A Stylometric Glance at Novels in Euskara

Dominika Werońska

Jagiellonian University, Doctoral School in the Humanities, Pałac Spiski, Rynek Główny 34, 31-010
Kraków, Poland
https://orcid.org/0000-0002-1053-9056
dominika.weronska@doctoral.uj.edu.pl

While Basque has been posited as possibly the oldest existing language on the European continent, it appears in written form only in the sixteenth century. The first Basque novel emerges over 300 years later and to this day the genre lacks exhaustive research. The article sets as its aim a stylometric analysis of selected twentieth- and twenty-first-century Basque novels, sourced from the online platforms Armiarma and Booktegi. These are analyzed based on the frequency of the most frequent words measured using cluster analysis and set against a backdrop of foreign novels translated into Euskara. The results show that the originals in Euskara remain distinct from translated works, pointing to the unique linguistic character of the Basque novel. Some linguistic patterns potentially responsible for this distinction are presented. The results are visualized on a map revealing the chronological evolution and the contribution of the Basque novel to the broader literary landscape.

Keywords: Basque literature / Euskara / text corpora / cluster analysis / stylometric map

Introduction

The Basque novel, though over 300 years old, remains relatively under-researched in terms of stylometry, a gap which this article aims to bridge through a stylometric analysis of selected Basque novels. The goal is to identify distinctive stylistic patterns through the application of various stylometric methods such as cluster analysis and the construction of bootstrap consensus trees. Prior to presenting the analysis, the linguistic and literary context of the Basque novel are outlined.

Before there was literature there was language: A few words on Basque

The Basque language, also known as Euskara, boasts a rich history spanning over two millennia (Jansen 7). Where Basque originated and how it came to exist on the European continent remains unknown. Despite numerous attempts to establish linguistic connections between Basque and other language families, including Celtic, African, Caucasian, and indigenous American languages, no definitive evidence has as yet emerged to support any hypothesis (8–9).

Attempts were made as early as the seventeenth century, when poet and historian Arnaut Oihenart suggested a connection between Basque and Iberian in his 1656 historiographical work *Notitia utriusque Vasconiae*, *tum Ibericae*, *tum Aquitanicae* (*A Survey of the Two Basconias, the Iberian and the Aquitaine*). While the theory was disproven, it was widely embraced in the nineteenth century by, for example, Spanish Jesuit and personal confessor of Queen Maria Anna of Neuburg, Manuel de Larramendi; according to his study *De la antiguedad, y universalidad del bascuenze en España* (*On the Antiquity and Universality of Basque in Spain*), Basque is a linguistic isolate and descendant of Iberian: "Now it is easily concluded that the Basque language was a universal language; […] that being the language of the first settlers, they would have to extend it along with the settlements they were founding, and […] would speak the same language, until foreigners came, who began to introduce their different languages." (Larramendi 43; my translation)

In his essay from 1818, Danish linguist Rasmus Rask noted a similarity between Basque and Greenlandic; he claimed that "verbs, in particular, have an extremely complex inflection, viz. eleven moods and six tenses in each of the first six moods, all of which seems rather to resemble Greenlandic than any language of the Gothic class" (Rask 90). Yet upon further deliberation, Robert L. Trask dismissed any notable similarity between the two languages, pointing instead to Basque's plausible North African origins in either Mauretania or Gaetulia (Trask, "Origins" 91).

In the twentieth century, research focused on finding a potential historical link between Basque and the Celtic languages (Trask, "Origins"; Trask, *History*) or Basque and Aquitanian (see Michelena), as well as on establishing a Proto-language which would have been a common ancestor for Proto-Basque and Proto-Indo-European (Tovar 1970). While Koldo Mitxelena (also known as Luis Michelena) could be credited with elucidating the most probable theory, the remaining studies were acknowledged to lack conclusive evidence.

At the same time, it must be admitted that Basque does have analogues in many languages across the globe. For example, like Hungarian, it is agglutinative and relies heavily on suffixation (Trask, *History* 201). Like Inuit, Mam, and Jacaltec, it is an ergative absolutive language with the subject of a transitive verb appearing in the ergative case (marked with the suffix *-k*) and the subject of an intransitive verb showing up in the absolutive case (unmarked). Finally, like Japanese and many Indo-Iranian languages, Basque follows subject-object-verb sentence order (89, 109).

However, establishing a conclusive linguistic relationship necessitates non-linguistic evidence and a sharing of more than just one or two coincidental features. Progress in this regard remains limited and the prospect of establishing a definitive genealogical link remote (Jansen 9).

Dialects of Basque

Spoken today by no more than 806,000 inhabitants of the Basque Country, Basque speakers are further divided according to various dialects. Most scholars assert that variants of the Basque language were spoken as early as the second millennium BC, and point to the mountainous terrain and to the low prestige of the language as the two main reasons for dialectal fragmentation (Trask, *History* 5). Mitxelena offers a different perspective, arguing that this linguistic fragmentation commenced in the Middle Ages, post the decline of the Roman Empire (Michelena 300).

While earlier classifications in the nineteenth and early twentieth centuries partitioned the Basque-speaking area into six to nine major dialects, including the now-obsolete Roncalese (Jansen 8–9), more recent investigations indicate that over the last five decades, the dialects have converged, with no more than five identifiable varieties (Zuazo 22). These comprise the central dialects of Bizcayan (spoken in the area of Bilbao), Gipuzkoan (spoken in Donostia, Tolosa, Bergara, and Zarautz), and Upper Navarrese (spoken in the high areas of Navarre), as well as the two peripheral dialects of Navarro-Lapurdian and Zuberoan (spoken in the French region). Furthermore, in the 1960s, the Royal Academy of the Basque Language succeeded in cultivating a standardized literary Basque, known as Euskera Batua, to confront pressures from state-supported French and Spanish entities. This standardized form of the language provided writers with an orthographic and morphological framework amenable for employment irrespective of their native dialects and strengthened Basque's resilience against external pressures. However, this also meant that individual varieties of Basque received less space to flourish as means of cultural expression, also in the domain of literature. The corpus selected for the present analysis comprises of works written from 1897 to 2022. Of the 32 works published before 1960, the majority were written in Guipuzkoan and Bizcayan.

In contrast, the majority of the works published from the 1970s to the twenty-first century predominantly employ Euskara Batua.

Basque literature as a small literature

Basque literature can be described as a small literature. The adjective *small* refers to both quantitative and qualitative aspects. Quantitatively, it encompasses measurable factors of the literary system such as the number of authors, yearly publications, sales, potential readers, and publishing houses. Qualitatively, the term reflects not only the size of a literary corpus but also its influence on literary production and reception, its targeted audience, internal and external perceptions of its scale, its political autonomy or lack thereof, and so on.

With about 300 authors, a readership of around 806,000 Basque speakers, and an annual publication range of 1,500 works, Basque literature can well be classified as a small literature in quantitative terms (Kortazar 11–12). For a comparison with English and Slovenian literature, refer to Table 1.

