Journal of French Language Studies (2020), **30**, 355–380
doi:10.1017/S0959269519000334

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

Some quantitative aspects of written and spoken French based on syntactically annotated corpora

Rafaël Poiret[a] and Haitao Liu[a,b,c,*]

[a]Zhejiang University, [b]Beijing Language and Culture University and [c]Guangdong University of Foreign Studies
*Corresponding author. Email: htliu@163.com

(Received 11 January 2019; revised 28 October 2019; accepted 21 November 2019; first published online 07 February 2020)

ABSTRACT

Based on two syntactically annotated corpora, and within the theoretical tradition of dependency grammar, the current study investigates the quantitative differences and similarities between written and spoken French. Our findings support the assumption that spoken and written French are two realizations of one language that do not differ in the syntactic categories, but in the frequency of these categories, and also in their organization in sentence. The subjects in spoken French are mostly pronouns, whereas in written French the subjects are mostly nouns and pronouns. Spoken and written French share many syntactic relations, but with different frequencies. For instance, dislocations are more diverse and frequent in spoken French. Spoken French and written French differ in the word order of vocative nominal phrases. Finally, written French is slightly more difficult to process than spoken French.

Keywords: French; written and spoken languages; quantitative aspects; dependency grammar; treebanks; POS; word order; dependency distance

1. INTRODUCTION

Linguists had long neglected spoken language until European and North American structuralists, such as Saussure and Sapir, pointed out the primacy of spoken language over written language. However, it's not until the 1970s that some serious attempts were made to describe features of spoken language (Gadet, 1996: 14). These attempts have yielded some authoritative works (Halliday, 1985; Blanche-Benveniste and Jeanjean, 1987; Blanche-Benveniste *et al.*, 1990; Blanche-Benveniste, 1997; Miller and Weinert, 1998). A comparative study of both spoken and written languages can probably better reveal the features of spoken language. This is crucial for theoretical and applied linguistics (Tannen, 1980).

Nowadays, with regard to languages like French or English, spoken and written tend to be two realizations of the same language (Halliday, 1985; Blanche-Benveniste, 1997; Morel and Danon-Boileau, 1998; Gadet, 1996, 2007a; Béguelin, 1998). Moreau (1977: 236) underlines that these two realizations do not distinguish between themselves in terms of the grammatical phenomena *per se*, but in the frequency of these grammatical phenomena.

According to Chafe and Tannen (1987: 387), the first linguistic quantitative comparison of spoken and written productions goes back to 1977, which is concerned with English. More recently, Liu, Niu and Liu (2012, 2013) have compared spoken and written Chinese based on Chinese syntactically annotated corpora. The quantitative researches on spoken French alone are numerous, covering the fields of phonology, prosody (Berns, 2015; Meinschaefer, Bonifer and Frisch, 2015; Avanzi, Gendrot and Lacheret, 2010; Brunetti, Avanzi and Gendrot, 2013), and grammar (Henry and Pallaud, 2003; Coveney, 2004; De Cat, 2005). Labbé (2003) has statistically conducted a comparative study into the coordination and subordination in written and spoken French, which has two limitations. First, the data, as the author himself recognized, are not representative enough. Labbé's spoken corpus is made up of interviews of sociologists, and his written corpus is literary texts mainly. Second, the author focused on the word classes and the word forms. This approach permitted him to conclude that grammatical words are used in spoken French to establish logical links between utterances, rather than to construct complex sentences in written French. Our research represents a straight continuation of Labbé's work. Using syntactically annotated corpora as materials, we will try to give a global quantitative account of differences and similarities between written and spoken French. We will focus on syntactic categories, namely, parts of speech and syntactic relations, and their organization in sentence, from the point of view of word order and dependency distance. Investigating the interaction between semantics, syntax and pragmatics on the one hand, and the interaction between discourse competence and cognition on the other, is essential to explain speakers' linguistic choices and preferences. Much promising progress has been achieved in this direction (e.g. Arnold, 2001; De Cat, 2011, 2012; Serratrice and De Cat, 2019). These aspects, however, go beyond the scope of our study. The question we ask might be synthesized as: what are the quantitative differences and similarities between written and spoken French from a quantitative syntactic perspective? To tackle this issue, we will investigate the following four aspects:

1. Parts of speech: How parts of speech are distributed in spoken and written French? What are the differences, between spoken and written French, in the syntactic roles occupied by these parts of speech?
2. Syntactic relations: How syntactic relations are distributed in spoken and written French? Can we observe evident differences in syntactic relations of spoken and written French?

https://doi.org/10.1017/S0959269519000334 Published online by Cambridge University Press

3. Word order: Is the word order of spoken French different from that of written French?
4. Comprehension difficulty: For spoken and written French, which is more difficult to process syntactically?

Our study tried to make the language materials as representative as possible, which is a crucial condition in order to ensure the scientific value of the findings.

2. MATERIALS AND DEFINITIONS

With the development of information sciences, and natural language processing in particular, many resources of written but also of spoken French are now available to researchers. We can not only record, store and transcribe hours of speech, but also automatically process large texts, and annotate them with syntactic information. It is on this kind of resources, or treebanks, that this study is based.

2.1. The syntactic annotation of the written and spoken treebanks

Nowadays, sentences in most treebanks are annotated with dependency relations. Here are the three properties that are generally seen as the kernel features of a dependency relation (Tesnière, 1959; Hudson, 1990, 2007):

1. It is a binary relation between two linguistics units.
2. It is usually asymmetrical, with one of the two units acting as the governor (G) and the other as dependent (D).
3. It is classified in terms of a range of general grammatical relations, as shown conventionally by a label on top of the arc linking the two units.

From a functional point of view, the dependency relationship is not between a Governing word and a Depending word, but between a Governing word and a complete subtree depending on it. A Governor is a terminal in the dependency tree and can have many such complements, which are non-terminal constituents. In Figure 1, the Governor *likes* has two complements, *Charles* and *little dogs*, and not just *Charles* and *dogs*. Each complement as a whole is characterized by one grammatical function; most morpho-syntactic features (case, agreement, word order) apply to the whole complement and not just to one word on it (Hellwig, 2003: 603). The dependency relations in Figure 1 are *subject*, *direct object* and *noun modifier*.

For the sake of comparative study, both the spoken and the written corpora should be annotated with the same annotation scheme. Otherwise, the consistency can hardly be guaranteed. Universal Dependencies (UD) is a collaborative project that aims at developing a cross-linguistically consistent annotation scheme for treebanks (Nivre *et al.*, 2016). As pointed out by Gerdes *et al.* (2018), in order to maximize parallelism between languages, UD made the controversial choice of using content words as governors, because content words are more consistent across languages

https://doi.org/10.1017/S0959269519000334 Published online by Cambridge University Press

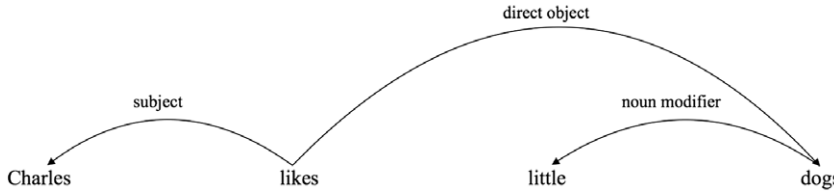

Figure 1. A sentence analysed with dependency relations.

than function words. This choice, however, goes against syntactic tradition, which defines syntactic functions by the distributional properties of words. Gerdes *et al.* (2018) also pointed out another weakness of the UD scheme: the relation of a word is labeled with both its category – a clause or a noun, etc. – and its syntactic function – a subject or an object, etc. "As an alternative to UD", Gerdes *et al.* (2018) proposed the Surface-syntactic Universal Dependencies (SUD) scheme. The authors developed a tool that converts UD to SUD and SUD to UD. These tools are freely distributed, and SUD treebanks are available on the Internet.[1] In the present study, we used two of the SUD treebanks. The two treebanks we used are the SUD Sequoia treebank and the SUD Spoken-French treebank, converted respectively from the UD Sequoia treebank (version 2.2) and the UD Spoken-French treebank (version 2.2). The UD Sequoia treebank is the result of the automatic annotation with manual correction of the Sequoia corpus (Candito and Seddah, 2012). The UD Spoken-French treebank is an automatic conversion with manual correction of the Rhapsodie treebank (Lacheret *et al.*, 2014). The SUD Sequoia treebank is used to investigate written French, while the SUD Spoken-French treebank investigates spoken French.

