ISSN 1798-4769
Journal of Language Teaching and Research, Vol. 9, No. 5, pp. 928-938, September 2018
DOI: http://dx.doi.org/10.17507/jltr.0905.05

A Corpus-based Computational Stylometric Analysis of the Word "Árabe" in Three Spanish Generación Del 98 Writers

Mohamed M. Mostafa
GUST, Kuwait;
University of Malaga, Spain

Nicolas Roser Nebot
University of Malaga, Spain

***Abstract*—Although the Generation of '98 writers represents a group of renown Spanish novelists, philosophers, essayists and poets active during the 1898 Spanish-American war, no previous studies have attempted to analyze the diverse linguistic and stylistic features employed by such writers. This study aims to use computational stylometry to detect hidden stylistic and linguistic patterns employed by three Generation of '98 writers, namely Pío Baroja, Vicente Blasco Ibáñez and Miguel de Unamuno. We employ a large corpus comprising 1,702,243 words representing nineteen works by the three writers. Several rigorous criteria were satisfied in designing the corpora such as authorship, genre, topic and register. Concordance, wordclouds, consensus trees, multidimensional and cluster analyses were performed to reveal the different stylistic and linguistic patterns used by the three writers. Although we focus solely on the use of the word "árabe", we show that computational stylometry techniques can be used to help detect hidden stylistic and linguistic patterns employed by different writers. This result is significant since it can help the reader navigate across various possibilities of expressions and terminologies employed by different writers.**

***Index Terms*—corpus linguistics, Generation of '98, stylometric analysis, concordance, consensus trees**

I. Introduction

Generación del 98 or Generation of '98 represents a group of Spanish novelists, philosophers, essayists and poets active during the 1898 Spanish-American war. Outstanding figures of this group include the novelists Pío Baroja (1872-1956), Vicente Blasco Ibáñez (1867-1928), and Ramón María del Valle-Inclán (1866-1936), the philosophers Miguel de Unamuno (1864-1936) and José Ortega y Gasset (1883-1955), and the poets Antonio Machado (1875-1939) and Manuel Machado (1874-1947), among others. The disastrous defeat of Spain in the 1898 war, which resulted in the loss of its last colonies, prompted many writers and philosophers to embark on a soul-searching journey aiming at identifying Spain's ills and problems. Ramsden (1974, p. 465) argued that the Generation of '98 is a generation of "protest against the social, moral and intellectual state of Spain." Padreira (1929, p. 315) states that this generation aimed to "responde a la necesidad surgida ante la crisis de ideales de toda Europa en los últimos años del siglo XIX/to respond to the need arising from the crisis of ideals throughout Europe in the last years of the nineteenth century."

Although the Generation of '98 had shown diverse linguistic and stylistic features, all writers and thinkers who belong to the group had in common a desire to restore the purity and authenticity of the medieval Golden Age in Spain (Entralgo, 1945). Restoring Spain's glory had drove the Generation of '98 writers and thinkers towards utopias, solitude, and individualism. In so doing they enabled the ordinary people to reassess their own values within the context of the modern world.

Computational stylometry plays a major role in the study of style, linguistic analysis and lexicography (Sinclair, 1984). Chapelle (2001, p. 38) states that this relatively new field of study has resulted in creating a "corpus revolution." Hultsijn (1992) argued that a corpus, as opposed to a dictionary, typically calls for deeper processing that enhances the learning process. Miangah (2012) argued that computational stylometry can be used successfully in determining collocations, sub-categorizations, word clusters, which ultimately can be used to validate linguistic hypotheses. Baker (2006) claims that computational stylometry techniques can help formulate new research questions, identify linguistic norms and outliers and remove bias. A major trend in computational stylometry has been the use of corpora in language learning (Miangah, 2012), technical writing (Noguchi, 2004), and translation (Frankenberg-Garcia, 2012).

This study is organized as follows. The next section reviews relevant literature. The following section deals with the methodology employed to conduct the analysis. The subsequent section presents empirical results. Finally, the article sets out some implications and deals with research limitations. In this section we also explore avenues for future research.

II. Literature Review

Several studies have used stylometric techniques to investigate stylistic and linguistic differences among authors and/or texts. For example, using computational stylometry, Botz-Bornstein and Mostafa (2017) analyzed and compared stylistic and linguistic differences in analytic and continental philosophical texts. The authors found that texts belong to each school are distinct stylistically. The authors also concluded that philosophical thought depends on language. Similarly, Nelson (2005) investigated the usage of the terms "global", "international" and "local" in a specialized English corpus. Results revealed a distinct pattern in usage since the term "global" often collocated with phrases like "business activities", whereas "international" collocated with phrases like "companies and institutions." On the other hand, the word "local" collocated mostly with "non-business" terms. The author argued that although the three terms belong semantically to the same class, the word "local" is used usually with non-business activities as compared to the other two terms. In a similar vein, Sayoud (2012) employed a computer-assisted stylometric analysis to investigate "author discriminability between the Holy Quran and prophet's *hadith*." The author concluded that the two texts are stylistically and linguistically distinct and cannot be written by the same "author."

Stylometric techniques have also been employed to detect traces of lexical idiosyncrasies and/or translators' fingerprints. For example, using computational stylometric methods, Rybicki and Heydel (2013) successfully determined the chapter in which one translator took over from the other in a corpus of Polish translations of Virginia Woolf's novels. Similarly, Forsyth and Lam (2014) investigated authorial discriminability in 144 letters written to Vincent van Gogh by his brother Theo when translated from the original French into English. Based on a corpus comprising Chinese translations of James Joyce's *Ulysses*, Wang and Li (2012) investigated translator's style and use of specific linguistic patterns. Results revealed that translators usually "leave some traces of lexical idiosyncrasies that may be detected by analyzing translation corpora." Similarly, Li, Zhang and Liu (2011) used stylometric techniques to detect stylistic differences among different translators of a classical Chinese novel. The authors argued that such differences in style might be attributed to the socio-political, cultural and ideological perspective taken by the translator. Other studies have investigated translators' words choice (Saldanha, 2011) and disfluencies (Straniero-Sergio & Falbo, 2012).