Literature	Authors	Publications (per year)	Male/Female Writers	Speakers (potential readership)	Publishing houses
Basque	≈ 300	≈ 1,500	82% / 18%	806,000	226
Slovenian	269	≈ 6,000	42% / 58%	2,100,000	1400
U.S.	54,010	4,000,000	49.5% / 50.5%	400,000,000 (natives) 1,500,000,000 non-native)	2840

Table 1: Statistics behind the literary market (a comparison of Basque, Slovenian, and American literature).

Basque literature also meets the qualitative conditions for a small literature. In terms of typology, the Basque literary canon belongs to the group of literatures in peripheral languages without an external reference such as Estonian, Latvian, or Welsh. As is the case with most literatures produced at the boundaries of a major language in a small language, Basque literature finds itself in a dialectic between ideology and literary autonomy (Kortazar 12). On the one hand, it is expected to commit to the national idea and reinforce the identity of its people; on the other hand, it seeks to be cosmopolitan and express the personality of its individual authors. The tension it encounters is integral to its survival, as language is closely tied with identity, and often results in a collective appropriation of the discourse with a preference for philologized language, as the use of language itself holds intrinsic literary value.

It is interesting to note that a similar observation was made by the Slovenian poet France Prešeren in a letter to his friend Matija Čop dated 5 July 1832, in which the poet remarked that "the tendency of our carmina and similar literary activities is no other than to cultivate our mother tongue" (qtd. in Juvan 185). The expectations put upon literature to be subservient to language are undoubtedly a characteristic trait of small literatures and a cause for tension, amplified by having to resist potential attacks of opposing forces on both political and cultural grounds. This is certainly the case with the Basque people who have, in the words of Mark Kurlansky, "stubbornly fought for their unique concept of a nation without ever having a country of their own" (Kurlansky 5).

The oppression faced by the Basques certainly took its toll on the development of their literature, and of their novel in particular. While the first known words in Basque were most probably engraved on a bronze relic already 2,100 years ago, and the first work of Basque literature, the collection of poems *Linguae Vasconum Primitiae* (*The First Fruits of the Basque Language*) by Bernard Etxepare, was printed in 1545, almost 300 years had to elapse before the first Basque novel appeared, and then it was only a proto-novel in dialogue form (Benzine). The work most scholars recognized as the first Basque novel was *Auñemendiko lorea* (*The Flower of Auñemendi*) by Domingo Agirre. It appeared only in 1897 and began a series of historical novels in the Romantic style soon overtaken by *costumbrista* prose. However, according to Jesús María Lasagabaster the novel never really had the chance to develop:

> It was always my opinion that the Basque novel, which was born so late, never had—and could never have had—the same kind of historic development Basque poetry had and was forced to jump instead, without logical progression or continuity, from these anachronistic novels of manners of blatant idealistic and romantic roots to experimentation and the avant-garde, without going first through realism, the current that—in European literatures at least—marked the coming of age and the establishing of the novel as a genre. (Lasagabaster 16)

Yet Lasagabaster claims that the lost ground has been recovered and that recovery is visible in the number of prizes awarded to Basque authors in recent years (16). While this may be the case, he is forced to admit that literature in Basque is "'contaminated' (in the etymological sense of the word)" (18) with linguistic and sociocultural influences current to France and Spain, the political places where it evolves.

A stylometric map of Basque literature

The idea of presenting a corpus of literature on a map is far from innovative. Critics of literature have often resorted to using elements of cartography to visualize texts in space, to outline the development of genres, or to illustrate the stylistic similarity between works (Brooks; Moretti; Bulson).

One early example is Cleanth Brooks's mapping of the setting and geography of William Faulkner's novels. In *William Faulkner: The Yoknapatawpha Country*, this American literary scholar explored the Mississippi writer's fictional county and the important role it played in so much of his work.

In *Novels, Maps, Modernity: The Spatial Imagination, 1850–2000*, Eric Bulson examined the depiction of place in nineteenth-century works of such writers as Charles Dickens and Herman Melville all the way through twentieth-century fiction. He likened the experience of reading a novel to examining a map, balancing between the immediate experience and a broader structural understanding.

Literary mapping has also been explored and popularized by Franco Moretti. In *Graphs, Maps, Trees: Abstract Models for a Literary History*, Moretti argues that these materialist concepts of form can often reveal more about literature than the texts themselves. An advocate of distant reading, that is, studying world literature devoid of direct textual analysis where one relies solely on studies done by other researchers, Moretti posits that seeing less is actually seeing more. He claims that maps possess "emerging" qualities which allow us to see aspects not visible at the lower level (Moretti 61).

In line with Moretti's distant reading, scholars within the discipline of stylometry have begun to produce graphical representations of literature in order to explore the relationships between individual works as well as entire literary canons (Rybicki, "Pierwszy rzut"; Rybicki, "Second Glance").

Stylometry, as the term suggests, measures authorial style using the statistical analysis of distinct features such as n-grams, grammatical categories, and most frequent words (MFWs). Methods relying on the frequency of the MFWs have proven particularly successful at resolving authorship issues, delineating the chronology of a set of works, or even distinguishing between different translators.

The following study applies stylometric methods such as cluster analysis and the construction of bootstrap consensus trees to a corpus of novels originally in Basque as well as to novels translated from various languages into the Basque language. The objective is to construct a literary map for these works and to explore: (1) the efficacy of stylometry in authorship attribution for a so-called small literature within a highly inflected, agglutinative language; (2) the potential visibility of chronological evolution of the Basque novel and the extent to which the purported lack of continuity and progression in the genre's development is reflected on the stylometric map; (3) the distinction between novels originally written in Euskara and those translated from other languages; (4) the capability of stylometry to identify features of translationese; and (5) the impact of culling on the results of stylometric analysis and the nature of this impact.

Method

Perhaps the most influential stylometric method is Burrows's Delta. To quote Jan Rybicki, this "frankeinsteinish" approach consists in chopping up the pieces of analyzed literature into individual words, selecting from among them the least significant ones (MFWs) which constitute a kind of connective tissue of the works, and calculating their frequency for each text (Rybicki, "Second Glance" 7). The frequencies are then ordered in a frequency table and compared using some type of distance measure. Evert et al. have shown the Cosine Delta variant to be the most successful in terms of authorship attribution, hence its use in the present study. Thanks to the Cosine Delta measure, a distance matrix for each pair of texts is created. The smaller the distance between each text, the greater the similarity between the texts analyzed. For greater reliability, the calculations can be repeated for different numbers of the MFWs.

The results obtained are processed with consensus network analysis and visualized using the open-source software Gephi (Bastian et al.), which makes it possible to create a map of the given texts.

In the following study, seven types of stylometric tests relying on Cosine Delta were conducted on three different corpora. Below is a more detailed description of each test with the results obtained.

The first three tests were a series of cluster analyses performed on a corpus of works originally in Basque. Tests IV and V consisted in using the function Oppose to detect translationese in works by an individual author, Nikolas Ormaetxea (a.k.a. Orixe). Tests VI and VII were conducted using cluster analysis and bootstrap consensus trees, respectively, on a mixed corpus of originally Basque novels and novels translated into Basque from other languages. Below is a more detailed description of each test with the results obtained.

