Cross-Language Stylometry
Picasso's Writings in Spanish and French

Nanette Rißler-Pipka

Abstract For multilingual corpora—and in this particular case for the Spanish and French writings of Pablo Picasso—we do not have an acceptable method for quantitative literary analyses. This paper discusses the existing possibilities to solve the problem of cross-language stylometry by comparing the results of single-language stylometric methods but does not present a new method. Further on, the hypothesis that Picasso's writings and poetry are characterized by a unique style is tested with a Spanish and French corpus, focusing on both the semantic and syntactic similarities and differences. In comparison to contemporary writers, the difference and distinctiveness can be shown by cross-language analyses.

Keywords Picasso, stylometry, part-of-speech, Romance literature

1. Introduction

The example of one of the most known artists who wrote half of his experimental texts in his native language, Spanish, and half of them in the language of his exile country, French, helps to illustrate one of the most urging problems for digital stylistics in Romance studies: cross-language analysis. If we don't want to use translations (and this is out of question for literary Romance studies), the only way to compare a multilingual text collection is to separate the texts by language and compare them to texts of contemporary writers of the same language. The research in stylometry done in this field so far is testing the functionality of different Deltas with different languages (Jannidis et al. 2015; Rybicki and Eder 2011). This is very important to know before trying stylometry with the same parameters (Deltas) on corpora in different languages. Though, it does not help treating the problem of stylistics for authors writing in two (or more) languages. Juola and Mikros started to think about "cross-linguistic stylometric features" (2016) and analyzed a corpus of tweets by bilingual Twitter users (Spanish and

Rißler-Pipka, Nanette: Cross-Language Stylometry. Picasso's Writings in Spanish and French, in: Hesselbach, Robert, et al. (Eds.): Digital Stylistics in Romance Studies and Beyond, Heidelberg: Heidelberg University Publishing, 2024, pp. 53–80. DOI: https://doi.org/10.17885/heiup.1157.c19366

English). The results, though, cannot easily be adopted on the question of literary style: They suggest "cross-linguistic similarities" (Juola and Mikros 2016) based on features like length of tweet and word-length. One might argue that this is due to the media Twitter that forces users to be short in every language. The similarities in French and Spanish texts by Picasso are probably proven by different style markers which are to be detected initially by close reading.

Since Picasso started to publish part of his writings (mostly his two plays) in the late 1930s and 1940s, readers and critics tried to catalogize or to cluster them into literary history (Leiris, 1966; Éluard, 1934; Sabartés, 2017). His exceptional style was compared to James Joyce, Paul Éluard, Guillaume Apollinaire, Federico García Lorca, Rafael Alberti, Góngora, Mallarmé and other authors known for innovative avant-garde-like style (Heydenreich 1979; Fernández Molina 1988; Béhar 1993; Goddard 2006; Michaël 2008; Rißler-Pipka 2015). Several papers have been published about stylometry on James Joyce (O'Sullivan, Eder, and Rybicki 2018; O'Sullivan 2014; Clement 2013), but fewer about the authors of avant-garde literature in Romance languages (Calvo Tello 2019). While using close reading as a method of stylistics it is possible to compare texts by examples of phrases, motives, etc. If necessary and well-argued this comparison may combine several languages, periods and genres, e.g. to speak about related works like the plays by Alfred Jarry and Picasso's prose poem on Guernica or about the baroque mannerist poems by Góngora and Picasso's twentieth-century avant-garde poetry (concrete examples in Rißler-Pipka 2015, 369–71. (Jarry) and 261–63 (Góngora)). Digital stylistics needs, out of mere technical and algorithmic reasons, a common ground for comparison: same language, same period, same genre. On the one hand that seems to be a disadvantage: if you want to compare the ideas Picasso took from Mallarmé to write his Spanish texts, the intertextual links are hidden in the difference of languages. Certainly, some influences are measurable (like the abandonment of punctuation which Picasso indeed borrowed from Mallarmé) and some are only traceable in close reading, because they represent the re-use of broader ideas (like the surrealist combination of disparate elements). On the other hand, stylometry has the advantage to free itself from the ever cited canon and to compare far more texts in a quantitative way (even if we know that selection process is still to be considered). Previous work (Rißler-Pipka 2019a; 2019b) showed rough stylistic resemblance between the French writings of Picasso and Raymond Roussel (*Nouvelles impressions d'Afrique* [1901]) and Ramón Gómez de la Serna for the Spanish writings. Roussel was named by Michaël (2012, 166–67) as possible model but both authors would not necessarily have been on the list of candidates for influences before the quantitative analysis (shown in Rißler-Pipka 2019a; 2019b). However, a direct comparison to the two-language corpus of Picasso's writings alone is still missing and will be discussed in this paper.

2. Hypothesis

Picasso wrote over 200 Spanish and over 300 French prose poems. His very particular style of writing is hard to describe and even harder to catalog as part of literary history. The mere label of *prose poem* can be doubted, but we are missing the accurately fitting description. I do not completely agree with the concept provided by Enrique Mallen (Mallen 2009; 2012; Mallen and Meneses 2019, for a discussion on that point see Rißler-Pipka 2015, 40–42) that Picasso repeats his cubist painting in his writing. Nevertheless, comparing Picasso's writing with the technique of montage (also cited as "rhizomatique" by Michaël 2012, 163) is very tempting. Similar to geometrical forms in cubism, Picasso uses words or grammatical elements in poetry to build new images out of the very same things. By close reading we come to the hypothesis that in both languages, Spanish and French, we find the same technique, pattern, and system. If we follow up this hypothesis, we have to compare the prose poems not only regarding their content and semantics, but also regarding the syntax, word distribution, and other stylistic markers. At this point cross-language stylometry comes up.

Still, using stylometry as a method means to compare quantitatively a set of texts based on style markers (features) like word frequencies, etc. and it is not possible to compare texts of different languages in the same experiment. All efforts done in this direction end up with translations which necessarily influence the results and make them less robust (Heydel and Rybicki 2012). There are more premises than language for authorship contribution with stylometry (e. g. genre, period, etc.), but even experimenting with genre- and/or period-mixes while checking the results for consistency, methods like stylometry with R (Eder et al. 2013) do not accept two different languages. We can demonstrate the language difference also with an example in Spanish and French: The number of words for expressing "I think", in Spanish *pienso* and French *je pense*, are different because in Spanish no personal pronoun is necessary and the form of the verb expresses the case. The only way to compare the very same author in two languages is to be aware of these differences when using tools like the R package *stylo*, with the same parameters to compare the results in the end. The evaluation and interpretation of the results have to take this into consideration.

When Rybicki and Eder wrote about "cross-language authorial fingerprint" (2011) and Eder about "Delta across languages" (2011) they tested style markers for different languages to check if the most frequent words (MFW) are the best feature for authorship attribution—even for other languages than English. They prove that for French MFW still are good style markers. However, this is not true for every genre: English novels get higher rates than English epic poetry. If we agree on comparing the results of stylometry on the French and Spanish corpus we should be very careful because of the given difference of number of words. If *yo* ('I') in Spanish is only used to underline an expression (e.g., *yo, también* 'Me too') but usually not necessarily in combination of

Table 1 Example for grammatical difference in Spanish and French

Spanish	POS Spanish	French	POS French
hago ('I do')	"o" = 1st pers. sing. = POS: vmi	*Je fais* ('I do')	Personal pronoun (subject) + verb (1st pers. sing.) = POS: pro + v
Dame el papel / *Damelo* ('Give me the paper' / 'Give it to me')	"a" = 3rd pers. sing. (Imperative); "me" = personal pronoun; "lo" = direct object = POS: vmm + pp + pp	*Donne-moi le papier* / *Donne-le moi* ('Give me the paper' / 'Give it to me')	Verb (3rd pers. sing.); personal pronoun; direct object = POS: v + pro + pro

verb and personal pronoun this may influence the number of words. In consequence, the MFW count in a French and Spanish corpus will differ enormously. Furthermore, in Spanish object pronouns (e.g., *lo*, *la* 'it') are attached to the verb (see Table 1). This makes it even more difficult to compare the results of a method that counts and calculates the distribution of MFW in a corpus. For stylometry we keep MFW as dominant style marker because they proved to be the feature that needs less presumptions than others (Evert et al. 2017; Byszuk and Eder 2019).

In order to know how these differences will affect the results and the comparability of the corpora we have to test it with different methods like frequency of MFW, but also content words or parts of speech (POS) sequences. For choosing one of these methods, it is necessary to look at the kind of differences and similarities we would like to prove or test in Picasso's writings.

From 1935 on, Picasso wrote many small and some larger pieces of prose poems and two French plays during the time of German occupation in Paris 1942–44 (Table 2).

Astonishingly, the total sum of words in his Spanish and French writings over the whole period of about 30 years is quite similar. Despite the grammatical differences between Spanish and French (as shown above), the sums of words are very close. In this calculation we did not count the plays and we need to point out that Picasso only wrote in French in 1938, 1941–54, and in 1956 but wrote one of his longest and most known fragment of poetry and drama, *El Entierro del Conde de Orgaz*, in 1957–59.

