Digital Studies /
Le champ numérique

Bordonaba-Plou, David, and Laila M. Jreis-Navarro. 2025. "Characterizing Asymmetries in the TenTen Corpus Family Membership: An Implicit Hierarchy in Multilingual Digital Tools." *Digital Studies/Le champ numérique* 15(1): 1–18. https://doi.org/10.16995/dscn.16533.

Open Library
of Humanities

Characterizing Asymmetries in the TenTen Corpus Family Membership: An Implicit Hierarchy in Multilingual Digital Tools

David Bordonaba-Plou, Departamento de Lógica y Filosofía Teórica, Universidad Complutense de Madrid, davbordo@ucm.es
Laila M. Jreis-Navarro, Departamento de Lingüística y Literaturas Hispánicas, Universidad de Zaragoza, ljreis@unizar.es

In this work, we examine the limitations of digital tools in facilitating cross-linguistic and cross-cultural research from a humanistic perspective. Our primary objective is to draw comparisons between the TenTen corpora, assessing their degree of similarity. In order to achieve this goal, we will conduct cluster analysis on the 43 corpora within the TenTen Corpus Family using a set of parameters that characterize this family membership. This analysis pinpoints the TenTen corpora that exhibit the most similar characteristics within the family, bringing to the surface an implicit hierarchy within the Sketch Engine platform, a multilingual digital tool environment. This hierarchy is structured into four distinct clusters, definable by size, number of functional tools, versions, and Part-of-Speech (PoS) tagging. The findings of the current study call for prudence when comparing the TenTen corpora, but also suggest a way of improving a multilingual environment; the examination and establishment of connections among the TenTen corpora are imperative for a comprehensive understanding of multilingualism in Digital Humanities.

Dans ce travail, nous examinons les limites des outils numériques dans la facilitation de la recherche interlinguistique et interculturelle d'un point de vue humaniste. Notre objectif principal est de comparer les corpus TenTen, en évaluant leur degré de similarité. Pour atteindre cet objectif, nous réaliserons une analyse de regroupement sur les 43 corpus de la famille des corpus TenTen en utilisant un ensemble de paramètres caractérisant cette appartenance familiale. Cette analyse identifie les corpus TenTen qui présentent les caractéristiques les plus similaires au sein de la famille, révélant une hiérarchie implicite au sein de la plateforme Sketch Engine, un environnement d'outils numériques multilingues. Cette hiérarchie est structurée en quatre groupes distincts, définis par la taille, le nombre d'outils fonctionnels, les versions et le marquage des parties du discours (PoS). Les résultats de l'étude actuelle appellent à la prudence lors de la comparaison des corpus TenTen, mais suggèrent également un moyen d'améliorer un environnement multilingue ; l'examen et l'établissement de connexions entre les corpus TenTen sont impératifs pour une compréhension complète du multilinguisme dans les Humanités Numériques.

1. Introduction

There is an increasingly meaningful connection between Digital Humanities (DH) and Multilingual DH (M-DH), putting language at the centre of its very definition. For a long time, the availability of high-quality resources and tools built for the English language has favoured specific practices in DH that take little account of cultural and linguistic diversity. However, as Nilsson-Fernàndez and Dombrowski defend, "to understand DH (or any discipline in the humanities by extension—be it literature, history, philosophy, etc.), it is essential to look beyond the scope of any single language" (Nilsson-Fernàndez and Dombrowski 2022, 83). Although one of DH desiderata is to break accessibility barriers (open-source programs, the FAIR principles) (GO FAIR 2016), it has been argued for several years that DH has not substantially improved accessibility for languages other than English. This, among other reasons, is why some authors have defended that M-DH should be core to DH (Fiormonte 2012).

M-DH is critical with the centrality of the English language in the scholarly practices, resources, and tools originated in the field, and, therefore, multilingualism has a strong link to the cultures and languages of the so-called Global South, but it is also central to Global North policies, as it is stated in the European Commission webpage, "The co-existence of many languages in Europe is a powerful symbol of the European Union's (EU) aspiration to be united in diversity, one of the cornerstones of the European project" (European Commission 2023). Several possible solutions have been proposed to overcome, or at least alleviate, this problem. For example, Spence and Brandao highlight the need for linguistic labelling in DH infrastructures (Spence and Brandao 2021), following Nilsson-Fernàndez and Dombrowski's (Nilsson-Fernàndez and Dombrowski 2022, 90) suggestion of applying the so-called "Bender Rule" (Bender 2019, 18) to DH. This rule states that all studies using a given language should always explicitly state which language is being used, even if this language is English, since failure to do so may create the impression that the study is language-independent. Therefore, we should distinguish between one-language-DH (be it English-DH, Māori-DH, Slovak-DH, and so on), and M-DH.

Multilingualism in DH is an issue that has many facets and that is concerning scholars in the field (Viola and Spence 2024). For example, several works denounce the prevalent Anglocentrism present in DH and the negative repercussions and injustices that this generates in the academic context for individuals who are not native English speakers (Fiormonte 2012; Galina 2013; Galina 2014; Mahony 2018). Another of these facets is inextricably related to cross-linguistic studies. Developing strong M-DH practices is essential to enhance cross-linguistic and cross-cultural studies. Freake, Gentil, and Sheyholislami argue that adopting a cross-linguistic perspective can enable us to discover similarities and differences in the linguistic phenomenon we are addressing (Freake, Gentil, and Sheyholislami 2011). In a similar vein, Raffaelli, Katunar, and Kerovec (Raffaelli, Katunar, and Kerovec 2019) contend that only by adopting this perspective will we be in a position to differentiate universal linguistic patterns from those that are language-specific. Besides, it helps to reframe concepts developed in humanistic disciplines before the availability of multilingual tools (Bordonaba-Plou and Jreis-Navarro 2023).