Moreover, stylometric techniques have been used to test stylistic and linguistic hypotheses such as the simplification and the normalization/conventionalization hypotheses. For example, Laviosa (2002, 2011) argued that lexical variety and lexical density are both lower in translated corpora as opposed to original texts corpora. Several studies in different languages have replicated this finding, including Chinese (Xiao, He, & Yue, 2010) and Spanish (Corpas-Pastor, 2008). Several authors have also used computational stylometry to investigate the normalization/conventionalization hypotheses in translated texts in a corpus. For example, Puurtinen (2003) found that translated corpora tend to conform to more conventional rather than creative target strings.

Parallel corpora have also been used extensively to detect stylistic and linguistic differences among different languages. For example, in a stylometric analysis of medical French and English corpora, Deleger, Merkel, and Zweigenbaum (2009) found that even a single English term like "lifelong" may be rendered into French by a whole phrase such as "qui dure toute la vie." In a similar vein, Simo (2011) investigated stylistic differences in "blood" metaphors between Hungarian and English languages. The author found remarkable difference in usage patterns, frequency and connotation and of blood-based metaphors cross-culturally. Other studies using parallel corpora include Schmied's (1998) study comparing stylometrically the German proposition "mit" to the English "with", Cosme and Gilquin's (2008) study also comparing the use of the French "avec" to the English "with" in the *Poitiers-Louvain Échange de Corpus Informatisé* and Perez-Guerra's (2012) study focusing on the translation of the English existential term "there" into Spanish.

From this brief literature review we find that although numerous studies have used computational stylometrics to investigate stylistic and linguistic patterns, virtually no studies have focused on examining the diverse linguistic and stylistic features employed by the Generation of '98 Spanish writers.

III. Methodology

A. Corpus

Cermak (2010) argued that a balanced corpus is essential in stylometric studies. In this study we employ a large corpus comprising 1,702,243 words representing nineteen works by the three writers as shown in Table 1. Several rigorous criteria were satisfied in designing the corpora such as authorship, genre, topic and register (Biber, 1993). The creation of our corpora was facilitated by the availability of vast amount of electronic texts online. The Baroja corpus included 384,957 words, the Ibáñez corpus included 1,118,883 words, while the Unamuno corpus included 198,403 words. The size of our corpora is larger in size than other corpora reported in published studies, including Ferrero's (2011) study (692, 751 words), Merakchi and Rogers' (2013) study (288, 306 words, and Grabowski's (2013) study (705, 460 words).

TABLE 1.
CORPORA SUMMARIES

Author **Corpus size**	**Baroja** 384,957 words	**Ibáñez** 1,118,883 words	**Unamuno** 198,403 words
Works included	El aprendiz de conspirador Los Caminos del Mundo Los Caudillos de 1830 Con la Pluma y con el Sable Los Contrastes de la Vida Las Furias Mala Hierba	La araña negra Arroz y tartana La Catedral Los enemigos de la mujer Entre naranjos La horda La maja desnuda	Abel Sánchez: Una Historia de Pasión Amor y Pedagogía Niebla La Tía Tula Tres novelas ejemplares y un prólogo
Medium	Written	Written	Written
Subject	Literature	Literature	Literature/Philosophy
Language	Spanish	Spanish	Spanish

B. Procedures

Having compiled the Generation of '98 Spanish writers corpora, we focused on preparing the texts for analysis. We started by transforming the original texts html format into plain text format. This step is a prerequisite needed by the text analysis software packages used. All statistical analyses were conducted using both the R Stylo package 0.5.2 (Eder, Rybicki, & Kestemont, 2013) and the AntConc 3.3.5 software (Anthony, 2012). These software packages were selected because of their extensive tools that can be used to handle clusters of words and lexical bundle analyses. Since they include powerful concordance and frequency generators, the packages can also identify hidden patterns in textual data.

IV. RESULTS

A. Frequency Lists

Analyzing a corpus usually stars with generating a word frequency list or simply an incremental count of words in a corpus. Some authors have argued that albeit its simplicity, such approach can provide useful insights regarding the topic analyzed (O'Leary, 201). Similarly, Barlow (2004, p. 207) stated that this step is probably "the most radical transformation of a text used in linguistic analysis." We started by creating frequency lists for the three corpora. Figure 1 shows an example of a histogram for the most frequent terms found in Miguel de Unamuno's corpus.

From figure 1 it is clear that Miguel de Unamuno's corpus is dominated by words such as "que" (7900 times), "los" (1656 times), "con" (1581) times, "por" (1463 times), etc. Romer and Wulff (2010) argued that frequency lists might be more useful compared to alphabetical order lists. This is because the latter usually results in creating a list of function words like "los", "las", or "les", which do not really provide much information about the essence of the corpus. Based on the word frequency lists a type to token ratio may be calculated. This allows for the creation of a lexical variety index in the corpus. It should be noted, however, that such an index is extremely sensitive to corpus length (Kenny, 2001). Figure 2 shows a wordcloud of Miguel de Unamuno's corpus. A wordcloud or a tag cloud is a visual device indicating the frequency of occurrence of a specific word in a document. The higher the frequency of a word, the larger will its presence in the wordcloud.