Picasso had moved to Paris, the European cultural capital at this period, in 1901 to pursue his career. He stayed, became very quickly a famous artist of the avant-garde and was forced to stay when the civil war in Spain began in 1936. Picasso, himself, thought much about communication across languages. His rather well-known prose poem on translation proves this aspect:

> Si je pense dans une langue et que j'écris "le chien court derrière le lièvre dans le bois" et veux le traduire dans une autre, je dois dire "la table en bois blanc

Table 2 Overview of the complete writings of Pablo Picasso in Spanish and French

SPANISH			FRENCH		
Date	Texts	Words	Date	Texts	Words
1935	82	25,377	1935	52	9,454
1936	35	6,896	1936	98	13,233
1937	11	1,581	1937	46	8,853
1938	0	0	1938	29	6,284
1939	2	61	1939	12	3,180
1940	61	19,376	1940–41	44	7,035
1941–54	0	0	1942–44	17 (+2 plays)	2,860 (+18,080)
1955	2	174	1945–46	0	0
1956	0	0	1947–61	48	5,619
1957–59	1	5,966	–	–	–
1958–68	6	1,374	–	–	–
Total	**200**	**60,805**	**Total**	**346 (+2)**	**56,518 (+ plays = 74,598)**

enfonce ses pattes dans le sable et meurt presque de peur de se savoir si sotte" (Picasso, October 28, 1935).

"If I think in a language and write: 'The dog runs after the rabbit in the wood' and want to translate it in another language, I have to say 'The table of white wood marks its paws in the sand and nearly dies of fear knowing himself that silly'" (translation by N. R.-P.).

Here, he sticks to French, but still tells us something about translation without translating one language to another—how is this possible? There are traces of the first sentence in the second one: *chien* 'dog'—*pattes* 'paws'; *bois* 'wood'—*bois* 'wood'; *court* 'runs'—*enfonce* 'marks') and some grammatical POS are repeated like the articles and the prepositions (*le, la, dans* 'it', 'in'). Both sentences have an abstract poetic meaning which is produced by using rather concrete expressions in weird combination.[1] However, we can draw at least two possible readings out of the translation-example by Picasso:

1. Translating from one language to another means, particularly in the case of poetry, new sense, new style, all in all it means a new creation.

1 It might be interesting to test the distribution of concrete and abstract words in Picasso's writings following the example of the ongoing study by Ryan Heuser (2020).

2. Translation produces new meanings, new poetry which might only hint to the *original* without representing the original.

Based on the results of a close-reading analysis of the two corpora in French and Spanish we cannot say that Picasso just translates what he is thinking from one language to another but that he is experimenting with the same thought and vocabulary in each language (Michaël 2008; Rißler-Pipka 2015, 383). The result is probably as ill-suited for comparison as the two sentences Picasso gave as an example for translation (see above). The only chance to find out is by testing his writings and comparing the corpora regarding the following hypotheses based on previous close-reading analyses:

1. Picasso uses the same system of deconstructing language in both languages, Spanish and French.
2. It is not translation from one language to another that influences his style or makes it inventory—but the transformation of ordinary language into poetic *Picasso-language*.
3. The effect of chaos and semantic non-sense is produced by a recognizable system of grammatical repetition.

3. Analyses

After formulating hypotheses during close reading, we want to operationalize them in a quantitative computational analysis. Beforehand, we don't know if exemplary findings like extraordinary grammatical constructions or vocabulary will hold up to a quantitative testing. The hypotheses indicate the features that should be analyzed. For the first hypothesis (1) Picasso uses *the same system* in both languages, the *system* needs to be described more precisely. In Picasso's case this can be an obsessively repeated set of *vocabulary* which is easily comparable between two roman languages like Spanish and French. Stylometry on the basis of MFW illuminates the inner structure of authorial style and permits a comparison of different authors. This would support also hypothesis (2) saying that *Picasso-language* is detectable by this method. For the third hypothesis (3) the grammatical repetition can be detected by POS tagging and the comparison of POS sequences (n-grams). The latter is only one possibility out of several but is precisely drawn out of the observation in close reading, arguing that Picasso prefers some POS sequences, like prepositions (*de*, *que*) and noun (Rißler-Pipka, 2015, 270; Rißler-Pipka 2019a and b).

3.1 Vocabulary

To give a first overview of the French and Spanish corpus of Picasso's writings, I created a word cloud of the most frequent content words using Voyant Tools (Sinclair and Geoffrey 2016) and applying a self-defined stop-word list. The effect is even stronger because Picasso does not write narrative texts but prose poems. Readers familiar with the texts recognize the typical motifs which supports the hypothesis that Picasso uses mostly a fixed set of vocabulary (Michaël 2008; Rißler-Pipka 2015, 383).

Figures 1 and 2 are not very surprising if you know some of Picasso's texts and works of art: they are dominated by different words describing color, light, night and body parts. The famous bullfight topic is hidden because *toro* ('bull') is much less used than *tarde* ('afternoon'). From the previous close reading and the historical/cultural context we know they belong together because the bullfights were always scheduled in the afternoon. Comparing both word clouds' direct translations in French and Spanish

Fig. 1 Word Cloud for the Spanish texts (Rißler-Pipka, CC BY).

Fig. 2 Word Cloud for the French texts (Rißler-Pipka, CC BY).

are striking, for example for the colors (*bleu*/*azul* 'blue' *vert*/*verde* 'green'), for body parts (*main*/*mano;* 'hand'), for sensations (*odeur*/*olor* 'smell'; *cris*/*gritos* 'cries') and motifs (*feu*/*fuego* 'fire'; *ailes*/*alas* 'wings'). However, there are some differences in frequency we can show when Picasso speaks about the same things in Spanish and French prose poems. For example, French *ciel* ('sky') is used more often than Spanish *cielo* ('sky'), but *azul* ('blue') more often than *bleu* ('blue') because they belong together and are part of the very same semantic field. In direct comparison via translation or comparison of the bare numbers of most frequent content words this difference would not become visible (for specific semantic concepts in both languages see Meneses and Mallen 2019).

3.2 Stylometry

We want to look deeper into stylometry to be able to compare Picasso internally: what happens when he switches from one language to another? What are the main findings when comparing Picasso and his contemporaries in Spanish and French literature (1890–1960; poems or prose poems in avant-garde literature). For previous work I created a corpus of French and Spanish literature for the period of Picasso's life and writing (Rißler-Pipka 2019 a and b). For this study I used a reduced text collection of a group of authors known to be stylistically nearer to Picasso. The prose poems by Picasso were also sliced chronologically into equal sized parts to make them comparable to the length of common anthologies of poetry. The two plays *Le désir attrapé par la queue* (1941) and *Les quatre petites filles* (1948) were left out.

— Spanish corpus: 31 documents, sample size: 3,000–8,000, ∑ 157,863 tokens
— French corpus: 20 documents, sample size: 6,000–30,000, ∑ 290,131 tokens

According to the sample size and previous experiments and recommendation the chosen parameters for the initial analyses are: 2,500 MFW and the Wurzburg cosine delta (see Jannidis et al. 2015). The composition of the corpora is definitely not completely well balanced and the selection should be more precisely discussed. In the French corpus, Picasso's writings (apart of the plays) are split into seven more or less equal sized parts. The only other author who needed splitting up was Apollinaire: the well-known poetry anthologies *Calligrammes* and *Alcools* were split up in two and combined with his two erotic novels to have the prose genre represented as well. Éluard is represented in the French corpus with two texts (*Poèmes 1913–1940* and *La capitale de la douleur*). Unfortunately, for each of the other authors (Laurémont, Lourys, Segalèn, Mallarmé, Toulet) I found only one text per author and neither of these was long enough to be split up. More texts were either not available in a digital form or not suitable for comparison. For the Spanish corpus, the disproportionate number of texts by Picasso is even more striking because the sample size needed to be smaller in order to make it comparable to the other texts of the collection. Therefore, the Spanish corpus consists of ten texts by Picasso, six texts by Machado, four texts by Lorca, two by Alberti and two by Gómez de la Serna. Furthermore, seven authors are only represented with one single text.

One difficulty was being aware of the genre mixture in the avant-garde literature which might have negative influence toward the author signal. Another problem was the selection and accessibility of digital texts available for this period. However, for the rather detailed question if Picasso is using the same technique in two different languages, the representativeness of the corpus is not as important as for more advanced research questions.