The study of languages based on digitized corpora, known as Corpus Linguistics, is also part of DH. Several authors (see, for example, Hockey 2004; Hugues, Constantopoulos, and Dallas 2016; Le Deuff 2018) have related the emergence of contemporary DH to the success of the web and the increase in size and accessibility of corpora. The use of digital corpora is not exclusive to linguistics, but it is also relevant in disciplines such as corpus-assisted discourse studies (Freake, Gentil, and Sheyholislami 2011; Taylor 2013; Nardone 2018) as well as in humanistic disciplines where the use of corpora, although not traditionally employed, is becoming more frequent, for example, experimental philosophy of language (Sytsma et al. 2019; Bordonaba-Plou 2023) or corpus philology (Faulkner 2023; Jreis-Navarro 2024).

To address the need for multilingual, readily searchable corpora, Sketch Engine (Kilgarriff et al. 2014) and the TenTen Corpus Family (Jakubícek et al. 2013; Sketch Engine 2025a) have provided the academic community with software equipped with various tools and resources. They offer tools such as word sketches and concordancers alongside a collection of web corpora in diverse languages, filling a crucial gap in linguistic research resources. However, although there has been a great deal of investment and effort in developing multilingual materials and tools, much work remains. In a previous study, we argued that there is a sense of linguistic injustice directly related to the development and implementation of multilingual tools, referred to as the “paradox of Anglocentric multilingualism” (Bordonaba-Plou and Jreis-Navarro 2024, 130). This paradox states that studies conducted in English enjoy distinct advantages over those in other languages. These advantages stem from higher quantity and diversity of results and greater precision in the digital tools performance, enhancing the researchers’ ability to substantiate or refute research hypotheses. We showed the differences in the output generated by various searches in several languages using Sketch Engine tools and different corpora of the TenTen Corpus Family. This poses the following question: is the TenTen corpora really a corpus family?

In this work, we will examine some of the shortcomings of digital tools in supporting cross-linguistic and, thus, cross-cultural research from a humanistic perspective. More specifically, the work aims to establish comparisons between the TenTen corpora to determine their level of similarity. In doing so, it will be possible to identify those corpora of the TenTen Corpus Family that share the most characteristics.

The remainder of the paper is structured in five sections. The first section describes the TenTen Corpus Family. The second one explains the criteria by which they can be identified as a family, and, then, it expounds the parameterization of the corpora and details the procedure used to conduct the comparison, cluster analysis. The third section describes the similarity results between the TenTen corpora. The fourth one discusses their implications for cross-linguistic studies, leading to some conclusions related to M-DH practices in the fifth section.

2. The TenTen Corpus Family: An overview

The TenTen Corpus Family includes corpora with three distinguishable characteristics. Firstly, all these corpora share a common basis: they are web-based collections of diverse texts gathered using web crawling techniques outlined in the works of Sharoff (Sharoff 2006) and Baroni and colleagues (Baroni et al. 2009). Secondly, they are similar in size, as indicated by the term "TenTen," denoting a target corpus size of over 10 billion words per language. Thirdly, the available analytical tools are consistent across the family; these tools are seven: Word Sketch, Thesaurus, Keywords, Wordlist, N-grams, Concordance, and Text type analysis. However, as we will argue, some of these characteristics represent desiderata rather than actual characteristics. Starting with size, of all the corpora that make up the family, there is a significant disparity in size when we compare them (see **Figure 1**). In other words, just a quick glance and it is evident that in the TenTen corpora there are "high resource and low resource languages" (Bender 2019), or, rather, low resource languages, medium resource languages, and one high resource language, English.

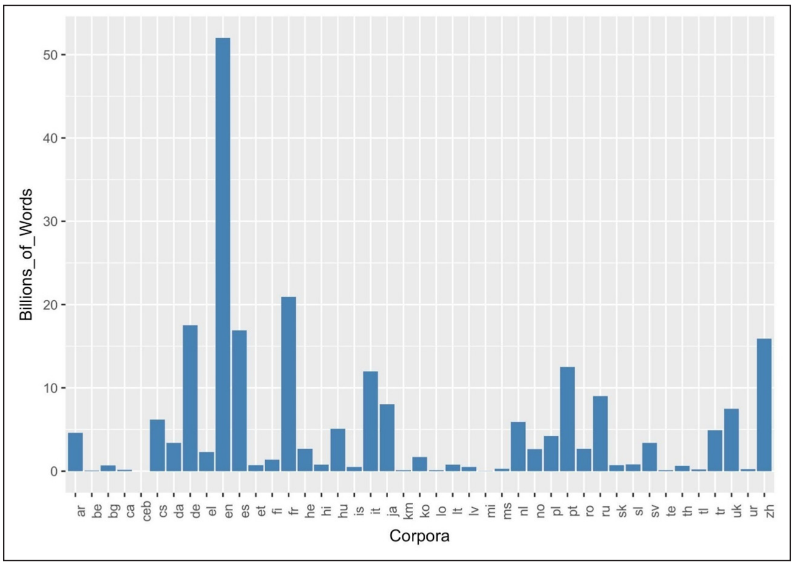

Figure 1: Sizes of the TenTen Corpus Family members.

Clearly, of all the corpora, only some reach the minimum of the 10-billion-word size that defines the TenTen corpora. Also, the seven tools provided for analysis by Sketch Engine are not functional in all of them. Even more important than all of the above is that the performance of the tools varies significantly from one corpus to another, which is related to the Part-of-Speech (PoS) tagger available for each. These asymmetries affect any comparative study to be carried out using this platform and multilingual resources, which, to date, is the best available.