Test I

The first test was a cluster analysis performed in R using the stylo packet.

Corpus

Works selected for the cluster analysis were downloaded from Armiarma (37 works) and Booktegi (20), two online platforms which provide access to PDF versions of literary works in Basque, spanning from the eighteenth to the twenty-first century. Selection criteria included a minimum length of 25 pages (50 KB) and the inclusion of another work by the same author in the study. Consequently, 57 texts by 20 authors met these criteria.

The corpus ranged from Azkue's 1897 novel *Batxi Guzur* to Ortega's novels from 2022. Prominently featured authors, with at least four novels each, comprised Domingo Agirre, Jean Barbier, Txema Arinas, and Pedro Urruzuno. In contrast, authors such as Tomas Agirre, Itxaro Borde, and Manuel Etxeita contributed only two texts each.

The works, initially in PDF format, were converted to text files for analysis in R.

Method

The first cluster analysis was performed using Cosine Delta on 100 MFWs. Culling was not used, meaning that all words were considered for the analysis.

Results

Figure 1 presents a dendrogram illustrating the results of the cluster analysis. The root of the tree is on the right and branches extend to the left, culminating in the individual works. The *x*-axis represents the distance or dissimilarity between clusters, with larger distances implying less similarity. Encircled works signify instances of misattribution. As can be seen from the dendrogram, 48 of the 57 works are grouped correctly by author, yielding an authorship attribution accuracy of 84%.

Notably, this is below the 94% accuracy achieved by Eder in his analysis of 66 English novels (Eder 54). Misattributions include works by Azkue, Agirre, Borde, Orixe, Erkiaga, and Etxeita. Considering the dendrogram's sensitivity to the number of variables (MFWs), a bootstrap consensus tree was applied in Test 2 to enhance reliability.

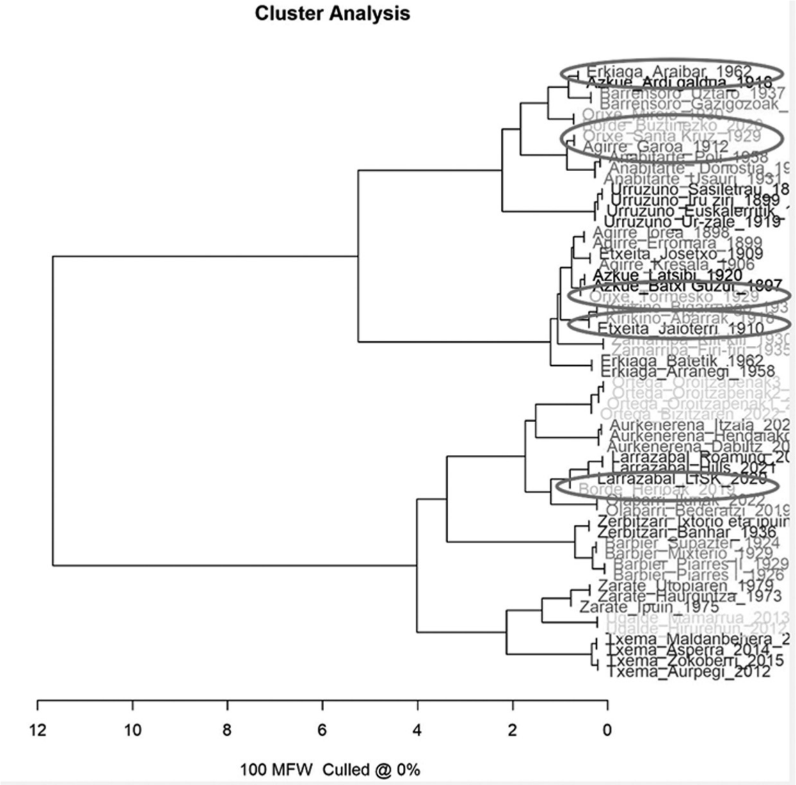

Figure 1: Results of Test I (cluster analysis performed on 57 novels originally in Basque).

Test II

The second test was a series of cluster analyses performed with the bootstrap method in R using the Stylo packet.

The bootstrap consensus tree is a revised, more reliable approach than cluster analysis (Eder 62). It is a statistical technique which involves repeatedly resampling the dataset to improve reliability by filtering out local disturbances. Links indicate consensus strength derived from multiple snapshot dendrograms instead of stylometric distances between individual texts. The corpus remained the same as for Test 1.

Method

Multiple cluster analyses were conducted on the corpus using a range of 100 to 1,000 MFWs. Culling was not employed. Cosine Delta measure was used for each cluster analysis. From the multiple bootstrap samples, a consensus tree was constructed (see Figure 2).

Results

The tree in Figure 2 represents the most stable and recurring relationships found across the various bootstrap samples. The tree indicates that authorship attribution was largely successful, with 50 out of 57 works attributed correctly (almost 88%). Seven works were misattributed, as identified by the circles in the tree.

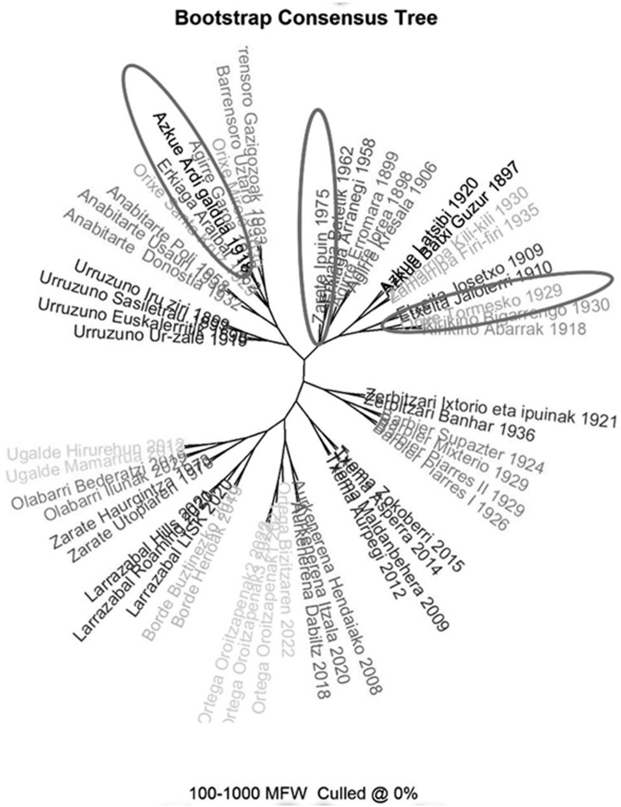

Figure 2: Results of Test II (a bootstrap consensus tree on 57 Basque novels, without culling).

As can be seen in Figure 2, the number of authors with misattributed works had decreased to Orixe, Agirre, Azkue, and Erkiaga. Additionally, Zarate's *Ipuin antzeko alegi mingotsak* (*Bitter Fables in the Guise of Stories*) clustered with Erkiaga's works. In order to decrease dimensionality with the hope of improving discrimination, culling was employed in Test III.