The tokenization strategies are slightly different. In the Sequoia SUD treebank, compound words with hyphen-like *sous-préfet* ('sub-prefect'), *auto-financement* ('self-financing') or *savoir-faire* ('know-how') were treated as one token, whereas in the Spoken-French SUD treebank, such compound words — *chef-d'oeuvre* ('work of art'), *rond-point* ('roundabout'), *mi-temps* ('first half') — were split into two different tokens. Compound proper nouns like *Alsace-Lorraine* or *Reuilly-Diderot* are similarly treated in the two treebanks. Because this kind of lexical unit is easy to recognize automatically with the presence of the hyphen, we modified the tokenization in the Spoken-French treebank congruously. Grammatical compound words like *grâce à* ('thanks to') are annotated with the Universal Dependencies *fixed* relation systematically in the Sequoia treebank but unsystematically in Spoken-French treebank. With the list of the grammatical compound words of the Sequoia treebank, and the list of grammatical compound words of the Orféo project (Debaisieux, Benzitoun and Deulofeu, 2016), we completed and merged the annotation of grammatical compound words in our two treebanks. We also merged the annotation of parts of speech (POS). In Sequoia, *avoir* ('to have'), *être* ('to be') and the causative *faire* ('to

[1]https://gitlab.inria.fr/grew/SUD

https://doi.org/10.1017/S0959269519000334 Published online by Cambridge University Press

make') were annotated as auxiliary, whereas in Spoken-French, in addition to *avoir* and *être*, modal verbs like *pouvoir* ('to be able'), *vouloir* ('to want') and *devoir* ('to have') were also treated as auxiliaries. Following the majority of French grammars (Le Goffic, 1993; Jones, 1996; Grevisse and Goosse, 2008; Riegel, Pellat and Rioul, 2016), we annotated only *avoir* and *être* as auxiliaries. The written treebank contains 28,987 tokens and the spoken treebank 28,960. Before presenting how the components of our written and spoken treebanks are organized, we have to define the notions of written and spoken languages.

2.2. Definitions of written and spoken languages

The difference between written and spoken languages can vary throughout the ages. In ancient China, the language used by public servants to write official texts was very distinct from the language they used in daily conversation. The difference can also vary from one language to another. Arabic-speaking communities nowadays are in a situation of diglossia, because there is an important disparity between classical Arabic and spoken Arabic (Halliday, 1985: 41–42). Whether French is in a situation of diglossia or not is still a matter of debate (Coveney, 2002, 2011; Gadet, 2007b; Massot, 2010; Massot and Rowlett, 2013; Zribi-Hertz, 2011), but the reality of grammatical variation is undisputed. Whereas written French is codified and fixed, spoken French is less controlled and more unstable. The division between spoken and written French, though, remains unclear (Gadet, 1996: 16–17). As Koch and Oesterreicher (2001) have pointed out, the opposition between the phonic and graphic media is dichotomous; whereas spoken and written are not polar opposites, instead the relationship between them forms a continuum. The opposite ends of this continuum are defined by a set of parameters that are themselves gradable. They characterize two communicative situations, immediacy (Fr. *immédiat*) and distance (Fr. *distance*). These parameters are shown in Table 1 below. We also added the translations of the original French terms (written in brackets and in italics).

In practice, the phonic medium is closely related to immediacy and the graphic medium is closely related to distance (Gadet, 2007b: 48).

2.3. The composition of the written and spoken treebanks

Based on the notions of immediacy and distance, opposed to the graphic and phonic media, we established two treebanks based on different genres. Each genre corresponds to one or more corpus. These resources have all been presented in diverse kinds of publications. Table 2 and Table 3 below present the relevant information.

The genres composing the spoken treebank are transcriptions of more or less immediate spoken French, whereas the genres composing the written treebank are texts of more or less distant written French. For instance, the genre of professional report in the written treebank verifies the parameters of preparation, weak emotionality and spatio-temporal separation. On the other hand, the genre of political debate in the spoken treebank verifies the parameters

https://doi.org/10.1017/S0959269519000334 Published online by Cambridge University Press

Table 1. Communicative situations parameters (Koch and Oesterreicher, 2001: 586)

Immediacy	Distance
Private communication (Fr. *communication privée*)	Public communication (Fr. *communication publique*)
Close interlocutor (Fr. *interlocuteur intime*)	Unknown interlocutor (Fr. *interlocuteur inconnu*)
Strong emotionality (Fr. *émotionalité forte*)	Weak emotionality (Fr. *émotionalité faible*)
Acting and situational anchoring (Fr. *ancrage actionnel et situationnel*)	Actional and situational detachment (Fr. *détachement actionnel et situationnel*)
Reference anchoring in the situation (Fr. *ancrage référentiel dans la situation*)	Referential detachment in the situation (Fr. *détachement référentiel de la situation*)
Spatio-temporal co-presence (Fr. *coprésence spatio-temporelle*)	Spatio-temporal separation (Fr. *séparation spatio-temporelle*)
Intense communicative cooperation (Fr. *coopération communicative intense*)	Close communicative cooperation (Fr. *coopération communicative intime*)
Dialogue (Fr. *dialogue*)	Monologue (Fr. *monologue*)
Spontaneous communication (Fr. *communication spontanée*)	Prepared communication (Fr. *communication préparée*)
Thematic freedom (Fr. *liberté thématique*)	Thematic fixation (Fr. *fixation thématique*)
etc.	etc.

Table 2. Composition of the written French treebank

Genre	Description and references of each subcorpus
Parliamentary debate	Sentences from parliamentary debates at the European Parliament Europarl (Koehn, 2005)
Narration	Entries from French Wikipedia about famous social or political affairs FrWiki (Villemonte de La Clergerie *et al.*, 2008)
Print media	Articles from the French regional newspaper *L'Est Républicain* Annodis (http://www.cnrtl.fr/corpus/estrepublicain)
Professional report	Documents from European Medicines Agency that are essentially public evaluation reports concerning drugs EMEA from OPUS Corpus (Tiedemann, 2009)

of spontaneity, strong emotionality, and spatio-temporal co-presence. This is to say, this study will focus on the relationship between genres of a rather distant written French and of a rather immediate spoken French. We can now investigate on how spoken and written French differ from each other syntactically.

https://doi.org/10.1017/S0959269519000334 Published online by Cambridge University Press

Table 3. Composition of the spoken French treebank

Genre	Description and references of each subcorpus
Interview and conversation	- Discussing about people's daily life in the city of Paris CFPP2000 (Branca-Rosoff *et al.*, 2012) - Asking passer-by for directions Avanzi (Avanzi, 2012) - Interviewing people about their life, their childhood, adolescence and career Lacheret (Lacheret, 2003), PFC (Durand, Laks and Lyche, 2009) - Interviewing a personality of a city about this city Eslo (Eshkol-Taravella *et al.*, 2012)
Monologue	- Describing a movie scene Rhapsodie-Movie (Lacheret *et al.*, 2014) - Students presenting their own academic orientation Rhapsodie-Professional (Lacheret *et al.*, 2014)
Conversation transmitted on radio	- Debates between scientific specialists about a specific news items Rhapsodie-Broadcast (Lacheret *et al.*, 2014) - Political debate Rhapsodie-Broadcast (Lacheret *et al.*, 2014) - Soccer match commentaries Rhapsodie-Broadcast (Lacheret *et al.*, 2014) - Dialogue about literature, classical music Rhapsodie-Broadcast (Lacheret *et al.*, 2014) - Home shopping show Rhapsodie-Broadcast (Lacheret *et al.*, 2014) - Interviews of celebrities on their career and on familiar topics Rhapsodie-Broadcast (Lacheret *et al.*, 2014) Mertens (Mertens, 1987)

3. PARTS OF SPEECH AND SYNTACTIC ROLES

3.1. The distribution of the POS in both corpora

There is a great difference in POS between spoken and written productions, as has been emphasized by Halliday (1985), who distinguished between them in terms of informational density. Written language has a higher lexical density, whereas spoken language has a higher grammatical density (Halliday, 1985: 64, cited by Gadet, 1996: 23). Our data presented in Figure 2 confirm Halliday's finding. This graph displays the percentage of each POS occurrence to the total number of words (ignoring the category X that describes these words that cannot be assigned to any POS). The percentage of lexical words in written texts is higher (54.61%) than the percentage of grammatical words (45.16%).[2] On the contrary, lexical words account for 43.61% and grammatical words 56.09% in the spoken corpus.

Additionally, the fact that the writer is not under time pressure during the production results in a higher proportion of lexical noun phrases in the written language than in the spoken language (Mazur-Palandre, 2015). In her study on

[2]Determiners, pronouns, prepositions, coordinative and subordinative conjunctions, auxiliary verbs, interjections and particles are considered as grammatical words, and nouns, verbs, adjectives, proper nouns, adverbs and numerals as lexical words.

https://doi.org/10.1017/S0959269519000334 Published online by Cambridge University Press

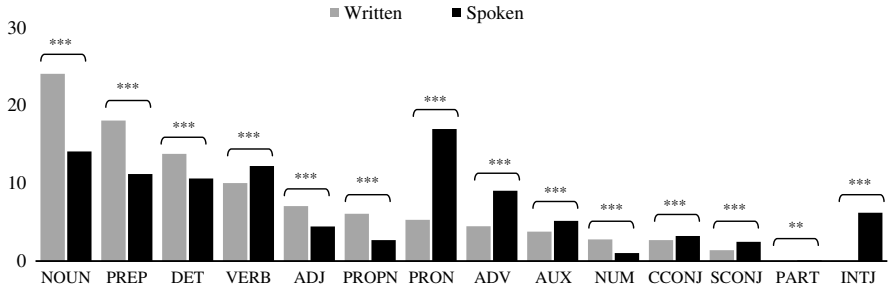

Figure 2. The distribution of POS in written and spoken French[3].

around 120 French speakers and writers, Mazur-Palandre (2015) found that the mean number of lexical noun phrases per clause is higher in the written texts as opposed to the spoken texts. As the author puts it, it corroborates the idea that the modality of production impacts the specific characteristics of production (e.g. Berman and Verhoeven, 2002; Fayol, 1997; Jisa, 1998; Ravid *et al.*, 2002). In the written corpus, the proportion of nouns to the total number of words reaches 24.12% (6,991 ex.). In spoken French, nouns are scarcer, accounting for only 14.12% (4,090 ex.).[4] The difference is significant ($Z(1) = 759.48$, $p < 0.001$). In contrast, the percentage of verbs is 10.05% (2,914 ex.) in written French, and 12.24% (3,546 ex.) in spoken French ($Z(1) = 61.83$ $p < 0.001$). The difference of pronouns frequency is even more significant ($Z(1) = 1{,}773.9$, $p < 0.001$).

