Investigating the Relation between Syntactic Complexity and Subgenre Distinction

A Case Study on Two Contemporary French Authors

Robert Hesselbach

Abstract The purpose of this article is to explore the ways in which the analysis of the syntactic complexity of a text's sentences can help distinguish between works belonging to different literary subgenres written by the same author. Based on the considerations of an earlier study (Hesselbach 2019), syntactic complexity is understood here as an array of qualitative as well as quantitative features. Applying this method to a corpus of two contemporary French authors and their novels (1979–2002), comprising both crime fiction (*roman policier*) and 'high literature' (*littérature blanche*), the results of this study show that syntactic complexity has very little influence on genre distinction, at least for the two subgenres examined. In fact, a very stable distribution of results can be observed in both qualitative and quantitative terms. Furthermore, the evidence suggests that the degree of syntactic complexity is more likely to appear as an author-related characteristic.

Keywords syntactic complexity, (sub)genre distinction, French novel, Yasmina Khadra, Jean Echenoz

1. Introduction

The present study was conducted within the framework of the research group *CLiGS (Computational Literary Genre Stylistics)* at the University of Würzburg/Germany, one of whose objectives was to develop computer-based methods to identify

Hesselbach, Robert: Investigating the Relation between Syntactic Complexity and Subgenre Distinction. A Case Study on Two Contemporary French Authors, in: Hesselbach, Robert, et al. (Eds.): Digital Stylistics in Romance Studies and Beyond, Heidelberg: Heidelberg University Publishing, 2024, pp. 235–259.
DOI: https://doi.org/10.17885/heiup.1157.c19374

genre-specific characteristics working with Romance, and more precisely French and Spanish, literature. Genre distinction from a digital point of view has been the subject of a number of recent research publications, as is evident from papers on stylometric analysis of ancient Greek literary texts (Gianitsos et al. 2019) or German novels using word frequencies and character tetragrams (Hettinger et al. 2016), human versus machine genre classification of Spanish novels (Calvo Tello 2021), sentiment analysis for Spanish American novels (Henny-Krahmer 2018), or topic modeling on French Classical and Enlightenment drama (Schöch 2017). This paper is concerned with a subject which can be situated at the intersection of literary studies and linguistics: the relation between syntactic complexity and (sub)genre distinction. In this context, it can be noted that very different research questions are associated with the concept of syntactic complexity on the one hand and the task of genre distinction on the other hand. Nevertheless, an attempt will be made at this point to identify certain aspects of syntactic complexity, as already defined in an earlier study (cf. Hesselbach 2019), on the basis of a corpus of two contemporary French authors and their novels, comprising both crime novels (*roman policier*) and 'high literature' (*littérature blanche*). Subsequently, it may be possible to make statements as to whether certain subgenres of French authors exhibit common features in terms of syntax.

The next section will first discuss which approaches already exist to describe syntactic complexity and which (qualitative as well as quantitative) perspective is used in this study. Section 3 then presents the actual empirical study of two contemporary French authors (Yasmina Khadra and Jean Echenoz), focusing first on the description of the corpus-based method before presenting the results and situating them in the research context. In the last section, after a brief synopsis, further possible research perspectives are presented.

2. Approaches to *Syntactic Complexity*

If one tries to approach the concept of *syntactic complexity*, one will find that there are various ways to approach the issue from a conceptual perspective. The first problem concerns the level of the linguistic system: does the expression refer to the complexity of phrases or entire sentences? In this article we refer only to *syntactic complexity* in the sense of sentence complexity, although looking at the complexity of phrases from a stylistic point of view also opens up interesting research perspectives. Another fundamental question is whether one refers to quantitative or qualitative aspects of complexity. The shortest complex sentence in Spanish as a pro-drop language can consist of only two words, as in Sp. *oigo cantar* 'I hear (somebody) singing' whereas simple sentences can turn out to be fairly long, as the French example in (1) illustrates:[1]

(1) En chemin, dans le crépuscule, elle *nomma* le palais de justice, la sous-préfecture, la mairie, la prison, la maison natale de Frédérick Lemaître. (E_Eq_po, 17)[2]
‘On the way, in the twilight, she named the courthouse, the sub-prefecture, the town hall, the prison, the birthplace of Frédérick Lemaître.’

The above sentence shows no kind of coordination nor subordination and consists of 23 words. This should make clear that one can make different statements regarding the *type* (qualitative aspects) and *extent* (quantitative aspects) of a construction when speaking of *syntactic complexity*. For this reason, both qualitative and quantitative aspects will be taken into consideration in this article. After a brief overview of different qualitative and quantitative approaches to (measure) *syntactic complexity*, section 3 presents the method and results of the study presented here.

2.1 Qualitative Aspects

Taking a closer look at (not only) French grammars and manuals, complexity is usually understood as a syntactic hierarchical relationship between two sentences/clauses, so that consequently a distinction is made between the *complex* and the *simple* sentence. While examples (2) and (3) are each syntactically simple sentences, (4) and (5) are correspondingly complex sentences (examples taken from Kiesler 2013, 613; originally in: Dubois et al. 1994, s.v. *parataxe*):

(2) Cet homme *est* habile.
‘This man is clever.’

(3) Il *réussira*.
‘He will succeed.’

[1] Since finite verbs can be regarded as a quantitatively relevant feature of *syntactic complexity*, they are italicized in all given examples, unless another linguistic unit is to be emphasized.

[2] The information at this point refers to texts of the corpus examined. The pattern “author_title_subgenre” is used, so that this refers to the 17th sentence randomly taken from a novel by Jean Echenoz with the title *L’Équipée malaise*, which can be assigned to the subgenre crime novel (*policier* = po).

(4) Cet homme *est* habile et il *réussira*.
'This man is clever and he will succeed.'

(5) Cet homme *réussira*
parce qu'il *est* habile.
'This man will succeed because he is clever.'

Example (4) represents a paratactic structure due to the coordination of the two main clauses, whereas (5) represents a hypotactic construction of main and subordinate clauses, where the hierarchical distinction is made clear by indenting the dependent subordinate clause.[3] Nevertheless, these representations in grammars and manuals are rather prototypical. In actual language use, and thus also in the literary production of novels, hybrid forms of these construction possibilities occur, as Kiesler (2013) points out based on a French corpus. The author presents a typology of complex sentences which distinguishes between (multiple) homogeneous and heterogeneous structures and serves as a basis for this study, as will be explained below (a–g). An example is given from Kiesler (2013), as well as from the corpus of novels analyzed here:

(a) homogeneous parataxis (ho-pa)[4]

A *homogeneous parataxis* (Fr. *phrase parataxique homogène*) is understood to be those cases in which only two main clauses are combined and neither of the two has its own dependent subordinate clauses, as illustrated by examples (6) and (7). It becomes clear that coordination can be carried out both syndetically (i.e. through a conjunction), as shown in example (6), and asyndetically (i.e. without any conjunction), as in (7):

(6) Je me *plaisais* et je *cherchais* à plaire. (Kiesler 2013, 617)
'I liked myself and I liked to please.'