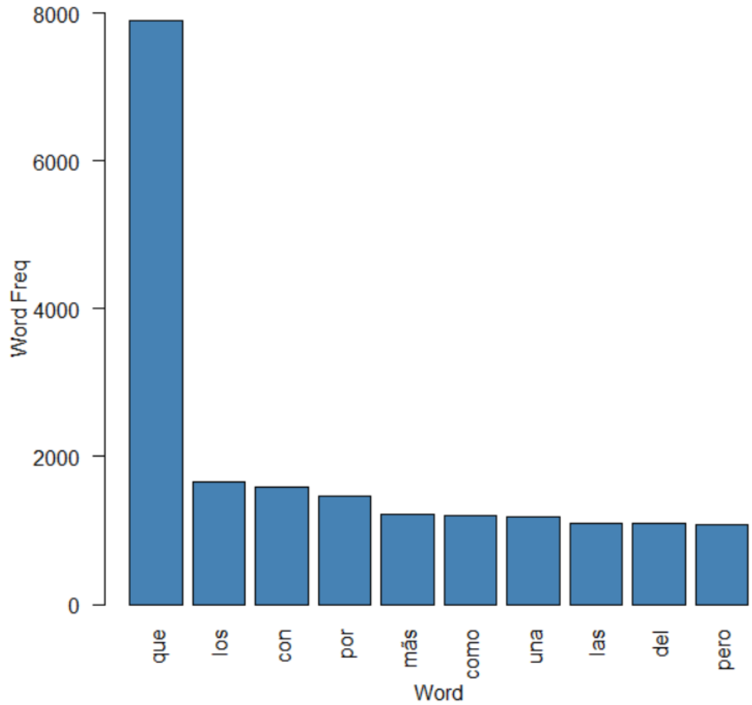


Figure 1. Most frequently used words in Miguel de Unamuno's corpus

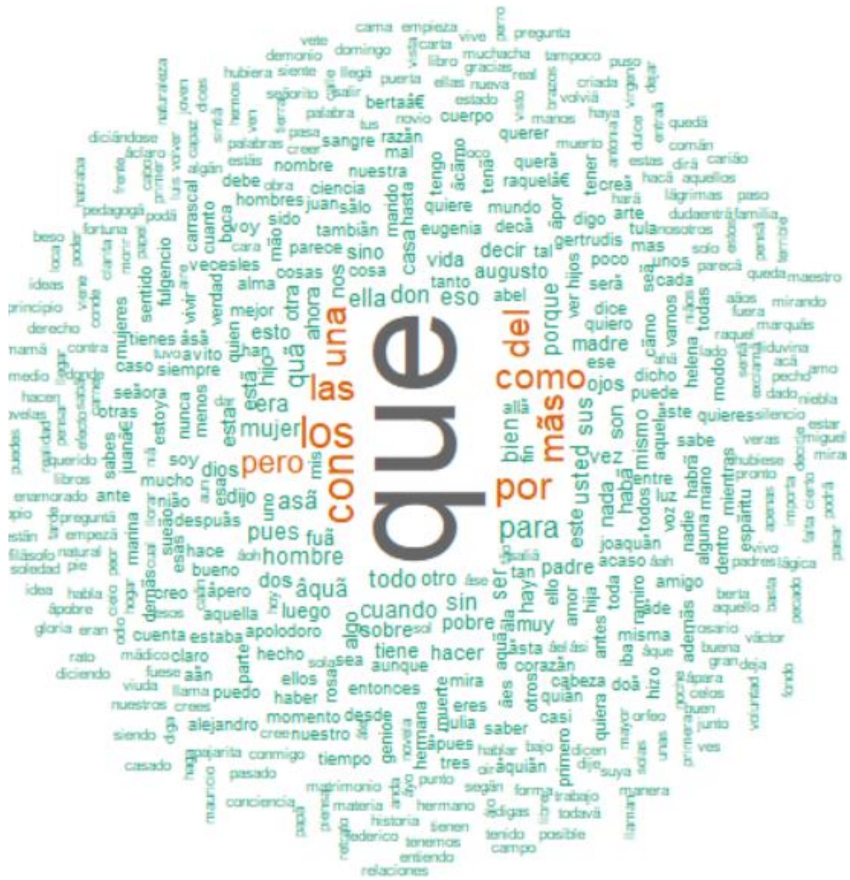


Figure 2. Wordcloud for Miguel de Unamuno's corpus

B. Concordances

Some authors have noted that the context in which a word is used in a corpus makes the reader aware of several linguistic issues, such as frequency-issues, phraseology, register and pragmatics. Such issues are generally not well-documented by traditional dictionaries (Aston, 1999). Contextual word use in a corpus is known as concordance. Barnbrook (1996, p. 65) noted that concordance aims to "place each word back in its original context, so that the details

of its use and behavior can be properly examined". Key word in context (KWIK) is usually used to present a certain term.

Figure 3 shows three examples of concordances produced for the term "árabe" in Baroja's 384,957 words corpus, Ibáñez's 1,118,883 words corpus, and in Unamuno's 198,403 words corpus. From this figure, we clearly see that the search term appears in the middle of the screen, whereas the context is displayed to the left and to the right of the term. This technique saves quite a lot of time going back and forth across the corpus in an effort to determine the contextual relevance of a particular term. The search word or the "node" is read vertically not horizontally. Atkins, Fillmore, and Johnson (2003) argued that such method can help us detect "(1) the syntactic contexts in which the node occurs, (2) the semantic properties of the node's syntactic companions, and (3) the membership of the node in classes of semantically similar words."

(a)

itas esperaban un español, moreno y lánguido, con aire de árabe. A pesar de esta primera impresión, Ribero siguió vi
:s, los necesarios para un hombre que podía vivir como un árabe del Desierto en una tienda de campaña. Sólo tenía e
:ntanas herméticamente cerradas. Antes de llegar al barrio árabe nos detuvimos en una casa baja y muy larga, con ce
u despacho y ha mandado al dragomán que lo traduzca al árabe, y me ha dicho que venga usted conmigo. Fuimos a
e a Chiaramonte y le pedí que me dejara una preciosa jaca árabe que tenía. --Sí, ya lo creo. Le pondré la mejor silla y
a casa. Al día siguiente se habló en Alejandría de la jaca árabe, montada por un oficial de marina inglesa, como de
uilé dos borriquillos y un criado o _zami_: fuimos al barrio árabe y pasamos por la puerta de la Columna. La columna
ás extraña jerigonza que imaginarse puede, una mezcla de árabe y de castellano arcaico que sonaba a algo muy raro.

(b)

lra, y allí, contemplando con el mismo arrobamiento que un árabe soñador las tornasoladas vedijas de azulado hum
cosmopolitismo y allí se codea, el ruso con el brasileño y el árabe con el yanqui. Ese es el París de los cafés, el París
:cción. Al volver un ángulo, apareció Bullier, con su fachada árabe alumbrada por hileras de llameante gas, encerrad
La _falla_ es la fiesta popular por excelencia: una costumbre árabe, transformada y mejorada a través de los siglos h
florido en la del Perdón y la de los Leones; la arquitectura árabe extiende sus graciosos arcos de herradura en el _t
·óximos a desaparecer en el olvido, se salvaban siguiendo al árabe invasor en sus conquistas. Aristóteles reinaba en l
tíficos, etc. ¿Y esto quién lo hizo sino España, aquella España árabe-hebreo-cristiana de los Reyes Católicos? El Gran C
ıcia al descuartizarse el cuerpo joven y robusto de la España árabe, cristiana y hebrea. Tiene usted razón, don Antolín
s. Aquí, por donde ha pasado el arte romano, el bizantino, el árabe, el mudejar, el gótico y el Renacimiento, todas las
os tenerlo todo. En la Sala Capitular, mezcla de arquitectura árabe y gótica, admiraban los visitantes la doble fila de a
on violentas contorsiones. Dudaba entre romper una ánfora árabe, próxima, ó abalanzarse sobre aquella cabeza incli

(c)

muerde, ladra. —Ah, pues haz lo que dice el refrán árabe: «Si vas a detenerte con cada perro que te salga

Figure 3. KWIK concordance for the word "árabe" as used in the corpora of Pío Baroja (a), Vicente Blasco Ibáñez (b) and Miguel de Unamuno (c)

From Figure 3, we see that the word "árabe" has been used eight times in Baroja's 384,957 words corpus. We present here the eight occurrences with their Arabic translation.

