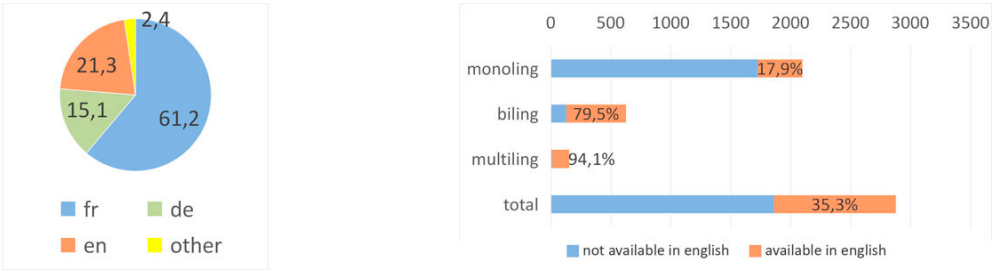
Below, you may find the results for the websites in the TLD .lu (Luxemburg) (Table A35, Table A36, and Figure A18).

Table A35. Primary languages (lu).

Languages	Total	Percent
fr	1760	61.2%
de	434	15.1%
en	614	21.3%
other	68	2.4%

Table A36. English language availability (lu).

	Monoling	Biling	Multiling	Total
not available in English	1723	128	9	1860
available in English	375	497	144	1016
total	2098	625	153	2876
English percentage	17.9%	79.5%	94.1%	35.3%



(a) (b)

Figure A18. Visualization of the results for domain lu. (**a**) Primary languages pie chart and (**b**) English language availability bar chart.

B.19. Malta

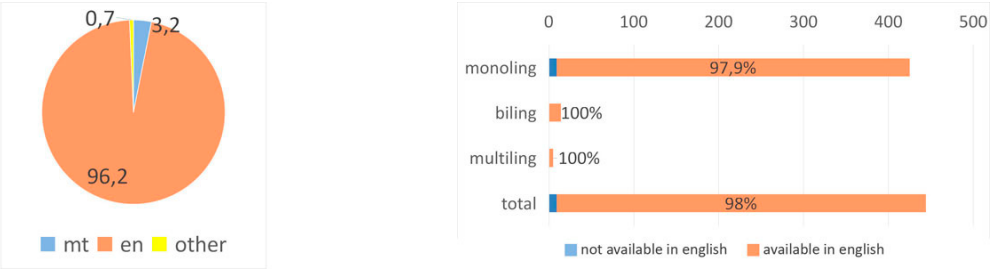
Below, you may find the results for the websites in the TLD .mt (Malta) (Table A37, Table A38, and Figure A19).

Table A37. Primary languages (mt).

Languages	Total	Percent
mt	14	3.2%
en	427	96.2%
other	3	0.7%

Table A38. English language availability (mt).

	Monoling	Biling	Multiling	Total
not available in English	9	0	0	9
available in English	416	14	5	435
total	425	14	5	444
English percentage	97.9%	100.0%	100.0%	98.0%



(a) (b)

Figure A19. Visualization of the results for domain mt. (**a**) Primary languages pie chart and (**b**) English language availability bar chart.

B.20. Netherlands

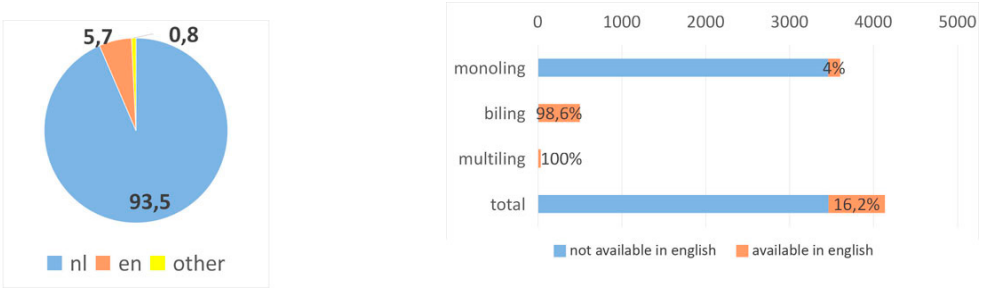
Below, you may find the results for the websites in the TLD .nl (Netherlands) (Table A39, Table A40, and Figure A20).