Sketch Engine tools can process collocations and word combinations of a searched word in a corpus (Word Sketch), generate lists of words belonging to the same category (Thesaurus), identify the words that are specific to one corpus (Keywords), generate frequency lists of various kinds (Wordlist), identify the most frequently used multiword expressions (N-grams), find examples of use in context (Concordances), and show statistical results of metadata analysis (Text type analysis). However, as can be seen in the section "Tools for Text Analysis" (Sketch Engine 2025c) on the Sketch Engine website, the smooth performance of the tools is exemplified with English language corpora, and they do not work as smoothly with other languages, such as Arabic or Spanish (Bordonaba-Plou and Jreis-Navarro 2023; Bordonaba-Plou and Jreis-Navarro 2024).

Nevertheless, this study will not address any particular language. Instead, it will take all the TenTen corpora and study their similarities as members of this family in terms of general and quantifiable criteria. At the time of developing the present study (on October 14, 2023), the TenTen Corpus Family has 43 members of varying characteristics, with English at the top. Each member corpus may have one or more versions. Generally, the number of versions has to do with their seniority in the family and with the degree of development of the corpus; thus, enTenTen (English) has 7 versions, itTenTen (Italian) has 3, while isTenTen (Icelandic) has 1. Corpora with 3 or more versions tend to be larger, have PoS Tagging, and, therefore, 6 or more functional tools. Most of the TenTen members have a language-specific tagset or one common to a group of members, as is the tagset for Indian languages.

Sketch Engine has an explicit language support statement on its website. In this statement, it is said that "the features available for each language and sometimes even for each corpus differ" (Sketch Engine 2025b). The platform distinguishes between preloaded corpora, which is the case for the TenTen Family, and user corpora, which are created by users uploading their own data. The list of languages provides detailed information on the level of support each language receives and the collection of preloaded corpora that correspond to it. For example, in the case of Arabic (see Bordonaba-Plou and Jreis-Navarro 2025), along with the arTenTen (now processed by CAMeL tools), there are many other corpora with different purposes and PoS tagging, such as the Arabic Learner corpus (ACL; tagged with Stanford CoreNLP) or the KSUCCA (Classical Arabic; tagged with MADA tools). Therefore, specific information about the features available for each member of the TenTen Family is provided on the detail page of each corpus.

3. Methodology: Similarity criteria and cluster analysis

Our goal, as stated above, is to test whether there are degrees of kinship within the TenTen family. The first step in carrying out the comparison between the 43 family corpora was to devise a way to characterize them. We have characterized each member of the TenTen Corpus Family following specific criteria. We have used four different parameters: size, tools available for each corpus, number of versions, and the availability of a specific tagset (see Appendix). Each parameter is defined as follows:

1. Size: the size of the last corpus version in billion words.
2. Tools: the availability (Y) or not (N) of the seven functional tools.
    a. Abbreviations for tools: Word Sketch (WS), Thesaurus (Th), Keywords (K), N-grams (N-g), Concordance (C), and Text type analysis (T).
    b. Total number of tools (No.).
3. Versions (V): number of versions of each TenTen corpora.
4. Specific tagset (S-t): this characteristic accounts for the Part-of-Speech (PoS) tagging and takes into account three different cases:
    a. The member has a language-specific tagset, for example, arTenTen (Arabic).
    b. The member has a shared dataset and an adapted tagset, for example, MULTEXT-East, which is a multilingual dataset shared by the following languages: roTenTen (Romanian), ruTenTen (Russian), slTenTen (Slovenian), and ukTenTen (Ukrainian). Here is also included ptTenTen (Portuguese), which uses the French FreeLing PoS tagset.
    c. The member has no PoS tagging, for example, miTenTen (Māori), or it is not specified in the corresponding Sketch Engine corpus webpage, for example, beTenTen (Belarusian).

Then, in order to make it possible to compare the different corpora, we codified the values of the different parameters in numerical values. Other authors apply a similar method; for example, Kern and colleagues comment that in "cases of dictionaries with discrete categories ('negative,' 'positive,' 'neutral'), labels were replaced with numerical values to allow quantitative analyses on the dictionaries" (Kern et al. 2021, 6).

The first parameter is the size of each corpus in billions of words. The second parameter is the number of accessible tools in each corpus (values between 3 and 7). The third parameter is the number of versions (values between 1 and 7). The fourth and last parameter is the type of tagset used by the corpus, acquiring value 1 when the corpus is an instance of Case A, value 2 when it is an instance of Case B, and value 3 when it is an instance of Case C. At the end, each corpus is described numerically by means of the following four-tuple:

<size, no. of tools, no. of versions, tagset>

For example, the arTenTen is depicted employing the following four-tuple:

<4.7, 7, 3, 1>

Then, we used cluster analysis to compare the different four-tuples (i.e., to measure the similarity between the different TenTen corpora). Putting it simply, cluster analysis "is the art of finding groups in data" (Kaufman and Rousseeuw 2005, 1). More precisely, cluster analysis is a set of different statistical procedures whose objective is to organize items in groups based on their similarity, minimizing the differences between the members of the group or cluster, and maximizing the differences between the groups or clusters.

Cluster analysis includes two main methods: partitioning and hierarchical methods. *Partitioning methods* construct *k* clusters, where *k* is determined by the user. Among these methods, the most used ones are the centre-based methods, for example, *k*-means (Rokach 2024, 30–31) and *k*-medoids (Kaufman and Rousseeuw 2005, 40–41; Rokach 2024, 47–51), which are based on the choice of *k* representative objects (or centres) in the data set and the assignment of each of the remaining objects to these centres, creating in this way the clusters.