1. "... ironía burlona, que el poco éxito de mi amigo Ribero entre las damas dependía de que era rubio, con un tipo común de suizo o de francés, y las señoras y señoritas esperaban un español, moreno y lánguido, con aire de árabe. A pesar de esta primera impresión, Ribero siguió visitando la casa y se hizo amigo de todos."

"ومن سخرية القدر أن النجاح الضئيل الذي حققه صديقي ريبيرو مع السيدات قد اعتمد على كونه أشقر، وهو ما يشيع بين السويسريين أو الفرنسيين، على حين كانت السيدات والفتيات ينتظرن فتى إسباني يتصرف بتلقائية وأسمر اللون مع ملامح عربية. وعلى الرغم من هذا الانطباع الأول، فقد واصل ريبيرو زيارة المنزل وتكوين صداقات مع الجميع."

2. "...los necesarios para un hombre que podía vivir como un árabe del Desierto en una tienda de campaña."

"وهي الضروريات التي تتيح لرجل أن يعيش في خيمة كعربي في الصحراء."

3. "Recorrimos la calle de los Francos y fuimos por una callejuela de casas blancas, con puertas y ventanas herméticamente cerradas. Antes de llegar al barrio árabe nos detuvimos en una casa baja y muy larga, con celosías pintadas de verde."

"نزلنا في شارع الفرنجة، ومضينا عبر زقاق من بيوت بيضاء، مع أبواب ونوافذ مغلقة بإحكام. وقبل أن نصل إلى الحي العربي، توقفنا بمنزل منخفض البناء، وإن كان بالغ الامتداد."

4. "El coronel ha leído su despacho y ha mandado al dragomán que lo traduzca al árabe."

"قرأ الكولونيل رسالته، وأرسل بمترجم لترجمتها إلى العربية."

5. "Si aceptas, si encuentras bien la idea, te proclamarán general en jefe y presidente de la Junta; yo seré tu segundo y mandaré la caballería. ..... Se lo dije a Chiaramonte y le pedí que me dejara una preciosa jaca árabe que tenía. —Sí, ya lo ... Le pondré la mejor silla y arneses, y yo iré también con un caballo muy bonito."

"إذا قبلت ذلك، أي إذا راقتك الفكرة، فسوف يعلنوك قائداً عاماً ورئيساً للمجلس العسكري. وسوف أكون أنا ذراعك اليمنى وأتولى إرسال كتيبة الخيالة... لقد أخبرت شيارامونتي وطلبت منه أن يترك لي مهره العربي الأصيل..."

6. "Al día siguiente se habló en Alejandría de la jaca árabe, montada por un oficial de marina inglesa…"

"في اليوم التالي كان هناك حديث في الإسكندرية عن المهر العربي الذي يمتطيه ضابط بحرية انجليزي..."

7. "... Alquilé dos borriquillos y un criado o zami: fuimos al barrio árabe y pasamos por la puerta de la Columna."

"استأجرت اثنين من البغال، وخادماً، ومضينا إلى الحي العربي ومررنا من باب العمود."

8. "…su suegro y él en la más extraña jerigonza que imaginarse puede; una mezcla de árabe y castellano arcaico que sonaba a algo muy raro."

"جدا غريب شيء وكأنها بدت التي القديمة والقشتالية العربية من خليط؛ تخيلها يمكن رطانة بأغرب تحدث."

From the same figure, we note that word "árabe" has been used eleven times in Ibáñez's 1,118,883 words corpus. We present here the eleven occurrences with their Arabic translation.

1. "Se sentó en un banco de piedra, y allí contemplando con el mismo arrobamiento que un árabe soñador las tornasoladas vedigas de azulado humo que su cigarro arrojaba en el espacio,"

"جلس على كرسي من حجر، وهناك تأمل الأمر كما يتأمل حالم عربي دخان سيجاره الأزرق وهو ينير الفضاء..."

2. "un confuso cosmopolitismo y allí se codea, el ruso con el brasileño y el árabe con el yanqui. Ese es el París de los cafés, el París de los teatrillos desvergonzados, de las bandadas de cocottes, de los restaurants que admirarán a Gargantúa; el París que dice al mundo entero con acento dictatorial…"

"إنها كوزموبوليتانية مشوشة حيث يلتقي الروسي بالبرازيلي، والعربي بالأمريكي. إنها باريس المقاهي، باريس المسارح المشينة، أشجار جوز الهند، والمطاعم التي تثير شهية كل من لديه نهم للطعام، باريس التي تخاطب العالم كله بلهجة دكتاتورية."

3. "Cuando llegaron a la terminación de la avenida del Observatorio, vieron que la concurrencia en el bulevar iba engrosando y que todos marchaban en la misma dirección. Al volver un ángulo, apareció Bullier, con su fachada árabe alumbrada por hileras de llameante gas, encerrado en vasos de colores…"

"وعندما وصلوا إلى نهاية طريق المرصد، أبصروا حشداً كبيراً يسير في نفس الاتجاه. وعندما استداروا عند المنعطف ظهر بولييه، بواجهته العربية المضاءة بزخارف من قناديل الغاز المشتعل، والمغطاة بمشكاوات ملونة .. "

4. "La falla es la fiesta popular por excelencia: una costumbre árabe, transformada y mejorada a través de los siglos hasta convertirse en caricatura audaz."

"الخلا هو العيد الشعبي بلا منازع . إنه تقليد عربي تبدل وتحسن مع مرور القرون ليصبح بمثابة كاريكوتور جريء."

5. "la arquitectura árabe extiende sus graciosos arcos de herradura en el trifono, que corre por todo el ábside tras el altar mayor, siendo obra de Cisneros, que quemaba los libros de los musulmanes y restablecía ..."

"تمد العمارة العربية أقواسها الأنيقة المتدلية على شكل حدوة حصان في التريتون، الذي يمتد عبر الجزء العلوي من المذبح الرئيسي، وهو من عمل سيسنيروس الذي أحرق كتب المسلمين واستعاد..."