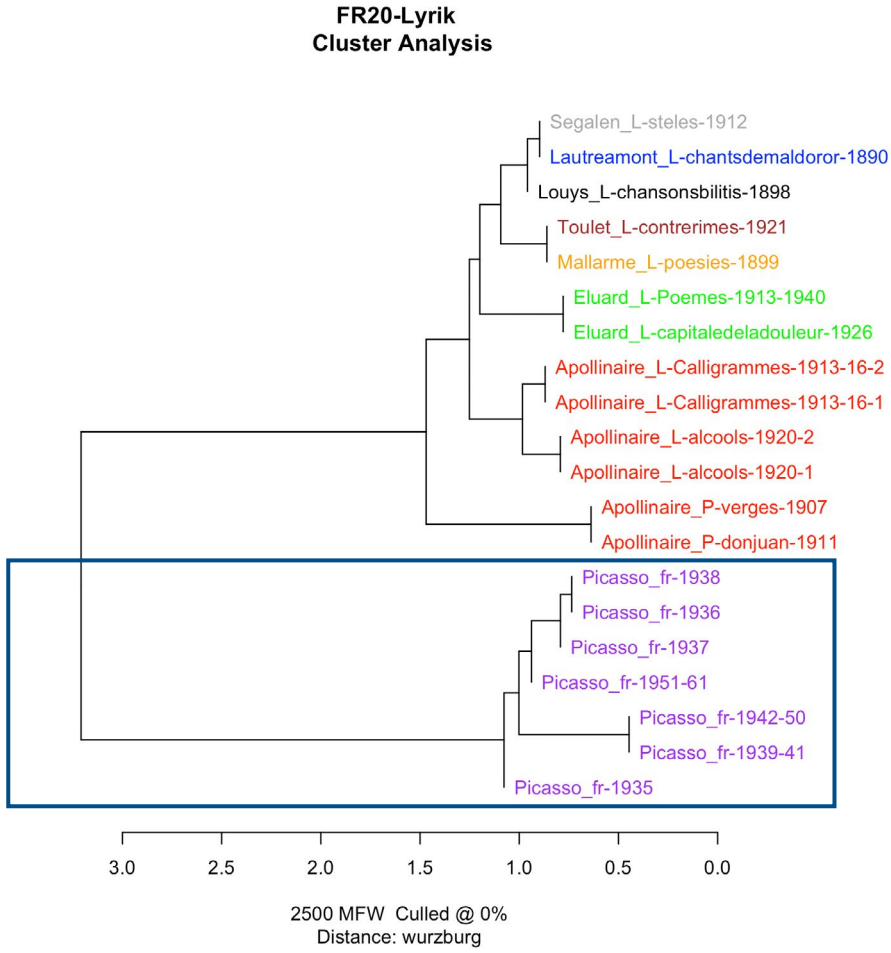

Fig. 3 Dendrogram for Cluster Analysis of the French texts (Rißler-Pipka, CC BY).

In Figure 3, Picasso's French writings take an own branch in the dendrogram and are not clustered with the surrealistic poetry or prose (or prose poetry) of the same period. The genre signal shows here as much effect as the author signal. Apollinaire is neatly divided in prose (*Les exploits d'un jeune Don Juan*; *Les onze mille verges*) and poetry (*Calligrammes; Alcools*) as well as the single authors who are more or less separated along the genre criteria. Picasso's writings are nearly all lined up according to the dates of creation: Picasso_Poesia1, 2 and 5 (1935–36); 4 and 3 (1935); 7 and Parchemin (1940) and the rest 6, 8 and the Entierro (1936–1957) are clustered

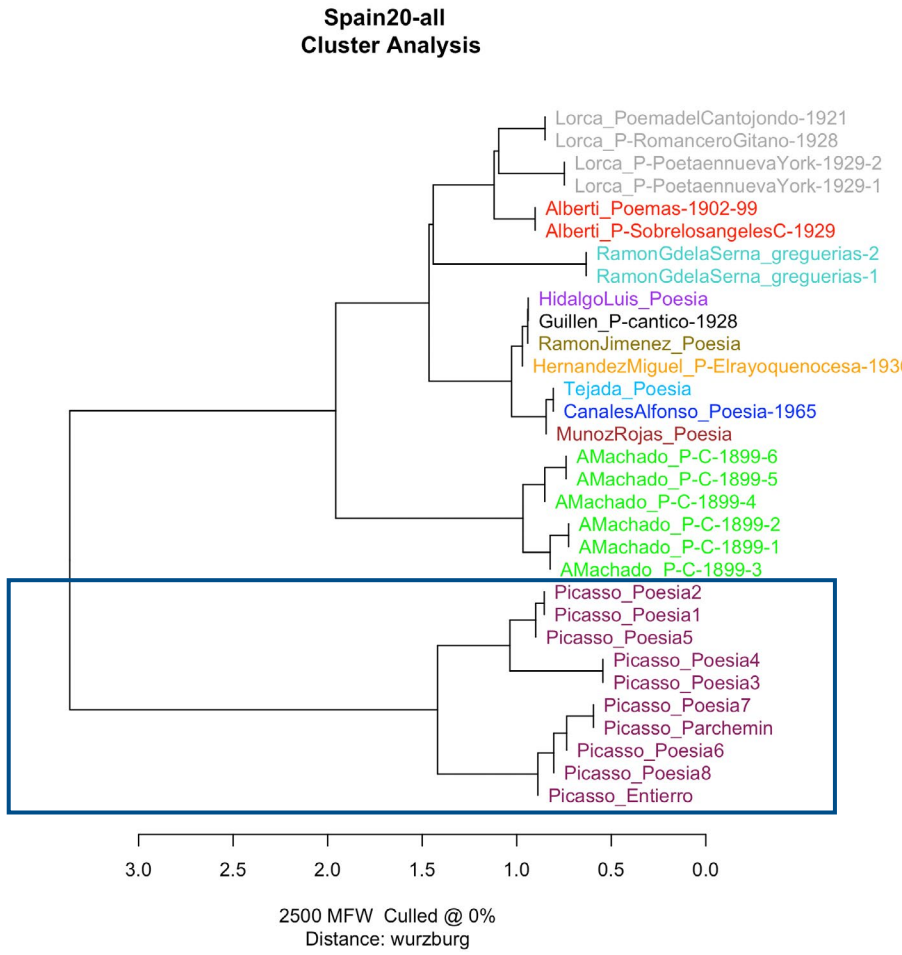

Fig. 4 Dendrogram for Cluster Analysis of the Spanish texts (Rißler-Pipka, CC BY).

together. In direct comparison to the Spanish corpus, we observe a similar effect as shown in Figure 4.

From previous analysis, I would expect that Picasso might cluster together with Ramón Gómez de la Serna but again (cf. Figure 4) his writings are completely separated from the rest of the corpus. Another outlier is—as expected—Antonio Machado because his poetry represents more the style of *generación 98* than surrealism and avant-garde. Regarding the motifs and subjects, we know from close reading that there is a similarity between Picasso and the upper group (Lorca, Gómez de la Serna, Alberti and particularly

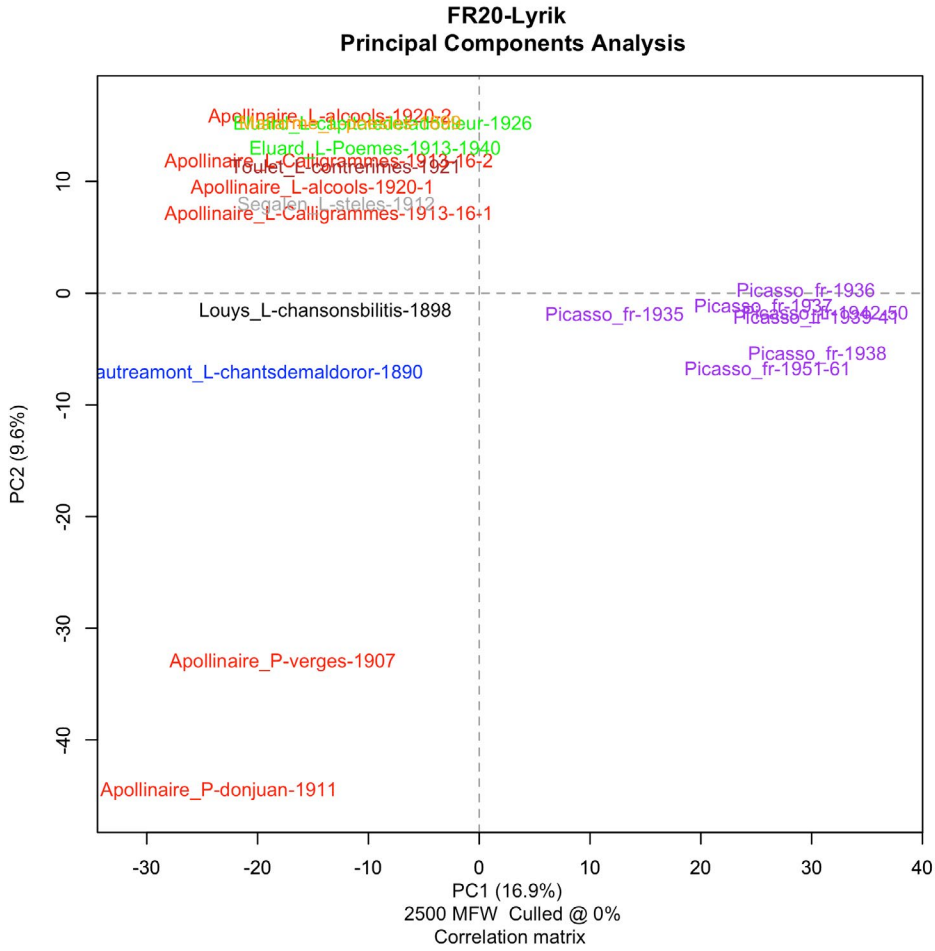

Fig. 5 Principal component analysis of the French texts (Rißler-Pipka, CC BY)

Miguel Hernández). They all use the technique of montage and the surreal description of the visible and palpable reality, which is at the same time abstract and concrete. This is probably more on the content level and that is the reason why the signal is not as strong as the stop-words which are responsible for the distribution of the MFW. The open question is accordingly: Why does Picasso's style also differ on the level of stop-words/MFW?