3.2. The syntactic roles occupied by the POS

The most prominent syntactic roles in French are subject and object. Figure 3 below shows that the nominal subjects are six times less frequent in spoken French (6.91%) than in written French (42.81%). It corroborates the findings of Blanche-Benveniste (1994, cited by Gadet, 1996: 24). In contrast, in spoken French, the percentage of pronominal subjects (91.22%) is twice as much as that in written French (45.91%). In other words, in written French, nouns and pronouns respectively account for about 50% of subjects, whereas in spoken French, the majority of subjects are pronouns.

Figure 4 shows the distribution of POS occupying the role of object. As in the case of subjects, there are more nominal objects in written than in spoken French, and more pronominal objects in spoken than in written French. But the difference in the percentages of nominal objects is not so striking as the difference in the percentages of pronominal objects. The percentage of nominal objects in written French (69.96%) is only 1.4 times as much as that in spoken French (50.88%), while the percentage of pronominal objects in spoken French (25.62%) is twice as much as that in written French (12.14%). The percentages of subordinating conjunction introduced clauses occupying the role of object are similar (around 12%) in written and spoken French.

[3]$p < 0.01$ **; $p < 0.001$ ***

[4]These figures are not very far from what Labbé found in his corpus: 19.3% in written, 13.8% in the spoken corpus.

https://doi.org/10.1017/S0959269519000334 Published online by Cambridge University Press

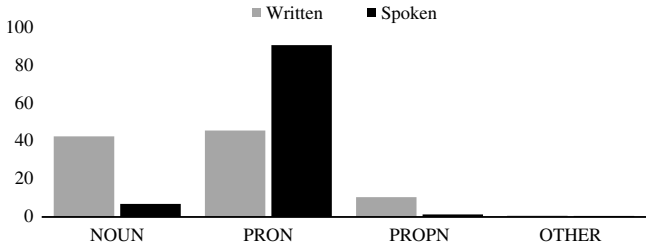

Figure 3. The distribution of POS occupying the role of subject.

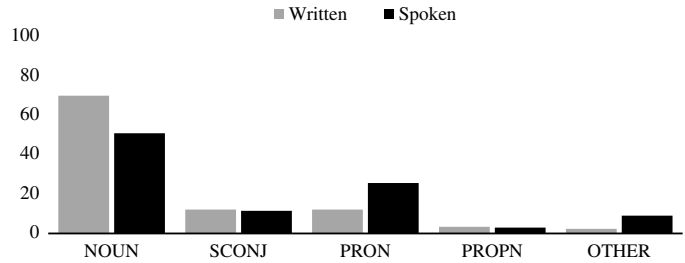

Figure 4. The distribution of POS occupying the role of object.

These results reflect one of the Preferred Argument Structure constraints, which is that lexical noun phrases rarely occupy the subject role of a transitive clause (Du Bois, 1987). As stated by Mazur-Palandre (2015), 'accumulating the production of lexical noun phrases and putting them in subject position seems to be much too costly. To avoid such a cognitive burden, lexical nouns are preferentially in non-subject position'. This is consistent with our findings for spoken French, in which the most frequent clause pattern may not be a SVO, but a VO pattern (Blanche-Benveniste *et al.*, 1990; Blanche-Benveniste, 1995; François, 1974; Jeanjean, 1980; Lambrecht, 1987; Ashby and Bentivoglio, 1993). In sum, written French verifies an SVO clause pattern, whereas spoken French verifies a VO clause pattern.

4. SYNTACTIC RELATIONS

We can distinguish micro-syntactic relations (we call syntactic functions), which describe strong cohesion between words (such as *subject* and *direct object*), from macro-syntactic relations, which describe the relation of non-governed elements (such as *discourse* and *parataxis*). Some relations are paradigmatic (such as *conjunct* and *disfluency*). Table 4 below shows differences in the distributions of syntactic relations in spoken and written French.

4.1. An overview of the syntactic relations' frequency

4.1.1 The two absent relations: discourse and disfluency

The *discourse* and *disfluency* relations only occur in the spoken treebank. The *discourse* relation accounts for 6.69% (1,937 ex.) of the relations and the

https://doi.org/10.1017/S0959269519000334 Published online by Cambridge University Press

Table 4. Percentages of each relation in the written and spoken treebanks, significance of the frequency difference and effect size[5]

Relation	Written	Spoken	Significance	Effect size
noun modifier	23.48% (6,806 ex.)	9.46% (2,741 ex.)	$p < 0.001$	0.3864
subject	6.45% (1,871 ex.)	11.99% (3,472 ex.)	$p < 0.001$	0.1936
prep. and sub. conj	18.1% (5,248 ex.)	11.35% (3,288 ex.)	$p < 0.001$	0.1917
dislocation	0.01% (4 ex.)	0.72% (208 ex.)	$p < 0.001$	0.1499
determiner	13.67% (3,963 ex.)	9.5% (2,751 ex.)	$p < 0.001$	0.1308
copula	1% (291 ex.)	2.46% (712 ex.)	$p < 0.001$	0.1147
conjunct	3.42% (992 ex.)	1.85% (537 ex.)	$p < 0.001$	0.0991
passive auxiliary	1.18% (343 ex.)	0.5% (145 ex.)	$p < 0.001$	0.0761
verb modifier	8.82% (2,558 ex.)	10.96% (3,173 ex.)	$p < 0.001$	0.0718
parataxis	0.23% (68 ex.)	0.68% (197 ex.)	$p < 0.001$	0.0692
direct object	4.43% (1,285 ex.)	5.69% (1,647 ex.)	$p < 0.001$	0.0576
expletive	0.65% (187 ex.)	1.10% (320 ex.)	$p < 0.001$	0.0487
appositional modifier	0.83% (240 ex.)	0.46% (132 ex.)	$p < 0.001$	0.0467
oblique object	3.1% (899 ex.)	2.55% (739 ex.)	$p < 0.001$	0.0332
ellipsis	0.08% (24 ex.)	0.02% (7 ex.)	$p < 0.01$	0.0283
coordinating conjunction	2.72% (789 ex.)	3.09% (894 ex.)	$p > 0.01$	0.0220
tense auxiliary	1.61% (466 ex.)	1.45% (421 ex.)	$p > 0.05$	0.0130
open clausal comp.	1.83% (531 ex.)	1.72% (497 ex.)	$p > 0.05$	0.0083
vocative	0.12% (34 ex.)	0.1% (28 ex.)	$p > 0.05$	0.0060
causative	0.05% (14 ex.)	0.04% (13 ex.)	$p > 0.05$	0.0047
discourse	0 ex.	6.69% (1,937 ex.)		
disfluency	0 ex.	3.94% (1,140 ex.)		

The effect size indicates the size of difference between spoken and written French relations frequency. The four relations where the effect size is the highest are *noun modifier* (0.3864), *subject* (0.1936), *preposition and subordinating conjunction* (0.1917), and *dislocation* (0.1499).

disfluency relation, 3.94% (1,140 ex.). Of the 22 relations of the spoken treebank listed in Table 4, *discourse* and *disfluency* relations are the sixth and eighth most frequent relations. This means that they play a significant role in spoken French. The spoken conception is defined by parameters of spatio-temporal co-presence, intense communicative cooperation and acting, and situational anchoring. Discourse particles (*bon*, *eh*, *bah*), which punctuate the speech are material effects of these conceptional characteristics (1).

[5]For sake of space and clarity, *fixed* and *dep* relations are not shown in Table 4. The *fixed* relation is meant to describe multi-word expressions, and *dep* is a default relation.

https://doi.org/10.1017/S0959269519000334 Published online by Cambridge University Press

(1) **bah** honnêtement pas vraiment (Spoken / Interview and conversation / CFPP)
'well honestly not really'

The spoken modality implies that the speaker undergoes time pressure on the one hand, and on the other hand that the production is invisible and impermanent. As a result, the speaker cannot modify his production (Gadet, 2007b: 49; Mazur-Palandre, 2015: 29). Hesitations, repetitions (2a) and reformulations (2b) described by the *disfluency* relation are material effects of these medial characteristics.

(2a) c'est **c'est c'est** surtout l'hôpital qui m'attire (Spoken / Interview and conversation / CFPP)
'it' it's it's especially hospital that attracts me'
(2b) là je viens de faire mes **des** vaccins par exemple (Spoken / Interview and conversation / CFPP)
'I just made my some vaccines for example'

4.1.2 Subjects and copula

The percentage of subjects in spoken French (11.99%, 3,472 ex.) is almost twice as that in written French (6.45%, 1,871 ex.), and the difference is significant ($Z(1) = 479.73$, $p < 0.001$). This is not astonishing due to the higher frequency of verbs in the spoken treebank, and consistent with the assumption that the spoken language is a mode of action (Halliday, 1985: 81): an action implying a process, the verb and an agent, prototypically the subject. Apart from these general principles, specific grammatical phenomena tend to explain the preference of spoken French for subjects, for instance the frequently used illocutionary units which imply a subject, like *je vois* ('I see') or *je pense* ('I think'). And the important proportion of subjects in spoken French has to be associated with the frequency of attributives, described with the *copula* relation.