6. "Los filósofos griegos, próximos a desaparecer en el olvido, se salvaban siguiendo al árabe invasor en sus conquistas. Aristóteles reinaba en la famosa Universidad de Córdoba. Nacía el espíritu caballeresco entre los árabes españoles, apropiándoselo después los guerreros del Norte…"

"لقد حافظ الفتح العربي على الفلاسفة الإغريق ، الذين سرعان ما اختفوا في غياهب النسيان، وهكذا ساد أرسطو جامعة قرطبة الشهيرة. وهكذا ولدت روح الفروسية بين العرب الأسبان ، وهي الروح التي اقتبسها محاربو الشمال فيما بعد..."

7. "Sociedad nueva, con cultivos, industrias, ejércitos, conocimientos científicos, etc. ¿Y esto quién lo hizo sino España, aquella España árabe-hebreo-cristiana de los Reyes Católicos? El Gran Capitán enseñó al mundo el arte de guerrear moderno; Pedro Navarro fue un ingeniero asombroso…"

"إنه مجتمع جديد مع زراعة وصناعة وجيوش ومعرفة علمية وما إلى ذلك. لكن من فعل ذلك غير إسبانيا العربية ـ العبرية ـ المسيحية تحت إمرة الملوك الكاثوليك ؟ لقد علم القبطان العظيم العالم فن الحرب الحديثة، وكان بدرو نافارو مهندساً مرموقاً."

8. "Pero antes de morir los Reyes Católicos ya empieza la decadencia al descuartizarse el cuerpo joven y robusto de la España árabe, cristiana y hebrea. Tiene usted razón, don Antolín: por algo se llamaban Católicos aquellos reyes. Establece la Inquisición doña Isabel con su fanatismo de hembra."

"ولكن وقبل وفاة الملوك الكاثوليك ، بدأ عصر الانحطاط بتشويه الجسد الفتي القوى لإسبانيا العربية والمسيحية والعبرية . أنت على حق يا دون أنطولين: لقد كان هناك سبب لتسمية أولئك بالملوك الكاثوليك. لقد تأسست محاكم التفتيش على يد دونا إيزابيل بتعصبها الأنثوي."

9. "Aquí, por donde ha pasado el arte romano, el bizantino, el árabe, el mudejar, el gótico y el Renacimiento, todas las artes de Europa…"

"هنا تغلغل الفن الروماني والبيزنطي والعربي والمدجن والقوطي ، فضلاً عن فنون عصر النهضة، في كافة فنون أوربا."

10. "En la sala capitular, mezcla de arquitectura árabe y gótica, admiraban los visitantes la doble fila de arzobispos toledanos pintados en la pared con mitras y báculos de oro. Gabriel llamaba la atención sobre don Cerubruno, el prelado medieval llamado así por su enorme cabeza."

"وفي الصالة الرئيسية، وهي مزيج من العمارة العربية والقوطية، أبدي الزوار إعجابهم بالصفين المزدوجين لرؤساء أساقفة طليطلة المرسومين على جدار القصر المطلي بالذهب. ولفت غابرييل الانتباه إلى صورة دون سيروبرونو، أسقف القرون الوسطى ، الذي سمى كذلك لضخامة رأسه."

11. "Violentas contorsiones. Dudaba entre romper una ánfora árabe, proxima, o abalanzarse sobre aquella Cabeza."

"تردد بين كسر القارورة العربية المجاورة ، أو الانقضاض على تلك الرأس."

Finally, we note that word "árabe" has appeared just once in Unamuno's 198,403 words corpus. We present here this occurrence with its Arabic translation.

1. "Ah, pues haz lo que dice el refrán árabe: Si vas a detenerte con cada perro que te salga él ladrar al camino; nunca llegarás al fin de él."

"آه ، إعملي إذاً بما يقوله المثل العربي :'إن توقفت مع كل كلب يخرج لينبح عليك في الطريق، فلن تصل أبداً إلى نهايته.'"

From the concordance analysis, we see that the word "árabe" has been used only twenty times across a large corpus comprising 1,702,243 words representing nineteen works by the three Generation of '98 Spanish writers. Baroja's usage of the word focuses on some stereotypical characteristics like tents in the desert, the Arabic quarter in a city, or the Arab horse. Ibáñez's 1,118,883 words corpus uses the word "árabe" to signify the important intellectual contribution played by the Arabs in the history of Spain. Thus, the word is used in some historical and architectural contexts. The writer also highlighted the role of the Arabs in preserving the Greek's philosophical heritage in Spain and their influence on Spanish culture and music. Finally, Unamuno only used the word "árabe" only once. Not surprisingly, he used it within a philosophical context by referring to an Arab proverb in his philosophical novel "Niebla" or "Fog."

C. Multidimensional Scaling and Principal Component Analyses

Borg and Groenen (1997) noted that multidimensional scaling (MDS) can be used to visually detect complex patterns in high-dimensional datasets. MDS shows the structure of distance-type data in a two-dimensional graph by arranging points in space based on similarities between different objects. Following Cha, Kim, and Lee (2009), we used MDS to map the relationships among sub-corpora through the construction of a low k-dimensional space based on perceived similarities or dissimilarities among the set of sub-corpora. The alternating least squares approach to scaling (ALSCAL) algorithm (Zsoka, Szerenyi, Szechy, & Kocsis, 2013) is used in this study since this algorithm has been shown to optimally compute the Euclidean distances between objects in the k-dimensional space.

Figure 4 shows the resulting MDS for the three Generation of '98 writers corpora. From this figure, we see that the bottom left hand corner includes all the novels written by P ó Baroja (El aprendiz de conspirador, Los Caminos del Mundo, Los Caudillos de 1830, Con la Pluma y con el Sable, Los Contrastes de la Vida, Las Furias and Mala Hierba). The uper left hand corner includes all the novels written by Vicente Blasco Ib áñez (La ara ña negra, Arroz y tartana, La Catedral, Los enemigos de la mujer, Entre naranjos, La horda, and La maja desnuda). Finally, the right-hand corner is dominated by Miguel de Unamuno's works (Abel S ánchez: Una Historia de Pasi ón, Amor y Pedagog á, Niebla, La T á Tula, and Tres novelas ejemplares y un pr ólogo). This result confirms the fact that although the three authors belong to the Generation of '98, every author had shown different and distinct linguistic and stylistic features. The principal components graph shown in Figure 5 demonstrates basically the same results.