Test III

The third test was a series of cluster analyses performed using the bootstrap method and culling. Culling allows users to set a threshold for how frequently a feature must appear in the corpus to be considered in the analysis (Eder et al. 111). Features not meeting the specified occurrence threshold across the samples are excluded. For example, at 100% culling rate, only words appearing in at least 100% of the samples are included. Culling has been shown to improve discrimination leading to more accurate clustering. The corpus remained the same as for Tests I and II.

Method

The attribution test was run several times with different vectors of MFWs and with different settings for culling. Cosine Delta was performed using 100 to 2,000 MFWs, with increments of 100. The culling was performed from 0–100% with an increment of 20.

Results

The tree in Figure 3 represents the most stable and recurring relationships found across the various bootstrap samples. With 53 out of 57 works attributed correctly (almost 93%), authorship attribution was the most successful of all three tests.

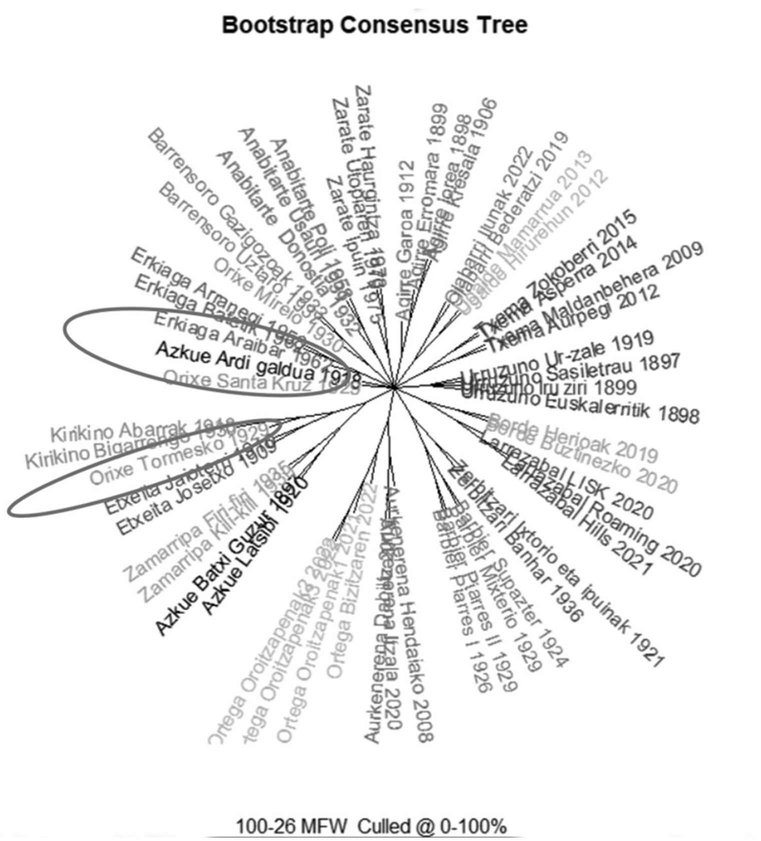

Figure 3: Results of Test III (a bootstrap consensus tree on 57 Basque novels, with culling).

The four works which were not clustered according to their authors comprised Resurreción Maria Azkue's *Ardi galdua* (*The Lost Sheep*), Eusebio Arriaga's *Araibar zalduna* (*Sir Araibar*), as well as Orixe's *Tormes'ko itsu-mutila* (*Lazarillo de Tormes*) and *Mireio*. These four works were consistently misattributed across the three tests. Hence, they warranted further literary investigation relying on traditional literary scholarship.

The first of these works, *Ardi galdua*, was written by Resurreción Maria Azkue, a priest, poet, and academic (Olaziregi 147). As the first head of Euskaltzaindia, Azkue worked on finding a literary standard for the Basque language. He proposed Gipuzkera Osotua, a codified and expanded form of the Gipuzkoan dialect, as the possible contender for euskara batua. *Ardi galdua* was his first work written in Gipuzkera

Osotua. Erkiaga's *Araibar zalduna,* which clusters with *Ardi galdua*, is also written in the Guipuzkoan dialect. While Erkiaga experimented with Navarrese and other dialects, *Araibar zalduna* was his attempt at Guipuzkoan. Incidentally, Agirre's *Garoa* (*Fern*), which in Test II clustered with the two works just mentioned, was also written in Gipuzkera Osotua. Agirre, a Carmelite priest who formed part of Euskaltzaindia, happened to be a friend of Azkue. While he wrote *Kresala* (*Saltpeter*) and *Auñemendiko lorea* in the Basque language of Bizkaia, both about life at sea, *Garoa*, focusing on the agricultural world, was written in Guipuzkoan (Olaziregi 145). It would seem, then, that stylometry was indicating dialectal variations. This could imply its potential utility in dialectometric studies.

The remaining texts for which authorship attribution was unsuccessful were all Orixe's works. Although all three were procured from the section of the Armiarma website containing works originally in Basque, only *Santa Kruz apaiza* (*The Priest Santa Cruz*) was actually originally written in Basque. The remaining two were translations, namely of *Mireio* from Provencal French and of *Lazarillo de Tormes* from Spanish. This aligns with Rybicki's observation that stylometry often groups texts by author rather than by translator (Rybicki, "Stylometric" 203–204). It is possible that Orixe's translatorial signal was too weak for stylometric analysis to recognize him. At the same time, cluster analysis indicated anomalies in authorship, revealing some works as translations. This raises the question of stylometry's potential in detecting translationese.

Tests IV & V

A potentially suitable method for addressing the question of characteristic textual features of translationese is the Oppose function. This function is typically used to identify which features are most characteristic of different authors (Eder et al. 117). In the following two tests, it is employed to differentiate between translated and original texts. Oppose operates by dividing a group of texts into two sets, the primary and the secondary one, and quantifying word frequencies within each. Once it identifies words with significant frequency disparities between sets, Oppose applies statistical measures, such as z-scores or Craig's Zeta, to allocate a distinctiveness score to each term. It then produces a list of words ranked by their scores, indicating which words are characteristic of the primary set (positive scores: words preferred) and which are characteristic of the secondary set (negative scores: words avoided).

Test IV uses the Oppose function to contrast Orixe's translation of *Mireio* from French (primary set) with his work in Basque (secondary set). Test V compares Orixe's translation of *Lazarillo de Tormes* from Spanish (primary set) with his original work (secondary set). The results of these tests are presented in Figures 4 and 5.