In spoken French, the percentage of the *copula* relation is 2.46% (712 ex.), in written French, it is 1% (291 ex.). The difference of frequency is significant ($Z(1) = 176.71$, $p < 0.001$). As in any attributive construction, the complement of copula can either be a nominal phrase (3), an adjective, a prepositional phrase, a pronoun, a proper noun or an adverb.

(3) c'est **un fauteil crapaud** un véritable (Spoken / Interview and conversation / PFC)
'it's an easy chair a real one'

In written French, 42.61% (124 ex.) of the subjects of this relation (when there are no auxiliaries or semi-auxiliaries) are pronouns, and 36.77% (107 ex.) are nouns. In spoken French, 83.01% (591 ex.) are pronouns, and 8.57% (61 ex.) are nouns.

4.1.3 The modifying and argumental relations

Noun modifiers can be adjectives (*un pays* ***formidable*** 'a great country'), prepositional phrases (*la fin* ***de la guerre*** 'the end of the war'), nominal phrases (*activité* ***théâtre*** 'theatre activity'), etc. The percentage of the *noun modifier* relation is higher in the written treebank (23.48%, 6,806 ex.) than in the spoken

https://doi.org/10.1017/S0959269519000334 Published online by Cambridge University Press

treebank (9.46%, 2,741 ex.), and the difference of frequency is significant ($Z(1) = 1730.8$, $p < 0.001$). This is probably due to the great number of nouns and nominalizations in written French (Gadet, 1996: 23). In (4), *traitement* ('treatment') is a nominalization of the verb *traiter* ('to treat'), the instrument role is realized by the prepositional phrase *par Aclasta* ('by Aclasta'). Similarly, *renouvellement* ('turnover') is a nominalization of *renouveler* ('to renew'). The patient role is realized by the adjective *osseux* ('bony').

(4) Le traitement [par Aclasta] réduit rapidement la vitesse [de renouvellement [osseux]], à partir de taux [post-ménopausiques] [élevés]. (Written / Professional report / Emea)
'Aclasta treatment rapidly reduces the rate of bone turnover from high postmenopausal levels'

This statement is corroborated by the fact that in the written corpus, nominal noun modifiers account for 4.43% of all the nouns, whereas in the spoken corpus they occupy 1.37%. On the contrary, the *verb modifier* relation is significantly more frequent in the spoken treebank than in the written treebank ($Z(1) = 65.996$, $p < 0.001$). The distribution of nouns and verb modifiers has much to do with the distribution of nouns and verbs presented in subsection 3.1. Verbs are more frequent in the spoken French corpus, and as a result, the *verb modifier* relation is also more frequent. Verbal verb modifiers account for 2.68% of all the verbs in written, and their proportion is lower in the spoken corpus with 1.52%. All these results lead to the same conclusion that written French has a greater preference for modification than does spoken French.

The z-test indicates that the objects are significantly ($Z(1) = 44.694$, $p < 0.001$) more frequent in the spoken treebank (5.69%, 1,647 ex.) than in the written treebank (4.43%, 1,285 ex.). The percentage of the verbs to appear with the realization of an object in the spoken treebank is 44.16%, while 43.34% appear in the written treebank. On the contrary, oblique objects[6] are significantly ($Z(1) = 15.629$, $p < 0.001$) more frequent in the written treebank (3.1%, 899 ex.), than in the spoken treebank (2.55%, 739 ex.). The percentage of the verbs to appear with the realization of an oblique object in the written treebank is 27.66%, and 20.05% in the spoken treebank. The difference of the frequency of *open clausal complement* relations (*Les parents ne semblent pas* ***connaître les dangers*** 'Parents seem not being aware of the dangers'; *étant considérée* ***comme accidentelle*** 'being considered as accidental') is not significant ($Z(1) = 1.1245$, $p > 0.05$).

4.1.4 The grammatical relations

Chafe (1979, cited by Redeker, 1984:44) reported that the written language has more passives than the spoken language, which is confirmed by our data: 1.18% (343 ex.)

[6]In our treebanks, the *oblique object* relation subsumes prepositional arguments which can be pronominalized by a dative clitic (*parler à Marie* 'talk to Mary' ~ *lui parler* 'speak to her') and all other prepositional arguments (*penser à Marie* 'think of Mary'). The terminological choice to call 'oblique objects' rather than 'indirect objects' these prepositional arguments was made by the producers of the treebanks, and we maintained it for sake of consistency.

https://doi.org/10.1017/S0959269519000334 Published online by Cambridge University Press

of the relations in the written treebank are *passive auxiliary* relations, whereas only 0.5% (145 ex.) in the spoken treebank. The difference in frequency is significant ($Z(1) = 80.336$, $p < 0.001$). The *determiner* and the *preposition and subordinating conjunction* relations' frequencies in the written treebank are significantly higher than in the spoken treebank. The analysis on the noun part of speech and the *noun modifier* relation above explains the reason for these distributions: the more the nouns, the more the determiners, and the more the noun modifiers, the more the prepositions (i.e. *la fin* ***de la guerre*** 'the end of the war', *mes études* ***de médecine*** 'my medical studies'). There are fewer *tense auxiliary* relations in spoken French (1.45%, 421 ex.) than in written French (1.61%, 466 ex.), however, the difference is not significant ($Z(1) = 2.283$, $p > 0.05$). The next section presents the distributions of two phenomena often discussed in studies on spoken language, namely dislocation and parataxis.

4.2. Dislocation and parataxis

4.2.1 Dislocations

Dislocation is a common phenomenon in spoken French (Larsson, 1979; Campion, 1984; Barnes, 1985; Lambrecht, 1994, 2001; Blasco-Dulbecco, 1999; De Cat, 2002, 2007; Prévost, 2003; Avanzi, 2012). It is impossible to do justice here to all syntactic and pragmatic aspects that have been discussed in the abundant literature on the subject. We recommend the reader to De Cat (2007) for a thorough study on the interaction between prosody, cognition, pragmatics and syntax at play in this phenomenon. We will limit ourselves to give an overview of how dislocation is represented in our written and spoken corpora, based on a broadly accepted definition and taxonomy. We can define dislocations as grammatical constructions that serve to mark a constituent as denoting the topic (or theme) with respect to which a given sentence expresses a relevant comment (e.g. Dik, 1978; Gundel, 1988; Lambrecht, 1981, 1994, 2001). The referent of the dislocated constituent has to be accessible in the hearer's short-term memory (Lambrecht, 2001: 1075; Horváth, 2018: 51). When the dislocated constituent is placed on the left side of the sentence, it is a left dislocation; on the right, it is a right dislocation.[7] Based on this definition, four criteria can be used to identify a dislocation (Lambrecht, 2001: 1050):

(i) extra-clausal position of a constituent
(ii) possible alternative intra-clausal position
(iii) pronominal co-indexation
(iv) special prosody

These four criteria can only be met simultaneously in typical cases. In fact, only the first one is necessary to identify a dislocation. For example, a dislocated constituent can neither have a possible alternative intra-clausal position, nor be co-indexed with a resumptive element in the clause, as shown in the example (5) from (Barnes, 1985: 101):

[7]De Cat (2007) data showed that left and right dislocations are equally distributed in spontaneous French.

https://doi.org/10.1017/S0959269519000334 Published online by Cambridge University Press

Table 5. Frequencies of different types of dislocations

Subrelation	Written	Spoken	Significance
dislocated subject	4 ex.	171 ex.	$p < 0.001$
dislocated direct object	0 ex.	11 ex.	
dislocated oblique object	0. ex.	4 ex.	
unlinked dislocation	0.ex.	22 ex.	

(5) Le métro, avec la carte orange, tu vas n'importe où.
'The subway, with Orange Card, you go anywhere.'

This kind of dislocated constituent is called unlinked dislocated constituents. Unlinked dislocations can be further classified (Barnes, 1985; Fradin, 1990; Stark, 1999; Horváth, 2018), or included in a broader category, which is hanging topic (Deulofeu, 1979; Berrendonner and Reichler-Béguelin, 1997). When the dislocated constituent is resumed by an element in the clause (a clitic or the pronoun *ça*), it can have one of the following syntactic roles: subject, direct object, oblique object or modifier. The presence of the pronoun in the clause is an important indicator distinguishing this dislocated constituent from a typical modifier of the clause. Compare the modifier *aujourd'hui* 'today' in the sentence *Aujourd'hui, Pierre ne travaille pas* ('Today, Pierre doesn't work') with the dislocated constituent *sur le pont* ('on the bridge'), resumed by the clitic *y*, in $[Sur\ le\ pont\ d'Avignon]_i$, *on* y_i *danse tout en rond* ('On the Avignon bridge, people dance all around') (Lambrecht, 2001: 1055). We did not find, in our corpus, dislocated sentences where the clitic has the function of modifier. We distinguished different subrelations of the relation *dislocation*, according to the syntactic role played by the co-indexed pronoun. Table 5 shows the distributions of these subrelations in our two corpora.

In the example below, the element that resumes the dislocated constituent within the clause is underlined. In the written treebank, the only subrelation that appears is dislocated subject. There are significantly ($Z(1) = 159.37$, $p < 0.001$) more dislocated subjects in the spoken corpus (6b) than in the written corpus (6a).