As for *hierarchical methods*, they build the clusters progressively based on the dissimilarity between the elements of the data set. There are two main hierarchical methods, the *agglomerative* (Kaufman and Rosseeuw 2005, 44; Rokach 2024, 89–90) and the *divisive* (Kaufman and Rosseeuw 2005, 44; Rokach 2024, 103–104). The former starts with each element of the data set as its own cluster, and then at each step amalgamates or joins two similar elements into a cluster until only one element remains. However, the latter works just the other way around. It begins with a single cluster containing all the elements of the data set, and in each following step a cluster is split, until there are *n* of them.

Cluster analysis is a methodology used in a multitude of different disciplines and areas of knowledge: astronomy, social sciences, geography, medicine, or history (see Kaufman and Rosseeuw 2005, 1–2), to name just a few. In the DH, clustering methods have been central to the field of stylometry (Eder 2017; Neal et al. 2017), although over time they have begun to be used to evaluate other DH resources. For example, Antonenko, Toy, and Niederhauser (Antonenko, Toy, and Niederhauser 2012) employ cluster analysis to investigate students' use of online learning environments. Kern and colleagues use cluster analysis to compare German polarity dictionaries for sentiment analysis (Kern et al. 2021).

In this work, we have used Agglomerative Clustering, one of the most common types of hierarchical cluster analysis (Baayen 2008, 138–48; Desagulier 2017, 276–82). Since not all the variables in the four-tuples are expressed in the same measurement units, we need to standardize the data because "by standardizing one attempts to give all variables an equal weight, in the hope of achieving objectivity" (Kaufman and Rousseeuw 2005, 11). To do this, we have used the *scale* function. We have also used the R function *agnes* (DataCamp 2025; see Rokach 2024, 101–103 for a detailed explanation) with *Euclidean distance* as the metric for calculating dissimilarities, and *Ward* as the clustering method because "it has the advantage of generating clusters of moderate size" (Desagulier 2017, 279). The result is depicted by a tree-based representation or *dendrogram.*

4. Results: Cluster dendrogram

As can be seen through the hierarchy (see **Figure 2**), the parameters divide the TenTen Corpus Family into four different clusters. Six of its members belong to a cluster, the purple cluster, that is clearly separated from the others, followed by a uni-member cluster, the blue cluster, and two other clusters, the red and green clusters, each including 16 members.

The members of the purple cluster (see **Table 1**) have no PoS tagging, are smaller than 1 billion words, and have only 1 version. The small size does not seem to be a particularly defining characteristic, as etTenTen (Estonian) belongs to the red cluster and is almost the same size as ltTenTen (Lithuanian), the biggest member of the purple cluster. The absence of PoS tagging that hinders the functionality of many tools, especially Word Sketch (WS) and Thesaurus (Th), is most significant in this separated cluster. The preliminary state of these corpora also coincides with the number of versions (one) and, therefore, their "lifetime" in the platform.

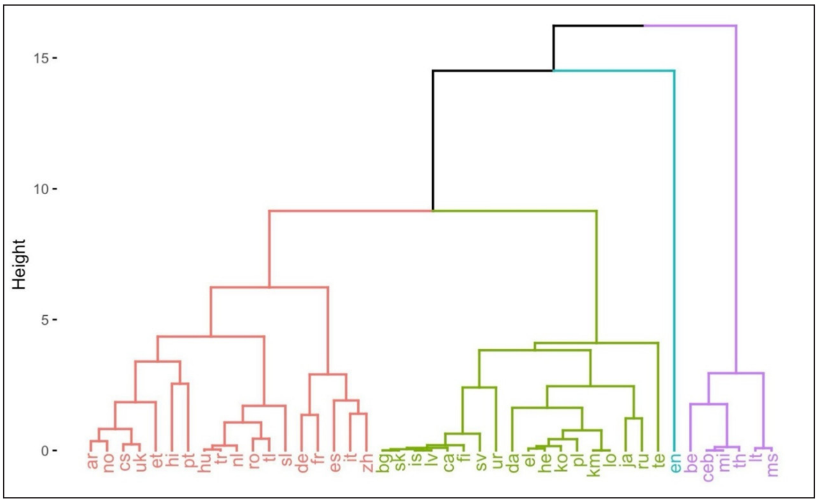

Figure 2: Cluster dendrogram of the TenTen Corpus Family.

Member	Size	Tools								V	S-t
		WS	Th	K	W	N-g	C	T	No.		
beTenTen	0,063	N	N	N	Y	Y	Y	N	3	1	3
cebTenTen	0,0045	N	N	Y	Y	Y	Y	N	4	1	3
ltTenTen	0,778	N	Y	Y	Y	Y	Y	N	5	1	3
miTenTen	0,011	N	N	Y	Y	Y	Y	N	4	1	3
msTenTen	0,296	N	N	Y	Y	Y	Y	Y	5	1	3
thTenTen	0,64	N	N	Y	Y	Y	Y	N	4	1	3

Table 1: Purple cluster characteristics. (Table 1 is a portion of the information contained in Table A1 [see Appendix], where we have characterized all the TenTen Family corpora.)

As might have been expected, the enTenTen (English) stands alone in the blue cluster, with its huge size (52 billion words), its 7 functional tools, its 7 versions, and its specific tagset.

The members of the red cluster vary in size. Some of them have less than 1 billion words, for example, the tlTenTen (Tagalog), with 198 million words, while others have more than 10 billion words, for example, the esTenTen (Spanish), with almost 17 billion words, or the frTenTen (French), with nearly 21 billion words. They also vary significantly concerning the number of versions, as we find corpora with only one version, such as the slTenTen (Slovenian), and corpora with four versions, such as the deTenTen (German). For the other two characteristics, the number of tools and the tagset, there is more uniformity. They usually have specific tagsets, although there are some cases, such as the hiTenTen (Hindi) and the ptTenTen (Portuguese), where both have an adapted tagset. Finally, most of the corpora belonging to this cluster have all seven tools, the only exceptions being the hiTenTen (Hindi), the itTenTen (Italian), and the zhTenTen (Chinese).