[3] Within his theory of *Junktion* ('linkage') Raible (1992) refers to the fact that, at this stage of clause-linkage, there is basically freedom of position of the different clauses, so that no change of meaning is caused in the following correspondences of the given examples: (4') *Cet homme réussira et il est habile* and (5') *Parce que cet homme est habile, il réussira*. Restrictions on the freedom of position exist in cases where an element of the second sentence refers to the first sentence, cf. (4") *Cet homme est habile et* c'est pourquoi *il réussira* ('This man is clever and *that is why* he succeeds') versus **Cet homme réussira et c'est pourquoi il est habile* ('This man will succeed and that is why he is clever').

[4] Kiesler himself speaks of "phrases parataxiques simples" – "simple paratactic sentences" (2013, 616), but for the sake of a coherent terminology, I prefer to denominate this construction as *homogeneous parataxis*.

(7) Ses doigts *tripotent* dangereusement l'instrument de mort,
le *ramassent*. (K_Do_po, 100)
'His fingers are dangerously touching the instrument of death, picking it up.'

(b) heterogeneous parataxis (he-pa)

In contrast to the previous sentence type, the *heterogeneous parataxis* (Fr. *phrase parataxique hétérogène*) consists of a combination of two main clauses, at least one of which contains a dependent subordinate clause.

(8) Obélix ne *veut* pas finir son sanglier,
il *dit*
qu'il n'*a* plus faim! (Kiesler 2013, 614)
'Obelix doesn't want to finish his boar, he says he's not hungry anymore.'

(9) Il *vérifia* la monnaie déposée d'avance sur la table
et
plia son journal
sans quitter des yeux le secrétaire. (E_Me_po, 85)
'He checked the change left on the table and folded his newspaper without taking his eyes off the secretary.'

In example (8) the second main clause (*il dit*) contains a dependent subordinate clause with a finite verb (*qu'il n'a plus faim*). The following example (9) demonstrates that these dependent subordinate clauses can also be realized through infinite verbal structures (*sans quitter des yeux le secrétaire*).

(c) multiple homogeneous parataxis (mu-ho-pa)

A *multiple homogeneous parataxis* (Fr. *phrase parataxique multiple homogène*) presents a coordinated sequence of at least three main clauses, none of which has a dependent subordinate clause, as can be observed in (10) and (11):

(10) Je *regardais*,
je *palpais*,
j'*apprenais* le monde, à l'abri. (Kiesler 2013, 614)
'I watched, I palpated, I learned about the world, sheltered.'

(11) Omar *vibre* des épaules,
passe une grosse langue sur ses lèvres
et
fait tinter ses bagues sur son comptoir. (K_Mo_po, 199)
'Omar shivers his shoulders, runs a big tongue across his lips and jingles his rings on his counter.'

In (10) a sequence of three main clauses (*je regardais*; *je palpais*; *j'apprenais le monde*) can be observed, all of which are linked without any conjunction. Example (11) basically works in a similar way and consists of three main clauses, but in this case the last one (*fait tinter ses bagues sur son comptoir*) is connected by the conjunction *et*.

(d) multiple heterogeneous parataxis (mu-he-pa)

Analogous to the properties of the *heterogeneous parataxis*, a *multiple heterogeneous parataxis* (Fr. *phrase parataxique multiple hétérogène*) is a string of three or more independent main clauses in which at least one dependent subordinate clause must occur, as (12)[5] and (13) illustrate:

(12) [alors il y *avait* des défilés]
[il y *avait* des tas de trucs]
et
[tout ça se *passait*
d'où je *travaillais*
parce que c'*était* juste en face] (Kiesler 2013, 614)
'so there were parades and there was a lot of stuff and it all happened from where I worked because it was right across the street'

(13) Les prières s'*émiettent* dans la furie des mitrailles,
les loups *hurlent* chaque soir à la mort,
et
le vent,
lorsqu'il se *lève*,
livre la complainte des mendiants au croassement des corbeaux.
(K_Hi_bl, 50)
'Prayers crumble in the fury of the machine-gun fire, the wolves howl, and the wind, when it rises, delivers the lament of the beggars to the crows' caws.'

[5] In this example, the bracketing has been taken from Kiesler's original example.

In the case of (12) a string of three main clauses can be seen, the last of which contains two dependent subordinate clauses (*d'où je travaillais*; *parce que c'était juste en face*). Example (13), which is taken from Yasmina Khadra's novel *Les Hirondelles de Kaboul*, also contains three main clauses, the third of which (*le vent livre la complainte des mendiants au croassement des corbeaux*) again contains a dependent subordinate clause, more precisely an adverbial clause (*lorsqu'il se lève*).

(e) simple hypotaxis (sim-hy)

The first type of hypotactic constructions is the so-called *simple hypotaxis* (Fr. *phrase hypotaxique simple*). It consists only of a main clause and a subordinate clause dependent on that main clause, as can be seen in (14) and (15):

(14) Quand le chat n'*est* pas là,
les souris *dansent*. (Kiesler 2013, 619)
'When the cat's away, the mice dance.'

(15) Mais trois heures plus tard,
lorsque Georges *rentra* rue Oberkampf,
Véronique n'*était* toujours pas là. (E_Ch_po, 192)
'But three hours later, when Georges returned to rue Oberkampf,
Véronique was still not there.'

In example (14), the whole construction begins with the subordinate clause (*quand le chat n'est pas là*) which is followed by the main clause (*les souris dansent*), whereas in (15) the adverbial clause (*lorsque Georges rentra rue Oberkampf*) is inserted in the main clause (*mais trois heures plus tard Véronique n'était toujours pas là*).

(f) multiple homogeneous hypotaxis (mu-ho-hy)

Compared to the type of the *simple hypotaxis*, the *multiple homogeneous hypotaxis* (Fr. *phrase hypotaxique multiple homogène*) is also characterized by a main clause with a hypotactic structure, but in this case it has several dependent clauses which all belong to the same type of subordinate clause, as the following two examples illustrate:

(16) Geoffroy *a* un papa très riche
[qui lui *achète* tous les jouets
[qu'il *veut*.]] (Kiesler 2013, 619)
'Geoffroy has a very rich daddy who buys him all the toys he wants.'

(17) Il *subodorait* l'imminence d'une révolution
qui ne *pardonnerait* rien à ceux
qui ne *prendraient* pas le train en marche. (K_Re_bl, 140)
'He sensed the imminence of a revolution that would not forgive those
who did not get on board.'

Besides the main clause (*Geoffroy a un papa très riche*), example (16) reveals two dependent subordinate clauses (*qui lui achète tous les jouets*; *qu'il veut*), which can both be characterized as relative clauses. The following sentence taken from Yasmina Khadra's *À quoi rêvent les loups* also consists of one main (*Il subodorait l'imminence d'une révolution*) and two subordinated clauses (*qui ne pardonnerait rien à ceux*; *qui ne prendraient pas le train en marche*). Even if the second subordinate clause does not depend on the first subordinate clause, it still belongs to the same type of subordinate clause, namely a relative clause.

