Alicante Journal of English Studies 30 (2017): 151-184

Syntactic complexity and language contact: A corpus-based study of relative clauses in British English and Indian English

Iván Tamaredo
Universidade de Santiago de Compostela
ivan.tamaredo@usc.es

ABSTRACT

The aim of the present paper is to test the claim that contact simplifies language (cf. Kusters, 2008) by comparing the domain of relative clause formation in British English, a L1 variety, and Indian English, a L2 variety. According to Hawkins (1999), the processing cost of relativizing a noun phrase increases down the Accessibility Hierarchy (Subject > Direct Object> Indirect Object > Oblique > Genitive> Object of Comparison) proposed by Keenan and Comrie (1977). Subject relative clauses are thus easier to process than direct object relatives, and so on. The results of a corpus study of the British and Indian components of the *International Corpus of English* show that the Accessibility Hierarchy has an indirect effect on the production of relative clauses in British English and Indian English: whereas the distribution of relative clauses with respect to the hierarchy is very similar in both varieties, the number of complex relatives, i.e., with coordination or further embedding, decreases in the lower positions in Indian English. These results thus suggest that language contact plays a significant role in relative clause use and accounts for certain differences between L1 and L2 varieties of English in this grammatical domain.

Keywords: syntactic complexity, language contact, relative clauses, British English, Indian English, variation

1. Introduction

Language and dialect contact is a pervasive situation in the world nowadays. Bilingualism and multilingualism are the norm, while monolingualism is actually a restricted phenomenon (Valdés, 2012). As argued, for instance, by Trudgill (2009: 109), the evolution of language may be very different under the influence of contact:

> I have argued that language contact involving widespread adult language learning leads to an increase in simplification including loss of morphological categories. […] And I have argued, more hypothetically, that small community size and isolation may promote the spontaneous growth of morphological categories; and that they may also promote the growth of irregularity, redundancy, and low transparency.

Thus, studying the effects of contact on language variation and change is an issue of utmost importance.

The present contribution has two main goals, which are reflected in the title: (1) to propose a metric of syntactic complexity of relative clauses on the basis of the preferences of speakers, since the "locus of contact is the language processing apparatus of the individual multilingual speaker" (Matras, 2009: 3); and (2) to assess the effects of contact on relative clause complexity. To do so, relative clauses will be analysed in two varieties of English, British English (BrE) and Indian English (IndE), since they represent two different types of varieties: BrE is a native variety, and IndE is a L2 variety that developed under contact conditions and with a strong exonormative pressure. Two sets of results will be presented. Firstly, the distribution of relativizers and the different positions of the preposition in relative clauses that relativize a prepositional complement NP will be commented on. Second, the frequency of simple and complex relative clauses will be examined in the two varieties at hand, in order to discover contact effects.[1]The inclusion of relativizer choice and preposition placement in the study is motivated by the intrinsic interest that variation with respect to these different structural options has for a study of relative clause formation. However, it serves an additional function: previous research on relative clauses in English has focused mostly on the factors underlying the selection of relativizers and the placement of the preposition in prepositional complement relatives. Therefore, this analysis can be used to compare the present results with what has been found in earlier studies, and to test their validity and generalizability.

The article is structured as follows. Section 2 introduces the metric used here to quantify the syntactic complexity of relative clauses, and reviews the effects of language contact found in previous research. Section 3 deals with the formation of relative clauses in Standard English and in Indian English. Next, section 4 describes the data and the methodology of the study, followed by the results in section 5. Section 6 focuses on the discussion of the results, and, finally, section 7 presents some conclusions.

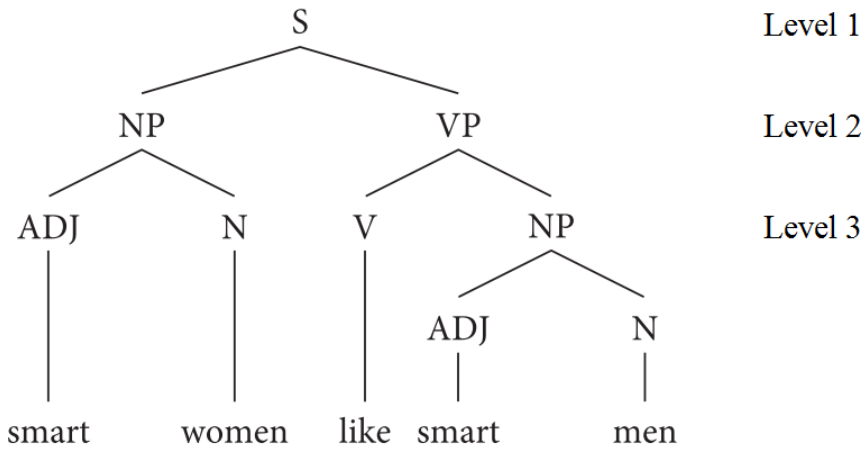
2. Syntactic complexity

Measuring grammatical complexity is definitely a complicated task. There is yet no agreement as regards the best way to determine the complexity of grammatical features, and less so if we move to the level of entire linguistic systems, such as phonology, morphology, or syntax (and even less so at the level of whole languages or dialects). Quantifying syntactic complexity is particularly difficult due to the abstract nature of the objects to be measured, which in this case comprise schematic rules and constructions (Dahl, 2009: 62-63). Previous operationalizations of syntactic complexity include, among others,

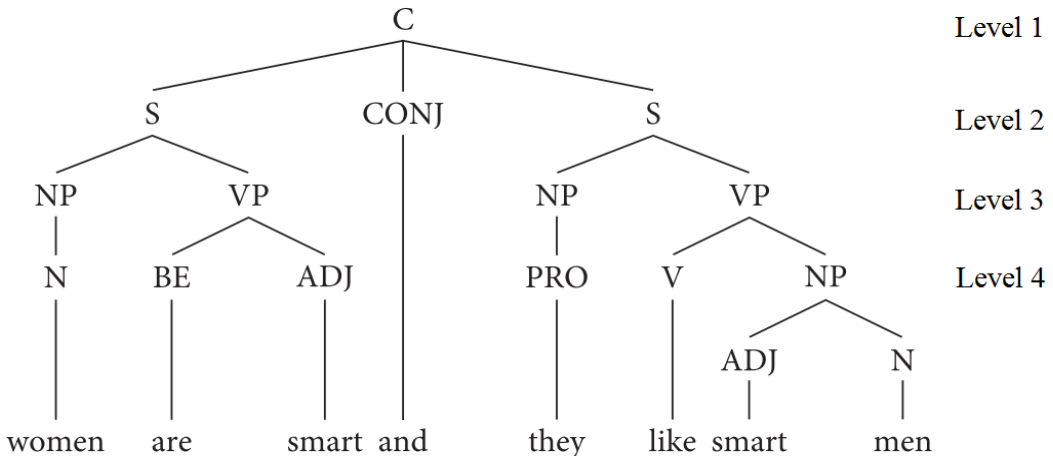
- the number of rules that operate in the syntax of a language; the more, the more complex (cf. Szmrecsanyi and Kortmann, 2012: 9),
- the degree of clausal embedding in a structure or the upper limit allowed by the grammar of a language; the more embedded clauses, the more complex (Karlsson, 2009: 192),
- and the number of phrasal nodes that a syntactic unit (e.g., a phrase or a clause) dominates; the more nodes, the more complex (Szmrecsanyi, 2004: 1033).[2]

Givón (2009: 4-5) argues that complexity can be measured as the level of hierarchical organization of a system. In the case of languages, this means that syntactic complexity increases as linguistic elements are hierarchically grouped into phrases, clauses, and sentences. A simple transitive clause can have a 3-level hierarchical organization, as in example (1); coordinated clauses add an extra level of structure, as in example (2); and embedded clauses contain a 5-level hierarchy, as in example (3) (examples adapted from Givón, 2009: 4-5).