The members of the green cluster are also of varying sizes, from the smallest, the KhTenTen (Khmer), with 103 million words, to the biggest, the ruTenTen (Russian), with 9 billion words. However, neither reaches 10 billion words; let us remember, this is one of the defining characteristics of the family and the one that gives it its name. Most have 6 tools, except the teTenTen (Telugu), with 4 tools. The tool missing for all of them is the Text type analysis (T), which usually provides genre and topic classification. The members have 1 or 2 versions, except for the daTenTen (Danish), which has 3 versions. Finally, all of them have a specific tagset, except for the urTenTen (Urdu), which uses a "Unified Parts of Speech (POS) Standard in Indian Languages."

5. Discussion: Asymmetries

These asymmetries leave us with many questions: could any purple cluster members be compared with any other cluster members, and even between them? To what degree is the blue/English cluster comparable to others? Are the red and green cluster members the most comparable? According to Sketch Engine, "All TenTen corpora are prepared according to the same criteria and can be regarded as comparable corpora" (Sketch Engine 2025a). The comparability of the TenTen Family members is then defined as follows: first, by opposing them to parallel corpora (i.e., corpora consisting of the same texts translated into different languages); and second, the texts "belong to the same domain with the same metadata," for example, the same Wikipedia article in different languages (Sketch Engine 2025d).

As seen in the second section of this study, there are three identifying characteristics of the TenTen Corpus Family: compilation mode, size, and tools. After the analysis, we have been able to verify that the only common characteristic is that the texts are gathered using web crawling techniques. Our first insight is that the comparability between the TenTen corpora will depend on whether the corpora are in the same cluster. It can be stated that the members of the red and green clusters seem the most comparable (similar number of functional tools and PoS tagging) as long as the comparison does not take into consideration genre and topic classification, which is the most essential feature of the seventh tool (T: Text type analysis); this asymmetry is not a fundamental one. However, the corpora comprising the purple cluster and having no

PoS tagging could not be comparable because they are not suitable for cross-linguistic studies. That is, it will be possible to carry out the studies, but it will be likely that we will encounter practical problems.

Is English incomparable to any other corpus since it is its own cluster? We think the answer is no because it shares some characteristics with the red and green cluster corpora, such as the number of tools or specific tagsets. However, the fact that it has been defined as a cluster in itself is not a trivial matter because being the most developed corpus of the whole family implies advantages over other languages. We show that testing the performance of tools is important because their functionality may be more theoretical than practical (Bordonaba-Plou and Jreis-Navarro 2023; Bordonaba-Plou and Jreis-Navarro 2024; Bordonaba-Plou and Jreis-Navarro 2025). For example, when comparing two languages, the same tool can produce erroneous collocates and keywords in one language but not in the other. Even more importantly, the results of the statistical significance markers, such as the MI-score (Hunston 2002, 71; Baker 2006, 101) or the Log Dice (Gablasova, Brezina, and McEnery 2017), can be of different reliability. All this will only cast doubt on the coherence and consistency of the research.

6. Conclusions

In short, comparability between corpora has to do with the particular language and the specific research purposes involved in each study, and the hierarchical structures that have emerged in our analysis can lead the way. If, as the "Bender Rule" and its DH advocates highlight, digital tools need to be specific about their language dependency, multilingual digital tools and resources need to state their language hierarchy, if any; implicit asymmetries enhance linguistic injustice and cultural inequalities. When carrying out cross-linguistic studies comparing the TenTen corpora in Sketch Engine, researchers can make use of an explicit language hierarchy, not just to justify handicaps in their research results, but also to support their right to publish their research results as a way of improving a multilingual environment. Once the asymmetries are acknowledged and a preliminary path (characterization and clustering) has been traced, the problem has been signalled.

The criteria posed in the present study configure this preliminary path, and more parameters should be included through collaborative work (e.g., examining, in practice, the tools' performance comparing different TenTen Family members). Other interesting issues to be addressed: Does Word Sketch produce functional collocation lists in every TenTen corpus? What percentage of these collocates are accurate? And so on. Examining and establishing connections among the TenTen corpora is crucial for comprehending the concept of multilingualism. Within the TenTen Corpus Family, the multilingual aspect is primarily integrated through many different language corpora gathered with web crawling techniques and subsequently processed with language technologies. Despite these advancements, there is a lack of genuine dialogue. Therefore, the multilingual family remains a promise.

Advocating multilingualism in digital resources and tools is not about translating specific tools from one language into other languages, nor it is about applying universalistic and language-agnostic approaches. Translation in this field is often synonymous with speed but not quality results, where decisions are based on economic aspects (limited budgets, market and academic competitions, etc.), taking advantage of existing methodologies to "get results quickly"; see, for example, this statement on The Penn Arabic treebank:

> [W]e considered both using a traditional Arabic grammar style and using the Penn Treebank style. [...] As speed was important to the project, we chose to take advantage of methodologies already in place for treebanks of other languages at Penn. [...] [W]e were able to take advantage of the existing understanding of how to manipulate treebank structures and get results quickly. (Maamouri et al. 2004)

As for universalistic approaches, although they tend to offer standardization in contrastive analysis, they do not seem to provide better results, at least in linguistic annotation (see Bordonaba-Plou and Jreis-Navarro 2025). The role of DH in this field of research is to fill in the blanks left by other disciplines to provide the humanistic approach. The focus of DH in multilingual technology should not be dictated by huge amounts of raw data and fast results, the increase of which will further widen the gaps between linguistic and cultural clusters, but by finding points of confluence. Multilingual-DH need to determine how linguistic justice should be provided in the digital arena, enhancing the existing tools with supervised machine learning and annotation standards that are the result of effective cross-linguistic and intercultural dialogue.