(6) *dislocated subject*
   a. **Faire s'exprimer les enfants à travers cette activité**, c'est important. (Written / Print media / Annodis)
   'Let children express themselves throughout this activity, it's important'
   b. et **moi** je suis allé en Ethiopie […] (Spoken / Interview and conversation / Lacheret)
   'and me I went to Ethiopia […]'

The *dislocated direct object* (7), *dislocated oblique object* (8) and *unlinked dislocation* (9) subrelations occur only in the spoken treebank:

(7) *dislocated direct object*
tel tel et tel cas on les verrait pas en hô~ en hôpital privé (Spoken / Interview and conversation / CFPP)
'such such and such case we won't see them in a private hospital'

https://doi.org/10.1017/S0959269519000334 Published online by Cambridge University Press

Table 6. Proportions of different types of parataxis

Subrelation	Written	Spoken	Significance
associated illocutionary unit	2 ex.	126 ex.	$p < 0.001$
quoting direct speech	2 ex.	37 ex.	$p < 0.001$
incidental clause	50 ex.	34 ex.	$p > 0.05$
incised clause	9 ex.	0 ex.	
juxtaposition	5 ex.	0 ex.	

(8) *dislocated oblique object*
je vois euh **moi** la fac ça m'a fait beaucoup de bien (Spoken / Interview and conversation / CFPP)
'I see eeh me college it did me a lot of good'

(9) *unlinked dislocation*
bah f~ déj~ déjà **les teintes** bon faut savoir que tu as une base euh pff (Spoken / Interview and conversation / PFC)
'well first of all tints you must know you have a basis eeh pff'

Dislocations are not peculiar to spoken French, as pointed out by Blanche-Benveniste (1991). However, it may be more frequent in spoken than in written French. Dislocations are not only more frequent in spoken French, but they are also more diverse in usage. Besides *dislocation*, the *parataxis* phenomenon is also more frequent in spoken French (Gadet, 1991: 110).

4.2.2 Parataxis

According to the Universal Dependencies annotation scheme, the *parataxis* relation is meant to describe two clauses or sentences placed side by side without any explicit coordination or subordination. This relation describes a heterogeneous set of clausal junctions. It is more relevant to reach a conclusion from the frequency of each of these subtypes than from the relation *parataxis* alone. Table 6 below shows the percentages of these different subtypes of parataxis in both treebanks.

The difference in frequency of associated illocutionary units in the two corpus is significant ($Z(1) = 120.12$, $p < 0.001$). Associated illocutionary units are idiomatic expressions that punctuate the speech of a speaker (*écoute* 'listen', *tu vois* 'you see', *on dirait* 'it seems'). In this respect, they may be found in written-to-be-spoken genres like political discourse (see 10a). The spoken conception is typically dialogical, which implies more involvement of the locator in the communicative situation. This may explain the high frequency of associated illocutionary units, which are material effects of these conceptional characteristics in spoken discourse (10b).

(10) *associated illocutionary unit*
a. Permettez-moi enfin de vous dire que notre Parlement s'est, **je crois**, très largement retrouvé dans les propos que vous avez tenus. (Written / Parliamentary debates / Europarl)

https://doi.org/10.1017/S0959269519000334 Published online by Cambridge University Press

'Finally, let me tell you that our Parliament, I believe, widely agree with what you have declared.'

b. les policiers sont arrivés en raison du du du vacarme **je p~ je pense** (Spoken / Monologue / Rhapsodie-Movie)
'Policemen came because of the din I think'

As Redeker (1984: 44) puts it, 'speakers and listeners in a typical conversational situation tend to be more involved in their communication than writers and readers'. For Chafe (1979), this involvement results in an important usage of direct speech. And indeed, we found a significant difference ($Z(1) = 31.41$, $p < 0.001$) of *quoting direct speech* between the frequency of the written (11a) and the spoken treebanks (11b):

(11) *quoting direct speech*

a. Il est inconcevable que la Comission puisse dire **"cela n'est pas très important pour nous"** […] (Written / Political discourse / Europarl)
'It is incredible that the Commission can say "this is not very important to us"'

b. je me disais **j'irai peut-être à Vire** (Spoken / Interview and conversation / PFC)
'I told to myself I will maybe go to Vire'

Writers prefer to report speech with *incised clause* (12). This relation is absent from the spoken treebank.

(12) *incised clause*

Jean-Claude Méry, **expliquait-il**, lui avait mis "le couteau sous la gorge". (Written / Narration / FrWiki)
'Jean-Claude Méry, he explained, "put a gun to its head"'

Simple juxtapositions of two independent illocutionary clauses have only been found in the written treebank (13).

(13) *juxtaposition*

Frégates de Taïwan : **l'ancien directeur adjoint de la Société générale témoigne, Sud Ouest, 13 mars 2002** (Written / Print media / Annodis)
'Taiwan frigates: Former Deputy Director of General Society in Taiwan testifies, Sud Ouest, March 13, 2002'

The difference of *incidental clause* relations (14) frequency between the two treebanks is not significant ($Z(1) = 3.0476$, $p > 0.05$).

(14) *incidental clause*

a. […] il est assez incroyable de se trouver dans cette salle – **je ne puis guère parler d'assemblée à ce moment précis** – et de devoir constater que […] (Written / Parliamentary debates / Europarl)
'[…] it is quite incredible to be in this room – I can not speak of an assembly in this actual moment – and to see that […]'

b. alors que Heinze **c'est quand même assez extraordinaire hein** c'est le patron de la défense (Spoken / Soccer match commentaries / Rhapsodie-Broadcast)
'whereas Heinze it's quite extraordinary he's the boss of defense'

https://doi.org/10.1017/S0959269519000334 Published online by Cambridge University Press

Table 7. Percentages of word order in each relation

	G→D		D←G	
Relation	Written	Spoken	Written	Spoken
vocative	14.7% (5 ex.)	39.3% (11 ex.)	85.3% (29 ex.)	60.7% (17 ex.)
oblique object	88.3% (794 ex.)	74.6% (551 ex.)	11.7% (105 ex.)	25.4% (188 ex.)
direct object	89.3% (1,148 ex.)	79.5% (1,309 ex.)	10.7% (137 ex.)	20.5% (338 ex.)
subject	4.5% (85 ex.)	2.7% (94 ex.)	95.5% (1,786 ex.)	97.3% (3,377 ex.)
copula	94.2% (274 ex.)	95.5% (680 ex.)	5.8% (17 ex.)	4.5% (32 ex.)

This section presented an overview of the distributions of POS and syntactic relations in both corpora, the next section will give more details about word order.

5. WORD ORDER

5.1. The distributions of word order in some syntactic relations

As defined earlier, a relation is a labeled asymmetrical link between two linguistic units: G—r→D where G is the Governor, D the Dependent, and r the label of the relation. According to Tesnière (1959: 22), if the dependent precedes the governor, the order is governor-final; if the dependent follows the governor, the order is governor-initial. This is defined as the dependency direction of a dependency relation (Liu, 2010). Dependency direction is a useful concept in comparative studies on the syntax of different languages or different genres (Liu, Zhao and Li, 2009; Liu, 2010; Jiang and Liu, 2015). Table 7 shows the percentages of word order for each relation in the two treebanks. Many relations in spoken and written French do not present much difference in terms of word order because the order is fixed, determined by grammatical rules, such as the relations of *determiner*, *expletive*, *open clausal complement*, *tense auxiliary* and *causative*. We also excluded the relations of *apposition*, *conjunction*, *parataxis* and *disfluency* from Table 7, because they are orthogonal to syntax. The *dislocation* relation's occurrences are too scarce in the written treebank (4 ex.) to reach any conclusions. The relations *verb modifier* and *noun modifier* describe large arrays of linguistic facts, consequently they are also not considered here.

Table 8 shows that *vocative*, *direct object*, *oblique object* are the first three relations that give rise to the most obvious difference in terms of word order. If vocatives in the written treebank are placed before the head of a clause (85.3%) (15a) in most cases, the word order is more variable in the spoken treebank (60.7%) (15b).

(15a) **Madame la Présidente**, le président de groupe M. Barón Crespo s'est aussi adressé à moi. (Written / Parliamentary debates / Europarl)
'Madam President, the group president Mr. Barón Crespo also addressed me'

https://doi.org/10.1017/S0959269519000334 Published online by Cambridge University Press

Table 8. Constructions of the oblique object relation

Written		Spoken	
VERB→PREP	78.64% (707 ex.)	VERB→PREP	70.37 % (520 ex.)
PRON←VERB	9.79% (88 ex.)	PRON←VERB	24.09% (178 ex.)
ADJ→PREP	6.45% (58 ex.)	VERB→ADV	2.17% (16 ex.)
VERB→PRON	1.33% (12 ex.)	ADJ→PREP	0.54% (4 ex.)
PREP←VERB	1.22% (11 ex.)	PREP←VERB	0.54% (4. ex)
...		...	

(15b) **Emmanuelle** est-ce que vous avez déjà fait sortir un amant ou une maîtresse par la fenêtre en catastrophe ? (Spoken / Conversation on radio / Rhapsodie-Broadcast)
'Emmanuelle, have you ever brought a lover out the window in a panic?'

The distributions of word order in the *copula* and *subject* relations are similar in the written and spoken treebanks, that is, subjects are rarely governor final (4.5% in written, 2.7% in spoken).