(g) multiple heterogeneous hypotaxis (mu-he-hy)

In case the dependent subordinate clauses of a hypotactic construction belong to different classes, Kiesler speaks of a *multiple heterogeneous hypotaxis* (Fr. *phrase hypotaxique multiple hétérogène*). The two sentences in (18) and (19) can be described as such:

(18) Agnan,
[qui *est* le premier de la classe […],]
a dit
[que ce *serait* dommage de ne pas avoir arithmétique,
[parce [qu'il *aimait* ça]
et
[qu'il *avait* bien fait tous ses problèmes.]]] (Kiesler 2013, 620)
'Agnan, who is at the top of the class, said that it would be a shame
not to have arithmetic, because he liked it, and that he had done all his
problems well.'

(19) Lorsque Georges *fut* entré dans ce passage,
quatre personnes au moins s'y *engagèrent* également,
qui toutes s'*intéressaient* à lui. (E_Ch_po, 84)
'When George had entered this passage, at least four people entered as
well, all of whom were interested in him.'

In the example taken from Kiesler (2013) a total of four dependent subordinate clauses can be identified, which depend on a main clause. Whereas the first one (*est le premier de la classe*) presents a relative clause modifying the noun *Agnan*, there are two parallelly coordinated *that*-clauses (*que ce serait dommage de ne pas avoir arithmétique*; *qu'il avait bien fait tous ses problèmes*) and one adverbial clause (*parce qu'il aimait ça*). The first subordinate clause (*lorsque Georges fut entré dans ce passage*) in example (19) again shows an adverbial clause (with temporal meaning), whereas the second subordinate clause (*qui toutes s'intéressaient à lui*) can again be described as a relative clause, which defines the noun *personnes* more precisely.

After this typology of complex sentences has been presented, a note must be made about the distinction of simple syntactic constructions. While in the area of complex syntactic structures, Kiesler's typology of complex sentences is used for the analysis, in the field of simple sentences, only those with a finite verbal element and constructions without any finite verb are distinguished. Examples (20) and (21) can be characterized as a *simple sentence* (*simple*), whereas (22) and (23) both represent a *simple sentence without finite verb* (*simple wv*).[6]

(20) Je ne *peux* pas accepter. (E_Ch_po, 119)
'I can't accept.'

(21) Les trois gamins se *sont* déportés sur une autre voiture. (K_Mo_po, 81)
'The three kids fled to another car.'

(22) Rien à faire. (K_Re_bl, 193)
'Nothing to do.'

(23) Et le type? (E_Eq_bl, 40)
'And the guy?'

Thus, these nine types of sentences are used for the qualitative analysis of the sentences of the research corpus (cf. sections 3.1 and 3.2).[7] In the following section, on the other

[6] On the question of whether such constructions can be called *sentences*, Trabant, among others, comments with reference to Bloomfield: "Der Satz (sentence) [sic] wird nämlich deswegen von Bloomfield hervorgehoben, weil er die sprachliche Form ist, die als abgeschlossene Äußerung (*utterance*)—also als Text—auftreten kann. So ist z.B. nicht nur die Äußerung *Poor John ran away* ein Satz, sondern auch *Poor John!* oder *John!*, eben weil sie die abgeschlossene Äußerungen—Texte—ausmachen können" (1981, 8), Eng. "The sentence is emphasized by Bloomfield because it is the linguistic form which can appear as a closed utterance—that is, as text. Thus, for example, not only the utterance *Poor John ran away* is a sentence, but also *Poor John!* or *John!* because they can constitute the completed utterance—text".

[7] These include two different simple sentences, four different types of parataxis, and three types of hypotaxis, so that one can assume the following order from simple to complex: *simple sentence without finite verb* (*simple wv*)—*simple sentence* (*simple*)—*homogeneous parataxis* (*ho-pa*)—*heterogeneous parataxis* (*he-pa*)—*multiple homogeneous parataxis* (*mu-ho-pa*)—*multiple heterogeneous*

hand, the quantitative perspective of studies already carried out will be discussed and an own vector model will be presented.

2.2 Quantitative Aspects

In the past, a large number of quantitative approaches to syntax have been taken, so only a short excerpt will suffice at this point. Using French, Koch (1995) has described a method by which he determines the number of dependent subordinate clauses per main clause and describes the result as *complexité quantitative* ("quantitative complexity"). If one takes another look at examples (18) and (19) and applies Koch's approach, the ratio of the *complexité quantitative* for (18) can be described as 1:4, while for (19) it can be represented as 1:2.

(18) Agnan,
        [qui *est* le premier de la classe […],]
    a dit
        [que ce *serait* dommage de ne pas avoir arithmétique,
            [parce [qu'il *aimait* ça]
        et
        [qu'il *avait* bien fait tous ses problèmes.]]] (Kiesler 2013, 620)
    'Agnan, who is at the top of the class, said that it would be a shame
    not to have arithmetic, because he liked it, and that he had done all his
    problems well.'

(19) Lorsque Georges *fut* entré dans ce passage,
    quatre personnes au moins s'y *engagèrent* également,
        qui toutes s'*intéressaient* à lui. (E_Ch_po, 84)
    'When George had entered this passage, at least four people entered as
    well, all of whom were interested in him.'

This method works very well for complex sentences which only have one main clause. However, for complex constructions containing several main clauses (see above), this method can only be used for the individual main clauses (and the corresponding dependent subordinate clauses).

However, a simple and intuitive way to describe syntactic complexity—not only for linguists—is certainly the measurement of the sentence length which several

parataxis (*mu-he-pa*)—*simple hypotaxis* (*sim-hy*)—*multiple homogeneous hypotaxis* (*mu-ho-hy*)—*multiple hypotaxis* (*mu-he-hy*).

scholars refer to: length can be measured in words (i.e. Sowinski 1999; Szmrecsányi 2004, 1032–33), through the number of immediate constituents (Altmann, Best and Popescu 2014, 94; Altmann and Köhler 2000, 192), through the number of syllables (Best 2005, 300;[8] Fucks 1968, 87) or—for example for sign languages—through the number of characters (Jing 2001). Defining sentence length by counting the words and applying this approach to the examples (18) and (19), the sentence in (18) with 33 words would be more complex than (19) comprising 21 words. As mentioned before, length cannot be the only parameter for measuring complexity, as there can be long sentences without any subordination but also relatively short sentences containing several subordinate clauses. Hence, the degree of embedding must be taken into consideration as well, and therefore represents the main characteristic of syntactic complexity for several scholars (i.e. Givón 2009; Givón and Shibatani 2009). Another way of measuring syntactic complexity is counting the number of nodes dominated (i.e. Johnson 1966; Ferreira 1991), on which Szmrecsányi comments: "Presupposing some notion of formal complexity, counting the number of nodes dominated is conceptually certainly the most direct and intuitively the most appropriate way to assess syntactic complexity. This is because the method reflects how the human parser is supposed to work" (2004, 1033). Szmrecsányi finally proposes a so-called *Index of Syntactic Complexity* (ISC), which is defined as follows:

$$\mathrm{ISC}(u) = 2 \times n\ (u, \mathrm{SUB}) + 2 \times n\ (u, \mathrm{WH}) + n\ (u,\mathrm{VF}) + n\ (u, \mathrm{NP})$$[9]

Even if this approach of taking several factors into account offers an interesting perspective, it remains unclear, among other things, on which scale the determined values are to be located. The author therefore also gives cause for consideration when introducing this formula: "I would, then, like to suggest the following formula –which, admittedly, is somewhat tentative and ad-hoc [!]– to establish (ISC)" (Szmrecsányi 2004, 1034).