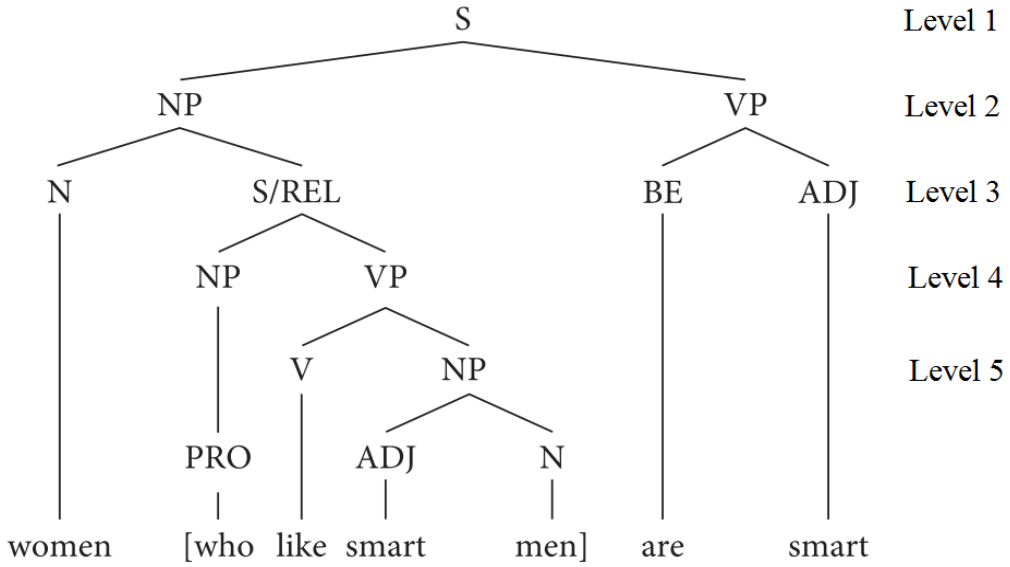
(1) Simple transitive clause:



(2) Coordinated clauses:



(3) Embedded relative clause:



As can be deduced from this brief review, metrics of syntactic complexity can differ in the nature of the objects that they measure and in the perspective from which they approach those objects: some try to estimate the complexity of the syntax of a language by counting the number of syntactic rules or by calculating the upper limit of clausal embedding allowed by the grammar, while others focus on individual structures and quantify the number of nodes or the degree of embedding that they contain. The former are systemic metrics of complexity and the latter are structural metrics (Dahl, 2004: 42-45). Systemic metrics measure the complexity of the rules of a grammar that produce the structures used by speakers, while structural metrics focus on the complexity of the structures that are the outputs of those rules. This is an important distinction because these two different types of metrics may provide diverging results. For instance, despite

the fact that, of all grammatical components, morphology and syntax have been traditionally considered the least vulnerable to the effects of language contact (Thomason and Kauffmann, 1988: 51-52), recent studies (e.g. Mukherjee and Gries, 2009; Schröter and Kortmann, 2016; Suárez-Gómez, 2017) have shown that contact languages or dialects tend to contain grammatical structures that are the result of transfer from one of the languages/dialects involved in the contact situation to the other(s). These innovations introduce new variants in the grammar and, therefore, increase its complexity from a systemic point of view by expanding the set of rules/constructions available. However, these innovative uses may result in simpler structures, i.e., with fewer phrasal or clausal nodes and fewer levels of hierarchical organization. In these cases (and many others), systemic and structural metrics provide opposite assessments of the complexity of the syntax of a language.

A further distinction should be made between absolute and relative complexity metrics (Kusters, 2008; Miestamo, 2008). On the one hand, in absolute metrics complexity is understood as an objective property of grammars, which are in turn conceptualized as autonomous entities independent from considerations related to language use. A language is considered to be more complex the more elements (e.g. phonemes, morphemes, or syntactic patterns) and the more connections between elements it has. Relative metrics, on the other hand, understand complexity as rooted in the preferences of language users. Therefore, a language is more complex if it is harder for its users to process or learn. As with the systemic/structural distinction, these two different types of metrics may provide opposite results. Using again as an example the transfer of grammatical features from the native language/dialect to another in contact situations, an innovation may result, from an absolute perspective, in a more complex grammar with more rules/constructions, and even in more complex structures, i.e., with more phrasal or clausal nodes and more levels of hierarchical organization. However, from a relative point of view, these innovations may be easier to process or learn for speakers due to the fact that they are patterns found in the speakers' native languages.

The approach adopted in the present paper in order to quantify the complexity of relative clauses is *structural* and *relative*. It is structural because it measures the complexity of individual instances of relative clauses; it is relative because the relative clauses that are considered as complex are the ones that can be characterized as difficult to process for speakers on the basis of independent evidence. The next section (§ 2.1) describes this approach to syntactic complexity.

2.1. Relative clause complexity

Relative clauses are characterized by having a gap in their structure, i.e., one element of the clause is 'missing' and has to be retrieved from a NP in the main clause in which it is embedded, called the head NP (Biber et al., 1999: 608; Huddleston and Pullum et al., 2002: 1034). The other major component is the relativizer, which can be overt, such as relative pronouns or adverbs that explicitly mark the function of the relativized NP, or covert, in which case we have a particle that does not overtly signal the function of the gap in the relative clause (e.g. *that* or the *zero* relativizer in English).[3] In relative

constructions, a dependency is established between the gap (or the word that subcategorizes for it in the clause) and the head NP, sometimes mediated by an overt relative pronoun. There is experimental evidence which shows that relative clauses, and other structures that involve a dependency between a gap and an antecedent NP, are difficult structures to process (cf. Hawkins, 1999, and references therein): as soon as a relative construction is encountered, the addressee has to store in memory all the information related to the antecedent until the gap is located, and at the same time the words and syntactic/semantic dependencies in the way from the antecedent to the gap must be correctly parsed. Therefore, as argued by Hawkins (1999), processing relative clauses becomes easier the simpler they are (i.e., the fewer phrasal nodes and levels of hierarchical organization they have) and the shorter the distance between the antecedent and the gap, because this means that the head NP has to be kept in memory for a shorter time and that there are fewer additional words and syntactic/semantic operations that have to be processed simultaneously with gap identification.