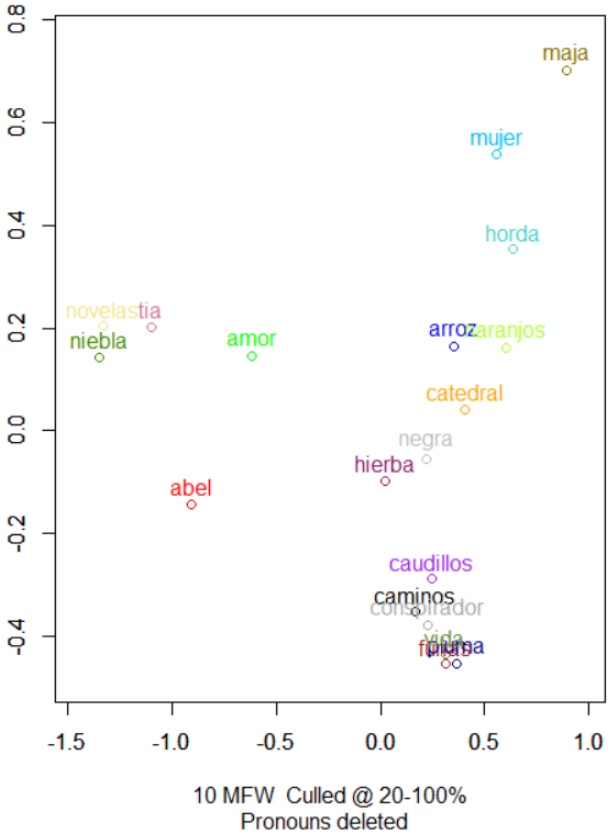


Figure 4. Multidimensional scaling (MDS) for the corpora used in the study

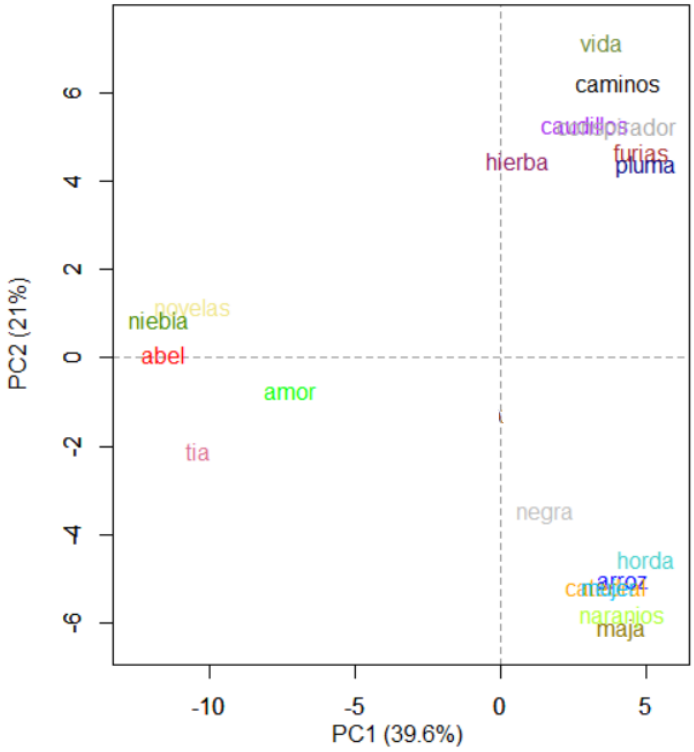
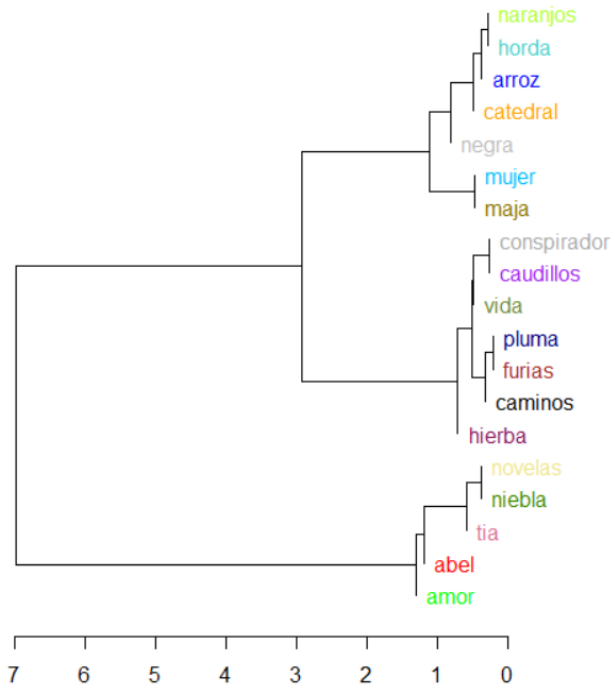