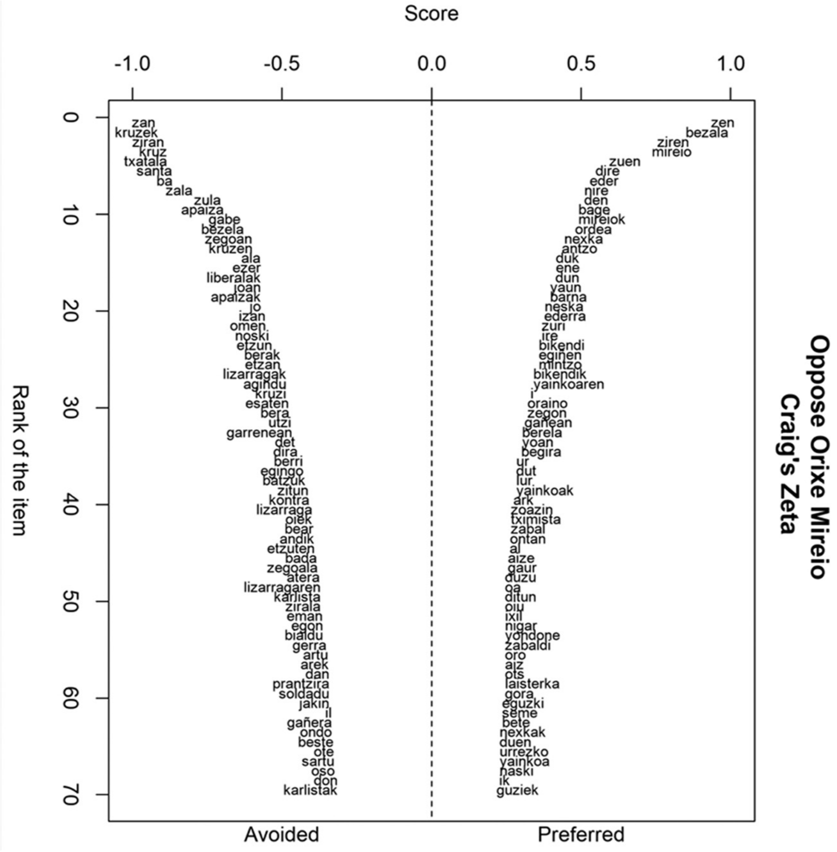

Figure 4: Results of Oppose on *Mireio* (words preferred) and *Santa Kruz apaiza* (avoided).

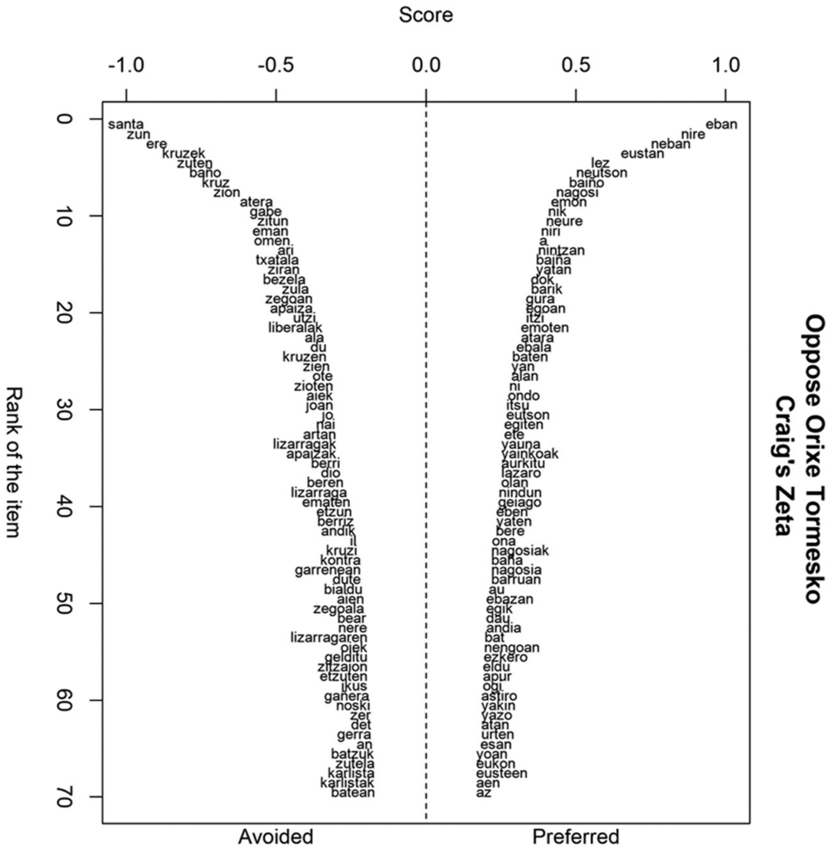

Figure 5: Results of Oppose on *Tormesko itsu-motila* (words preferred) and *Santa Kruz apaiza* (avoided).

The words presented on the right side of each Figure are the ones preferentially used in Orixe's translations. Meanwhile, the words on the left are preferred in *Santa Kruz apaiza.* The farther away we move from the center line, the stronger the association with the corresponding set of texts.

The words presented on the graphs belong to various grammatical categories. The bulk of these are nouns and verbs, but there are also adjectives, adverbs, prepositions, conjunctions, and interjections. Nouns mainly indicate the subject matter of the works analyzed. It is of little surprise that the words "kruz" 'cross,' "apaiza" 'priest,' "santa" 'holy,' "errege" 'king,' "bandera" 'flag,' "meza" 'Mass,' and "gerra" 'war' should abound in *Santa Kruz apaiza*, a novel about a priest and guerrilla fighter of the name Manuel Santa Kruz. In a similar vein, *Moreio*, a Provencal literary text about thwarted love, is only expected to contain words such as "nexka" 'girl,' "nigar" 'tear' (or 'cry'), "ark" 'chest,' "lore" 'flower,' and "izar" 'star,' while a Spanish picaresque novel about a young boy serving various masters in a hypocritical and corrupt society is bound to be filled with nouns such as "gura" 'desire' (or 'will'), "dirua" 'money,' "goseak" 'hunger,' "yazo" 'place,' and "gizagaxua" 'unfortunate man.' Yet even among the nouns, some indications of borrowings can be observed. Place names, for example, give away the language of the original work. In *Mireio*, the proper name "Mireio" is one of the preferred words ("Mireille" in French), as is "provenzako" 'Provencal.' Both hint at the origins and the setting of the story. Similarly, in *Santa Kruz apaiza*, the typical Basque last name Lizarraga appears with greater frequency, as does "euskaldunak" 'Basque speakers.' Finally, in Orixe's translation of *Lazarillo de Tormes* many words appear which seem to be closely connected to Spanish, such as "aiz" ('axe'; 'hacha' in Spanish), "oa" ('wave'; 'ola' in Spanish), "aen" ('breath,' 'air'; 'aliento' in Spanish), "abade" ('abbot'; 'abad' in Spanish), "yazo" ('place'; in Spanish, 'yacer' means to lie in a resting place, especially in the context of a burial). While these could be examples of Spanish influence, it is more likely that some of them entered Basque from Latin during Roman colonization (Trask, *History* 189, 259–61).