[8] Cf. Best (2005, 300): "Es spricht nun nichts dagegen, Satzlänge auch ganz anders zu messen: nach der Zahl der Silben pro Satz […], oder was anscheinend noch niemand versucht hat, nach der Zahl der Morphe. Auch noch kleinere Einheiten (Laut, Buchstabe) können für Satzlänge genutzt werden […]", Eng. "There is nothing to be said against measuring sentence length in a completely different way: according to the number of syllables per sentence […], or, which apparently no one has ever tried, according to the number of morphs. Even smaller units (phone, letter) can be used for sentence length".

[9] "Let u be the unit of linguistic data under analysis, let ISC(u) be the ISC of the unit of linguistic data under analysis, and let n(u,SUB) be the number of occurrences of SUB in the unit of linguistic data under analysis, etc. According to this formula, ISC of a given unit of data is twice the number of occurrences of subordinating conjunctions and WH-pronouns plus the number of occurrences of verb forms and noun phrases in that unit" (Szmrecsányi 2004, 1035).

The method developed in my dissertation (Hesselbach 2019) takes into account the most important quantitative characteristics of syntactic complexity, namely the sentence length (SL) measured in words, the number of finite verbs (FV) and the maximum depth of embedding (DE) of a syntactic construction. Instead of connecting the determined values in a quotient or a product,[10] they are represented as the following vector: $x = \begin{pmatrix} SL \\ FV \\ DE \end{pmatrix}$. Figure 1 shows this vector presentation schematically.

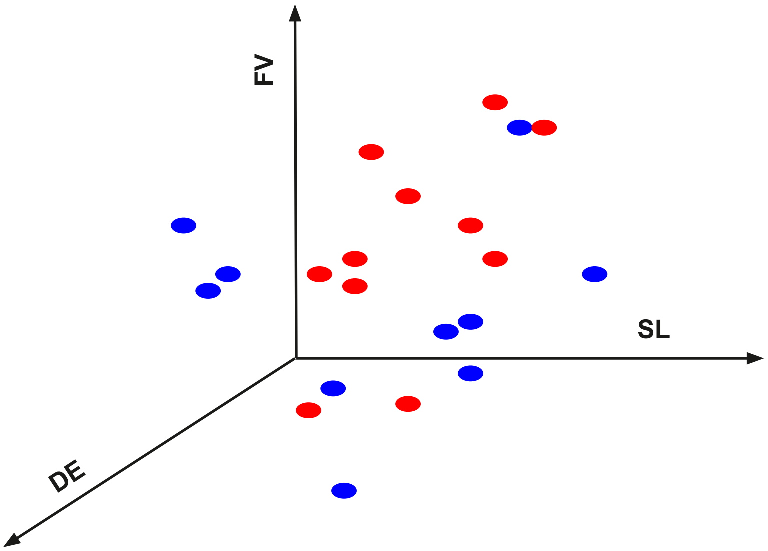

Fig. 1 Schematic illustration of the vector representation (Hesselbach, CC BY).

The advantage of this method—in agreement with Altmann (1978)—is that no mathematical operations are performed to describe syntactic complexity, but that the numerically measurable values of a construction, as just described, are only plotted as vectors.

[10] In this context Altmann speaks of a *sin* which is common in linguistics: "Die üblichste 'Sünde' in der Linguistik ist die Bildung von irgendwelchen Quotienten ohne Rücksicht auf die linguistischen und die mathematischen Aspekte des gegebenen Indexes" (1978, 91), Eng. "The most common 'sin' in linguistics is the formation of some quotients without regard to the linguistic and mathematical aspects of the given index".

The visualization then makes several things clear: On the one hand, the different subcorpora can be marked in color, so that a quick overview of the actual ratios can be gained visually. On the other hand, this method can also be used to compare the degree of complexity: the further away from the zero position a data point is located, the more complex the syntactic construction. In addition to determining the qualitative characteristics, this method will now be applied to our corpus of contemporary French novels.

3. A Case Study on Two Contemporary French Authors

Previous studies have focused on the application of this method to a stylistically heterogeneous corpus of modern European Spanish and French (Hesselbach 2019) and to a diachronic corpus of Spanish literature (Hesselbach, in prep.). In contrast, the study presented here focuses on two contemporary French authors and their novels and aims to examine whether determining syntactic complexity, as described above, can help to define subgenre distinction within an author's oeuvre. In the next step, the method and composition of the corpus is explained, before the results of the analysis are presented and interpreted in section 3.2.

3.1 Method

As mentioned before, the aim of the present study is to make a statement as to whether and to what extent the analysis of syntactic complexity can help to define (sub)genre distinctions within an author's oeuvre. Therefore, a corpus-based approach was chosen to investigate different texts. Hence, a corpus of contemporary French-language novels was compiled and digitized. It comprises eight novels written by two different authors (Jean Echenoz and Yasmina Khadra)[11] and covers the period from 1979 to 2002. It is important for the compilation of the corpus that four novels each can be ascribed to the subgenres (*roman*) *policier* and (*littérature*) *blanche*, as both authors have written novels in both subgenres. Table 1 shows the compilation of the corpus indicating the date of publication, the subgenre and the abbreviation (chosen for this study). Since the research question presented at the beginning aims at subgenre distinctions, the texts in this table are clustered due to their belonging to different subgenres (*roman policier* versus *littérature blanche*) and neither chronologically nor alphabetically.

[11] Even though Yasmina Khadra, an Algerian-born writer living in France, is part of this analysis, the regional variety affiliation is not considered significant for an investigation of syntactic complexity—in contrast to questions concerning the lexicon or idiomatic expressions, for example.