As a next step, I tried a principal component analysis (PCA) with the very same parameters in *stylo* and the same text collection. With this additional method I intended to get a better view of genre and author signal in comparison and will be able to visualize

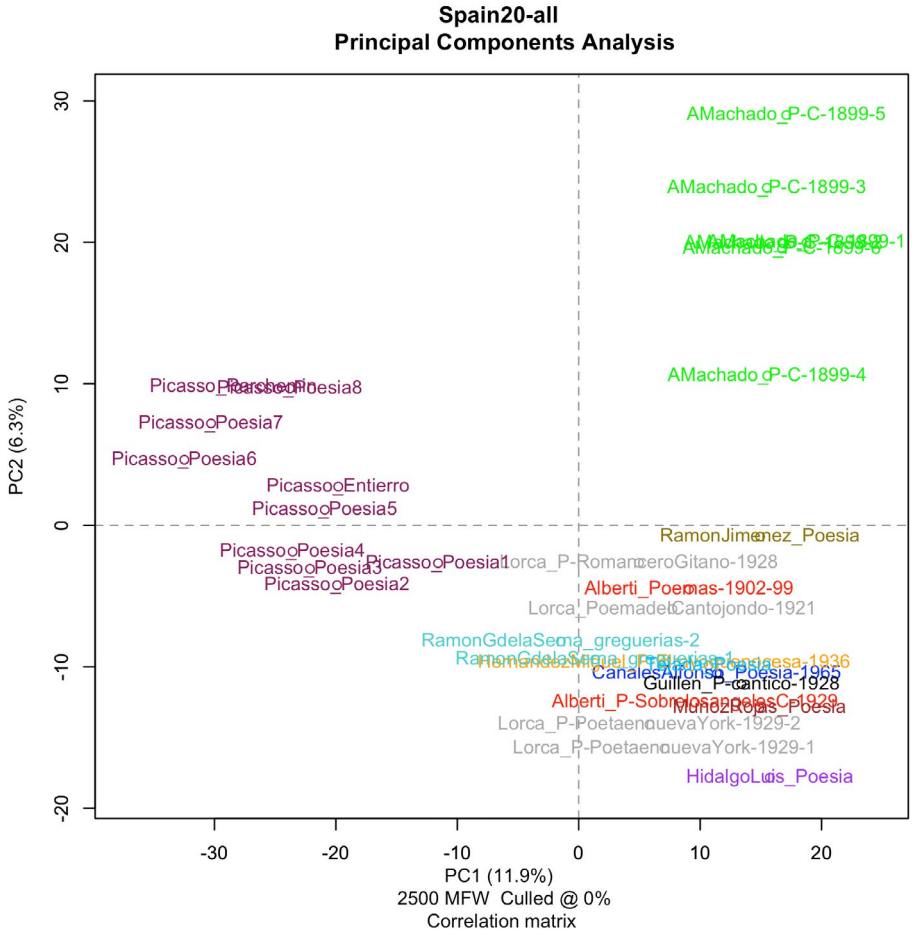

Fig. 6 Principal component analysis of the Spanish texts (Rißler-Pipka, CC BY).

the distance between the texts.[2] In Figure 5 we can also check the inner consistency of the two corpora. As we know from further research in stylometry (Eder 2010), the composition of the individual text collections may influence the results heavily.

In the upper part of the matrix (Figure 5) clusters the poetry—in the bottom the prose. It is very nice to see that Picasso as well as Louys and Lautréamont are marking

2 The complexity of the initial distance table is reduced by the PCA, this may produce different results (for a critical explanation see Craig 2004).

the frontier between both genres. All three of them write prose in a surrealistic and poetic way. In Lourys and Lautréamont this aspect is expressed in the word *chant* ('song', 'lyrics') in the title (*Les Chants de Maldoror* and *Chansons de Bilitis*) which also hints to a subgenre of poetry. Picasso does not care to give his writings a formal literary category. However, they are clearly nearer to the prose genre than to the prose poems by Éluard or Apollinaire, which contributes to a debate highly relevant in literary studies of that time and later on (Follet 1987). The analysis clusters two texts by Apollinaire together and separates them from the rest of the texts by the same author. This apparently confirms the genre difference: the upper cluster represents poetry and the lower prose (two erotic novels by Apollinaire). The effect may also be due to other indicators like a specific vocabulary (as it is to be expected in the erotic novels by Apollinaire)—here again the composition of the corpus is relevant and could be improved in this case. Even more than in the cluster analysis, the PCA shows Picasso as an outlier on the very right side of the matrix (compare Figures 4 and 5). Again, we observe a quite similar effect in the Spanish corpus (see Figure 6).

The PCA shows three different clusters but this time not necessarily due to genre: above, on the right side clusters the more traditional classic poetry by Antonio Machado, at the bottom the avant-garde/ Surrealism poetry and on the left side Picasso, isolated, clustering half in the more experimental poetry and half in not yet categorized part of the PCA matrix. In the Spanish corpus, there is no initial mixture in genre composition (poetry versus prose) as in the French corpus. The only prose-poets in this corpus are Picasso and Ramón Gómez de la Serna. Nevertheless, in both analyses, the PCA and cluster analysis, Gómez de la Serna is clustered together with the other avant-garde poetry by Lorca and Alberti. Consequently, we need to think of another style marker beyond genre that is responsible for the Picasso outlier and that also works for Spanish and French in quite the same way.

For both languages and in different statistical settings (dendrogram, PCA), Picasso proved to be stylistically an outlier in comparison to his contemporaries. This finding supports the hypotheses (1) and particularly (2) regarding the difference of everyday language, poetic language and *Picasso-language* and consequently the uniqueness of his style.

3.3 Part of Speech Repetition

Close reading Picasso we get the impression that he loves to repeat not only of motifs, topics, objects but also grammatical entities like POS. A simple example, chosen randomly out of a longer prose poem written in summer 1940 (see Table 3) shows how two pieces of text belong together without really repeating themselves word by word and without speaking about the same thing.

Table 3 Example for repetition of grammatical entities in Picasso's poetry [red color = word repetition; green color = repetition of stop-words; blue color = strong resemblance in word melody and sound; italic = repetition of grammatical constructions]

19 July1940	01 August 1940
"buscando dentro del oro ***derretido*** ***que*** *corre* la cortina ***sus*** *fuegos* fatuos sobre el filo del estoque por el amor del dios amor *su limosnita* la línea que sube"[a]	"que vomita la boca sin dientes del calor ***derretido*** ***que*** *extiende* ***sus*** *piernas* por encima del almohadón de las faldas del veneno píldora del amor la boca abierta a la sed del estoque que la enlaza azucena que sube"[b]

a "looking inside the melted gold which runs the curtain its vain fires over the wire of the sword of the love of the gods of love his donation the rising line" (translation by N. R.-P.).

b "which vomits (or vomiting) the mouth without teeth (because) of melted heat which spreads its legs over the pillow of the skirts of the poison pill of love (with) open mouth to the thirst of the sword which ties her lily which is rising" (translation by N. R.-P.).

The semantic meaning of these pieces of poetry is clearly not relevant in terms of understanding and detecting the literary context. Picasso uses a highly poetic language and every content word like *dientes* ('teeth'), *calor* ('heat'), *estoque* ('sword') has a symbolic meaning and is at the same time very ambiguous. Here, we could easily detect words or symbols belonging to the bullfight motif and analyse them regarding their symbolic meaning and context in Picasso's other pieces of art. For example, 'teeth' and 'heat' are very important in the bullfight context in his writings but can rarely be detected explicitly in his visual art. Therefore, one could argue that by writing prose poems Picasso adds the invisible to his visual art.

However, what is he doing inside his poetry that distinguishes his writings from the French and Spanish avant-garde poets and writers? It might be the repetition of POS elements. Looking at the example in Table 3, we observe that the construction of the sentence—or part of speech—is similar with tiny changes: *dentro* 'inside' / *dientes del oro* 'teeth of melted gold' / *calor derretido que corre* 'heat which runs' / *extiende [la cortina] sus fuegos* 'spreads [the curtain] its fires' / *pierna* 'leg') etc. Using the deductive method of close reading, we could now extrapolate from here to the conclusion that the repetition of POS elements is the unique characteristic which separates Picasso's writings from his contemporaries. This also underlines the findings out of stylometry (see 3.2) because a high frequent use of prepositions, for example, is also detectable in the MFW. In Figures 3–6 the position of Picasso as an outlier may be explained in combination with the POS frequencies.

The quantitative analysis of POS elements and cross-language comparison using stylometry may prove this hypothesis not only by the observation of recurring and randomly selected paragraphs but consistently by counting the repetition of POS elements and comparing them relatively in a whole corpus. The differences in languages between Spanish and French still do not allow us to compare the POS elements or