However, translationese often manifests itself in function words and syntax, areas prone to unnoticed errors because, as James Pennebaker notes, our brain is not wired to notice those "small, stealthy words" that account for less than 1‰ of our vocabulary while making up almost 60% of the words we use (Pennebaker ix). Morphology and syntax also seem more susceptible to translationese, given that they are more deeply ingrained than lexis and more closely tied to early cognitive development.

For example, in his translation of *Lazarillo de Tormes*, Orixe seems to overuse the past form of the verb *eduki* 'to have,' such as *eban*, *neban*, *ebala*, *ebazan*, and *ebanean*. This may be due to the fact that Spanish has a tendency to express narratives in the past tense. Meanwhile, Basque has multiple ways of expressing the past, and a literal translation from a language like Spanish may lead to an over-reliance on this tense.

Another example are possessive forms such as *nire* 'my,' *neure* 'my own,' and *niri* 'to me.' These tend to be more explicitly stated in Spanish than in Basque, as the latter often conveys possession by using a combination of verb forms and suffixes attached to nouns. Orixe's translation overuses these forms, no doubt because the author allows himself to be ruled by Spanish morphology and syntax.

As we can see, Oppose allows us to direct our attention to places where translationese might occur. While it does not answer the question of what translationese is, it provides a glimpse into the fabric of the language and suggests where the tears might be. The question which could be asked at the end of Tests IV and V is where Orixe's works would be positioned in a corpus comprising original Basque works and translations into Basque: would they align with the translations, assuming these form a distinct cluster in relation to original Basque texts?

Test VI

To examine the positioning of translated works within a mixed corpus, Basque novels were analyzed alongside literature translated into Basque.

Corpus

The literature was procured from Armiarma, spanning from Shakespeare's *Hamlet* to Steinbeck's short story cycle, *The Pastures of Heaven*. 49 works were added into the corpus from English (18), French (12), Spanish (6), Russian (5), German (3), Czech (1), Danish (1), Italian (1), Norwegian (1), and Portuguese (1). Prominent translators included Anton Garikano (who translated two works from English and two from German) and Jose Morales Belda (translator of Russian works).

Method

A series of cluster analyses ranging from 100 to 1,000 MFWs and employing Cosine Delta was visualized as a bootstrap consensus tree (Figure 6). The analysis was done without culling.

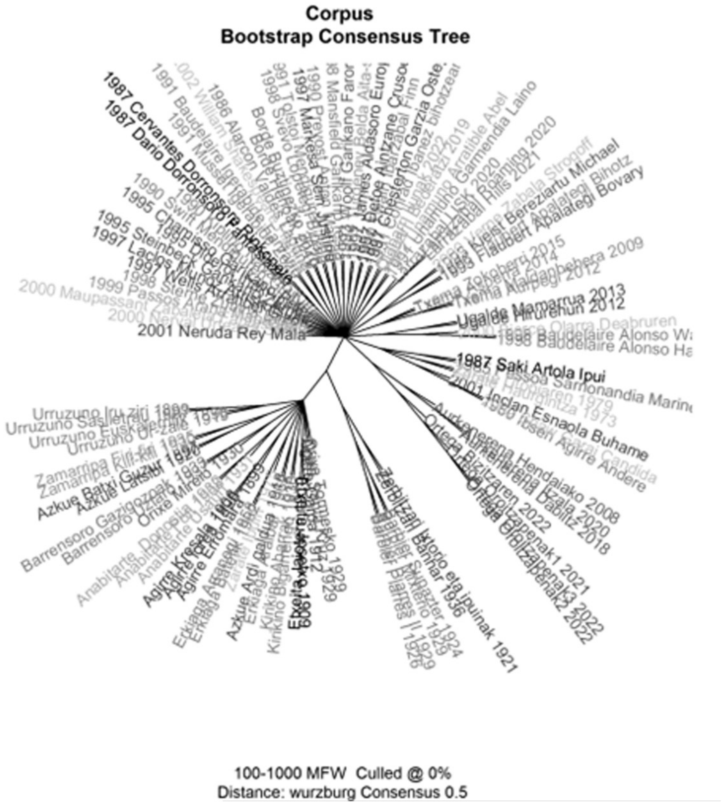

Figure 6: Stylometric analysis of Basque literature and translated literature (no culling).

Results

The bootstrap consensus tree revealed a division between original Basque prose and prose translated into Basque. Original Basque works predominantly clustered by author, showing minimal chronological progression, except for the distinct grouping of twenty-first-century texts. Notably, contemporary works by Borde, Larrazabal, and Olabarri clustered with translated literature. Azkue's *Ardi galdua* aligned with Erkiaga's *Araibal zalduna* among the Basque originals, as did Orixe's translations.

Test VII

Corpus

For the final analysis, a bootstrap consensus tree was conducted using the corpus from Test VI, encompassing original Basque novels and Basque translations from English, Spanish, French, Russian, German, Czech, Danish, Italian, Norwegian, and Portuguese. The attribution analysis was executed multiple times using varying vectors of MFWs and culling.

Method

The Delta measure was applied starting at 100 MFWs and increasing by increments of 100 up to 2,000 MFWs. Culling was also progressively applied starting from 0% to 100% in increments of 20. The data obtained was entered into Gephi and a map of literature in Basque was created using network analysis (see Figure 7).