5.2. The difference of word order in relations direct object and oblique object

5.2.1 Oblique objects

In the written treebank, 88.3% of the *oblique object* relations are governor-initial; whereas in the spoken treebank, the percentage is 74.6%. In order to better interpret these results, we have to further analyse the corresponding constructions of this relation. Table 8 shows a higher percentage of the PRON←VERB construction in spoken French (24.09%) than in written French (9.79%).

This may be due to the preference of spoken language for pronouns, as mentioned in section 3.1. Additionally, a grammatical rule imposes that clitic pronouns are placed before the verbs on which they depend (except if the verb is in the imperative modality). This increases significantly the proportion of governor-final *oblique object* relations. The same logic stands to explain the word-order difference in the *direct object* relation between spoken and written French.

5.2.2 Direct objects

The *direct object* relation presents differences in word order between spoken and written French. In written French, 89.3% of the *direct object* relations are governor-initial, while in the spoken treebank, 79.5%. Table 9 shows the most frequent constructions corresponding to the *direct object* relation with the percentages of frequency.

Table 9 indicates that the PRON←VERB construction is much more frequent in the spoken treebank (20.04%) than in the written treebank (10.66%). This section actually provides the evidence that spoken and written French are two systems of the same language. They share the same grammatical rules, namely that the objects

https://doi.org/10.1017/S0959269519000334 Published online by Cambridge University Press

Table 9. Constructions of the direct object relation

Written		Spoken	
VERB→NOUN	69.96% (899 ex.)	VERB→NOUN	50.46% (831 ex.)
VERB→SCONJ	11.05% (142 ex.)	PRON←VERB	20.04% (330 ex.)
PRON←VERB	10.66% (137 ex.)	VERB→SCONJ	9.47% (156 ex.)
VERB→PROPN	3.42% (44 ex.)	VERB→PRON	5.4% (89 ex.)
VERB→PRON	1.48% (19 ex.)	VERB→PROPN	2.98% (49 ex.)
...		...	

have to follow the verb they depend on, and that the subjects have to precede. If we observed a difference of word order in oblique object and direct object relations, this difference can be explained by the preference of spoken French for clitic pronouns. A real difference of word order between written and spoken French is manifested on the macro-syntactic level with vocative nominal phrases. It is insisted by some that spoken French is easier than written French. This belief may be rooted in the higher frequency of dislocations and parataxis in spoken French. Indeed, the speech is fragmented by dislocations and parataxis into short blocks, which could leave the impression of simple structures. In the next section, we try to cast some doubt in this lasting belief.

6. COMPREHENSION DIFFICULTY

6.1. The Mean Dependency Distance of the corpora

According to Halliday (1985), both spoken and written languages are complex but not in the same way: 'The complexity of the written language is static and dense. That of the spoken language is dynamic and intricate' (Halliday, 1985: 87). Halliday employs different criteria to evaluate their complexity. In terms of intricacy of movement, spoken language is more complex; but in terms of density of substance, written is more complex. What if we compare spoken and written language complexity in terms of the same criterion? Yngve (1960) described the depth of a sentence as 'the maximum number of symbols needed to be stored during the construction of a given sentence'. The depth of a sentence cannot exceed a certain threshold, which is nearly equal to the capacity of human working memory (Miller, 1956; Cowan, 2001, 2005). By introducing the Depth Hypothesis, Ygnve addressed the need of a universal metric for language comprehension difficulty. The principle of Early Immediate Constituent and the Dependency Locality Theory (Hawkins, 1994; Gibson, 1998, 2000) then established further a link between linear order and comprehension difficulty. They have been tested experimentally on different languages (Gibson, 1998; Hsiao and Gibson, 2003; Grodner and Gibson, 2005). Grodner and Gibson (2005) emphasized on the fact that 'the difficulty associated within integrating a new input item is heavily determined by the amount of lexical material intervening between the input item and the site of its target dependents'. In

https://doi.org/10.1017/S0959269519000334 Published online by Cambridge University Press

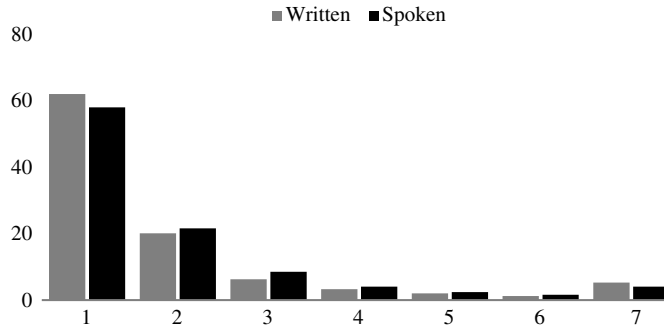

Figure 5. Dependency distances frequencies.

other words, the longer the distance between two words, the more the working memory is affected, the more the processing difficulty arises. The dependency distance of a sentence in a dependency treebank can be computed by the following method (Liu, 2008). For any dependency relation between the two words W_a and W_b, the dependency distance is the difference in their positions in the sentence: *a*-*b*. For adjacent relations, the dependency distance is 1. Figure 5 displays the distribution of DD in both treebanks. The DD in Figure 5 is the absolute value, that is |*a*-*b*|.

The written treebank has a higher percentage of adjacent dependencies (61.93%) than the spoken treebank (57.92%). The higher percentage of adjacent dependencies in written French may be caused by the large number of relations that impose the dependent and the governor to be close to each other: *determiner*, *tense auxiliary* and *preposition and subordinating conjunction*. The overall complexity of a sentence is measured by the mean dependency distance (MDD), which is defined as follows:

$$MDD(the\ sentence) = \frac{1}{n-1}\sum_{i=1}^{n-1}|DD_i| \quad [1]$$

In this formula, *n* is the number of words in the sentence and |DDi| is the absolute dependency distance of the i-th syntactic link of the sentence. In the sentence *Charles likes little dogs* (Figure 1), the distance of the three dependencies are respectively 1, 2 and 1. The DD of the root node is 0. Applying Formula [1], we can compute the MDD of this sentence, which is 4/3 = 1.33. We give two examples to show their different MDDs:

(16) Suzanne Sequin n'est plus. (Written / Print Media / Annodis)
'Suzanne Sequin is no more.'
MDD = 1.5

(17) donc on peut penser que c'est une tradition euh ici qui est représentée (Spoken / Interview and conversation / Interview classical music)
'so we can think that it is a tradition eh here which is represented'
MDD = 1.92

MDD can also be used as a complexity measure of a text or a collection of texts, computed with the following formula:

https://doi.org/10.1017/S0959269519000334 Published online by Cambridge University Press

Table 10. MDD of the syntactic functions

Relation	Written	Spoken
subject	2.58	1.36
noun modifier	−1.47	−1.01
oblique object	−1.42	−0.75
direct object	−2.03	−1.46
open clausal complement	−1.81	−1.56

$$MDD(the\ sample) = \frac{1}{n-s}\sum_{i=1}^{n-s}|DD_i| \quad [2]$$

In this formula, n is the number of words, s is the total number of sentences. The longer the MDD of a sentence, the more difficult the sentence is; and the longer the MDD of a text, the more difficult the text is. MDD has been proved to be an efficient index for studies on language typology and genre (Liu, Zhao and Li, 2009; Liu and Xu, 2012; Wang and Liu, 2017; Liu, Xu and Liang, 2017).

Using Formula [2], we computed the MDDs of the written and the spoken treebanks, which are rather similar: the MDD of spoken French is 2.1 and the MDD of written French is 2.13. Adjacent dependencies play an important role in minimizing dependency distance (Liu, 2008). Annotation scheme of the treebanks is also another factor that impacts this measure (Jiang and Liu, 2015; Yan and Liu, 2019), and that has to be taken into account in order to better interpret the results.

6.2. The Mean Dependency Distance of the relations

In this section, we use Formula [3] to compute the MDD of these major syntactic functions: *subject*, *direct object*, *oblique object*, *open clausal complement* and *noun modifier*.

$$MDD(relation) = \frac{1}{n}\sum_{i=1}^{n} DD_i \quad [3]$$

In Formula [3], n is the number of occurrences of this relation, and DDi is the distance of the i-th dependency that belong to this type. If the result is positive, this means that the relation tends to be governor-final. If it is negative, the relation tends to be governor-initial. The MDD of these relations are displayed in Table 10.

Table 10 shows that except for the subject, all other syntactic functions tend to be governor-initial. In addition, the MDD of these relations are greater in written French than in spoken French. For instance, the MDD of the subjects in written French is 2.58, and 1.36 in spoken French. However, the difference of both languages' MDD seems to be rather slight, which suggests no substantial difference in comprehension. In other words, we cannot firmly claim that spoken French is less difficult to process than written French.

https://doi.org/10.1017/S0959269519000334 Published online by Cambridge University Press

It would be noteworthy to investigate the reasons why the MDDs of spoken and written French treebanks are similar while the MDD of each syntactic function visually presents differences. It would also be interesting to study the MDD across the genres of French, as it has been previously done on Chinese (Liu, Zhao and Li, 2009) and English (Wang and Liu, 2017). In particular, this measure could help to pursue the investigation on the relationship between the genres and the media (Biber, 1988; Biber and Conrad, 2003). How different is the MDD of French written narrations, political discourses, scientific conferences, spontaneous conversations (online messages) from that of its spoken counterparts?