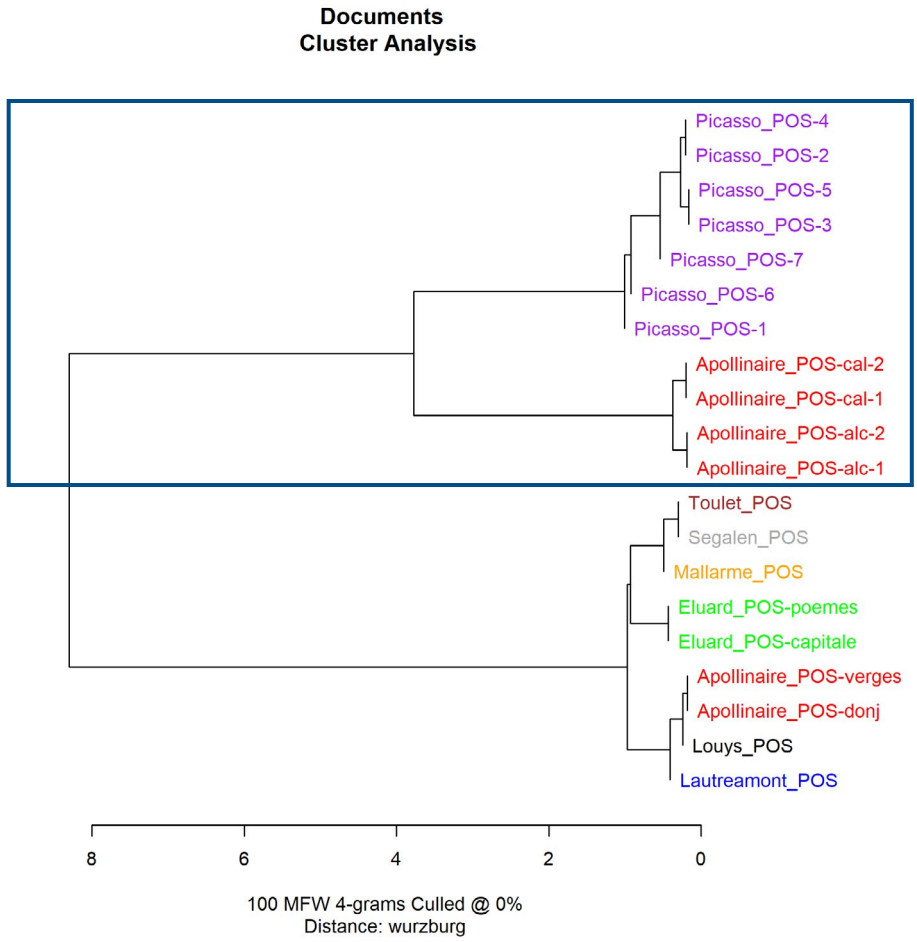

Fig. 7 Cluster analysis of POS 4-grams of the French texts (Rißler-Pipka, CC BY).

sequences directly even if we are freed of vocabulary differences. The reason is the same as for the MFW differences (see Table 1). Still, POS elements are less variable as MFW.

Jeremy Ochab tested POS sequences as style marker using the *stylo* package and comparing the most frequent POS ngrams in a corpus (Ochab 2017). Looking at the very same corpus of Spanish and French poetry including Picasso, the difference in language is visible via the number of POS tokens: in the Spanish corpus we count 157,863 word tokens and 181,905 POS tokens, while in the French corpus we count 290,131 word tokens and 316,073 POS tokens. I used the Standford POS tagger (Toutanova

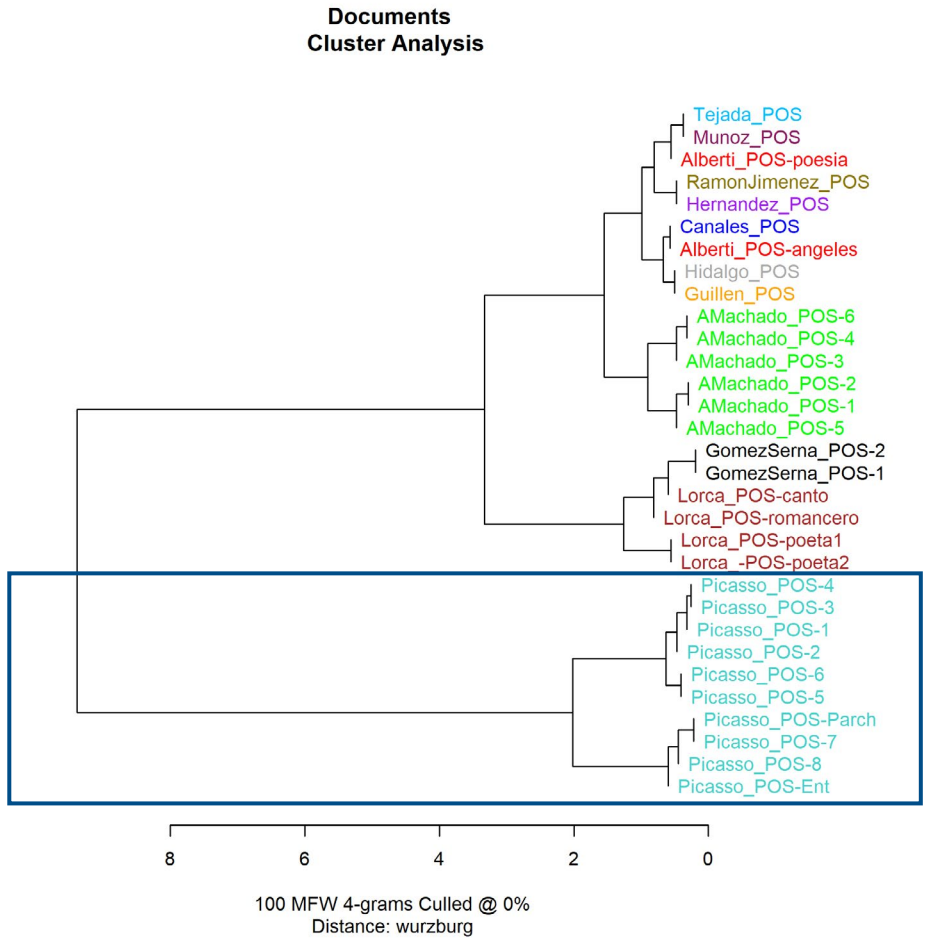

Fig. 8 Cluster analysis of POS 4-grams of the Spanish texts (Rißler-Pipka, CC BY).

et al. 2003) with different tagsets for Spanish and French. For the clustering experiment with *stylo*, I decided to test 4-grams of POS tags for each corpus with a maximum of one hundred MFW because there are not many more different 4-grams in the corpus.

At first sight, there is not much difference between the clustering with the MFW feature (Figures 3 and 4) or the POS 4-grams as a feature (Figures 7 and 8). Both cluster Picasso's writings for both languages separately. However, in the French corpus Picasso and the poetry by Apollinaire are clustered on the same branch and in the Spanish corpus Lorca and Gómez de la Serna are clustered together. All in all, the author

signal is still visible in the preferences regarding frequencies for POS sequences. The clustering of Apollinaire and Picasso does not necessarily mean that both prefer similar POS sequences but may also hint to a common characteristic that distinguishes both from the rest of the corpus. As we know, Picasso and Apollinaire both nearly never use punctuation in their poetry—this might be a simple explanation.

For a deeper look, I checked the most frequent POS 4-grams for their characteristics in each language according to different tagsets (using the Stanford NLP POS tagger) (Table 4).

Table 4 The three most frequent POS 4-grams in the French and Spanish corpus: nc / ncs = noun, p / sp = preposition, det / da = determiner, punc = punctuation, cs = conjunction, ao = adjective

French	Spanish
det nc p nc	ncs sp da ncs
nc p det nc	sp da ncs sp
p det nc p	da ncs sp ncs

The problem that disturbs the comparability here is not only the different grammar and syntax but also a different POS tagset used by the Stanford NLP POS tagger: punctuation is named differently and separated into the different kinds of punctuation in the Spanish tagset, but in the French tagset it is just all assembled as *punc* ('punctuation').[3] That means that two authors not using punctuation as Picasso and Apollinaire will most likely be clustered together simply because of this very striking style marker. Tagged or not, punctuation might influence the POS tagger whether to decide between ambivalent POS elements. The most difficult question is whether a word is used as noun or verb. Without punctuation, POS taggers are more prone to errors because of missing indicators. I discussed this for Picasso in an earlier study using the FreeLing POS tagger (Rißler-Pipka 2019b). Recently Byszuk et al. discussed the accurateness of detecting direct speech in multilingual corpora, pointing out that "while heavily relying on punctuation. The last one seems particularly important for misclassifications" (Byszuk et al. 2020). That means we have to look at the results doubtfully. Also, most recently Eder and Górski highlighted the limitations of POS for different languages (Eder and Górski 2023).

3 Mapping the tags to a universal tagset would be a solution here even if details may get lost. For me, this was not practicable at the time of the analysis. Plus, we have to keep in mind that all POS taggers coming out of the computational linguistics context are trained on contemporary everyday speech and we do not have any indication about their correctness when applying them on historic literary texts.

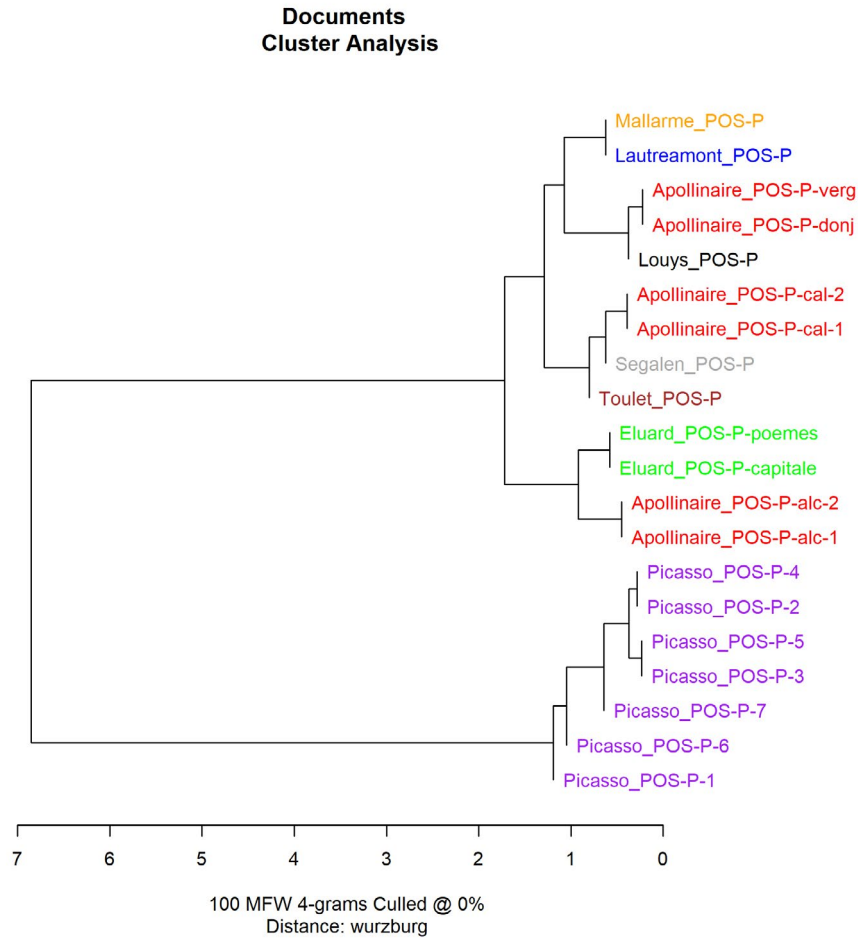

Fig. 9 Cluster analysis of POS 4-grams of the French texts without punctuation-tag (Rißler-Pipka, CC BY).