Table 1 Corpus of contemporary French novels

	Author	Title of novel	Year of publication	Subgenre	Abbreviation
1.	Echenoz, Jean	*Cherokee*	1983	*policier*	E_Ch_po
2.	Echenoz, Jean	*Le Meridien de Greenwich*	1979	*policier*	E_Me_po
3.	Khadra, Yasmina	*Double blanc*	1998	*policier*	K_Do_po
4.	Khadra, Yasmina	*Morituri*	1997	*policier*	K_Mo_po
5.	Echenoz, Jean	*L'Équipée malaise*	1986	*blanche*	E_Eq_bl
6.	Echenoz, Jean	*Nous trois*	1992	*blanche*	E_No_bl
7.	Khadra, Yasmina	*Les Hirondelles de Kaboul*	2002	*blanche*	K_Hi_bl
8.	Khadra, Yasmina	*À quoi rêvent les loups*	1999	*blanche*	K_Re_bl

From each of the texts, which were already available in digital form as plain-text documents, 200 sentences were randomly selected using *Python*,[12] so that the total corpus examined here has a size of (8×200 =) 1,600 sentences.[13] For each of these sentences, the sentence type, the sentence length, the number of finite verbs and the degree of maximum embedding depth—as described before—were determined.[14] Two example analyses are used to illustrate the procedure:

(12) [alors il y *avait* des défilés] [il y *avait* des tas de trucs]
et [tout ça se *passait*
d'où je *travaillais*
parce que c'*était* juste en face]

Sentence type	mu-he-pa	**Sentence length (in words)**	29
Number of finite verbs	5	**Depth of embedding**	2

(16) Geoffroy *a* un papa très riche
[qui lui *achète* tous les jouets
[qu'il *veut*.]]

Sentence type	mu-ho-hy	**Sentence length (in words)**	15
Number of finite verbs	3	**Depth of embedding**	2

[12] I would especially like to thank Daniel Schlör for his very valuable support in the application of *Python*.

[13] The research data can be accessed via the following link: https://doi.org/10.5281/zenodo.7458279.

[14] The sentence type as well as the degree of maximum embedding depth were determined manually, while the sentence length and number of finite verbs were analyzed automatically.

After having looked at the two example analyses, the results obtained by analyzing the entire corpus are now presented below.

3.2 Results

The aim of the study presented here is to use corpus data on two different subgenres of contemporary French novels to determine whether the description of syntactic complexity can help to define genre distinctions. In a first step, we will now take a closer look at the relationship between the different types of sentences, before the quantitative aspects of syntactic complexity are considered in section 3.2.2.

3.2.1 Qualitative Aspects

As explained above, both simple and complex sentences can be differentiated even more precisely. Table 2 shows the frequencies of the individual types of sentences for the two subgenres in question. For this purpose, the individual novels which can be assigned to the respective subgenre were combined and the individual frequencies of occurrence were summed up.

Table 2 Distribution of different types of sentences in the corpus

		Jean Echenoz		**Yasmina Khadra**	
		Subgenre		**Subgenre**	
Sentence type		*policier*	*blanche*	*policier*	*blanche*
Simple	simple wv	17	34	30	21
	simple	111	94	193	179
	Σ	128	128	223	200
Complex	ho-pa	61	52	44	50
	he-pa	46	54	25	25
	mu-ho-pa	27	27	13	7
	mu-he-pa	26	24	5	10
	Σ	160	157	87	92
	sim-hy	54	66	76	77
	mu-ho-hy	24	24	5	16
	mu-he-hy	34	25	9	15
	Σ	112	115	90	108
	Σ	400	400	400	400
	Σ	800		800	

If one reads the table from top to bottom, one can recognize the various frequencies and find that in both subgenres the *simple sentence* (with a finite verb) is by far the most common sentence type with 111 (*policier*) and 94 occurrences (*blanche*) for Echenoz and 193 (*policier*) and 179 (*blanche*) for Khadra. It is also noteworthy that the second most common type of sentence (in both subgenres) is the *simple hypotaxis* (with the exception of Echenoz' *policier* novels where the *homogenous parataxis* is predominant). This means that even in the field of complex sentences, the *simplest* types of sentences are very popular with both novelists.[15] If the individual sentence types are ordered according to their frequency, one can get a quick overview of the results in Table 3.

Table 3 Distribution of different types of sentences in the corpus

	Jean Echenoz		Yasmina Khadra	
	policier	*blanche*	*policier*	*blanche*
1.	simple (28 %)	simple (24%)	simple (48%)	simple (45%)
2.	ho-pa (15%)	sim-hy (17%)	sim-hy (19%)	sim-hy (19%)
3.	sim-hy (14%)	he-pa (14%)	ho-pa (11%)	ho-pa (13%)
4.	he-pa (12%)	ho-pa (13%)	simple wv (8%)	he-pa (6%)
5.	mu-he-hy (9%)	simple wv (9%)	he-pa (6%)	simple wv (5%)
6.	mu-ho-pa (7%) mu-he-pa (7%)	mu-ho-pa (7%)	mu-ho-pa (3%) mu-he-hy (3%)	mu-ho-hy (4%) mu-he-hy (4%)
7.	mu-ho-hy (6%)	mu-he-hy (6%) mu-ho-hy (6%) mu-he-pa (6%)	mu-he-pa (1%) mu-ho-hy (1%)	mu-he-pa (3%)
8.	simple wv (4%)	–	–	mu-ho-pa (2%)
9.	–	–	–	–

[15] Another result, which is initially not the focus of interest in the study presented here, is obtained by reading the table from left to right. This gives a general impression of the types of sentences across the genres, and it is shown that the simple sentence types, with a total of (128 +128 + 223 + 200 =) 679 (= 42%) cases, make up almost half of all sentences examined. The sum of the different paratactic constructions (160 + 157 + 87 + 92 = 496 = 31%) is similar to the values of the hypotactic structures (112 + 115 + 90 + 108 = 425 = 27%).

The distribution according to frequency clearly shows that the individual types of sentences are not only in a similar order, but also have comparable percentages for the most part. A clear difference can be seen not between the subgenres, but rather between the two authors: even though the simple sentence is the most common in both subgenres for the two authors, Khadra uses it almost twice as often (48 percent and 45 percent) as Echenoz (28 percent and 24 percent).[16] However, if we look at the characteristics for the particular subgenres for each author separately, we can see that no meaningful qualitative differences can be detected. It now remains to be examined whether this also applies to the corresponding quantitative characteristics.

3.2.2 Quantitative Aspects

In the description of the individual quantitative characteristics, they are first listed individually before they are finally combined in a vector diagram. First, (a) the extent of the sentence length is examined more closely, followed by an analysis of (b) the number of finite verbs and (c) the maximum degree of syntactic embedding.

(a) Sentence length

As described above, in this study the sentence length is determined by the number of words. The following box plot, Figure 2, gives an overview of the different values of sentence length in the samples of the individual texts. Note that the first four texts (on the left side of the box plot) belong to Echenoz's texts (with the subgenres *policier* and *blanche*), while the right half represents the results for both subgenres by Khadra.