Keenan and Comrie (1977, 1979), on the basis of typological data from a sample of about fifty languages, postulate that NPs are not all equally relativizable. They propose a so-called Accessibility Hierarchy (Keenan and Comrie, 1977: 66), which, as they suggest, represents the ease with which different NP positions can be relativized:

(4) The Accessibility Hierarchy
Subject (SU) > Direct Object (DO)> Indirect Object (IO) > Prepositional Complement (PCOMP) > Genitive (GEN) > Object of Comparison (OC)

The ease of relativizing a NP decreases down the hierarchy, with SU NPs being the easiest and OCs the hardest. According to Keenan and Comrie (1977: 67), languages differ with respect to which positions in the hierarchy they relativize, but this variation is constrained: if a NP position can be relativized, then all positions higher in the hierarchy must also be relativizable, but not the other way around. This means that, for instance, if a language allows relativization on IOs, then it must also allow it on DOs and SUs, but not necessarily on all other positions further down the hierarchy. Keenan and Comrie (1977: 88) also propose an explanation for the hierarchy in terms of processing difficulty: comprehension becomes more difficult as we go down the hierarchy, with relatives formed on lower positions being more difficult to process than those formed on higher ones. Hawkins (1994, 1999, 2004) suggests that relative clause complexity increases as we go down the Accessibility Hierarchy, with more phrasal nodes in the way from the head NP to the gap. As Hawkins (1999: 255) puts it,

> as the nodes increase there are more structural relations and cooccurrence requirements to compute and more morphosyntactic and semantic operations that apply, such as case assignment, θ-role assignment and thematic dependency computations. There will also be more terminal nodes to process in larger domains, with more words to recognize and process phonologically and morphologically.

It has to be borne in mind that all these processing operations take place simultaneously with the resolution of the head NP-gap dependency, which, as mentioned before, is a procedure that already burdens the human processor substantially. The order of the NP positions in the hierarchy can, therefore, be explained in the following terms (Hawkins, 1999: 253-254):

- A relativized SU is closer to the head NP than any other relativized positions, so it is the easiest to relativize.
- A relativized DO is separated from the head NP by at least a verb and a subject.
- A relativized IO presupposes the existence of a direct object and a subject in addition to the verb of the clause, and is thus separated from the head NP by at least those three elements.
- A relativized PCOMP is embedded in a prepositional phrase, and set apart from the head NP by at least the preposition, the verb, and the subject.[4]
- A relativized GEN NP is part of a possessive phrase, which creates extra syntactic depth and puts this position at the bottom of the hierarchy.[5]

Diessel and Tomasello (2005) and Diessel (2009) propose some further refinements to the Accessibility Hierarchy. Drawing on data from the domain of L1 acquisition,[6] they found that English-speaking children have more problems with transitive SU relative clauses (SU-TR) than with intransitive ones (SU-INT), a finding that they explain in terms of the additional referent that SU-TR relatives contain, i.e., a subject and an object "engaged in a transitive activity" (Diessel and Tomasello, 2005: 900). Furthermore, all SU relatives (both intransitive and transitive) were easier for children than DO, IO, PCOMP, and GEN relative clauses. Diessel and Tomasello (2005: 899) argue that this is because they are similar to simple sentences in that the initial NP, i.e., the head NP, is the one that expresses the subject, whereas in all other relative clauses the subject is expressed by another NP that is different from the head NP. DO, IO, and PCOMP relatives performed similarly: they all caused more problems than SU relatives, but were not different from one another, which, as suggested by Diessel and Tomasello (2005: 901), is due to the fact that they have a comparable structure with the same sequence of nouns and verbs, i.e., first the head NP, and then the subject NP, followed by the verb and the gap (NP [NP V …]$_{REL}$). Finally, GEN relative clauses were the most difficult for children because they are different from the rest: they "establish the link between the head noun and the relative clause by a genitive attribute, which even many adult speakers find difficult to process" (Diessel and Tomasello, 2005: 901).

2.2. Complexity and language contact

The other important goal of the present article, as stated in its title, is to analyse the effects of contact on language variation. Language contact (and concomitant L2 acquisition) has been shown to have a simplifying effect on grammar (cf., for instance,

the collection of papers in Miestamo et al., 2008). As argued by Trudgill (2009, 2011), among others, this is "due to the relative inability of adult humans to learn new languages perfectly" (Trudgill, 2009: 99): in the process of learning a language, adults simplify its grammar by decreasing its redundancy and opacity, regularizing paradigms, and eliminating semantic distinctions coded by different morphological categories. Many different studies have provided evidence to support the claim that language contact results in grammatical simplification due to the influence of L2 acquisition and use (McWhorter, 2001, 2007; Kusters, 2003, 2008; Parkvall, 2008; Sinnemäki, 2009). However, not all types of contact have this effect. Trudgill (2011) proposes a typology of contact situations and their influence:

- High-contact situations with short-term adult L2 acquisition result in grammatical simplification.
- High-contact situations involving long-term childhood bilingualism tend to lead to complexification due to "*additive* borrowing" (Trudgill, 2011: 27; italics in original), i.e., the incorporation of new features into the grammar derived from another language that coexist with the existing features, without substituting any of them.
- Low-contact situations result in complexification due to a spontaneous (i.e. non-borrowed) increase in morphological categories, redundancy, opacity, and irregularity.

Therefore, it is only in the first type of contact situations, i.e., those in which there is a high number of L2 users, that grammatical simplification takes place.

In the domain of varieties of English, previous research has demonstrated that those varieties with a history of language contact are less complex than low-contact ones. A series of studies conducted at the *Freiburg Institute for Advanced Studies* (Kortmann and Szmrecsanyi, 2009, 2011; Szmrecsanyi and Kortmann 2009a, 2009b; cf. Szmrecsanyi and Kortmann (2012) for a summary of these studies) revealed complexity differences between high- and low-contact varieties of English. High-contact varieties, i.e., high-contact L1s, indigenized L2s, pidgins, and creoles, have, in general, fewer grammatical features that "add contrasts, distinctions, or asymmetries without providing a communicative or functional bonus" (Szmrecsanyi and Kortmann 2012: 16) than low-contact varieties. They also present more features that result in fewer contrasts, distinctions, and asymmetries, which are easier to acquire and use for L2 speakers. In addition, L2 varieties show a lower degree of grammaticity, i.e., a lower frequency of grammatical markers, and less irregularity than L1 varieties. IndE is a L2 variety of English and an example of Trudgill's (2011) first type of contact situations, namely those characterized by short-term adult L2 acquisition and use (§ 3.1). Therefore, it is expected to display grammatical simplification vis-à-vis BrE.

Finally, and more relevantly for the purposes of the present paper, simplification effects due to language contact can also be found in the domain of relative clauses. As shown in a recent article by Suárez-Gómez (2017), IndE, Singapore English (SgE) and

Hong Kong English (HKE), all of them L2 Asian varieties of English, favour simpler relative structures more strongly than BrE. In Suárez-Gómez's study, relative clauses are more complex if they decrease transparency by containing elements that are not overtly expressed (i.e. *zero* relativizers) or that increase redundancy, for instance, via agreement (i.e. *wh*-pronouns), and by introducing discontinuities in the structure (i.e. relatives that are not adjacent to the head NP that they modify). Moreover, IndE, SgE and HKE speakers relativize the higher positions in Keenan and Comrie's Accessibility Hierarchy more frequently than BrE users, thus producing more SU relatives and fewer DO and, especially, PCOMP and GEN relatives than in the L1 variety. As mentioned in section 2.1, the difficulty of relativizing a NP increases as we move further down the hierarchy, so SU relatives are easier to process than DO, PCOMP and GEN ones.