7. CONCLUSION

Based on syntactically annotated corpora, our research quantitatively probed into the grammatical features of the genres of a rather distant written French and a rather immediate spoken French (we called written and spoken French). Confirming the lasting assumption that written and spoken French do not differ in the syntactic categories but in the frequencies of these categories, we showed that written and spoken French have different distributions of parts of speech and syntactic relations. The quasi totality of subjects in spoken French is pronouns, and more diverse dislocated sentences are more frequently used. A significant difference of word order between written and spoken French has been found in the placement of vocatives. The Mean Dependency Distance (MDD) is slightly higher in written French than in spoken French. The same difference in dependency distance is also found in syntactic functions, especially the subject.

ACKNOWLEDGEMENTS. We thank three anonymous reviewers and Dr Chunshan Xu for valuable suggestions and comments. This study was supported by the National Social Science Foundation of China (Grant No. 17AYY021) and the MOE Project of the Center for Linguistics and Applied Linguistics, Guangdong University of Foreign Studies.

REFERENCES

Arnold, J. E. (2001). The effect of thematic roles on pronoun use and frequency of reference continuation. *Discourse Processes*, **31.2**: 137–162.

Ashby, W. and Bentivoglio, P. (1993). Preferred argument structure in spoken French and Spanish. *Language Variation and Change*, **5.1**: 61–76.

Avanzi, M. (2012). *L'interface prosodie/syntaxe en français. Dislocations, incises et asyndètes*. Bruxelles: Peter Lang.

Avanzi, M., Gendrot, C. and Lacheret, A. (2010). Is there a prosodic difference between left-dislocated and heavy subjects? Evidence from spontaneous speech. Proceedings of the 5th Speech Prosody International Conference (SP'10). Chicago, United States, May 2010.

Barnes, B. K. (1985). *The Pragmatics of Left Detachment in Spoken Standard French*. Amsterdam: Benjamins.

Béguelin, M.-J. (1998). Le rapport écrit-oral. Tendances dissimilatrices, tendances assimilatrices, *Cahiers de Linguistique Française*, **20**: 229–253.

Berman, R. A. and Verhoeven, L. (2002). Cross linguistic perspectives on the development of text-production abilities: Speech and writing. *Written Language and Literacy*, **5.1–2**: 1–43.

Berrendonner, A. and Reichler-Béguelin, M.-J. (1997). Left-dislocation in French: Varieties, norm and usage. In: J. Cheshire and D. Stein (eds) *Taming the Vernacular. From Dialect to Written Standard Language*. London and New York: Longman, pp. 200–217.

Berns, J. (2015). Merging low vowels in metropolitan French. *Journal of French Language Studies*, **25.3**: 317–338.

Blanche-Benveniste, C. (1991). Les études sur l'oral et le travail d'écriture de certains poètes contemporains. *Langue Française*, **89**: 52–71.

Blanche-Benveniste, C. (1994). Quelques caractéristiques grammaticales des 'sujets' employés dans le français parlé des conversations. *Proceedings of the Conference Subjecthood and Subjectivity*. Paris/ London: Ophrys and Institut français du Royaume-Uni, pp. 77–107.

Blanche-Benveniste, C. (1995). Le semblable et le dissemblable en syntaxe. *Recherches sur le Français Parlé*, **13**: 7–32.

Blanche-Benveniste, C. (1997). *Approches de la langue parlée en français*. Paris: Ophrys.

Blanche-Benveniste, C. and Jeanjean, C. (1987). *Le français parlé: Transcription et édition*. Paris: Didier Erudition.

Blanche-Benveniste, C., Bilger M., Rouget C. and Van den Eyende K. (1990). *Le français parlé : Études grammaticales*. Paris: Éditions du CNRS.

Blasco-Dulbecco, M. (1999). *Les dislocations en français contemporain*. Paris: Honoré Champion.

Biber, D. (1988) *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, D. and Conrad, S. (2003). Register variation: A corpus approach. In: D. Schiffrin, D. Tannen, and Hamilton D. (eds), *Handbook of Discourse Analysis*, London: Blackwell, pp. 175–196.

Brunetti, L., Avanzi, M. and Gendrot, C. (2013). A quantitative study of sentence topic and its syntactic/ prosodic correlates on a French spoken corpus: Methodological and theoretical issues. *Proceedings of the Information Structure in Spoken Language Corpora Workshop*. Bielefeld, Germany, June 2013.

Campion, E. (1984). *Left Dislocation in Montréal French*. Ph.D. dissertation, University of Pennsylvania.

Chafe, W. (1979). Integration and involvement in spoken and written language. *Proceedings of the 2nd Congress of the International Association for Semiotic Studies.* Vienna, Austria, July 1979.

Chafe, W. and Tannen, D. (1987). The relation between written and spoken language. *Annual Review of Anthropology*, **16.1**: 383–407.

Coveney, A. (2002). *Variability in Spoken French*. Bristol: Intellect.

Coveney, A. (2004). The alternation between "l'on" and "on" in spoken French. *Journal of French Language Studies*, **14.2**: 91–112.

Coveney, A. (2011). A language divided against itself? Diglossia, code-switching and variation in French. In: F. Martineau and T. Nadasdi (eds), *Le français en contact*. Québec: Presses de l'Université Laval, pp. 51–85.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, **24**: 87–185.

Cowan, N. (2005). *Working Memory Capacity*. Hove, East Sussex, UK: Psychology Press.

De Cat, C. (2002). *French Dislocation*. Ph.D. dissertation, University of York.

De Cat, C. (2005). French subject clitics are not agreement markers. *Lingua*, **115.9**: 1195–1219.

De Cat, C. (2007). *French Dislocation. Interpretation, Syntax, Acquisition*. New York: Oxford University Press.

De Cat, C. (2011). Information tracking and encoding in early L1: Linguistic competence vs. cognitive limitations. *Journal of Child Language*, **38.4**: 828–860.

De Cat, C. (2012). Explaining children's over-use of definites in 'partitive contexts'. *First Language*, **32.1–2**: 137–150.

Deulofeu, J. (1979). Les énoncés à constituant lexical détaché. *Recherches sur le Français Parlé*, **2**: 75–108.

Dik, S. C. (1978). *Functional Grammar*. Amsterdam: North Holland.

Du Bois, J. W. (1987). The discourse basis of ergativity. *Language*, **63**: 805–855.

Fayol, M. (1997). *Des idées au texte : Psychologie cognitive de la production verbale orale et écrite*. Paris: Presse Universitaire de France.

Fradin, B. (1990). Approche des constructions à détachement: Inventaire. *Revue Romane*, **25.1**: 3–14.

François, D. (1974). *Français parlé. Analyse des unités phoniques et significatives d'un corpus recueilli dans la région parisienne*. Paris: S.E.L.A.F.

Gadet, F. (1991). Le parlé coulé dans l'écrit: Le traitement du détachement par les grammairiens du XXème siècle. *Langue Française*, **89.1**: 110–124.

Gadet, F. (1996). Une distinction bien fragile : Oral/écrit. *Revue Tranel (Travaux Neuchâtelois de Linguistique)*, **25**: 13–27.

Gadet, F. (2007a). La variation de tous les français. *Linx: Revue des Linguistes de l'Université Paris X Nanterre*, **57**: 155–164.

Gadet, F. (2007b). *La variation sociale en français*, 2nd edition. Paris: Ophrys.

Gerdes, K., Guillaume, B., Kahane, S. and Perrier, G. (2018). SUD or Surface-Syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD. *Proceedings of the Universal Dependencies Workshop 2018 (UDW'18)*. Brussels, Belgium, November 2018.

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition*, **68.1**: 1–76.

Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. In: *Image, Language, Brain*. A. Marantz *et al.* (eds). Cambridge, MA: The MIT Press, pp. 95–126.

Grevisse, M. and Goosse, A. (2008). *Le Bon Usage*, 14th edition. Bruxelles: De Boeck and Larcier.

Grodner, D. and Gibson, E. (2005). Some consequences of the serial nature of linguistic input. *Cognitive Sciences*, **29**: 261–290.

Gundel, J. K. (1988). Universals of topic-comment structure. In: M. Hammond, E. Moravcsik and J. Wirth (eds), *Studies in Syntactic Typology*. Amsterdam: Benjamins, pp. 203–239.

Halliday, M. (1985). *Spoken and Written Language*. Oxford: Oxford University Press.

Hawkins, J. A. (1994). *A Performance Theory of Order and Constituency*. Cambridge, England: Cambridge University Press.

Hellwig, P. (2003). Dependency unification grammar. In: V. Ágel *et al.* (eds), *Dependency and valency. An International Handbook of Contemporary Research*, Volume **1**. Berlin/New York: De Gruyter, pp. 593–635.

Henry, S. and Pallaud, B. (2003). Word fragments and repeats in spontaneous spoken French. *Proceedings of Disfluency in Spontaneous Speech Workshop (DiSS'03)*. Gothenburg, Sweden, September 2003.

Horváth, M. G. (2018). *Le français parlé informel. Stratégies de topicalisation*. Berlin: De Gruyter.

Hsiao, F. and Gibson, E. (2003). Processing relative clauses in Chinese. *Cognition*, **90**: 3–27.

Hudson, R. (1990). *English Word Grammar*. Oxford: Basil Blackwell.

Hudson, R. (2007). *Language Networks: The New Word Grammar*. Oxford: Oxford University Press.