The entire corpus comprises a total of 25,081 words, of which 15,273 belong to Echenoz (*policier*: 7,593; *blanche*: 7,680) whereas the other 9,808 words can be related to Khadra's novels (*policier*: 4,595; *blanche*: 5,213). The analysis clearly shows that the two authors do differ in terms of sentence length. Even though the value for the median of the individual text data is less than twenty words in all cases, a comparatively greater tendency toward longer sentences can be observed in Echenoz's texts regardless of the subgenre. The data in the box plot also provides the result that there are more significant differences between the two authors than between the subgenres: when looking at the results for the texts of Echenoz (*E_Ch_po*, *E_Me_po*, *E_Eq_bl*, *E_No_bl*) statistical outliers start at a limit of about 50 words. In contrast, this is already the case for Khadra with about 30 words. The longest sentence (132 words) can also be found in one of Echenoz's texts (*E_No_bl*). The following table shows again clearly

[16] The dominance of the *simple sentence* is highly significant for both authors in both subgenres when looking at the corresponding p-values (based on a χ^2-test) for Echenoz (*policier*: $p < 2.2e\text{-}16$; *blanche*: $p = 1.158e\text{-}12$) and Khadra (*policier*: $p < 2.2e\text{-}16$; Khadra, *blanche*: $p < 2.2e\text{-}16$).

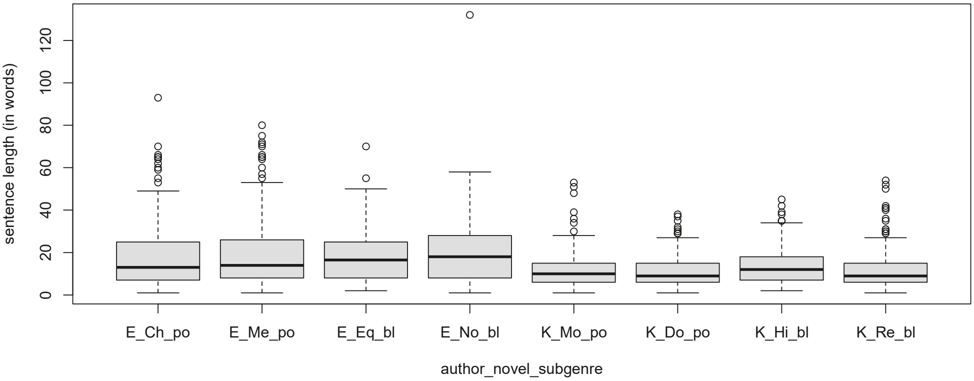

Fig. 2 Distribution of sentence length of the examined texts (Hesselbach, CC BY).

Table 4 Average sentence length (in words) by author

	Jean Echenoz	Yasmina Khadra
policier	19.0	11.5
blanche	19.2	13.0

that these differences manifest themselves between the two authors and not between the subgenres.[17]

As mentioned above, the analysis of sentence length as a quantitative criterion of syntactic complexity cannot help to distinguish genres, but it can help to distinguish authors. In the following, the feature of the number of finite verbs will be considered in more detail.

(b) Number of finite verbs

In a previous study it was found that scientific texts (in French) show an average of 1.89 finite verbs/sentence (Hesselbach 2019, 261). For narrative texts, (finite) verbs are of particular importance, since they are used to drive the plot of the story. It is therefore of great interest to see whether narrative texts (of different literary subgenres) show a value that differs from this previous result. The box plot in Figure 3 illustrates the ratios for the entire corpus.

[17] A two-sample t-test was performed to determine the statistical significance of the results: with p = 1.02e-27, it can be concluded that this distribution is highly statistically significant.

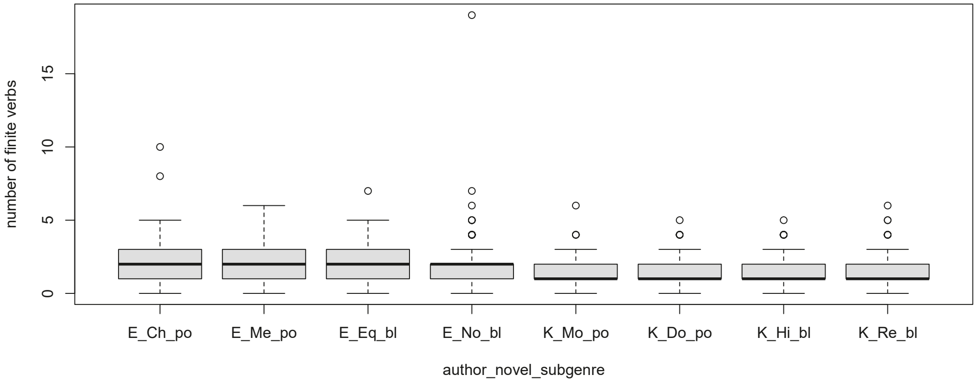

Fig. 3 Distribution of the number of finite verbs (per sentence) of the examined texts (Hesselbach, CC BY).

Table 5 Average number of finite verbs by author

	Jean Echenoz	Yasmina Khadra
policier	2.0	1.4
blanche	1.9	1.5

Again, it becomes clear from the box plot that the criterion of finite verbs also does not give any information about the affiliation of a novel to a certain subgenre. The analysis of the data shows that the distribution is very robust: out of a total of 1,600 sentences, there are only 16 sentences which can be considered statistical outliers. Table 5 shows the mean values for the texts analyzed, broken down by subgenre and author.

It is obvious that Echenoz's values are about 0.5 points higher than Khadra's and almost identical with the results of the previous study mentioned above. Once again, considering the results of Echenoz's texts, one can again ask the question of whether the number of finite verbs are rather an author-specific feature, since they are the only ones in the box plot which can be regarded as more complex (at least for *E_Ch_po*, *E_Me_po*, *E_Eq_bl*) compared to the data received from Khadra's novels, who tends to use less finite verbs in his fiction than Echenoz.[18] Finally, the maximum degree of syntactic embedding is determined in the following.

[18] Again, a two-sample t-test was performed: with p = 9.59e-17, it can be concluded that this distribution is highly statistically significant.

(c) Maximum degree of syntactic embedding

The last quantitative aspect deals with the depth of embedding of a syntactic construction, which for many authors constitutes the criterion par excellence for syntactic complexity (see section 2.2.). For each of the 1,600 sentences, the value for the deepest level of embedding was determined, so that the results can be seen in Figure 4.

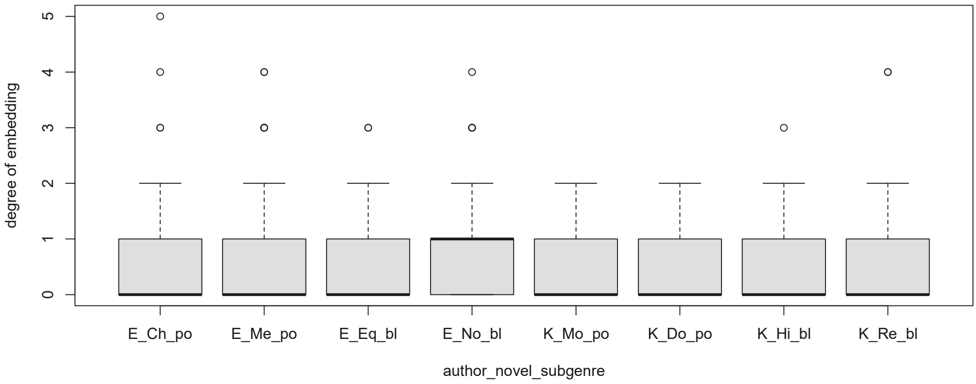

Fig. 4 Distribution of the maximum degree of embedding (per sentence) of the examined texts (Hesselbach, CC BY).