Table A39. Primary languages (nl).

Languages	Total	Percent
nl	3864	93.5%
en	237	5.7%
other	32	0.8%

Table A40. English language availability (nl).

	Monoling	Biling	Multiling	Total
not available in English	3457	7	0	3464
available in English	143	493	33	669
total	3600	500	33	4133
English percentage	4%	98.6%	100%	16.2%



(a) (b)

Figure A20. Visualization of the results for domain nl. (**a**) Primary languages pie chart and (**b**) English language availability bar chart.

B.21. Poland

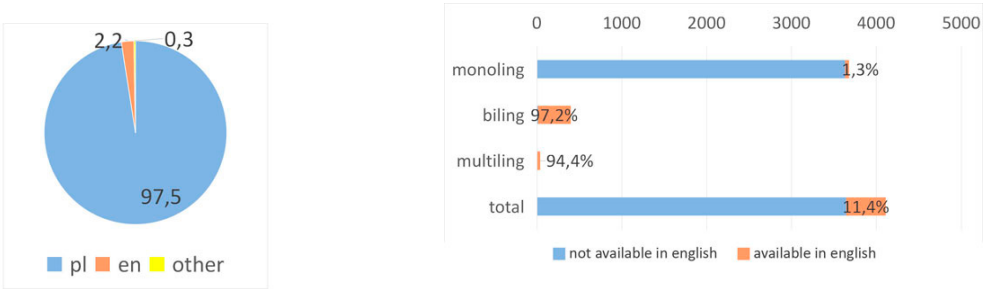
Below, you may find the results for the websites in the TLD .pl (Poland) (Table A41, Table A42, and Figure A21).

Table A41. Primary languages (pl).

Languages	Total	Percent
pl	4008	97.5%
en	90	2.2%
other	12	0.3%

Table A42. English language availability (pl).

	Monoling	Biling	Multiling	Total
not available in English	3629	11	2	3642
available in English	47	387	34	468
total	3676	398	36	4110
English percentage	1.3%	97.2%	94.4%	11.4%



(a) (b)

Figure A21. Visualization of the results for domain pl. (**a**) Primary languages pie chart and (**b**) English language availability bar chart.

B.22. Portugal

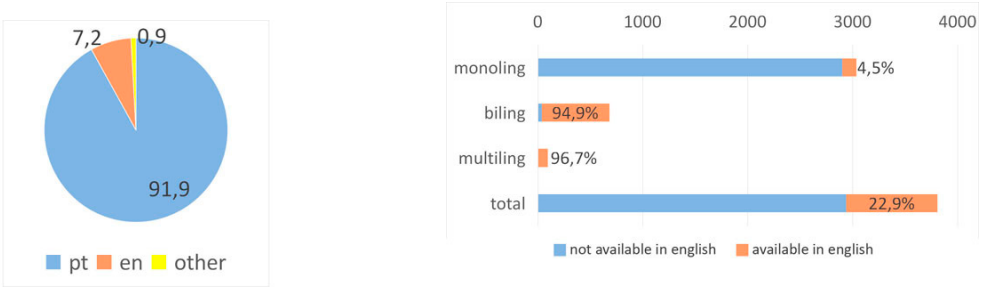
Below, you may find the results for the websites in the TLD .pt (Portugal) (Table A43, Table A44, and Figure A22).

Table A43. Primary languages (pt).

Languages	Total	Percent
pt	3502	91.9%
en	274	7.2%
other	33	0.9%

Table A44. English language availability (pt).

	Monoling	Biling	Multiling	Total
not available in English	2900	35	3	2938
available in English	136	646	89	871
total	3036	681	92	3809
English percentage	4.5%	94.9%	96.7%	22.9%



(a) (b)

Figure A22. Visualization of the results for domain pt. (**a**) Primary languages pie chart and (**b**) English language availability bar chart.