The literature reviewed in this section points to the conclusion that simplification dominates in those languages or dialects affected by language contact and short-term adult L2 acquisition, and that it also affects the domain of relative clauses. The following section focuses on this grammatical domain in the two varieties of English that are at the center of the present paper, i.e., IndE, an L2, and BrE, an L1, as well as on a hypothesis related to the distribution of relativization strategies in the two varieties.

3. Relative clauses in English

As mentioned in section 2.1, relativizers can be overt or covert. In English, overt relativizers consist of *wh*-pronouns,[7] i.e., *who*, *whom*, *whose*, and *which*, and covert relativizers include *that* and *zero* (Huddleston and Pullum et al., 2002: 1034). *Wh*-pronouns are more commonly found in formal written contexts (although *who* is also frequently used in spoken language, mostly in SU relatives with animate antecedents; cf. Cheshire, Adger and Fox, 2013), while *that* and *zero* are more frequent in speech and informal texts (Biber et al., 1999: 612). Relative clauses can also be classified with respect to the relation between the relative clause and the head NP into restrictive and nonrestrictive: restrictive relatives delimit the set of entities that the head NP can refer to, while the nonrestrictive type merely adds additional information about the antecedent without limiting the set of possible referents (Huddleston and Pullum et al., 2002: 1034-1035; Denison and Hundt, 2013: 140).

The choice of relativizer in English is determined, among other factors, by the animacy of the antecedent, the relation between the relative clause and the head NP, and the function of the gap in the clause (Quirk et al, 1985: 1247-1248; Biber et al., 1999: 609). In nonrestrictive relatives, *zero* is not possible and *that* is very infrequent; there is variation between *who/whom/whose*, used mostly with human antecedents, and *which*, which is mainly restricted to nonhuman referents (Quirk et al., 1985: 1258). In restrictive relatives, on the other hand, there is more competition between the different relativizers (Quirk et al., 1985:1249-1252; Biber et al., 1999: 612-620; Huddleston and Pullum et al., 2002: 1054-1056):

- With human antecedents, *who* is favoured in SU relatives (although *that* is also frequent) and *that/zero* in the rest. *Whom* and *whose* are restricted to DO/PCOMP and GEN relatives respectively.
- With nonhuman antecedents, *which* is in competition with *that/zero*, the former being the preferred variant in written and formal language. Additionally, in DO and PCOMP positions, *which* is favoured when the head NP is complex, i.e., when the antecedent noun is separated from the relativizer by complex phrases or clauses, or when it is realized by a demonstrative pronoun. *That/zero*, on the other hand, are more common than *which* in spoken and informal language, and, again in DO and PCOMP positions, when the head NP is either simple or realized by an indefinite pronoun.
- The choice between *that* and *zero* is governed by several factors. *Zero* is not allowed in SU relatives or when it is not adjacent to the subject of the relative clause. It is preferred, however, when the gap is a DO or a PCOMP, and when the subject of the relative clause is realized by a personal pronoun.[8] Finally, *that* is more common than *zero* in formal discourse.

Further variation can be found in PCOMP relatives, since there are three different structural options with respect to the position of the preposition (Quirk et al., 1985: 1252-1253, 1259; Biber et al., 1999: 624-625; Huddleston and Pullum et al., 2002: 1052; Hoffmann, 2005): it can move with the relativizer to the beginning of the relative clause, an operation known as pied-piping; it can be left stranded in its original position while the relativizer moves to the beginning of the relative clause; or it can be deleted. Pied-piping only occurs with *wh*-pronouns, and it is associated with formal written contexts. It is favoured in restrictive relatives and disfavoured in nonrestrictive ones. Preposition stranding can occur both with *that*/*zero* relativizers and *wh*-pronouns, although it is more frequently found with the former two (more often with *zero* than with *that*). It is associated with speech and informal written contexts. Finally, preposition deletion occurs exclusively with *that* and *zero* (again, more frequently with the latter), and it can be found both in speech and in written language.

3.1. Relative clauses in Indian English

English was introduced in India in the 17th century after it was colonized by the British. Nowadays, there are approximately 100 million speakers of English in the country, of which only 250,000 are native speakers (Sharma, 2010: 523). Most users, therefore, speak English as L2 or L3, making IndE a non-native variety. English, together with Hindi, is one of the co-official languages of India, although its use is restricted to certain domains of life, such as the government, administration, politics, higher education, the legal system, business and the media (Schneider, 2007). It also functions as an "interethnically neutral link language" (Schneider, 2007: 167), but it is not a marker of identity. The current status of English in India has been described as a "steady state" (Mukherjee, 2007: 158) in which both progressive and conservative forces are at play. The most important progressive forces are the linguistic innovations

that distinguish IndE from other varieties, but also certain developments allow us to entertain the possibility that it could become more established in India in the future. Among these are the increase in the number of literary works written in English by Indian authors, and the recent inclusion in the syllabus of a compulsory English subject in primary education. On the other hand, there are also conservative forces that hinder the spread of English in the country and make it difficult for IndE to become a carrier of Indian identity. First of all, teachers of English follow predominantly a British norm, which prevents the establishment of an Indian standard. Secondly, and most importantly, many Indians consider the innovative features characteristic of IndE as grammatical errors that must be avoided, and instead hold native varieties as the appropriate and correct ones. In terms of Trudgill's (2011) typology of contact situations, IndE is an example of the first type: those characterized by short-term adult L2 acquisition and use.

The most detailed and comprehensive description of relativization strategies in IndE to date is Suárez-Gómez (2014), who focuses on the choice of relativizer in IndE, HKE and SgE restrictive adnominal relative clauses in the spoken component of the *International Corpus of English* (ICE). She found that:[9]

- *who* is favoured with human antecedents in SU relatives, and there are also some cases of *zero* in these contexts;
- *which* is the preferred option with nonhuman referents in SU relatives;
- we find competition between *zero* and *whom* in nonsubject position with human antecedents;
- there is variation between *that*, *zero*, and *which* in nonsubject position with nonhuman referents;
- and in PCOMP relatives, there are very similar frequencies of pied-piping and preposition stranding constructions, and only one instance of preposition deletion.

This distribution shows that spoken IndE is more similar to spoken BrE in the 1950s as described in Quirk (1957), with a preference for *wh*-pronouns over *that* and *zero* and a high frequency of pied-piping constructions in PCOMP relatives. IndE seems to have been unaffected by more recent developments in spoken BrE, which in the 1990s showed a higher frequency of *that* than in the 1950s, and a decrease in the use of *wh*-pronouns (Tottie, 1997). On the other hand, the IndE preference for *wh*-pronouns may also be a reflection of the influence of the substrate languages, particularly Hindi, since relative clauses in this language may be formed by means of relative pronouns, resulting in structures that are very similar to English *wh*-relatives (Suárez-Gómez, 2017).

3.2. Expected distribution of relative clauses in IndE and BrE

Taking into account what was mentioned in sections 2.2 and 3.1, simplification processes are hypothesized to have affected the domain of relative clauses in IndE. IndE

is predominantly a L2 variety of English, i.e., a high-contact variety with many adult L2 speakers, and this is expected to be reflected in its syntax, so that IndE contains simpler relative clauses in comparison with BrE, an L1variety. Simpler relative clauses include, as mentioned in section 2.1, those in which the relativized position is higher in the Accessibility Hierarchy, and those with fewer nodes and levels of hierarchical organization. In the next section (§ 4), this definition of relative clause complexity will be specified in more detail, together with the data retrieval process and the methodology used in the present study.