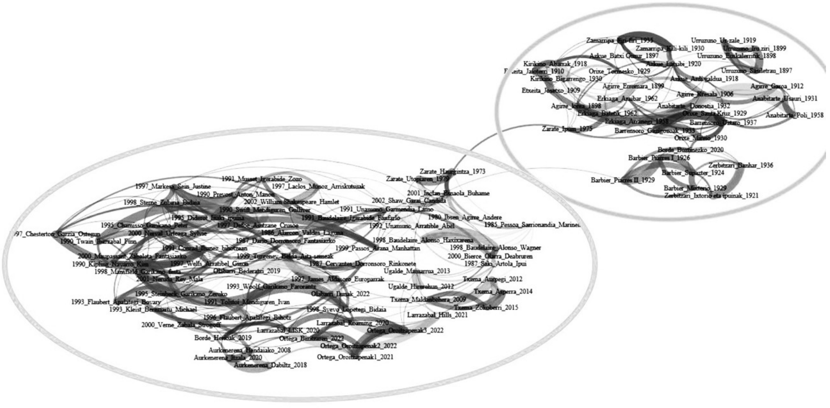
Zamarripa_Firi-firi_1935
Urruzuno_Ur-zale_1919
Zamarripa_Kili-kili_1930
Urruzuno_Iru ziri_1899
Azkue_Batxi Guzur_1897
Urruzuno_Euskalerritik_1898
Kirikino_Abarrak_1918
Azkue_Latsibi_1920
Etxeita_Jaioterri_1910
Orixe_Tormesko_1929
Urruzuno_Sasiletrau_1897
Kirikino_Bigarrengo_1930
Azkue_Ardi galdua_1918
Agirre_Erromara_1899
Agirre_Garoa_1912
Etxeita_Josetxo_1909
Agirre_Kresala_1906
Erkiaga_Araibar_1962
Anabitarte_Usauri_1931
Agirre_lorea_1898
Anabitarte_Donostia_1932
Erkiaga_Batetik_1962
Orixe_Santa Kruz_1929
Anabitarte_Poli_1958
Erkiaga_Arranegi_1958
Barrensoro_Uztaro_1937
Zarate_Ipuin_1975
Barrensoro_Gazigozoak_1933
Orixe_Mireio_1930
Borde_Buztinezko_2020
Barbier_Piarres I_1926
Zerbitzari_Banhar_1936
Barbier_Supazter_1924
Barbier_Piarres II_1929
Barbier_Misterio_1929
Zerbitzari_Ixtorio eta ipuinak_1921
Zarate_Haurgintza_1973
1991_Musset_Igerabide_Zozo
Zarate_Utopiaren_1979
1997_Markesa_Sein_Justine
1997_Laclos_Munoz_Arriskutsuak
1990_Prevost_Anton_Manon
2001_Inclan_Esnaola_Buhame
1998_Sterne_Zubiria_Bidaia
2002_William Shakespeare_Hamlet
2002_Shaw_Garai_Candida
1990_Swift_Mendiguren_Gulliver
1991_Unamuno_Garmendia_Laino
1995_Diderot_Inaki_ipuina
1991_Baudelaire_Igarabide_Fanfarlo
1980_Ibsen_Agirre_Andere
1995_Chamisso_Garikano_Peter
1997_Defoe_Aintzane_Crusoe
1992_Unamuno_Arratible_Abel
1985_Pessoa_Sarrionandia_Marinel
97_Chesterton_Garzia_Ostegun
2000_Nerval_Urteaga_Sylvie
1986_Alarcon_Valdes_Laguna
1990_Twain_Iturzabal_Finn
1987_Dario_Dorronsoro_Fantasiazko
1998_Baudelaire_Alonso_Haxixarena
1991_Conrad_Ibanez_bihotzean
1999_Passos_Arana_Manhattan
1998_Baudelaire_Alonso_Wagner
2000_Maupassant_Zabaleta_Fantasiazko
1999_Turgenev_Belda_Aita-semeak
2000_Bierce_Olarra_Deabruren
1990_Kipling_Navarro_Kim
1997_Wells_Arratibel_Gizon
1987_Cervantes_Dorronsoro_Rinkonete
1987_Saki_Artola_Ipui
1998_Mansfield_Garikano_festa
Olabarri_Bederatzi_2019
Ugalde_Mamarrua_2013
2001_Neruda_Rey_Mala
1997_James_Aldasoro_Europarrak
Txema_Aurpegi_2012
1993_Woolf_Garikano_Farorantz
Ugalde_Hiruerehun_2012
1995_Steinbeck_Garikano_Zeruko
Olabarri_Ilunak_2022
Txema_Asperra_2014
1993_Flaubert_Apalategi_Bovary
1991_Tolstoi_Mendiguren_Ivan
Txema_Maldanbehera_2009
1993_Kleist_Berasiartu_Michael
1998_Svevo_Lopetegi_Bidaia
Txema_Zokoberri_2015
1996_Flaubert_Apalategi_Bihotz
Larrazabal_Hills_2021
2000_Verne_Zabala_Strogoff
Larrazabal_Roaming_2020
Larrazabal_LISK_2020
Ortega_Oroitzapenak3_2022
Borde_Herioak_2019
Ortega_Bizitzaren_2022
Ortega_Oroitzapenak2_2022
Aurkenerena_Hendaiako_2008
Aurkenerena_Itzala_2020
Ortega_Oroitzapenak1_2021
Aurkenerena_Dabiltz_2018

Figure 7: A network analysis of literature in Basque (with culling).

Results

The map obtained via network analysis illustrates the distribution of Basque novels. Two primary clusters are evident: the smaller, rightmost cluster seems to primarily contain original Basque works, while the leftmost cluster comprises mostly translated works. Some newer Basque novels (penned in the twenty-first century) group with the translated works.

A modest chronological pattern can be noted among recent novels, but no clear temporal sequence is evident in the cluster of original Basque works. This may reflect Lasagabaster's view of the Basque novel's developmental discontinuity and lack of logical progression. Yet at the same time, it is important to acknowledge that the Basque corpus may simply be influenced by stylistic elements that transcend simple chronological development. By these I mean the various dialects spoken within the Basque Country. More research is needed to determine that; it is worth noticing, though, that, compared to works predating the standardization of Basque, the recent novels written in unified or standardized Basque show a tendency toward chronological clustering.

Conclusions

Firstly, let me begin by making the obvious, yet nevertheless very important observation: stylometric methods such as cluster analysis effectively identify authorship in Basque literature, despite the language's complex inflectional and agglutinative structure. Although authorship attribution accuracy was initially not as high for the corpus as that for English texts, culling has helped to increase it to almost 93%.

Secondly, culling has shown to be useful for discriminating between texts in different dialects of the same language. By allowing users to exclude features which do not meet a specified occurrence threshold, culling leads to more accurate clustering.

Thirdly, in agreement with Moretti, it cannot be denied that visual representations such as maps, graphs, and trees reveal deeper patterns not discernible through text alone. For example, cluster analysis helped us notice that some of Orixe's works were in fact translations. A boostrap consensus tree not only corroborated these findings but also pointed to strong dialectal influence in the case of the novels *Ardi galdua*, *Araibar zalduna*, and *Garoa*. Furthermore, the dialectical signal was found to be stronger than the authorial one, as some works clustered according to dialect and apart from other works by the same author.

Fourthly, stylometric analysis of Orixe's works has revealed instances of translationese (from French and Spanish). The word lists obtained using the Oppose function give insights into Orixe's language use, showing that the frequency of his words changes when translating as opposed to writing. While it does not provide ready answers, the analysis suggests there may be identifiable patterns which distinguish Orixe's writing from his translations and points to various paths which future work in translation and comparative literature research could take.

Finally, the map of Basque works and translated literature has shown a unique character of pre-twenty-first-century Basque novel, which for the most part forms a separate cluster on the map. Twenty-first-century Basque works have been shown to cluster with translated literature. Euskara batua and its increasing popularity as well as ongoing globalization may be the underlying reasons for the loss of distinction between newer novels penned in Basque and the translations. This may be a consequence of what literary historians have viewed as a developmental discontinuity in the history of the Basque novel, or it may be an effect of dialect impact on the corpus, or perhaps of both.

To conclude, besides providing answers to a series of questions, the research presented in this article points toward a vast field of the unknown and invites further stylometric research of Basque literature. What exactly is the impact of dialect on the linguistic style of Basque authors? To what extent does it influence the stylometric results and to what extent are these results a reflection of the Basque novel's unique history? How strong is the potential linguistic and sociocultural contamination of Basque literature by French and Spanish? To what degree can this influence be evidenced by a stylometric comparison of Basque novels with translations from French and Spanish? How prominent is euskara batua in contemporary Basque literature? And last but not least, what would a map of all Basque works look like? For so-called small literatures, there exists not only the temptation but the very real possibility of mapping the entire terrain. Such a map could shed new light on both traditional literary studies and stylometric research itself.