B.23. Romania

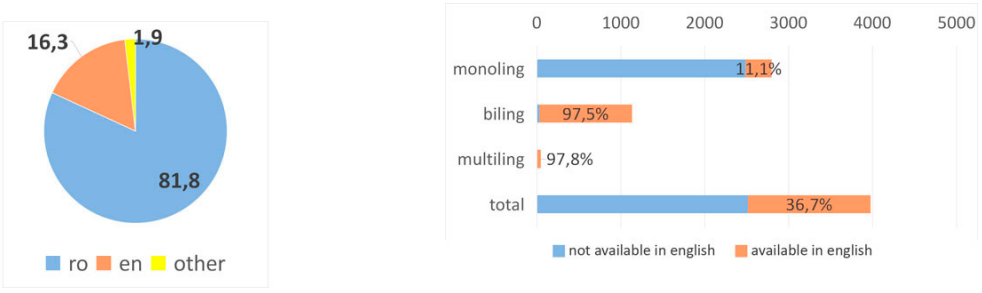
Below, you may find the results for the websites in the TLD .ro (Romania) (Table A45, Table A46, and Figure A23).

Table A45. Primary languages (ro).

Languages	Total	Percent
ro	3253	81,8%
en	648	16,3%
other	74	1,9%

Table A46. English language availability (ro).

	Monoling	Biling	Multiling	Total
not available in English	2486	28	1	2515
available in English	311	1104	45	1460
total	2797	1132	46	3975
English percentage	11.1%	97.5%	97.8%	36.7%



(a) (b)

Figure A23. Visualization of the results for domain ro. (**a**) Primary languages pie chart and (**b**) English language availability bar chart.

B.24. Slovakia

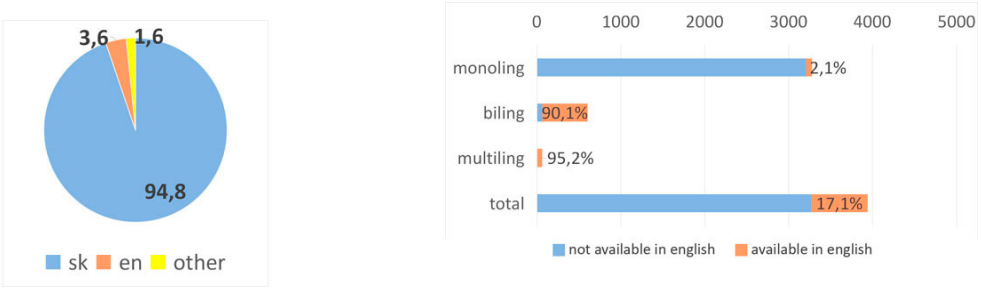
Below, you may find the results for the websites in the TLD .sk (Slovakia) (Table A47, Table A48, and Figure A24).

Table A47. Primary languages (sk).

Languages	Total	Percent
sk	3736	94.8%
en	143	3.6%
other	64	1.6%

Table A48. English language availability (sk).

	Monoling	Biling	Multiling	Total
not available in English	3207	60	3	3270
available in English	70	544	59	673
total	3277	604	62	3943
English percentage	2.1%	90.1%	95.2%	17.1%



(a) (b)

Figure A24. Visualization of the results for domain sk. (**a**) Primary languages pie chart and (**b**) English language availability bar chart.

B.25. Slovenia

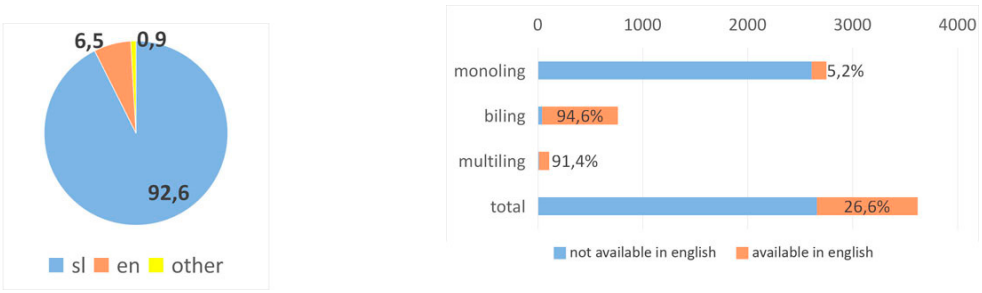
Below, you may find the results for the websites in the TLD .si (Slovenia) (Table A49, Table A50, and Figure A25).

Table A49. Primary languages (si).

Languages	Total	Percent
sl	3350	92.6%
en	236	6.5%
other	33	0.9%

Table A50. English language availability (si).