4. Data and methodology

The data of the study was extracted from the British (ICE-GB) and Indian (ICE-IND) components of ICE. ICE is a collection of corpora, each component of which contains one million words, 600,000 words of speech and 400,000 of written language, from a variety of English around the world. All the components were compiled following the same annotation template and design, so they are highly comparable. For the purposes of the present paper, 40 texts from each of the two ICE components mentioned above were selected (approximately 82,000 words from ICE-GB and 88,000 from ICE-IND): 10 texts were spoken informal (from the S1A category comprising private conversations), 10 texts were spoken formal (from the S2A category, unscripted speech from broadcast events), 10 were written informal (from the W1B category, social letters), and 10 were written formal (from the W2A category, academic writing). The data in the selected texts was produced in the 1990s: the ICE-GB texts range from the year 1990 to 1993, and those in ICE-IND from 1990 to 1998. Therefore, the differences found between the varieties are not expected to derive from diachronic effects.

All the instances of adnominal relative clauses in the selected texts introduced by a *wh*-pronoun, *that*, or *zero* were retrieved employing different methods:

- Relatives introduced by *wh*-pronouns and *that* in the texts taken from ICE-IND were retrieved using the concordance programme *WordSmith Tools 6*.
- *Zero* relatives in ICE-IND were manually extracted from the texts.
- Relative clauses in ICE-GB were retrieved using *ICECUP 3.1*.

Only instances found in valid data were used, i.e., those retrieved from extra-corpus material (marked <X></X>) were excluded.

The data was then analysed by means of a series of 'hierarchical configural frequency analyses' (HCFA), using Gries' (2004) *HCFA 3.2* script for *R* (R Core Development Team, 2015). HCFA is an extension of the chi-square test that allows the simultaneous analysis of more than 2 variables and that approaches the data in a more exploratory fashion (Hilpert, 2013: 56). This test compares the observed frequencies of the different configurations in a table against the frequency expected by chance: those that are found significantly more often than expected are called *types*, and those that have a significantly lower frequency than expected by chance are known as *antitypes*.

HCFA provides the global significance value of the table plus the configurations that are responsible for it. For example, consider Table 1, an extract of the results provided by the HCFA test for one of the analyses conducted in the present study, which is discussed in more depth below.

Variety	Restrictiv.	Text type	Relativized position	Relativizer	Freq	Exp	Cont. chisq	Obs-exp	Dec	Q
BrE	Restrictive	Sp. inf.	SU	wh-pronoun	32	338.817	0.1045	<	ns	0.002
IndE	Restrictive	Sp. inf.	SU	wh-pronoun	22	246.799	0.291	<	ns	0.002
BrE	Restrictive	Sp. inf.	SU	that	29	10.875	302.083	>	***	0.017
IndE	Restrictive	Sp. inf.	SU	that	0	79.215	79.215	<	ms	0.007
BrE	Restrictive	Sp. inf.	SU	zero	0	84.584	84.584	<	*	0.008
IndE	Restrictive	Sp. inf.	SU	zero	0	61.612	61.612	<	ns	0.006

Table 1: Sample of the results of a HCFA test analysing the distribution of relative clauses as a function of variety, restrictiveness, text type, relativized position, and relativizer

The first five columns in the table indicate the levels of each of the variables that are considered in the analysis: BrE and IndE for variety; restrictive and nonrestrictive for restrictiveness; spoken informal, spoken formal, written informal, and written formal for text type; SU, DO, PCOMP, and GEN for relativized position; and *wh*-pronoun, *that*, and *zero* for relativizer. The columns 'Freq' and 'Exp' provide the observed and expected frequencies respectively of each configuration in the table, and the 'Cont. chisq' column gives their chi-square values. 'Obs-exp' reflects the relation between the observed and expected frequencies: '<' means less observed frequency than expected, and '>' more than expected. The following column, 'Dec', states the significance level of each configuration ('ns' = not significant, 'ms' = marginally significant, '*' =significant at the 0.05 level, '**' = significant at the 0.01 level, '***' = significant at the 0.001 level). Finally, 'Q' stands for coefficient of pronouncedness, a measure of the size of the effect of each configuration (the higher, the stronger). As mentioned above, HCFA also provides global chi-square and significance values for the whole table, which in this case are χ^2 = 2010.34 (d.f. = 181) and $p < 0.001$, i.e., statistically significant.

In the extract of the results provided in Table 1, there are two significant configurations, and one that is marginally significant: restrictive *that* relatives with a SU gap in spoken informal texts are a type in BrE, i.e., they are significantly more frequent than expected by chance; restrictive *zero* relatives with a SU gap in spoken informal texts are an antitype in BrE, i.e., they are less frequent than expected by chance; and restrictive *that* relatives with a SU gap in spoken informal texts are a (marginally significant) antitype in IndE.

4.1. Variables included in the analysis

The following variables (and their levels) were explored by means of HCFA tests:

1. Variety: BrE vs. IndE.
2. Text type: spoken informal vs. spoken formal vs. written informal vs. written formal.
3. Restrictiveness: restrictive vs. nonrestrictive.
4. Relativizer: *wh*-pronoun vs. *that* vs. *zero*.
5. Relativized position: SU (in some cases divided into SU-INT and SU-TR; see below) vs. DO vs. PCOMP vs. GEN.
6. Preposition placement in PCOMP relatives: pied-piping vs. preposition stranding vs. preposition deletion.
7. Relative clause complexity: simple vs. complex.

Variables 1-4 and 6 are simple operationalizations of most of the dimensions of variation in relative clause formation identified in previous research and require no further explanation. As regards variable 5, relativized position, this is an operationalization of the Accessibility Hierarchy and its subsequent refinements. SU position is divided into SU-INT and SU-TR in those tests dealing with relative clause complexity, following Diessel and Tomasello's (2005) findings reviewed in section 2.1. Additionally, the IO position is excluded here because it is not distinguishable from PCOMP in terms of complexity according to Hawkins (1999: 253-254), and because English tends to assimilate IO to PCOMP in relative clause formation[10] (Keenan and Comrie, 1977: 72). Finally, no instances of OC relatives were found in the corpus, so this position is not included in the present study. Examples of relative clauses with each of the relativized positions can be found in (5)-(8).