WORKS CITED

Bastian, Mathieu, et al. "Gephi: An Open Source Software for Exploring and Manipulating Networks." *Proceedings of the International AAAI Conference on Weblogs and Social Media*, vol. 3, no. 1, 2009, pp. 361–362.

Benzine, Vittoria. "The Earliest Words Ever Written in Basque Have Been Found Engraved on a Revelatory 2,100-Year-Old Bronze Relic." *Artnet News*, 15 November 2022, news.artnet.com/art-world/hand-of-irulegi-earliest-basque-script-navarre-spain-2210963. Accessed 9 January 2024.

Brooks, Cleanth. *William Faulkner: The Yoknapatawpha Country*. LSU P, 1989.

Bulson, Eric. *Novels, Maps, Modernity: The Spatial Imagination, 1850–2000*. Routledge, 2017.

Collins, Roger. *The Basques*. Basil Blackwell, 1986.

Eder, Maciej. "Visualization in Stylometry: Cluster Analysis using Networks." *Digital Scholarship in the Humanities*, vol. 32, no. 1, 2017, pp. 50–64, https://academic.oup.com/dsh/article/32/1/50/2957386. Accessed 9 January 2024.

Eder, Maciej, and Jan Rybicki. "Do Birds of a Feather Really Flock Together, or How to Choose Training Samples for Authorship Attribution." *Literary and Linguistic Computing*, vol. 28, no. 2, 2013, pp. 229–236.

Eder, Maciej, et al. "Stylometry with R: A Package for Computational Text Analysis." *The R Journal*, vol. 8, no. 1, 2016, pp. 107–121, https://journal.r-project.org/archive/2016/RJ-2016-007/index.html. Accessed 9 January 2024.

Egurtzegi, Ander. "Metathesis of Aspiration as the Source of Anticipatory Voicelessness in Basque." *Journal of French Language Studies*, vol. 29, no. 2, 2019, pp. 265–279.

Evert, Stefan, et al. "Towards a Better Understanding of Burrows's Delta in Literary Authorship Attribution." *Proceedings of NAACL-HLT Fourth Workshop on Computational Linguistics for Literature*, edited by Anna Feldman et al., Kerrville (TX), The Association for Computational Linguistics, 2015, pp. 79–88, https://aclanthology.org/W15-0709/. Accessed 9 January 2024.

Jansen, Wim. *Beginner's Basque.* Hippocrene Books, 2007.

Juvan, Marko. *Worlding a Peripheral Literature*. Palgrave Macmillan, 2019.

Kurlansky, Mark. *The Basque History of the World.* Penguin, 2001.

Lasagabaster, Jesús María. "Introduction: Basque Literary History." *Basque Literary History*, edited by Mari Jose Olaziregi, translated by Amaia Gabantxo, Center for Basque Studies, University of Nevada, 2012, pp. 13–21.

Larramendi, Manuel de. *De la antiguedad y universalidad del bascuenze en Espaṇa: de sus perfecciones y ventajas sobre otras muchas lenguas*. Vol. 1, Por Eugenio Garcia de Honorato, 1728.

Michelena, Luis. "Lengua común y dialectos vascos. Berrarg." *Obras completas*, vol. 7, edited by Joseba A. Lakarra and Íñigo Ruiz Arzalluz, Universidad del País Vasco, 1981, pp. 291–313.

Moretti, Franco. *Graphs, Maps, Trees: Abstract Models for a Literary History*. Verso, 2005.

Mukarovsky, Hans G. "Outline of a Lexicostatistical Study of Basque and the Mande Languages, with a Note on Fula." *Euskalarien nazioarteko jardunaldiak*, Blbao, Euskaltzaindia, 1981, pp. 199–212, https://www.euskaltzaindia.eus/dok/ikerbilduma/6777.pdf. Accessed 9 January 2024.

Olaziregi, Mari Jose. "Worlds of Fiction: An Introduction to Basque Narrative." *Basque Literary History*, edited by Mari Jose Olaziregi, translated by Amaia Gabantxo, Center for Basque Studies, University of Nevada, 2012, pp. 137–200.

Pennebaker, James W. *The Secret Life of Pronouns*. Bloomsbury, 2011.

Rask, Rasmus. *Investigation of the Origin of the Old Norse or Icelandic Language*. Translated by Niels Ege, John Benjamins, 2013.

Rybicki, Jan. "A Second Glance at the Stylometric Map of Polish Literature." *Forum Poetyki*, no. 10, 2017, pp. 6–21, http://fp.amu.edu.pl/a-second-glance-at-a-stylometric-map-of-polish-literature/#sdfootnote1sym. Accessed 9 January 2024.

Rybicki, Jan. "Pierwszy rzut oka na stylometryczną mapę literatury polskiej." *Teksty drugie*, no. 2, 2014, pp. 106–128.

Rybicki, Jan. "Stylometric Translator Attribution." *The Translator and the Computer*, edited by Tadeusz Piotrowski and Łukasz Grabowski, Wydawnictwo Wyższej Szkoły Filologicznej we Wrocławiu, 2013, pp. 193–204.

Trask, R. L. *The History of Basque*. Routledge, 1997.

Trask, R. L. "Origins and Relatives of the Basque Language: Review of the Evidence." *Towards a History of the Basque Language*, edited by José Ignacio Hualde et al., John Benjamins, 1995, pp. 65–100.

Vennemann, Theo. *Europa Vasconica—Europa Semitica*. Edited by Patrizia Noel Aziz Hanna, de Gruyter, 2003.

Zuazo, Koldo. *Euskalkiak, herriaren lekukoak*. Donostia, Elkar, 2003.

Stilometrični pogled na baskovski roman

Ključne besede: baskovska književnost / baskovščina / besedilni korpusi / klastrska analiza / stilometrija

Čeprav velja baskovščina za verjetno najstarejši jezik na evropski celini, se v pisni obliki pojavi šele v šestnajstem stoletju. Prvi baskovski roman je izšel šele dobrih 300 let pozneje in žanr do danes ni bil izčrpno literarnovedno raziskan. Članek si za cilj zastavi stilometrično analizo izbranih baskovskih romanov 20. in 21. stoletja, pridobljenih s spletnih platform Armiarma in Booktegi. Ti so analizirani na podlagi pogostosti najpogostejših besed s klastrsko analizo (ang. *cluster analysis*), sledi primerjava s tujimi romani, prevedenimi v baskovščino. Rezultati kažejo, da se baskovski izvirniki jasno razlikujejo od prevedenih del, kar dokazuje edinstven jezikovni značaj baskovskega romana. Predstavljenih je nekaj jezikovnih vzorcev, ki bi lahko bili razlog za to razlikovanje. Vizualizacije rezultatov razkrivajo kronološki razvoj baskovskega romana in njegov prispevek k širši literarni krajini.

1.01 Izvirni znanstveni članek / Original scientific article
UDK 821.361.09:004
DOI: https://doi.org/10.3986/pkn.v47.i2.05