	Monoling	Biling	Multiling	Total
not available in English	2608	41	9	2658
available in English	143	722	96	961
total	2751	763	105	3619
English percentage	5.2%	94.6%	91.4%	26.6%



(a) (b)

Figure A25. Visualization of the results for domain si. (**a**) Primary languages pie chart and (**b**) English language availability bar chart.

B.26. Spain

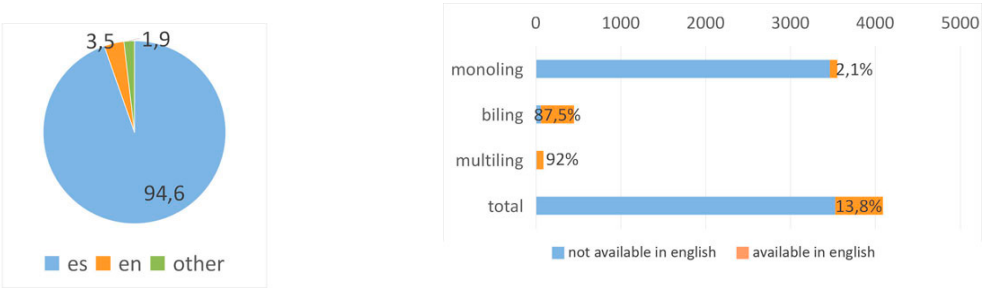
Below, you may find the results for the websites in the TLD .es (Spain) (Table A51, Table A52, and Figure A26).

Table A51. Primary languages (es).

Languages	Total	Percent
es	3868	94.6%
en	144	3.5%
other	76	1.9%

Table A52. English language availability (es).

	Monoling	Biling	Multiling	Total
not available in English	3461	56	7	3524
available in English	91	392	81	564
total	3552	448	88	4088
English percentage	2.6%	87.5%	92%	13.8%



(a) (b)

Figure A26. Visualization of the results for domain es. (**a**) Primary languages pie chart and (**b**) English language availability bar chart.

B.27. Sweden

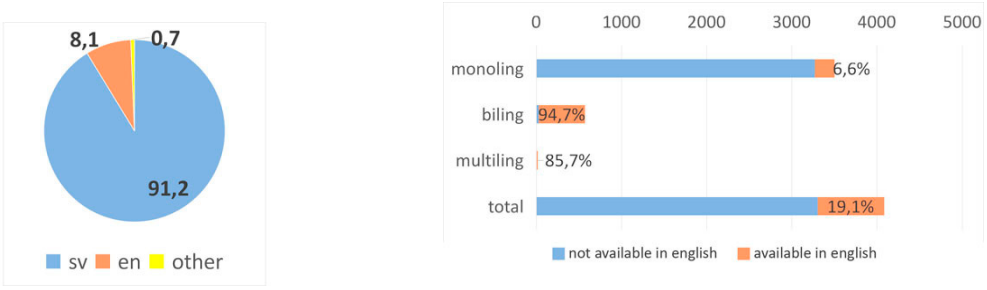
Below, you may find the results for the websites in the TLD .se (Sweden) (Table A53, Table A54, and Figure A27).

Table A53. Primary languages (se).

Languages	Total	Percent
sv	3726	91.2%
en	331	8.1%
other	27	0.7%

Table A54. English language availability (se).

	Monoling	Biling	Multiling	Total
not available in English	3270	30	2	3302
available in English	230	540	12	782
total	3500	570	14	4084
English percentage	6.6%	94.7%	85.7%	19.1%



(a) (b)

Figure A27. Visualization of the results for domain se. (**a**) Primary languages pie chart and (**b**) English language availability bar chart.

B.28. United Kingdom

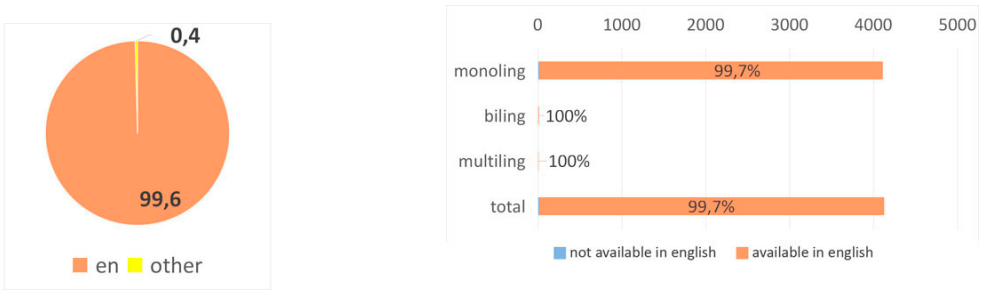
Below, you may find the results for the websites in the TLD .uk (United Kingdom) (Table A55, Table A56, and Figure A28).