(5) Just one and a half months ago <,>uhm<,> I had my aunty with us <,>uhm<,> my aunty [who is above seventy-five] (*SU-INT*) (ICE-IND:S1A-004#162:1:B)
(6) The line [which attracted me] (*SU-TR*) is this <,> smile is our instrument for winning <,> soul (ICE-IND:S1A-001#157:1:B)
(7) […] one of the things [that I felt] (*DO*) when I was studying dance <,> was I very much enjoyed the work [that I was involved in] (*PCOMP*) (ICE-GB:S1A-001 #31:1:B)
(8) Losing a husband <,,> losing a father <,,> a loving friend [whose smile could charm a heart of stone] (*GEN*) (ICE-IND:S2A-006#22:1:A)

Variable 7 refers to the complexity of relative clauses and has two levels: simple vs. complex. Complex relatives include those with coordination, as in example (9), and/or further embedding, as in (10):

(9) His study of Vico, [who denied the knowability of Nature and asserted that of History] (ICE-GB: W2A-003#30:1)
(10) […] the opportunity [that has arisen through the group [that we're working with now]] (ICE-GB:S1A-001#38:1:B)

As seen in examples (1), (2), and (3) in section 2, coordinated and embedded clauses are longer than simple ones, with more phrasal nodes and more levels of hierarchical organization: a hypothetical simple clause can already have a 3-level hierarchy and 9 phrasal nodes; coordination may generate a structure with a 4-level hierarchy and 16 nodes; and a sentence with an embedded clause may contain, at least, a 5-level hierarchy and 14 nodes. Therefore, the number of phrasal nodes that has to be parsed increases if the relative clause is coordinated or contains extra dependent clauses. Longer and more hierarchically embedded relativization domains make the process of gap identification more difficult and, as a consequence, relative clauses become harder to process.[11]

5. Results

A total of 637 instances of relative clauses in BrE and 464 in IndE were identified in the corpus during the retrieval process and selected for further analysis.[12] Two sets of results are provided in this section based on this data. The first set deals with issues related to relativizer choice and preposition placement in PCOMP relatives. The second set focuses on complexity effects.

5.1. Relativizer choice and preposition placement

Table 2 shows the distribution of relavizers in BrE as a function of restrictiveness, text type, and relativized position.

		RESTRICTIVE			**NONRESTRICTIVE**		
		wh-pro	***that***	***zero***	***wh-pro***	***that***	***zero***
SP. INF.	***SU***	32	**29 T**	**0 A**	16	0	0
	DO	**0 A**	**38 T**	8	2	0	0
	PCOMP	3	**14 T**	**15 T**	7	0	0
	GEN	0	0	0	0	0	0
SP. FOR.	***SU***	**9 A**	8	**0 A**	**69 T**	0	0
	DO	**1 A**	9	4	6	0	0
	PCOMP	**1 A**	1	1	1	0	0
	GEN	0	1	0	0	0	0
WR. INF.	***SU***	31	8	1	15	0	0
	DO	**2 A**	8	**42 T**	6	0	0
	PCOMP	**0 A**	3	**23 T**	3	0	0
	GEN	1	0	0	0	0	0
WR. FOR.	***SU***	81	34	**0 A**	26	**0 A**	0
	DO	12	7	11	4	0	0
	PCOMP	**33 T**	1	0	4	0	0
	GEN	4	0	0	2	0	0
TOTAL		476			161		

Table 2: Relativizer choice in BrE

Table 2 can be interpreted as follows:

- the number in each cell is the raw frequency of examples in the corpus of a specific configuration (e.g. *wh*-pronouns in SU position in spoken informal texts in restrictive relatives);
- the numbers in boldface represent configurations that are statistically significant: the ones followed by 'T' are types, and those with an 'A' are antitypes (see section 4).

Table 2 does not provide percentages because of the complexity of the data: there are four variables that are hierarchically organized, and three of them have more than two levels. With this kind of data, it is very difficult to decide out of which total the percentages should be calculated. For instance, taking as an example the number of *wh*-pronouns in SU position in spoken informal texts in restrictive relatives, i.e., 32, we could calculate the following percentages, among others, depending on the focus of the study:

- 41.56% out of the total instances of SU relatives in spoken informal texts (77),
- 15.24% out of the total instances of *wh*-pronouns in restrictive relatives (210),
- 8.62% out of the total instances of *wh*-pronouns in both restrictive and nonrestrictive clauses (371),
- 6.72% out of the total instances of restrictive relatives (476),
- 19.51% out of the total instances of relative clauses in spoken informal texts (164).

Since what is of interest here is the general distribution of relativizers in BrE as a function of each and all of the other variables in Table 2, the data can be better summarized in graphical form. Figure 1 is a visual representation of Table 2.

Figure 1: Relativizer choice in BrE

Figure 1 plots the proportion of cases of each configuration in relation to the rest, on the basis of the four variables selected: restrictiveness, text type, relativizer, and relativized position. The first division in Figure 1 is that between restrictive and nonrestrictive relative clauses (vertical axis), so that each of the two largest boxes represents the proportion of restrictive and nonrestrictive relatives. Within each of them, we first have a division into text types, and then into relativized positions (horizontal axis): there are four rows that represent the four text types distinguished here and, within each of them, there are four smaller rows, one per relativized position. On the vertical axis there is another division, that between *wh*-pronouns, *that*, and *zero* relativizers. The size of each of the partitions in the plot reflects the proportion of cases of a specific configuration in comparison with the rest. For instance, the black box on the top left corner of Figure 1 represents the proportion of restrictive relative clauses occurring in spoken informal texts with a *wh*-pronoun as a relativizer and with a SU relativized position.

The distribution presented in Table 2 and Figure 1 is statistically significant (χ^2 = 2010.34, d.f. = 181, $p < 0.001$). The picture that emerges from this distribution is a complex one. In restrictive clauses, *that* is favoured in BrE in spoken (both informal and formal) texts, since it is the most frequent option: it is in fact a type in SU, DO, and PCOMP positions in spoken informal texts. *That* occurs more often in SU and DO positions, although it is not an infrequent choice in PCOMP relatives (especially in spoken informal texts). *Zero* is the most frequent relativizer in restrictive relatives in written informal texts. It is favoured in DO and PCOMP positions and strongly disfavoured, or even forbidden, in SU position:[13] It is a type in PCOMP relatives in spoken informal texts and in DO/PCOMP in written informal ones, and an antitype in SU in spoken informal/formal and written formal texts. *Wh*-pronouns are favoured in restrictive relatives in written formal texts (type in PCOMP position in written formal texts), and disfavoured in the other three text types (antitype in DO in spoken informal texts, in SU/DO/PCOMP positions in spoken formal ones, and in DO/PCOMP relatives in written informal texts). They occur more commonly overall in SU position, although they are also very frequent in PCOMP position in written formal texts, where they are a type. *Wh*-pronouns are the most common option in GEN relatives, which are very infrequent overall. There is, however, one case with *that* and a stranded preposition *of*:

(11)[…] he is carrying this famous letter [that the whole world is waiting to see the contents of] (ICE-GB:S2A-008 #134:3:A)

In nonrestrictive relative clauses, the only available option in BrE are *wh*-pronouns, since there are no instances of *that* or *zero* in the data. They occur more commonly in SU position and in spoken formal texts, where they are a type.