Figure 5. Principal components analysis (PCA) for the corpora used in the study

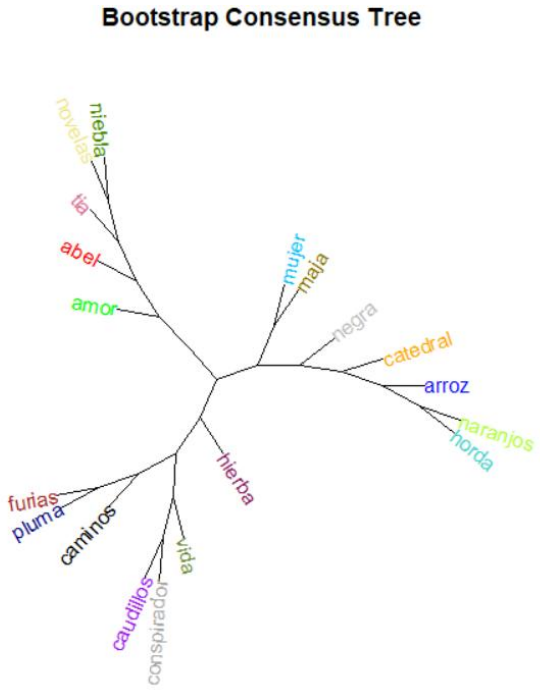
D. Cluster Analysis and Consensus Trees

To cluster sub-corpora of the three Generation of '98 Spanish writers, the Ward's method was used. This method generates a set clusters based on proximity of sub-corpora, which allows the detection of which sub-corpora were the most similar. A dendogram showing how the sub-corpora clusters are formed is shown in Figure 6. From this figure, we can detect three clearly distinguished clusters. The first cluster comprises all five novels by Miguel de Unamuno (Abel Sánchez: Una Historia de Pasión, Amor y Pedagogía, Niebla, La Tía Tula, and Tres novelas ejemplares y un prólogo). The second cluster includes all seven novels by Pío Baroja (El aprendiz de conspirador, Los Caminos del Mundo, Los Caudillos de 1830, Con la Pluma y con el Sable, Los Contrastes de la Vida, Las Furias and Mala Hierba), whereas the third cluster includes all seven novels by Vicente Blasco Ibáñez (La araña negra, Arroz y tartana, La Catedral, Los enemigos de la mujer, Entre naranjos, La horda, and La maja desnuda). This result confirms the results of other statistical techniques used such as the PCA and the MDS. Thus, it seems that each of the three authors shows different and distinct linguistic and stylistic features.

Finally, a Delta-normalized bootstrapped cluster analysis was used to generate a consensus tree (Hoover, 2004). This tree shows distances between the three Generation of '98 Spanish writers sub-corpora. In this study we used the similarity between sequences of most-frequent-word frequencies (MFW) to generate the bootstrapped consensus tree shown in Figure 7. Burrows (2002) has shown that bootstrapping can alleviate several problems attributed to the original Delta-normalized method. Following Rybicki and Heydel (2013), personal pronouns were removed to avoid possible false attributions. From Figure 7 it is clear that three branches are formed for the three authors. Each branch includes all the novels by the relevant author, which again confirms the distinct stylistic and linguistic patterns used by each author.



10 MFW Culled @ 20-100%
Pronouns deleted Classic Delta distance
Figure 6. Hierarchical cluster analysis for the corpora used in the study



10-10 MFW Culled @ 20-100%
Pronouns deleted Classic Delta distance Consensus 0.5
Figure 7. Bootstrap consensus tree for the corpora used in the study

V. Conclusions and Implications

By performing important tasks such as determining word clusters, concordances, sub-categorizations, stylometry plays a major role in the study of style and linguistic patterns. Stylometry can also be used to validate linguistic and stylistic hypotheses. In this study, we used corpora of nineteen works representing three Spanish Generation of '98

writers to investigate their stylistic and linguistic differences. Our corpora were designed to satisfy several rigorous criteria such as genre, register, authorship, and topic. We argue that our computational stylometric approach might help in obtaining context-specific information regarding syntactic and semantic usage of the term "árabe" by Spanish Generation of '98 writers.

Translators can exploit the corpora used in this study in several ways. For example, they can refer to concordances to find a suitable translation for a particular term. Miangah (2012) argued that "adjectives that collocate with nouns have been proven to be very useful in understanding the context." This is particularly true when traditional dictionaries do not suggest a suitable translation. Boulton (2012) shows the inherent limitations of traditional dictionaries as opposed to corpora using the following French example «Je suis paralysé entre **le brûlot** et la chanson d'amour.» A dictionary offers the following possible meanings for the term "**le brûlot**", "fire ship", "pamphlet", or "gnat". However, the author used a large corpus to show that a good translation would be "rebel", "revolutionary", or "protest". In fact, this is what Renaud, the faous French singer, is famous for. Thus, Wright (1993, p. 70) noted that "documents must speak 'the language' of the target audience and should resemble other texts produced within that particular language community and subject domain. These considerations frequently require that translators move beyond merely correct strategies in terms of lexical and grammatical content in order to account for stylistically appropriate solutions." This is probably true since a translator is basically a text producer. Thus, in the first place, a translator should be able to envisage how words are used and how they relate to other words in a particular context.

REFERENCES

[1] Anthony, L. (2012). AntConc (Version 3.3.5). Tokyo, Japan: Waseda University.
[2] Aston, G. (1999). Corpus use and learning to translate. *Rivista dell'Associazione Italiana di Anglistica*, 12, 289-314.
[3] Baker, M. (2006). Using corpora in discourse analysis. Continuum: London and New York.
[4] Barlow, M. (2004). Software for corpus access and analysis. In J. Sinclaire (Ed.), *How to use corpora in language teaching*, John Benjamins: Amsterdam, 205-221.
[5] Barnbrook, G. (1996). Language and computers. Edinburgh University Press: Edinburgh.
[6] Biber, D. (1993). Using register diversified corpora for general language studies. *Computational Linguistics*, 2, 219-241.
[7] Borg, I., & Groenen, P. (1997). Modern multidimensional scaling. Berlin, Germany: Springer.
[8] Botz-Bornstein, T., & Mostafa, M. (2017). A corpus-based computational analysis of philosophical texts: Comparing analytic and continental philosophy. *International Journal of Social and Humanistic Computing*, 2, 230-246.
[9] Boulton, A. (2012). Beyond concordancing: Multiple affordances of corpora in university language degrees. *Procedia - Social and Behavioral Sciences*, 34, 33-38.
[10] Burrows, J. (2002). The Englishing of Juvenal: Computational stylistics and translated texts. *Style*, 36, 677-699.
[11] Cha, J., Kim, S., & Lee, Y. (2009). Application of multidimensional scaling for marketing-mix modification: A case study on mobile phone category. *Expert Systems with Applications*, 36, 4884-4890.
[12] Chapelle (2001). Computer applications in second language acquisition: Foundations for teaching, testing, and research. Cambridge University Press, Cambridge, UK.
[13] Cosme, C., & Gilquin, G. (2008). Free and bound propositions in a contrastive perspective: The case of with and avec. In F. Meunier and S. Granger (Eds.). *Phraseology: An interdisciplinary perspective.* John Benjamins: Amsterdam, 259-274.
[14] Deleger, L., Merkel, M., & Zweigenbaum, P. (2009). Translating medical terminologies through word alignment in parallel text corpora. *Journal of Biomedical Informatics*, 42, 692-701.
[15] Eder, M., Rybicki, J., & Kestemont, M. (2013). R Stylo package, version 0.5.2.
[16] Entralgo, L. (1945). La generación del noventa y ocho. Madrid.
[17] Forsyth, R., & Lam, P. (2014). Found in translation: To what extent is authorial discriminability preserved by translators? *Literary and Linguistic Computing*, 29, 199-217.
[18] Frankenberg-Garcia, A. (2012). Learners' use of corpus examples. *International Journal of Lexicography*, 25, 273-296.
[19] Hoover, D. (2004). Testing Burrow's Delta. *Literary and Linguistic Computing*, 19, 453-475.
[20] Hultsijn, J. (1992). Retention of inferred and given word meanings: Experiments in incidental vocabulary learning. In P. Arnaud and H. Bejoint (Eds.), *Vocabulary and Applied Linguistics*, Macmillan: London.
[21] Kenny, D. (2001). Lexis and creativity in translation: A corpus-based study. St. Gerome: Manchester, UK.
[22] Laviosa, S. (2002). Corpus-based translation studies: Theories, findings, applications. Rodopi: Amsterdam and New York.
[23] Laviosa, S. (2011). Corpus linguistics and translation studies. In V. Viana, S. Zyngier & G. Barnbrook (Eds.). *Perspectives on corpus linguistics.* John Benjamins: Amsterdam and Philadelphia.
[24] Li, D., Zhang, C., & Liu, K. (2011). Translation style and ideology: A corpus-assisted analysis of two English translations of Hongloumeng. *Literary and Linguistic Computing*, 26, 153-166.
[25] Miangah, T. (2012). Different aspects of exploiting corpora in language learning. *Journal of Language Teaching and Research*, 3, 1051-1060.
[26] Nelson, M. (2005). Semantic associations in business English: A corpus-based analysis. *English for Specific Purposes*, 25, 217-234.
[27] Noguchi, J. (2004). A genre analysis and mini-corpora approach to support professional writing by nonnative English speakers. *English Corpus Studies*, 11, 101-110.
[28] O'Leary, D. (2011). Blog mining-review and extensions: "From each according to his opinion". *Decision Support Systems*, 51, 821-830.
[29] Padreira, A. (1929). La generacion del 98, *Revista de las Espanas*. 4, 315-320.