As can be seen from the illustration, there are only extremely few sentences with a depth of embedding of 3 (and more) in the entire corpus, namely a total of only 10, which are visible here as statistical outliers. The data clearly shows—at least for this corpus—that the two authors rely rather on syntactically *flat* constructions, regardless of the literary subgenre. This is interesting in so far as syntactically demanding, i.e. complex, constructions are not necessarily a characteristic of *littérature blanche*, as one might expect. To verify this on the basis of larger datasets would certainly be a rewarding research project. A closer look at the results reveals that Echenoz's sentences again display a more noticeable complexity feature, as can be seen in Table 6.[19]

After having taken a closer look on those complexity features and the results of this study, it has become clear that Echenoz's sentences generally display a greater

[19] Again, a two-sample t-test was performed: with p = 1.29e-12, it can be concluded that this distribution is highly statistically significant.

Table 6 Average maximum degree of embedding by author

	Jean Echenoz	Yasmina Khadra
policier	0.6	0.3
blanche	0.6	0.4

degree of complexity than those of Khadra, and that syntactic complexity must be understood here as an author-specific, rather than a subgenre-specific, feature.

(d) Vector representation

In this subsection, the individually determined quantitative values are now to be put in relation to each other, so that—as described in 2.2.—a vector representation can be generated. Table 7 shows the average quantitative value of the syntactic complexity as a vector for the two investigated subgenres.

Table 7 Average vector values by author and subgenre

	Jean Echenoz	Yasmina Khadra
policier	$\begin{pmatrix} 19.0 \\ 2.0 \\ 0.6 \end{pmatrix}$	$\begin{pmatrix} 11.5 \\ 1.4 \\ 0.3 \end{pmatrix}$
blanche	$\begin{pmatrix} 19.2 \\ 1.9 \\ 0.6 \end{pmatrix}$	$\begin{pmatrix} 13.0 \\ 1.5 \\ 0.4 \end{pmatrix}$

Therefore, it can be concluded that the two subgenres do not differ fundamentally in terms of quantitative characteristics, as the main differences can be found between the two authors. If one now wants to visualize the results obtained, the individual sentences can be highlighted in color and then plotted as vectors. Blue marked data points represent the sentences from the crime novels (*roman policier*), whereas the red points reflect the sentences from both authors referring to the *littérature blanche*. The following plots in Figure 5, which were generated with *R*,[20] give a very detailed impression of the relation between the two subgenres.

The representations clearly show that the overwhelming majority of the data can be found in the compacted cluster and that only individual, extremely complex sentences, which are furthest away from the zero position, become visible as outliers.

[20] The three-dimensional version of the plot shown here can be seen on the following website: https://rpubs.com/RobertHesselbach/985657.

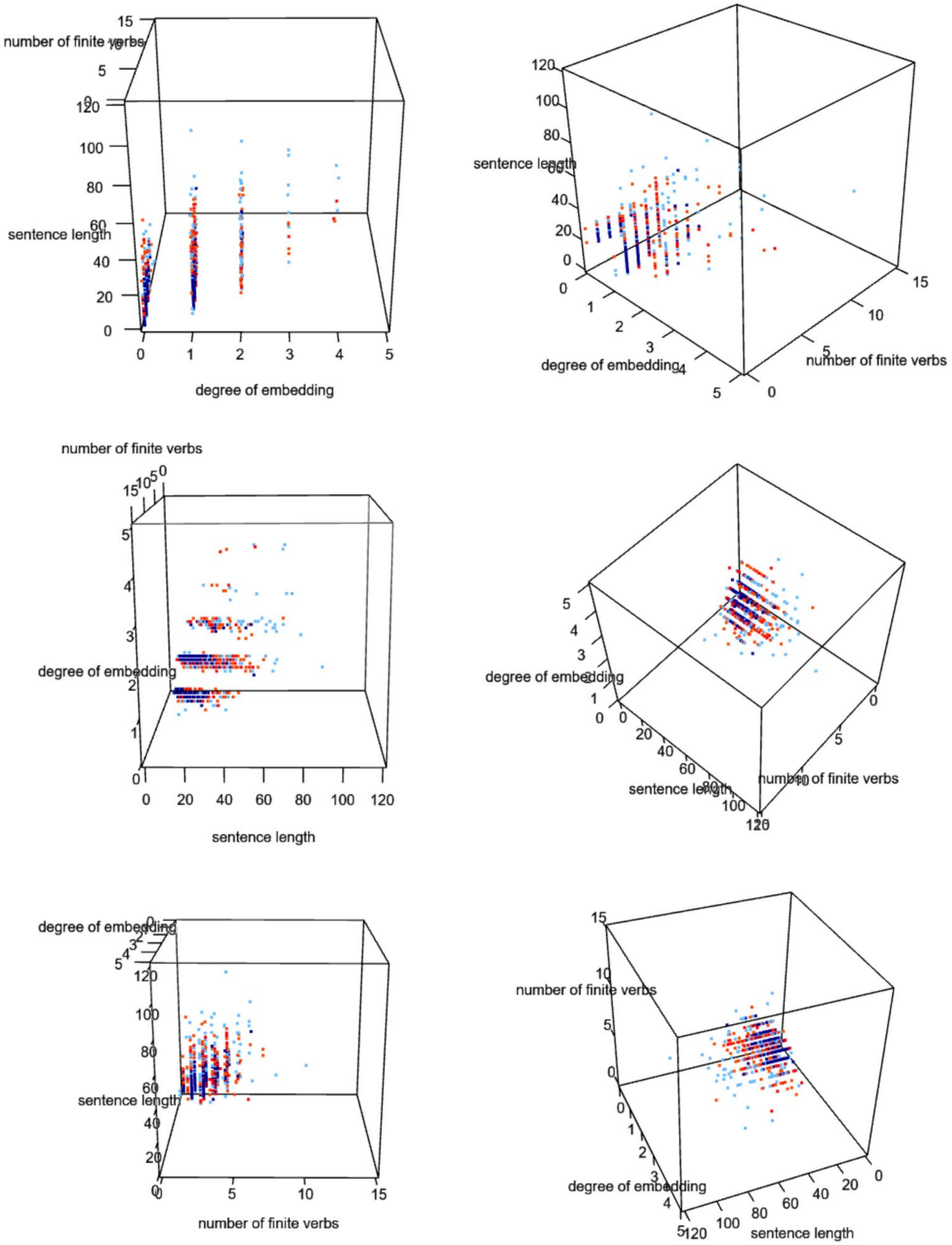

Fig. 5 Visualization of syntactic complexity in vector space for Echenoz's and Khadra's data (Hesselbach, CC BY).