As regards IndE, Table 3 and Figure 2 show the distribution of relativizers in this variety (χ^2 = 2010.34, d.f. = 181, $p < 0.001$):[14]

		RESTRICTIVE			NONRESTRICTIVE		
		wh-pro	***that***	***zero***	***wh-pro***	***that***	***zero***
SP. INF.	***SU***	22	**0 A**	0	9	0	0
	DO	5	4	7	2	0	0
	PCOMP	3	0	7	0	0	0
	GEN	0	0	0	0	0	0
SP. FOR.	***SU***	21	12	1	**55 T**	1	0
	DO	3	**12 T**	8	7	0	0
	PCOMP	3	3	6	1	0	0
	GEN	1	0	0	0	0	0
WR. INF.	***SU***	17	4	0	**31 T**	2	0
	DO	3	6	**26 T**	6	0	0
	PCOMP	4	1	4	6	0	0
	GEN	0	0	0	0	0	0
WR. FOR.	***SU***	63	12	**0 A**	31	1	0
	DO	7	5	6	1	1	0
	PCOMP	21	0	5	5	0	0
	GEN	2	0	0	1	0	0
TOTAL		304			160		

Table 3: Relativizer choice in IndE

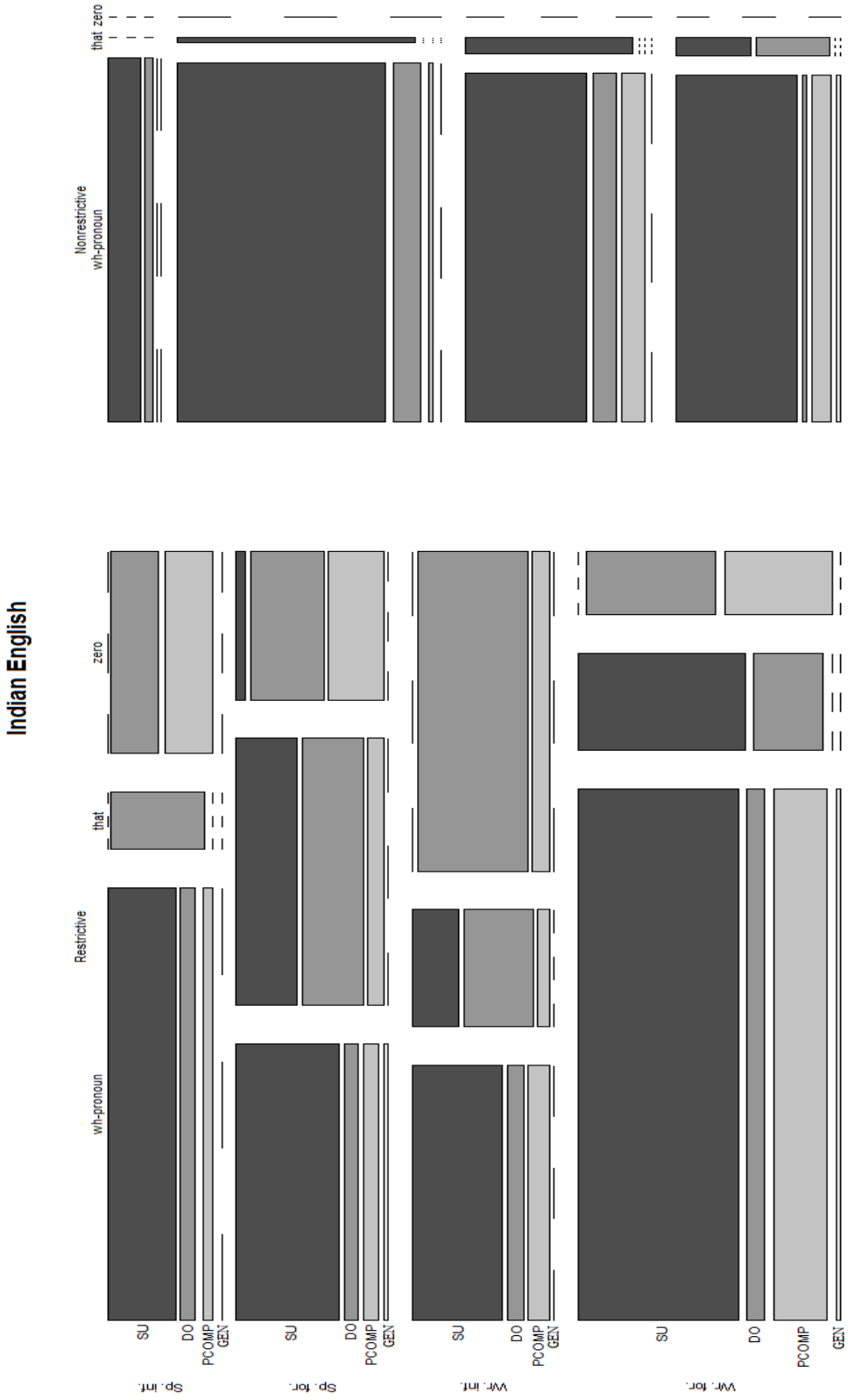


Figure 2: Relativizer choice in IndE

In IndE restrictive relative clauses, *wh*-pronouns are preferred in spoken informal and written formal texts. They are more frequent in SU position, although no statistical types or antitypes were found by the HCFA test in this case. There is competition between *wh*-pronouns and *that* in spoken formal texts, with a similar number of cases of both relativizers. *That* is disfavoured in spoken informal texts, with only 4 cases attested in the corpus (it is an antitype in SU position in this text type), and it occurs more frequently in DO position overall, becoming a type in spoken formal texts. There is also competition in written informal texts, in this case between *wh*-pronouns and *zero*, though the latter is more frequent. *Zero* is preferred in DO and PCOMP positions (it is a type in DO position in written informal texts), and it is very infrequent in SU relatives[15] (it is an antitype in SU position in written formal texts).

In IndE nonrestrictive relative clauses, *wh*-pronouns are the default choice. They occur more commonly in SU position and in spoken formal and written formal texts, where they are types.[16]

With respect to preposition placement in PCOMP relatives, Table 4 and Figure 3 show the distribution of the three different strategies, pied-piping, stranding, and deletion, as a function of restrictiveness and text type in both BrE and IndE (χ^2 = 197.94, d.f. = 40, $p < 0.001$):

		BRITISH ENGLISH		**INDIAN ENGLISH**	
		Restrictive	***Nonrestrictive***	***Restrictive***	***Nonrestrictive***
SP. INF.	***Pied-piping***	**3 A**	2	2	0
	Stranding	14	5	0	0
	Deletion	**15 T**	0	5	0
SP. FOR.	***Pied-piping***	1	0	2	1
	Stranding	2	1	3	0
	Deletion	0	0	**7 T**	0
WR. INF.	***Pied-piping***	**0 A**	2	4	**6 T**
	Stranding	14	1	4	0
	Deletion	12	0	1	0
WR. FOR.	***Pied-piping***	**33 T**	4	**21 T**	5
	Stranding	**1 A**	0	0	0
	Deletion	**0 A**	0	3	0
TOTAL		110		64	