Jeanjean, C. (1980). *Les formes sujets de type nominal: Étude sur le français contemporain*. PhD dissertation, Provence University.

Jiang, J. and Liu, H. (2015). The effects of sentence length on dependency distance, dependency direction and the implications — Based on a parallel English-Chinese dependency Treebank. *Language Sciences*, **50**, 93–104.

Jisa, H. (1998). Relative clauses in French children's narrative text. *Journal of Child Language*, **25**: 623–652.

Jones, M. A. (1996). *Foundations of French Syntax*. Cambridge: Cambridge University Press.

Koch, P. and Oesterreicher W. (2001). Langage parlé et language écrit. *Lexikon der romanistischen Linguistik*, Volume **1**. Tübingen, Max Niemeyer Verlag.

Labbé, D. (2003). Coordination et subordination en français oral. *Proceedings of the 4ème Journées de l'ERLA*, Brest, France, November 2003. In: D. Banks (ed.) (2007). *La coordination et la subordination dans le texte de spécialité*. Paris: L'Harmattan, 161–182.

Lambrecht, K. (1981). *Topic, Antitopic and Verb Agreement in Non-Standard French*. Amsterdam: Benjamins.

Lambrecht, K. (1987). On the status of SVO sentences. In: R. S. Tomlin (ed.), *Coherence and Grounding in Discourse*. Amsterdam: Benjamins, pp. 217–261.

Lambrecht, K. (1994). *Information Structure and Sentence Form. Topic, Focus and the Mental Representations of Discourse Referents*. Cambridge: Cambridge University Press.

Lambrecht, K. (2001). Dislocation. In: M. Haspelmath *et al.* (eds), *Language Typology and Language Universals: An International Handbook*, Volume **2**, Berlin: De Gruyter, pp. 1050–1078.

Larsson, E. (1979). *La dislocation en français*. Lund: Gleerup.

Le Goffic, P. (1993). *Grammaire de la phrase française*. Paris: Hachette.

Liu, H. (2008). Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, **9.2**: 159–191.

Liu, H. (2010). Dependency direction as a means of word order typology: A method based on dependency treebanks. *Lingua*, **120.6**: 1567–1578.

Liu, H., Zhao, Y. and Li, W. (2009). Chinese syntactic and typological properties based on dependency syntactic treebanks. *Poznań Studies in Contemporary Linguistics*, **45.4**: 509–523.

Liu, H. and Xu, C. (2012). Quantitative typological analysis of Romance languages. *Poznań Studies in Contemporary Linguistics*. **45.4**: 597–625.

Liu, H., Xu, C. and Liang, J. (2017). Dependency distance: A new perspective on syntactic patterns in natural languages. *Physics of Life Reviews*, **21**, 171–193.

Liu, B., Niu, Y., and Liu, H. (2012). Word class, syntactic function and style: A comparative study based on annotated corpora. *Applied Linguistics*, **4**: 134–142, in Chinese.

Liu, B., Niu, Y. and Liu, H. (2013). A comparative study of style-related differences in syntactic functions of part of speech. *Language Teaching and Linguistic Studies*, **5**: 97–104, in Chinese.

Massot, B. (2010). Le patron diglossique de variation grammaticale en français. *Langue Française*, **4**: 87–106.

Massot, B. and Rowlett, P. (2013). Le débat sur la diglossie en France : Aspects scientifiques et politiques. *Journal of French Language Studies*, **23.1**: 1–16.

Mazur-Palandre, A. (2015). Overcoming Preferred Argument Structure in written French: Development, modality, text type. *Written Language and Literacy*, **18.1**: 25–55.

Meinschaefer, J., Bonifer, S. and Frisch, C. (2015). Variable and invariable liaison in a corpus of spoken French. *Journal of French Language Studies*, **25.3**: 367–396.

Miller, G. (1956). The magical number seven plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, **63**: 81–97.

Miller, J. and Weinert, R. (1998). *Spontaneous Spoken Language*. Oxford: Clarendon Press.

Moreau, M.-L. (1977). Français oral et français écrit: Deux langues différentes?. *Français Moderne*, **45.3**: 204–242.

Morel, M.-A., and Danon-Boileau, L. (1998). *Grammaire de l'intonation. L'exemple du français*. Paris: Ophrys.

Nivre, J., De Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C. D., McDonald R., Petrov, S., Pyysalo S., Silveira N., Tsarfaty, R. and Zeman, D. (2016) Universal Dependencies v1: A Multilingual Treebank Collection. *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'16)*. Portorož, Slovenia, May 2016.

Prévost, S. (2003). Détachement et topicalisation: Des niveaux d'analyse différents. *Cahiers de Praxématique*, **40**: 97–126.

Ravid, D., van Hell, J., Rosado, E. and Zamora, A. (2002). Subject NP patterning in the development of text production: Speech and writing. *Written Language and Literacy*, **5**: 69–94.

Redeker, G. (1984). On differences between spoken and written language. *Discourse Processes*, **7.1**: 43–55.

Riegel, M., Pellat J.-C. and Rioul, R. (2016). *Grammaire méthodique du français*, 6th edition. Paris: PUF.

Serratrice, L. and De Cat, C. (2019). Individual differences in the production of referential expressions: The effect of language proficiency, language exposure and executive function in bilingual and monolingual children. *Bilingualism: Language and Cognition*. 1–16. DOI: 10.1017/S1366728918000962

Stark, E. (1999). Antéposition et marquage du thème (topic) dans les dialogues spontanés. In: C. Guimier (ed.), *La thématisation dans les langues. Actes du colloque de Caen 1997*. Bern: Peter Lang, pp. 337–358.

Tannen, D. (1980). Spoken and written language and the oral/literate continuum. *Proceedings of the 6th Annual meeting of the Berkeley Linguistics Society*. Berkeley, USA, February 1980.

Tesnière, L. (1959). *Éléments de syntaxe structurale*. Paris: Klincksieck.

Wang, Y. and Liu, H. (2017). The effects of genre on dependency distance and dependency direction. *Language Sciences*, **59**: 135–147.

Yan, J. and Liu, H. (2019). Which annotation scheme is more expedient to measure syntactic difficulty and cognitive demand? *Proceedings of First Workshop on Quantitative Syntax*. Stroudsburg, PA: Association for Computational Linguistics. pp. 16–24.

Yngve, V. (1960). A model and a hypothesis for language structure. *Proceedings of the American Philosophical Society*, **104**: 444–466.

Zribi-Hertz, A. (2011). Pour un modèle diglossique de description du français: Quelques implications théoriques, didactiques et méthodologiques. *Journal of French Language Studies*, **21.2**, 231–256.

CORPORA

Avanzi, M. (2012). *L'interface prosodie/syntaxe en français parlé. Dislocations, incises, asyndètes*. Bruxelles: Peter Lang.

Branca-Rosoff, S., Fleury, S., Lefeuvre, F. and Pires, M. (2012). Discours sur la ville. Corpus de Français Parlé Parisien des années 2000 (CFPP2000). Technical report.

Candito, M., and Seddah, D. (2012). Le corpus Sequoia : Annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical. *Proceedings of the Conférence sur le Traitement Automatique des Langues Naturelles (TALN'12)*. Grenoble, France, June 2012.

Debaisieux, J. M., Benzitoun, C. and Deulofeu, H.-J. (2016). Le projet ORFEO : Un corpus d'études pour le français contemporain. *Revue Corpus*, **15**: 91–114.

Durand, J., Laks, B. and Lyche, C. (eds) (2009). *Phonologie, variation et accents du français*. Paris: Hermès.

Eshkol-Taravella, I., Baude, O., Maurel, D., Hriba, L., Dugua, C. and Tellier, I. (2012). Un grand corpus oral "disponible" : le corpus d'Orléans 1968–2012. *Traitement Automatique des Langues*, **53.2**: 17–46.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. *Proceedings of the Machine Translation Summit X*. Phuket, Thailand, September 2005.

Lacheret, A. (2003). *La prosodie des circonstants en français parlé*. Paris: Peeters.

Lacheret, A., Kahane, S., Beliao, J., Dister, A., Gerdes, K., Goldman, J.-P., Obin N., Pietrandrea P. and Tchobanov, A. (2014). Rhapsodie: Un treebank annoté pour l'étude de l'interface syntaxe-prosodie en français parlé. *Proceedings of the 4th Congrès Mondial de la Linguistique Française (CMLF'14)*. Berlin, Germany, July 2014.

Mertens, P. (1987). *L'intonation du français : De la description linguistique à la reconnaissance automatique*. PhD dissertation, Louvain University.

Tiedemann, J. (2009). News from OPUS-A collection of multilingual parallel corpora with tools and interfaces. *Recent Advances in Natural Language Processing*, Volume **5**. Amsterdam/Philadelphia: Benjamins, pp. 237–248.

Villemonte de La Clergerie, É., Hamon, O., Mostefa, D., Ayache, C., Paroubek, P. and Vilnat, A. (2008). Passage: From French parser evaluation to large sized treebank. *Proceedings of the International Conference on Language Resources and Evaluation (LREC'08)*. Marrakesh, Morocco, May 2008.

Cite this article: Poiret R and Liu H (2020). Some quantitative aspects of written and spoken French based on syntactically annotated corpora. *Journal of French Language Studies* **30**, 355–380. https://doi.org/10.1017/S0959269519000334