To check whether the punctuation is responsible for the clustering of Picasso and Apollinaire, I used the same parameters for the cluster analysis but deleted all *punc* ("punctuation") in the whole French corpus (Figure 9). That does not mean that the result is not influenced anymore by punctuation but for the most frequent POS 4-grams it could change the clustering.

Now again Picasso's prose poems are clustered on a single branch. Astonishingly Apollinaire is exactly divided by his anthologies *Alcools* (1913) / *Calligrammes* (1913–16) and his erotic novels *Les exploits d'un jeune Don Juan* (1911) / *Les onze mille verges* (1907).

Zooming into the results of the most frequent POS tags in the two corpora, the Spanish and the French, we might better understand the difference Picasso makes by using a very rare style of POS repetition. Comparing the tables of frequencies for both corpora represented as a heatmap, we can see which POS 4-grams Picasso uses far more often than his contemporaries and what is the combination most common for other writers, but not used by Picasso.

The visualization (Figure 10) of the frequency table is a simple csv-excel-sheet—an output of *stylo*—showing the ranked POS 4-grams on the left-hand column. For the heatmap the statistical values for each text in the corpus (here the Spanish collection) are transformed accordingly into colors (from dark blue = 0 to dark red = high frequency). As explained according to the corpus composition, Picasso's texts form a larger part and are easily detectable in the visualization.

In many texts Picasso shows a very distinctive preference for dominant POS 4-grams which are very frequently used, but only by himself. On the one hand, the most frequent POS sequences like *ncs sp da ncs* ('noun, preposition, determiner, noun') of the overall corpus are used by Picasso. Also, in the lower rows Picasso uses many POS 4-grams much more often than other authors in the corpus. On the other hand, some sequences are not at all used by Picasso (the green lines in the Picasso-block) which is not surprising because they are indicating punctuation: for example, rows 18 and 25 of the table: *sp da ncs fp* ('preposition, determiner, noun, full-stop') and *sp da ncs fc* ('preposition, determiner, noun, comma'). Even if Picasso's texts are overrepresented in the whole corpus, this cannot explain or excuse the exceptional preference for some PoS4-grams. Moreover, through the visualization we could now detect some texts by Picasso in which he focused on the repetition of the very same POS sequence and a development over time. The texts are in chronological order (from 1 to 8) apart from the first two, *El Entierro del Conde de Orgaz* (1957–59) and *Carnet Parchemin* (1940). From those two, the *Carnet Parchemin* significantly deviates and differs from others texts by Picasso. The *Carnet* consists of just one very long text dated 7 November 1940—a time when Picasso was immobile in the city of Paris occupied by Nazi Germany during World War II (1940–44). Apart from this long prose poem the rest of the *Carnet Parchemin* is lost and as we know from the dates in Table 2, Picasso nearly stopped writing in his native language Spanish from 1938 to 1957, only the year 1940 marks an exception with 61 texts (19,376 words). The manuscript of the *Carnet* (which is not freely accessible, but an example can be seen in López Sánchez 2015; Rißler-Pipka 2015, 39; Bernadac and Piot 1989, 243–51) shows the writing method of Picasso in an exceptional way: the number of insertions, arrows and deletions makes the manuscript nearly indecipherable. As we now recognize by looking at the most frequent POS sequences, the impression of repetition and preferences in grammatical order followed up to details of every line (not sentence) makes even more sense in light of this context. The nearly obsessive character of insertion and working

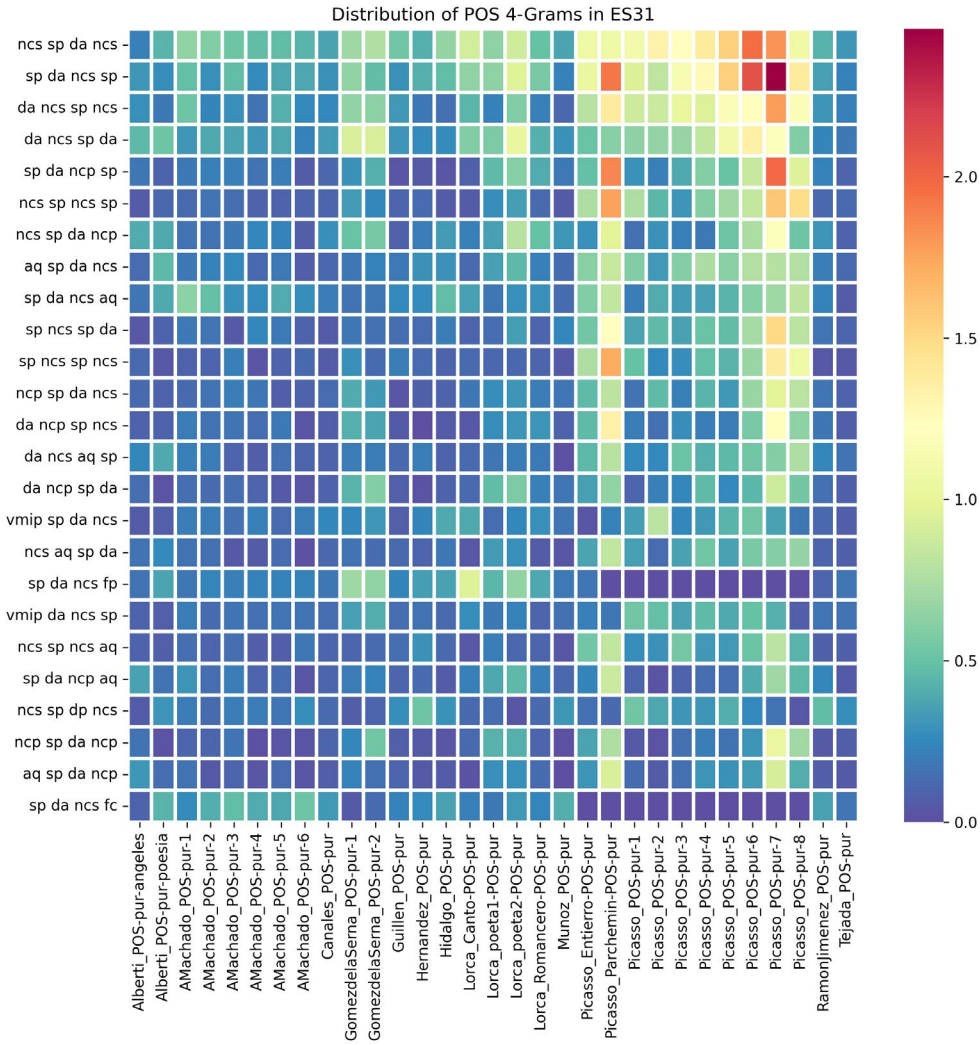

Fig. 10 Heatmap of the table of frequencies for the cluster analysis of the Spanish corpus of PoS4-grams as shown in Figure 8 (Rißler-Pipka, CC BY).

on the texts shows that he is not adding any content words but rather repeating POS sequences.