Appendix

Member	Size	Tools								V	S-t
		WS	Th	K	W	N-g	C	T	No.		
arTenTen	4,6	Y	Y	Y	Y	Y	Y	Y	7	3	1
beTenTen	0,063	N	N	N	Y	Y	Y	N	3	1	3
bgTenTen	0,705	Y	Y	Y	Y	Y	Y	N	6	1	1
caTenTen	0,182	Y	Y	Y	Y	Y	Y	N	6	1	1
cebTenTen	0,0045	N	N	Y	Y	Y	Y	N	4	1	3
csTenTen	6,2	Y	Y	Y	Y	Y	Y	Y	7	3	1
daTenTen	3,4	Y	Y	Y	Y	Y	Y	N	6	3	1
deTenTen	17,5	Y	Y	Y	Y	Y	Y	Y	7	4	1
elTenTen	2,3	Y	Y	Y	Y	Y	Y	N	6	2	1
enTenTen	52	Y	Y	Y	Y	Y	Y	Y	7	7	1
esTenTen	16,9	Y	Y	Y	Y	Y	Y	Y	7	2	1
etTenTen	0,725	Y	Y	Y	Y	Y	Y	Y	7	4	1
fiTenTen	1,4	Y	Y	Y	Y	Y	Y	N	6	1	1
frTenTen	20,9	Y	Y	Y	Y	Y	Y	Y	7	3	1
heTenTen	2,7	Y	Y	Y	Y	Y	Y	N	6	2	1
hiTenTen	0,79	Y	Y	Y	Y	Y	Y	N	6	3	2
huTenTen	5,1	Y	Y	Y	Y	Y	Y	Y	7	2	1
itTenTen	12	Y	Y	Y	Y	Y	Y	N	6	3	1
isTenTen	0,518	Y	Y	Y	Y	Y	Y	N	6	1	1
jaTenTen	8	Y	Y	Y	Y	Y	Y	N	6	1	1
kmTenTen	0,103	Y	Y	Y	Y	Y	Y	N	6	2	1
koTenTen	1,7	Y	Y	Y	Y	Y	Y	N	6	2	1
loTenTen	0,105	Y	Y	Y	Y	Y	Y	N	6	2	1
ltTenTen	0,778	N	Y	Y	Y	Y	Y	N	5	1	3
lvTenTen	0,53	Y	Y	Y	Y	Y	Y	N	6	1	1
miTenTen	0,011	N	N	Y	Y	Y	Y	N	4	1	3
msTenTen	0,296	N	N	Y	Y	Y	Y	Y	5	1	3

(Contd.)

Member	Size	Tools								V	S-t
		WS	Th	K	W	N-g	C	T	No.		
nlTenTen	5,9	Y	Y	Y	Y	Y	Y	Y	7	2	1
noTenTen	2,63	Y	Y	Y	Y	Y	Y	Y	7	3	1
plTenTen	4,2	Y	Y	Y	Y	Y	Y	N	6	2	1
ptTenTen	12,5	Y	Y	Y	Y	Y	Y	Y	7	3	2
roTenTen	2,7	Y	Y	Y	Y	Y	Y	Y	7	2	1
ruTenTen	9	Y	Y	Y	Y	Y	Y	N	6	2	1
skTenTen	0,715	Y	Y	Y	Y	Y	Y	N	6	1	1
slTenTen	0,829	Y	Y	Y	Y	Y	Y	Y	7	1	1
svTenTen	3,4	Y	Y	Y	Y	Y	Y	N	6	1	1
teTenTen	0,126	N	N	Y	Y	Y	Y	N	4	1	1
thTenTen	0,64	N	N	Y	Y	Y	Y	N	4	1	3
tlTenTen	0,198	Y	Y	Y	Y	Y	Y	Y	7	2	1
trTenTen	4,9	Y	Y	Y	Y	Y	Y	Y	7	2	1
ukTenTen	7,5	Y	Y	Y	Y	Y	Y	Y	7	3	1
urTenTen	0,245	Y	Y	Y	Y	Y	Y	N	6	1	2
zhTenTen	15,9	Y	Y	Y	Y	Y	Y	N	6	2	1

Table A1: Characteristics of the 43 corpora of the TenTen Family.

Acknowledgements

An earlier version of this work was presented at the *Global Digital Humanities Symposium 2024* (April 18–19), organized by Michigan State University. We thank the symposium audience and two anonymous referees for their fruitful comments.

Funding

This publication is the result of the project *Intuiciones y Filosofía Experimental del Lenguaje* (IFEL), PID2023-150396OA-I00, founded by MCIU/AEI/10.13039/501100011033/FEDER, UE. This work has also been carried out within the framework of the “Clarisel” Research Group, with the financial support of the Departamento de Ciencia, Tecnología y Universidad of the Gobierno de Aragón, and the European Social Fund. This publication is the result of the project PID2021-122872NB-C21, *Transformaciones del espacio magrebí en perspectiva histórica* (TRAMAGHIS), which, together with the project PID2021-122872NB-C22, *Tránsitos y migraciones en el norte de África: análisis diacrónico de la población y su entorno* (DIANA), is integrated in the coordinated research project *Tránsitos y transformaciones en el espacio y la población magrebíes* (MAGNA II). Both have been funded by MICIN/AEI/10.13039/501100011033 and FEDER *Una manera de hacer Europa*.

Competing interests

The authors have no competing interests to declare.

Contributions

Authorial

Authorship in the byline is by alphabetical order. Author contributions, described using the NISO (National Information Standards Organization) CrediT taxonomy, are as follows:

Author name and initials:

David Bordonaba-Plou (DB)
Laila M. Jreis-Navarro (LJ)

Authors are listed in descending order by significance of contribution. The corresponding author is DB.