Table A55. Primary languages (uk).

Languages	Total	Percent
en	4108	99.6%
other	17	0.4%

Table A56. English language availability (uk).

	Monoling	Biling	Multiling	Total
not available in English	14	0	0	14
available in English	4094	11	6	4111
total	4108	11	6	4125
English percentage	99.7%	100%	100%	99.7%



(a) (b)

Figure A28. Visualization of the results for domain uk. (**a**) Primary languages pie chart and (**b**) English language availability bar chart.

References

1. Costales, A.F. The internationalization of institutional websites: The case of universities in the European Union. *Transl. Res. Proj.* **2012**, *4*, 51–60.
2. Nunes-da-Cunha, I.; Martinez, F.M.; Fernandez-Llimos, F. A Global Comparison of Internationalization Support Characteristics Available on College of Pharmacy Websites. *Am. J. Pharm. Educ.* **2019**, *83*, 6592. [PubMed]

3. Schlesinger Wass, E. *Addressing the World: National Identity and Internet Country Code Domains*; Rowman & Littlefield Publishers: Lanham, MD, USA, 2003; p. 16.
4. Strover, S. A Review of: "Addressing the World: National Identity and Internet Country Code Domains". *Inform. Soc.* **2006**, *44*, 59–60. [CrossRef]
5. McLuhan, M. *The Gutenberg Galaxy*; University of Toronto Press: Toronto, ON, USA, 1963.
6. Khiabany, G. Globalization and the Internet: Myths and realities. *Trends Commun.* **2003**, *11*, 137–153. [CrossRef]
7. Internetworldstats.com. Top Ten Internet Languages in the World—Internet Statistics. Available online: https://www.Internetworldstats.com/stats7.htm (accessed on 3 February 2020).
8. Internetworldstats.com. World Internet Users Statistics and 2020 World Population Stats. Available online: https://www.Internetworldstats.com/stats.htm (accessed on 3 January 2020).
9. Giddens, A. *The Consequences of Modernity*; John Wiley & Sons: Hoboken, NJ, USA, 2013.
10. Block, D. Globalization, transnational communication and the Internet. *Int. J. Multicult. Soc.* **2004**, *6*, 13–28.
11. Latouche, S. *The Westernization of the World: The Significance, Scope and Limits of the Drive towards Global Uniformity*; Polity Press: Cambridge, UK, 1996.
12. Nederveen Pieterse, J. Globalization as hybridization. *ISS Working Paper Ser. Gen. Ser.* **1993**, *152*, 1–18.
13. Robertson, R. Globalization: Time-Space and Homogeneity-Heterogeneity. In *Global Modernities*; Sage Publications: Thousand Oaks, CA, USA, 1995; pp. 25–44.
14. Jiang, W. The relationship between culture and language. *ELT J.* **2000**, *54*, 328–334. [CrossRef]
15. Gazzola, M. Managing multilingualism in the European Union: Language policy evaluation for the European Parliament. *Lang. Policy* **2006**, *5*, 395. [CrossRef]
16. Crystal, D. *English as a Global Language*; Cambridge University Press: Cambridge, UK, 2012; pp. 125–127.
17. Kim, C.W. On the globalization of English: Some internal factors. *World Engl.* **2019**, *38*, 128–133. [CrossRef]
18. Graddol, D. Global English, global culture? In *Redesigning English*; Routledge: Abingdon-on-Thames, UK, 1997; pp. 193–246.
19. Dor, D. From Englishization to imposed multilingualism: Globalization, the Internet, and the political economy of the linguistic code. *Public Cult.* **2004**, *16*, 97–118. [CrossRef]
20. Fishman, J.A. The new linguistic order. *Foreign Policy* **1998**, 26–40. [CrossRef]
21. Edwards, J. *Language, Society, and Identity*; Basil Blackwell: Oxford, UK, 1985.
22. European Union. EU languages | European Union. Available online: https://europa.eu/european-union/about-eu/eu-languages_en (accessed on 5 February 2020).
23. Wodak, R.; Boukala, S. (Supra) national identity and language: Rethinking national and European migration policies and the linguistic integration of migrants. *Ann. Rev. Appl. Linguist.* **2015**, *35*, 253–273. [CrossRef]
24. Khun, T. Grand theories of European integration revisited: Does identity politics shape the course of European integration? *J. Eur. Public Policy* **2019**, *26*, 1213–1230. [CrossRef]
25. Mongeon, P.; Paul-Hus, A. The journal coverage of Web of Science and Scopus: A comparative analysis. *Scientometrics* **2016**, *106*, 213–228. [CrossRef]
26. Martín-Martín, A.; Orduna-Malea, E.; Thelwall, M.; Delgado López-Cózar, E. Google Scholar, Web of Science, and Scopus: A systematic comparison of citations in 252 subject categories. *J. Inform.* **2018**, *12*, 1160–1177. [CrossRef]
27. Wei, C.Y.; Kolko, B.E. Resistance to globalization: Language and Internet diffusion patterns in Uzbekistan. *New Rev. Hypermed. Multimed.* **2005**, *11*, 205–220. [CrossRef]
28. Phillipson, R. English-only Europe? In *Challenging Language Policy*; Psychology Press: London, UK, 2003.
29. Phillipson, R. Language policy and education in the European Union. *Lang. Policy Polit. Issues Educ.* **2008**, *1*, 255–265.
30. Hillier, M. The role of cultural context in multilingual website usability. *Electron. Commer. Res. Appl.* **2003**, *2*, 2–14. [CrossRef]
31. Mueller, M.L. The battle over Internet domain names: Global or national TLDs? *Telecommun. Policy* **1998**, *22*, 89–107. [CrossRef]
32. What is PHP? Available online: https://www.php.net/manual/en/intro-whatis.php (accessed on 3 February 2020).
33. About MariaDB Server. Available online: https://mariadb.org/about/ (accessed on 3 February 2020).
34. Common Crawl. Available online: https://commoncrawl.org/about/ (accessed on 3 February 2020).