Apart from this rather detailed point of view on one particular prose poem by Picasso, the clustering of POS 4-grams of the Spanish corpus also supports the hypothesis

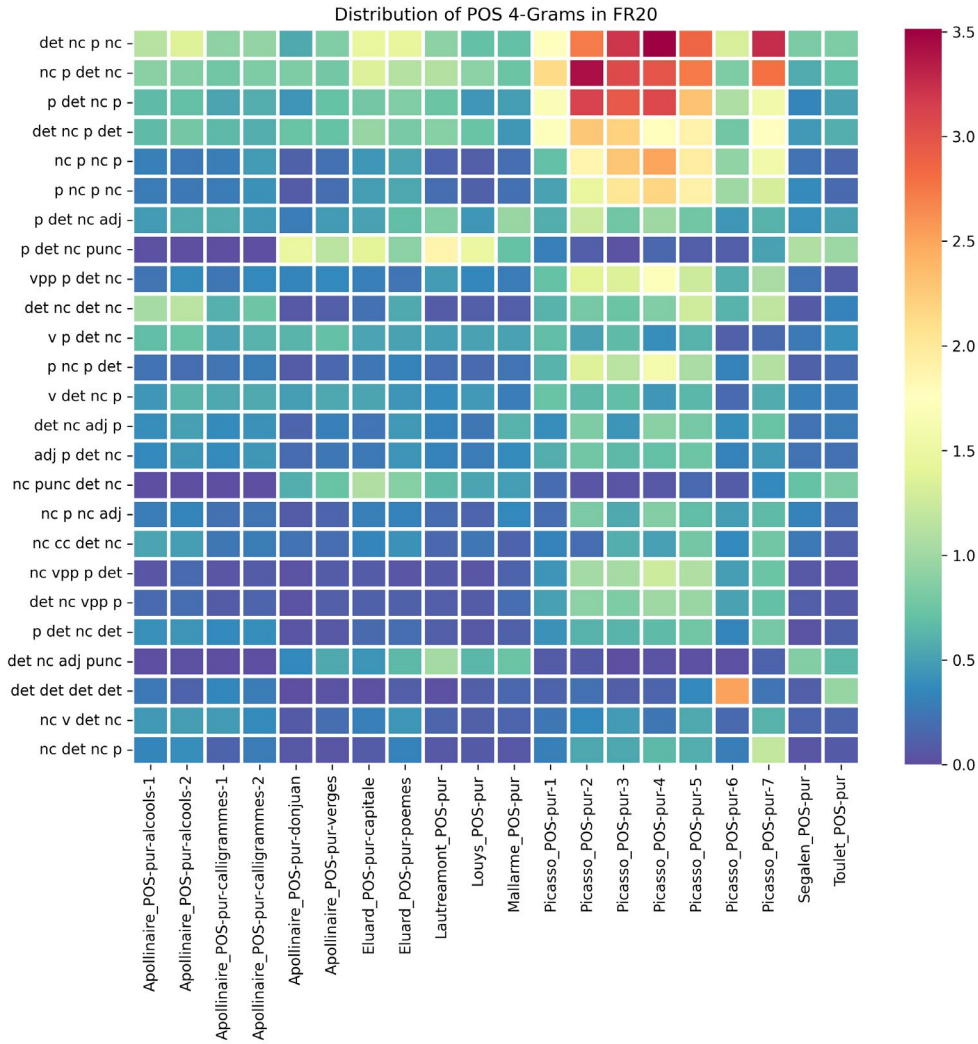

Fig. 11 Heatmap of the table of frequencies for the cluster analysis of the French corpus of POS 4-grams as shown in Figure 7 (Rißler-Pipka, CC BY).

that Picasso uses a striking system of grammatical repetition which is not comparable to other—even poetic—language (Rißler-Pipka 2015; but also Mallen 2012). In order to test if this is also true for the French corpus in which Apollinaire as a predecessor shows close resemblance in style with Picasso, I made the same visualization with the French texts (Figure 11).

As described for the Spanish corpus heatmap (Figure 10) we see the most frequent POS 4-grams ranked in the left-hand column and the heatmap shows their relative frequency for each text of the corpus. The texts by Apollinaire are easily detectable because of the dark blue brackets indicating POS 4-grams with punctuation (p det nc punc; nc punc det nc; det nc adj punc)—not used neither by Apollinaire in the four parts of his poetry anthologies (*Calligrammes* 1+2; *Alcools* 1+2) nor in most of Picasso's text. Picasso's texts present the most significant pattern. This is similar to the Spanish corpus (Figure 10). Picasso seems to have a clear preference for a set of POS sequences, which he repeats again and again. According to the heatmap for the Spanish corpus, Picasso's French texts are in chronological order beginning in 1935 up to 1961. The first year when Picasso began to write prose poems, he started writing in his native language, Spanish. The 52 French prose poems of 1935 were rather short (9,454 words all in all) while he was writing in the same year 82 prose poems in his native language Spanish (25,377 words). In the foreign language he had to find his personal style as a writer and started rather normally with very long 'sentences.' A style that can be compared to his surrealist friends, but not to the noun-preposition-noun-style he elaborated during his writer-period. As an example, we could have a look at the prose poem dated November 5, 1935:

> […] mais l'esprit de sel ne compte ses blessures sur le dos de l'amande qui se rit des regards courroucés de la foule et ne tient sa raison à la chaîne que pour mieux se moquer devant sa glace de la poche d'argent que dans sa main fait la monnaie rendue à l'ennemi […] (Picasso, November 5, 1935).

> "[…] but the spirit of the salt does not count its wounds on the back of the almond which laughs about the angry looks of the crowd and does not hold its reason at a chain which for better mockery in front of its mirror of the pocket of money which in its hand made the money given back to the enemy […]" (translation by N. R.-P.).

The verbs are highlighted (in green) in the citation because of the rather normal (= frequent) use of verbs which is not at all typical for Picasso's style from 1936 on or for his Spanish writings. As a counterexample we can read the prose poem dated February 6, 1938:

> […] entouré par les dents de la mâchoire du soleil plantées dans sa chair le carré de l'arène rempli d'eau soutenu par des cierges grelotte ses poils et secoue la cendre tombée sur la nappe chiffonnée déjà pleine de taches après le déjeuner collée au plafond de la grêle de flèches du clairon fixé au mât de l'épée qui […] (Picasso, February 6, 1938).

> "[…] surrounded by the teeth of the jawbone of the sun planted in its flesh the square of the arena filled with water supported by candles rings its hair and shakes the ash fallen out of the crumpled tablecloth already full of stains after the meal sticked to the ceiling of the hail of arrows of the horn attached to the pole of the sword which […]" (translation by N. R.-P.).

Here, active verbs are transformed to adverbs or adjectives describing the nouns, which are linked together, by prepositions and conjunctions (highlighted in red). Nobody or nothing does anything anymore. In this rather long paragraph, we only find two active verbs (highlighted in green). This striking preference for an endless description means a preference for certain POS sequences. The absence of verbs is rather unusual also in comparison to his contemporaries in Spanish and French avant-garde. By observing the frequencies and distances regarding every text we can also detect a development over time inside of Picasso's writings—a phenomenon we observed in a similar way in the Spanish corpus. After the first year of struggling and searching, 1935, Picasso found his typical style also in the foreign language French.

4. Conclusion

Summing up the discussion of this paper and others, it remains uncertain, if we can really compare Picasso's method of writing—or rather building texts out of a laboratory of words—to modern information capturing with the help of algorithms like the artist and art historian López Sánchez suggested already in 2015:

> […] tiene mucho que ver la visión crítica que Picasso hace de sus 'metadatos' con la labor que ejerce un analista de Big-data hoy día, pues ambos saben (quizás Picasso de manera inconsciente y por supuesto de una manera mucho más rudimentaria) que gestionando datos, analizándolos y extrayéndoles información cuando es necesario, es la mejor manera para generar el conocimiento (López Sánchez 2015).

> "[…] has much to do with the critical vision of Picasso regarding his 'metadata' and with the work a Big-data analyst would do today. We all know (might it be because Picasso does it unconsciously or on purpose in a much more rudimental way) that managing data, analyzing and extracting information out of them when necessary is the best way to gain and create knowledge" (translation by N. R.-P.).

Picasso was not a modern data scientist analyzing his metadata to get an illuminating knowledge graph in the end. The striking difference is the effect of destroying the grammatical logic of a language system to invent and to create something new which might lead to some knowledge about the construction of sense via syntax and semantics. Certainly, his aim was not to come to a deeper knowledge about the things and themes he was speaking about. Picasso writes about basic things (like food) and old things (like myths) and varies the perspective by new combinations in the very same system: how many combinations are possible sticking to the same vocabulary and some preferred POS 4-grams? What looks like a simple algorithmic or combinatoric question does not lead to a poetic machine (invented by Raymond Queneau at the same time, see Wolff 2016). Neither does it mean a constant variation without following any system or rule as Michaël suggests: "Il varie les écritures, passe de la prose au vers, change la disposition graphique sur la page, joue sur la linéartité et la polysémie" (2012, 169; "He varies different kinds of writing, switches from prose to verse, changes the graphical order of the page, plays with linearity and polysemy;" translation by N. R.-P.). The impression of constant variation is true, but the reason is not constant invention but a "mirror game and a game of repetition" ("jeu de miroir et de répétition") as Michaël (ibid., 179) also adds.

By stylometric analyses in a wider cross-language sense by comparing two text collections of two languages with one common author (Picasso), and by testing the hypothesis that Picasso uses a recognizable style in both languages that differs from his contemporaries, we were able to support the impression gained by close reading in various studies and to find out about the reasons behind this. This study is part of explaining the rules of the game Picasso is playing with his reader and spectator. The preferred POS 4-grams and the MFW might only represent a tiny part of the rules to be followed up. The three parts analyzed in a quantitative way, vocabulary, stylometry and POS 4-grams, all help to understand what makes the prose poems by Picasso in both languages, French and Spanish, recognizable and outstanding. It helps us to understand how he evokes the impression of a constant metamorphosis. Picasso keeps a rather fixed and small set of vocabulary (content words) in both languages but constantly changes the smaller grammatical units, while keeping some preferred and very often used stop-words and POS 4-grams.

ORCID®

Nanette Rißler-Pipka https://orcid.org/0000-0002-0719-9003

References

Béhar, Henri. 1993. "Picasso au miroir d'encre." In *L'artiste en représentation*, edited by Réné Démoris, 199–213. Paris: Desjonquères.

Bernadac, Marie-Laure, Christine Piot. "Critical Comments, Annex", in : Picasso, Pablo. Picasso: Collected Writings. Edited by Marie-Laure Bernadac. 1. publ. London: Aurum, 1989.