Conceptualization: DB, LJ
Methodology: DB, LJ
Formal Analysis: DB, LJ
Investigation: DB, LJ
Resources: LJ
Writing – Original Draft Preparation: DB, LJ
Writing – Review & Editing: DB, LJ
Visualization: DB
Supervision: DB, LJ

Editorial

Section and Translation Editor

Davide Pafumi, The Journal Incubator, University of Lethbridge, Canada

Production Editor

Christa Avram, The Journal Incubator, University of Lethbridge, Canada

Copy and Layout Editor

A K M Iftekhar Khalid, The Journal Incubator, University of Lethbridge, Canada

References

Antonenko, Pavlo D., Serkan Toy, and Dale S. Niederhauser. 2012. “Using Cluster Analysis for Data Mining in Educational Technology Research.” *Educational Technology Research and Development* 60: 383–398. Accessed January 26, 2025. https://doi.org/10.1007/s11423-012-9235-8.

Baayen, R. Harald. 2008. *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge University Press.

Baker, Paul. 2006. *Using Corpora in Discourse Analysis*. Continuum.

Baroni, Marco, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. “The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora.” *Language Resources and Evaluation* 43: 209–226. Accessed January 26, 2025. https://doi.org/10.1007/s10579-009-9081-4.

Bender, Emily M. 2019. “The #BenderRule: On Naming the Languages We Study and Why It Matters.” *The Gradient*, September 14. Accessed January 26, 2025. https://thegradient.pub/the-benderrule-on-naming-the-languages-we-study-and-why-it-matters/.

Bordonaba-Plou, David, ed. 2023. *Experimental Philosophy of Language: Perspectives, Methods and Prospects*. Springer.

Bordonaba-Plou, David, and Laila M. Jreis-Navarro. 2023. “Light in Assessing Color Quality: An Arabic-Spanish Cross-Linguistic Study.” In *Experimental Philosophy of Language: Perspectives, Methods and Prospects*, edited by David Bordonaba-Plou, 151–170. Springer.

———. 2024. “Linguistic Injustice in Multilingual Technologies: The TenTen Corpus Family as a Case Study.” In *Multilingual Digital Humanities*, edited by Lorella Viola and Paul Spence, 129–144. Routledge.

———. 2025. “Are the TenTen Corpora Really a Corpus Family? On Linguistic Tagging and Corpora Members’ Kinship Degrees.” *International Journal of Humanities and Arts Computing* 19 (1): 49–64. Accessed March 3, 2025. https://doi.org/10.3366/ijhac.2025.0344.

DataCamp. 2025. “agnes: Agglomerative Nesting (Hierarchical Clustering).” Rdocumentation. Cluster version 2.1.4. Accessed February 21. https://www.rdocumentation.org/packages/cluster/versions/2.1.4/topics/agnes.

Desagulier, Guillaume. 2017. *Corpus Linguistics and Statistics with R Introduction to Quantitative Methods in Linguistics*. Springer.

Eder, Maciej. 2017. “Visualization in Stylometry: Cluster Analysis Using Networks.” *Digital Scholarship in the Humanities* 32 (1): 50–64. Accessed January 26, 2025. https://doi.org/10.1093/llc/fqv061.

European Commission. 2023. “About Multilingualism Policy-European Education Area.” 2023. Education.ec.europa.eu. February 24. Accessed January 26, 2025. https://education.ec.europa.eu/focus-topics/improving-quality/multilingualism/about-multilingualism-policy.

Faulkner, Mark. 2023. “Corpus Philology: Using the Dictionary of Old English to Get Bigger Data for Old English Spelling Variation.” *Digital Scholarship in the Humanities* 38 (4): 1508–1521. Accessed January 26, 2025. https://doi.org/10.1093/llc/fqad064.

Fiormonte, Domenico. 2012. “Towards a Cultural Critique of the Digital Humanities.” *Historical Social Research* 37 (3): 59–76. Accessed January 27, 2025. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-378525.

Freake, Rachelle, Guillaume Gentil, and Jaffer Sheyholislami. 2011. “A Bilingual Corpus- Assisted Discourse Study of the Construction of Nationhood and Belonging in Quebec.” *Discourse & Society* 22 (1): 21–47. Accessed January 27, 2025. https://doi.org/10.1177/0957926510382842.

Gablasova, Dana, Vaclav Brezina, and Tony McEnery. 2017. “Collocations in Corpus-Based Language Learning Research: Identifying, Comparing, and Interpreting the Evidence.” *Language Learning* 67 (S1): 155–179. Accessed January 27, 2025. https://doi.org/10.1111/lang.12225.

Galina, Isabel. 2013. “Is There Anybody Out There? Building a Global Digital Humanities Community.” *Red Humanidades Digitales*, July 19. Accessed January 27, 2025. http://humanidadesdigitales.net/blog/2013/07/19/is-there-anybody-out-there-building-a-global-digital-humanities-community/.

———. 2014. “Geographical and Linguistic Diversity in the Digital Humanities.” *Literary and Linguistic Computing* 29 (3): 307–316. Accessed January 27, 2025. https://doi.org/10.1093/llc/fqu005.

GO FAIR. 2016. “FAIR Principles.” Go-fair.org. Accessed January 26, 2025. https://www.go-fair.org/fair-principles/.

Hockey, Susan. 2004. “The History of Humanities Computing.” In *A Companion to Digital Humanities*, edited by Susan Schreibman, Ray Siemens, and John Unsworth, 1–19. Blackwell.

Hugues, Lorna, Panos Constantopoulos, and Costis Dallas. 2016. “Digital Methods in the Humanities: Understanding and Describing Their Use Across the Disciplines.” In *A New Companion to Digital Humanities*, edited by Susan Schreibman, Ray Siemens, and John Unsworth, 150–170. Blackwell.