35. CDX Server API. Available online: https://github.com/webrecorder/pywb/wiki/CDX-Server-API (accessed on 3 February 2020).
36. Common Crawl Index Server. Available online: https://index.commoncrawl.org/ (accessed on 3 February 2020).
37. Publicsuffix.org. Available online: https://publicsuffix.org/list/public_suffix_list.dat (accessed on 3 February 2020).
38. Global WHOIS Search. Available online: https://www.whois365.com/en/listtld/europe (accessed on 3 February 2020).
39. GitHub. Available online: https://github.com/CLD2Owners/cld2 (accessed on 3 February 2020).
40. Cochran, W.G. Note on an approximate formula for the significance levels of z. *Ann. Math. Stat.* **1940**, *11*, 93–95. [CrossRef]
41. SurveyMonkey: Sample Size Calculator. Available online: https://www.surveymonkey.com/mp/sample-size-calculator/ (accessed on 3 February 2020).
42. SurveyMonkey: About Us. Available online: https://www.surveymonkey.com/mp/aboutus/?ut_source=footer (accessed on 3 February 2020).
43. GitHub. Available online: https://github.com/patrickschur/language-detection (accessed on 3 February 2020).
44. Roussos, P.L.; Tsaousis, G. *Statistics in Behavioural Sciences Using SPSS*; TOPOS: Athens, Greece, 2011.
45. Eurostat, Your Key to European Statistics. Available online: https://ec.europa.eu/eurostat/tgm/refreshTableAction.do?tab=table&plugin=1&pcode=tec00001&language=en (accessed on 15 April 2020).
46. EU Vocabularies: 7206 Europe. Available online: https://op.europa.eu/en/web/eu-vocabularies/th-concept-scheme/-/resource/eurovoc/100277?target=Browse (accessed on 3 February 2020).
47. The Publications Office of the European Union. Available online: https://op.europa.eu/en/web/about-us/who-we-are (accessed on 3 February 2020).