[30] Perez-Guerra, J. (2012). A contrastive analysis of (English) "there" and (Spanish) "hay" existential sentences. *Languages in Contrast*, 12, 139-164.
[31] Puurtinen, T. (2003). Genre-specific features of translationese? Linguistic differences between translated and non-translated Finnish children's literature. *Literary and Linguistic Computing*, 18, 389-406.
[32] Ramsden, H. (1974). The Spanish "Generation of 1898": The history of the concept. *Bulletin of the John Rylands Library*, 56, 463-491.
[33] Romer, U., & Wulff, S. (2010). Applying corpus methods to written academic texts: Explorations of MICUSP. *Journal of Writing Research*, 2, 99-127.
[34] Rybicki, J., & Heydel, M. (2013). The stylistics and stylometry of collaborative translation: Woolf's Night and Day in Polish. *Literary and Linguistic Computing*, 28, 708-717.
[35] Saldanha, G. (2011). Style of translation: The use of source language words in translations by Margaret Jull Costa and Peter Bush. In A. Kruger, K. Wallmach, & j. Munday (Eds.). *Corpus-based translation studies: Research and applications*. Continuum: London, 237-258.
[36] Sayoud, H. (2012). Author discrimination between the Holy Quran and prophet's statements. *Literary and Linguistic Computing*, 27, 427-444.
[37] Schmied, J. (1998). Differences and similarities of close cognates: English with and German mit. In S. Johanson and S. Oksefjell (Eds.). *Corpora and cross-linguistic research: Theory, method, and case studies*. Rodopi: Amsterdam, 255-275.
[38] Simo, J. (2011). Metaphors of blood in American English and Hungarian: A cross-linguistic corpus investigation. *Journal of Pragmatics*, 43, 2897-2910.
[39] Sinclair, J. (1984). Lexicography as an academic subject. In R. Hartmann (Ed.). *LEXeter 83 Proceedings*. Max Niemeyer Verlag: Tubingen, 3-12.
[40] Straniero-Sergio, F. & Falbo, C. (2012). Breaking grounds in corpus-based interpreting studies. Peter Lang: Bern.
[41] Wang, Q., & Li, D. (2012). Looking for translator's fingerprints: A corpus-based study on Chinese translations of Ulysses. *Literary and Linguistic Computing*, 276, 81-93.
[42] Wright, S. (1993). The inappropriateness of the merely correct: Stylistic considerations in scientific and technical translation. In S. Wright and L. Wright, Jr. (Eds.). *Scientific and technical translation*. John Benjamins: Amsterdam, 69-86.
[43] Xiao, R., He, L., & Yue, M. (2010). In pursuit of the third code: Using the ZJU corpus of translational Chinese in translation studies. In R. Xiao (Ed.). *Using corpora in contrastive and translation studies*, Cambridge Scholars: Newcastle, 182-214.
[44] Zsoka, A., Szerenyi, Z., Szechy, A., & Kocsis, T. (2013). Greening due to environmental education? Environmental knowledge, attitudes, consumer behavior and everyday pro-environmental activities of Hungarian high school and university students. *Journal of Cleaner Production*, 48, 126-138.

Mohamed M. Mostafa is a doctoral student in Spanish/Arabic Translation at the University of Malaga, Spain. He has also earned an MA in Translation Studies from the University of Portsmouth, UK, an MS in Applied Statistics from the University of Northern Colorado, USA, an MSc in Functional Neuroimaging from Brunel University, UK, an MSc in Social Science Data Analysis from Essex University, UK, an MBA and a BSc at Port Said/Suez Canal University, Egypt. He was employed at universities in the USA, Portugal, Egypt, Cyprus, Turkey, France, Jordan, United Arab Emirates, Bahrain and Kuwait. His current research interests include data mining, social networks analysis, artificial intelligence applications and translation studies.

Nicolás Roser Nebot obtained a doctorate degree from the Autonomous University of Madrid in 1997. His thesis title is "Politics and Religion: the Islamic Concept". He has worked as a professor of Arabic language and translation at the University of Malaga since 1991, where he becomes Senior Lecturer in 2001. He specializes in the field of Specialized Translation Arabic/Spanish/Arabic. He was granted an award of Excellence in 2001 for this PhD thesis from the Autonomous University of Madrid. He is an expert in the topic of the political theory of Islam, the didactics of Arabic as a foreign language, and the translation of the authoritative Islamic texts as well as in the translation of manuscripts and classical Arabic texts. He has supervised, and is currently supervising, several doctoral theses on subjects related to his research.