Hunston, Susan. 2002. *Corpora in Applied Linguistics*. Cambridge University Press.

Jakubíček, Miloš, Adam Kilgarriff, Vojtěch Kovář, Pavel Rychlý, and Vít Suchomel. 2013. “The TenTen Corpus Family.” In *Proceedings of the 7th International Corpus Linguistics Conference, CL 2013*, 125–127. Accessed February 21, 2025. https://www.sketchengine.eu/wp-content/uploads/The_TenTen_Corpus_2013.pdf.

Jreis-Navarro, Laila M. 2024. “The Use of *Nafs* ‘Soul’ for Self-Referencing in Al-Maqqarī’s *Nafḥ al-ṭīb* and the Evolution of the ‘Divided-Self.’” *Journal of Semitic Studies* 69 (2): 777–802. Accessed January 26, 2025. https://doi.org/10.1093/jss/fgad044.

Kaufman, Leonard, and Peter J. Rousseeuw. 2005. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley.

Kern, Bettina M. J., Andreas Baumann, Thomas E. Kolb, Katharina Sekanina, Klaus Hofmann, Tanja Wissik, and Julia Neidhardt. 2021. “A Review and Cluster Analysis of German Polarity Resources for Sentiment Analysis.” In *3rd Conference on Language, Data and Knowledge (LDK 2021)*, edited by Dagmar Gromann, Gilles Sérasset, Thierry Declerck, John P. McCrae, Jorge Gracia, Julia Bosque-Gil, Fernando Bobillo, and Barbara Heinisch, 1–17. OASICS. Accessed January 27, 2025. https://doi.org/10.4230/OASIcs.LDK.2021.37.

Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. “The Sketch Engine.” *Lexicography* 1 (1): 7–36. Accessed January 27, 2025. https://doi.org/10.1007/s40607-014-0009-9.

Le Deuff, Olivier. 2018. *Digital Humanities: History and Development*. Blackwell.

Maamouri, Mohamed, Ann Bies, Tim Buckwalter, and Wigdan Mekki. 2004. “The Penn Arabic Treebank: Building a Large-Scale Annotated Arabic Corpus.” In *NEMLAR Conference on Arabic Language Resources and Tools* 27, 466–467. Accessed March 3, 2025. https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/nemlar2004-penn-arabic-treebank.pdf.

Mahony, Simon. 2018. “Cultural Diversity and the Digital Humanities.” *Fudan Journal of the Humanities and Social Sciences* 11: 371–388. Accessed January 27, 2025. https://doi.org/10.1007/s40647-018-0216-0.

Nardone, Chiara. 2018. “‘Women and Work’: A Cross-Linguistic Corpus-Assisted Discourse Study in German and in Italian.” *Critical Approaches to Discourse Analysis across Disciplines* 10 (1): 167–186. Accessed February 21, 2025. https://www.lancaster.ac.uk/fass/journals/cadaad/wp-content/uploads/2018/06/10-Nardone.pdf.

Neal, Tempestt, Kalaivani Sundararajan, Aneez Fatima, Yiming Yan, Yingfei Xiang, and Damon Woodard. 2017. “Surveying Stylometry Techniques and Applications.” *ACM Computing Surveys* 50 (6): 1–36. Accessed March 3, 2025. https://doi.org/10.1145/3132039.

Nilsson-Fernàndez, Pedro, and Quinn Dombrowski. 2022. “Multilingual Digital Humanities.” In *The Bloomsbury Handbook to the Digital Humanities*, edited by James O’Sullivan, 83–92. Bloomsbury.

Raffaelli, Ida, Daniela Katunar, and Barbara Kerovec. 2019. “Introduction.” In *Lexicalization Patterns in Color Naming: A Cross-Linguistic Perspective*, edited by Ida Raffaelli, Daniela Katunar, and Barbara Kerovec, 1–19. John Benjamins.

Rokach, Lior. 2024. *Cluster Analysis: A Primer Using R*. World Scientific.

Sharoff, Serge. 2006. “Creating General-Purpose Corpora Using Automated Search Engine Queries.” In *WaCky! Working Papers on the Web as Corpus*, edited by Marco Baroni and Silvia Bernardini, 63–98. Gedit.

Sketch Engine. 2025a. “TenTen Corpus Family.” Sketchengine.eu. Accessed March 10, 2025. https://www.sketchengine.eu/documentation/tenten-corpora/.

———. 2025b. “Corpora by Language.” Sketchengine.eu. Accessed March 10, 2025. https://www.sketchengine.eu/corpora-and-languages/.

———. 2025c. “Tools for Text Analysis.” Sketchengine.eu. Accessed January 26, 2025. https://www.sketchengine.eu/tools-for-text-analysis/.

———. 2025d. “Glossary.” Sketchengine.eu. Accessed January 26, 2025. https://www.sketchengine.eu/guide/glossary/.

Spence, Paul J., and Renata Brandao. 2021. “Towards Language Sensitivity and Diversity in the Digital Humanities.” *Digital Studies/le champ numérique* 11 (1). Accessed January 27, 2025. https://doi.org/10.16995/dscn.8098.

Sytsma, Justin, Roland Bluhm, Pascale Willemsen, and Kevin Reuter. 2019. “Causation Attributions and Corpus Analysis.” In *Methodological Advances in Experimental Philosophy*, edited by Eugen Fischer and Mark Curtis, 209–238. Bloomsbury.

Taylor, Charlotte. 2013. “Searching for Similarity Using Corpus-Assisted Discourse Studies.” *Corpora* 8 (1): 81–113. Accessed January 27, 2025. https://doi.org/10.3366/cor.2013.0035.

Viola, Lorella, and Paul J. Spence, eds. 2024. *Multilingual Digital Humanities*. Routledge.