4. Conclusion

The goal of the study presented here was to determine whether the analysis of the syntactic complexity of the sentences of a text can help to distinguish between different subgenres or rather between authors writing in different subgenres. To achieve this, both quantitative and qualitative characteristics were evaluated using a corpus of two contemporary French authors. It is astonishing that the results suggest that there are no significant differences in both quantitative and qualitative terms. On the contrary, a different conclusion can be formulated: the high degree of congruence of both types of sentences and numerically ascertainable values describes an extremely robust distribution which bears witness to the fact that the degree of syntactic complexity for the narrative texts of both subgenres analyzed must be described as almost identical. The corpus analyzed here is admittedly too small to be able to make definitive statements about the distinction between *roman policier* and *littérature blanche* or even further statements with regard to other narrative genres. Nevertheless, further studies (on the same as well as on other genres) with larger amounts of data should provide a better understanding of the relationship between syntactic complexity and (sub)genre distinction.

With regard to syntactic complexity, however, it can be concluded from the available data that its extension seems to be an author-specific characteristic, as we have seen in the analysis of Jean Echenoz's texts, who relies on more complex sentences (in numeric terms) than Yasmina Khadra does. Finally, the combination of quantitative and qualitative results could open up interesting perspectives in the future with regard to the study of the style of individual authors.

Appendix

The research data is available on Zenodo: https://doi.org/10.5281/zenodo.7458279.

Acknowledgements

The research group was funded by the German Ministry for Education and Research (BMBF), and I would like to thank them for their support. A further thank you goes to Fotis Jannidis, Christof Schöch, Ulrike Henny-Krahmer, José Calvo Tello, Daniel Schlör, Steffen Pielström, Julian Schröter and Barbara Gersitz for being great colleagues at the University of Würzburg. I am also grateful to Christopher Geekie (Erlangen-Nürnberg) for his valuable advice.

ORCID®

Robert Hesselbach https://orcid.org/0000-0001-9758-8290

References

Altmann, Gabriel. 1978. "Zur Verwendung der Quotiente in der Textanalyse." In *Glottometrika*, edited by Gabriel Altmann, 91–106. Bochum: Brockmeyer.

Altmann, Gabriel, and Reinhard Köhler. 2000. "Probability Distributions of Syntactic Units and Properties." *Journal of Quantitative Linguistics* 7 (3): 189–200.

Altmann, Gabriel, Karl-Heinz Best, and Ioan-Iovitz Popescu. 2014. *Unified Modeling of Length in Language*. Lüdenscheid: RAM-Verlag.

Best, Karl-Heinz. 2005. "Satzlänge." In *Quantitative Linguistik*, edited by Gabriel Altmann, Reinhard Köhler and Rajmund G. Piotrowski, 298–304. Berlin/Boston: De Gruyter.

Calvo Tello, José. 2021. *The Novel in the Spanish Silver Age: A Digital Analysis of Genre Using Machine Learning*. Bielefeld: transcript.

Dubois, Jean et al. 1994. *Dictionnaire de linguistique et des sciences du langage*. Paris: Larousse.

Ferreira, Fernanda. 1991. "Effects of Length and Syntactic Complexity on Initiation Times for Prepared Utterances." *Journal of Memory and Language* 30 (2): 210–33.

Fucks, Wilhelm. 1968. *Nach allen Regeln der Kunst: Diagnosen über Literatur, Musik, bildende Kunst – die Werke, ihre Autoren und Schöpfer*. Stuttgart: Deutsche Verlags-Anstalt.

Gianitsos, Efthimios Tim, Thomas J. Bolt, Pramit Chaudhuri, and Joseph P. Dexter. 2019. "Stylometric Classification of Ancient Greek Literary Texts by Genre." *Proc. of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, 52–60. https://www.aclweb.org/anthology/W19-2507.pdf

Givón, Talmy. 2009. *The Genesis of Syntactic Complexity – Diachrony, ontogeny, neuro-cognition, evolution*. Amsterdam: John Benjamins.

Givón, Talmy, and Masayoshi Shibatani, eds. 2009. *Syntactic Complexity – Diachrony, Acquisition, Neuro-Cognition, Evolution*. Amsterdam: John Benjamins.

Henny-Krahmer, Ulrike. 2018. "Exploration of Sentiments and Genre in Spanish American Novels." *DH 2018*. https://dh2018.adho.org/exploration-of-sentiments-and-genre-in-spanish-american-novels/

Hesselbach, Robert. 2019. *Diaphasische Variation und syntaktische Komplexität – eine empirische Studie zu funktionalen Stilen des Spanischen mit einem Ausblick auf das Französische*. Berlin, Boston: De Gruyter.

Hesselbach, Robert. (in prep.). "Sobre la complejidad sintáctica de textos literarios del español a través del tiempo."

Hettinger, Lena, Isabella Reger, Fotis Jannidis, and Andreas Hotho. 2016. "Classification of Literary Subgenres." In *DHd 2016 Digital Humanities. Konferenzabstracts*, 154–58. Leipzig: Universität Leipzig. http://www.dhd2016.de/abstracts/vortr%C3%A4ge-049.html.

Jing, Zhuo. 2001. "Satzlängenhäufigkeiten in chinesischen Texten." In *Häufigkeitsverteilungen in Texten*, edited by Karl-Heinz Best, 202–10. Göttingen: Peust & Gutschmidt.

Johnson, Neal. 1966. "On the Relationship between Sentence Structure and the Latency in Generating the Sentence." *Journal of Verbal Learning and Verbal Behavior* 5 (4): 375–80.

Kiesler, Reinhard. 2013. "Pour une typologie des phrases complexes." *Zeitschrift für romanische Philologie* 129 (3): 608–28.

Koch, Peter. 1995. "Subordination, intégration syntaxique et 'oralité'." In *La subordination dans les langues romanes. Actes du colloque international, Copenhague 5.5.–7.5.1994*, edited by Hanne Leth Andersen and Gunver Skytte, 13–42. Copenhagen: Munksgaard.

Raible, Wolfang. 1992. *Junktion – eine Dimension der Sprache und ihre Realisierungsformen zwischen Aggregation und Integration*. Heidelberg: Winter.

Schöch, Christof. 2017. "Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama." *Digital Humanities Quarterly* 11 (2). http://www.digitalhumanities.org/dhq/vol/11/2/000291/000291.html.

Sowinski, Bernhard. 1999. *Stilistik: Stiltheorien und Stilanalysen*. 2nd ed. Stuttgart: Metzler.

Szmrecsányi, Benedikt. 2004. "On Operationalizing Syntactic Complexity." In *Le poids des mots. Proceedings of the 7th International Conference on Textual Data Statistical Analysis. Louvain-la-Neuve, March 10–12, 2004*, edited by Gérard Purnelle, Cédrick Fairon, and Anne Dister, 1032–39. Louvain-la-Neuve: Presses universitaires de Louvain.

Trabant, Jürgen. 1981. "Wissenschaftsgeschichtliche Bemerkungen zur Textlinguistik." In *Beiträge zur Linguistik des Französischen*, edited by Thomas Kotschi, 1–20. Tübingen: Narr.