Byszuk, Joanna, and Maciej Eder. 2019. "Feature Selection in Authorship Attribution: Ordering the Wordlist." In *Digital Humanities 2019: Conference Abstracts*. https://doi.org/10.34894/RCOIXS.

Byszuk, Joanna, Michał Woźniak, Mike Kestemont, Albert Leśniak, Wojciech Łukasik, Artjoms Šeļa, and Maciej Eder. 2020. 'Detecting Direct Speech in Multilingual Collection of 19th-Century Novels'. In Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages, 100–104. Marseille, France: European Language Resources Association (ELRA). https://aclanthology.org/2020.lt4hala-1.15.

Calvo Tello, José. 2019. "Delta Inside Valle-Inclán: Stylometric Classification of Periods and Groups of His Novels." In *Theorien von Autorschaft und Stil in Bewegung: Stilistik und Stilometrie in der Romania*, edited by Nanette Rißler-Pipka, 151–63. München: AVM edition.

Clement, Tanya. 2013. "Text Analysis, Data Mining, and Visualizations in Literary Scholarship." In *Literary Studies in the Digital Age*, edited by Kenneth M. Price and Ray Siemens. New York: Modern Language Association of America. https://dlsanthology.mla.hcommons.org/text-analysis-data-mining-and-visualizations-in-literary-scholarship/.

Craig, Hugh. 2004. "Stylistic Analysis and Authorship Studies". In *Companion to Digital Humanities*, edited by Susan Schreibman, Ray Siemens, and John Unsworth. Oxford: Blackwell Publishing Professional. http://www.digitalhumanities.org/companion/.

Eder, Maciej. 2010. "Does Size Matter? Authorship Attribution, Short Samples, Big Problem." *Digital Humanities 2010: Conference Abstracts*, 132–35.

Eder, Maciej. 2011. "Style-Markers in Authorship Attribution. A Cross-Language Study of the Authorial Fingerprint." *Studies in Polish Linguistics* 6: 99–114.

Eder, Maciej, Mike Kestemont, and Jan Rybicki. 2013. "Stylometry with R: A Suite of Tools." *Digital Humanities 2013. Conference Abstracts*, 487–89.

Eder, Maciej, and Rafał L. Górski. 2023. 'Stylistic Fingerprints, POS-Tags and Inflected Languages: A Case Study in Polish'. Journal of Quantitative Linguistics 30, no. 1: 86–103. https://doi.org/10.1080/09296174.2022.2122751.

Éluard, Paul. 1935. "Je parle de ce qui est bien," in: "Picasso 1930–1935," special issue, *Cahiers d'Art* 10, no. 7–10: 29–32.

Evert, Stefan, Thomas Proisl, Fotis Jannidis, Isabella Reger, Steffen Pielström, Christof Schöch, and Thorsten Vitt. 2017. "Understanding and Explaining Delta Measures for Authorship Attribution." *Digital Scholarship in the Humanities* 32: ii4–ii16. https://doi.org/10.1093/llc/fqx023.

Fernández Molina, Antonio. 1988. *Picasso Escritor*. Madrid: Prensa y Ed. Iberoamericanas.

Follet, Lionel. 1987. "Apollinaire entre vers et prose – de « L'Obituaire » à « La Maison des morts »." *Semen – Revue de sémio-linguistique des textes et discours* 3. https://doi.org/10.4000/semen.5523.

Goddard, Linda. 2006. "Mallarmé, Picasso and the Aethetics of the Newspaper." *Word & Image* 22 (4): 293–303.

Heuser, Ryan. 2020. "Abstraction: A Literary History." Talk given at King's College, Cambridge on February 18. https://ryanheuser.org/talks/kingscollege2020/.

Heydel, Magda and Jan Rybicki. 2012. *The stylometry of collaborative translation*. Digital Humanities 2012: Conference Abstracts. Hamburg: Hamburg University Press, pp. 212–14. https://www-archiv.fdm.uni-hamburg.de/dh2012/conference/programme/abstracts/the-stylometry-of-collaborative-translation/.

Heydenreich, Titus. 1979. "'Kilómetros y Leguas de Palabras…'. Pablo Picasso als Schriftsteller." *RZLG* 3: 154–68.

Jannidis, Fotis, Steffen Pielström, Christof Schöch, and Thorsten Vitt. 2015. "Improving Burrows' Delta – An Empirical Evaluation of Text Distance Measures." In *Digital Humanities 2015: Conference Abstracts*. Sydney: University of Western Sydney. http://dh2015.org/abstracts/index.php.

Juola, Patrick, and George K Mikros. 2016. "Cross-Linguistic Stylometric Features: A Preliminary Investigation." *JADT 2016: 13ème Journées internationales d'Analyse statistique des Données Textuelles*. https://jadt2016.sciencesconf.org/80692/Draft_Paper_style_jadt_2016_2.pdf.

Leiris, Michel. 1966. *Brisées*. Paris: Mercure.

López Sánchez, Santiago. 2015. "Picasso al margen: Un viaje por los metadato." In *II Taller Experimental de Investigación sobre Picasso y el Arte del siglo XX*. Málaga: Fundación Picasso Málaga. https://fundacionpicasso.malaga.eu/export/sites/default/cultura/fpicasso/portal/menu/portada/documentos/SANTIAGO_LxPEZ_Picasso_al_Margen.pdf.

Mallen, Enrique. 2009. "The Multilingual Poetry of Pablo Picasso." *Interdisciplinary Journal for Germanic Linguistics & Semiotic Analysis* 14 (2): 163–202.

Mallen, Enrique. 2012. "La poesía simpatética de Pablo Picasso." In *Picasso – Poesie – Poetik: Picassos Schaffen aus literatur-, sprach- und medienwissenschaftlicher Sicht = Picasso, his poetry and poetics*, edited by Nanette Rißler-Pipka, 105–40. Aachen: Shaker.

Mallen, Enrique, and Luis Meneses. 2019. "Adjoined Conceptual Domains in the Bilingual Poetry of Pablo Picasso." *Digital Studies/Le Champ Numérique* 9 (1): 20. https://doi.org/10.16995/dscn.320.

Meneses, Luis, and Enrique Mallen. 2019. "Semantic Domains in Picasso's Poetry." *Digital Scholarship in the Humanities* 34 (1): i123–i28. https://doi.org/10.1093/llc/fqy078.

Michaël, Androula. 2008. *Picasso Poète*. Paris: ENSBA.

Michaël, Androula. 2012. "Picasso écrivain, la réactualisation des préoccupations." In *Picasso – Poesie – Poetik: Picassos Schaffen aus literatur-, sprach- und medienwissenschaftlicher Sicht = Picasso, his poetry and poetics*, edited by Nanette Rißler-Pipka, 165–82. Aachen: Shaker.

Rißler-Pipka, Nanette. 2019. "In Search of a New Language: Measuring Style of Góngora and Picasso." In *Theorien von Autorschaft und Stil in Bewegung: Stilistik und Stilometrie in der Romania*, edited by Nanette Rißler-Pipka, 117–50. München: AVM edition. https://www.romanischestudien.de/index.php/rst/article/view/639.

Rißler-Pipka, Nanette. 2019a. "L'esthétique numérique de Picasso." *Philologie im Netz,* Beiheft 16: 39–58. http://web.fu-berlin.de/phin/beiheft16/b16t04.pdf.

Rißler-Pipka, Nanette. [2015] 2019b. *Picassos schriftstellerisches Werk: Passagen zwischen Bild und Text*. Bielefeld: Transcript. / OA version without images published in 2019: https://nbn-resolving.org/urn:nbn:de:hbz:467-14397.

Rybicki, Jan, and Maciej Eder. 2011. "Deeper Delta across Genres and Languages: Do We Really Need the Most Frequent Words?" *Literary and Linguistic Computing* 26 (3): 315–21. https://doi.org/10.1093/llc/fqr031.

Sabartés, Jaime. 2017. *Picasso: retratos y recuerdos*. Aguadulce: Confluencias editorial.

Sinclair, Stéfan, and Geoffrey Rockwell. 2016. *Voyant Tools*. Web. http://voyant-tools.org/.

Ochab, Jeremi K. "Stylometric Networks and Fake Authorships". *Leonardo* 50, Nr. 5 (Oktober 2017): 502–502. https://doi.org/10.1162/LEON_a_01279.

O'Sullivan, James, Katarzyna Bazarnik, Maciej Eder, and Jan Rybicki. 2018. "Measuring Joycean Influences on Flann O'Brien." *Digital Studies/Le Champ numérique* 8 (1): 6. https://doi.org/10.16995/dscn.288.

O'Sullivan, James. 2014. "Finn's Hotel and the Joycean Canon." *Genetic Joyce Studies* 14: 8.

Toutanova, Kristina, et al. 2003. "Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network." In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology,* vol 1. 173–80. Edmonton, Canada: Association for Computational Linguistics. https://doi.org/10.3115/1073445.1073478.

Wolff, Mark, 2016. 'Oulipian Code', http://markwolff.name/wp/digital-humanities-2/oulipian-code/.

For the software: https://nlp.stanford.edu/software/tagger.html.

For the corpora of digitized texts: Biblioteca Virtual Miguel de Cervantes, Wikisource and Mallen, Enrique. n.d. Picasso: Online Picasso Project. Enrique Mallen. https://picasso.shsu.edu/.