Table 4: Preposition placement in BrE and IndE

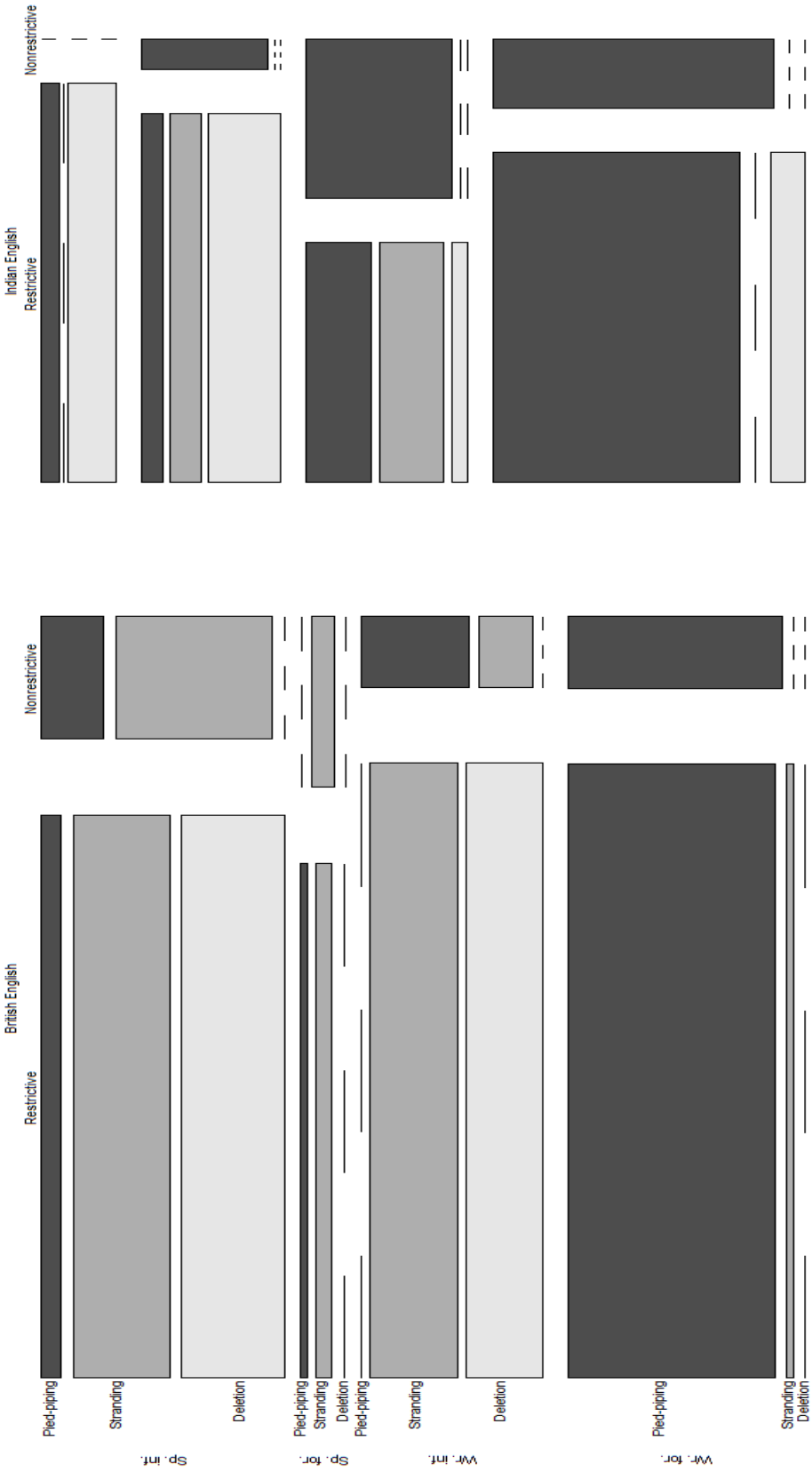


Figure 3: Preposition placement in BrE and IndE

In BrE, pied-piping is favoured in restrictive relatives in written formal texts, where it is a type, and disfavoured (antitype) in spoken informal and written informal texts. Preposition stranding is more common in restrictive clauses in spoken informal and written informal texts, although in this case it is not a statistically significant type, and it is infrequent in written formal ones, in which it becomes an antitype. Similarly, deletion is preferred in restrictive relative clauses in spoken informal (where it is a type) and written informal texts, and disfavoured in written formal ones (where it is an antitype). There are not many cases of nonrestrictive PCOMP relatives, so statistical significant results were not found, but a preference for pied-piping can be observed in written (both formal and informal) texts, while stranding is more common in spoken (again, both formal and informal) ones.

In IndE, pied-piping is more common in restrictive clauses in written formal texts, in which it is a type, and it is in competition with stranding in written informal ones. In spoken (both informal and formal) texts, on the other hand it is disfavoured, although it is not a statistically significant antitype. Stranding is only frequent in IndE restrictive relatives in written informal texts, where, as just mentioned, it is in competition with pied-piping. Deletion is more common in spoken texts, being a type in formal ones, and disfavoured in written (both informal and formal) texts. With respect to nonrestrictive relative clauses, pied-piping is the only variant attested in IndE in the data, and it is especially frequent in written informal texts, where it is a type.

5.2. Complexity effects

Tables 5 and 6 show the individual effects of relativized position ($\chi^2 = 5.95$, d.f. = 4, $p > 0.05$) and relative clause complexity ($\chi^2 = 6.34$, d.f. = 1, $p < 0.05$) respectively.

	BRITISH ENGLISH	**INDIAN ENGLISH**
SU	233 (48.95%)	152 (50%)
SU-INT	97 (20.38%)	82 (26.97%)
SU-TR	136 (28.57%)	70 (23.03%)
DO	142 (29.83%)	92 (30.26%)
PCOMP	95 (19.96%)	57 (18.75%)
GEN	6 (1.26%)	3 (0.99%)
TOTAL	476 (100%)	304 (100%)

Table 5: Frequency of relative clauses per relativized position in BrE and IndE[17]

	BRITISH ENGLISH	**INDIAN ENGLISH**
SIMPLE	332 (69.75%)	237 (77.96%)
COMPLEX	144 (30.25%)	67 (22.04%)
TOTAL	476 (100%)	304 (100%)

Table 6: Frequency of simple and complex relative clauses in BrE and IndE

Even though the global distribution of simple and complex relative clauses in BrE and IndE is statistically significant, we do not find any significant types or antitypes with respect to the variables relativized position and relative clause complexity, which means that they do not have individual effects on the distribution in the varieties at hand. We find that, in both varieties, there are more instances of relative clauses in SU position (SU-INT + SU-TR), then in DO, PCOMP, and, finally, in GEN position. In IndE, there are also more cases of SU-INT than SU-TR relatives, whereas BrE shows the opposite distribution: SU-TR > SU-INT. With respect to relative clause complexity, simple relatives are much more frequent than complex ones.

Statistically significant results are found when we investigate the conjoined effect of both variables. Table 7 and Figure 4 show the interaction of relativized position and relative clause complexity in BrE and IndE ($\chi^2 = 36.05$, d.f. = 13, $p < 0.001$).

	BRITISH ENGLISH		**INDIAN ENGLISH**	
	Simple	***Complex***	***Simple***	***Complex***
SU	149 (31.31%)	84 (17.64%)	112 (36.84%)	40 (13.16%)
SU-INT	65 (13.66%)	32 (6.72%)	61 (20.06%)	21 (6.91%)
SU-TR	84 (17.65%)	**52 T** **(10.92%)**	51 (16.78%)	19 (6.25%)
DO	107 (22.48%)	35 (7.35%)	69 (22.70%)	23 (7.56%)
PCOMP	74 (15.55%)	21 (4.41%)	53 (17.43%)	**4 A** **(1.32%)**
GEN	2 (0.42%)	4 (0.84%)	3 (0.99%)	0 (0%)
TOTAL	476 (100%)		304 (100%)	

Table 7: Conjoined effect of relativized position and relative clause complexity in BrE and IndE

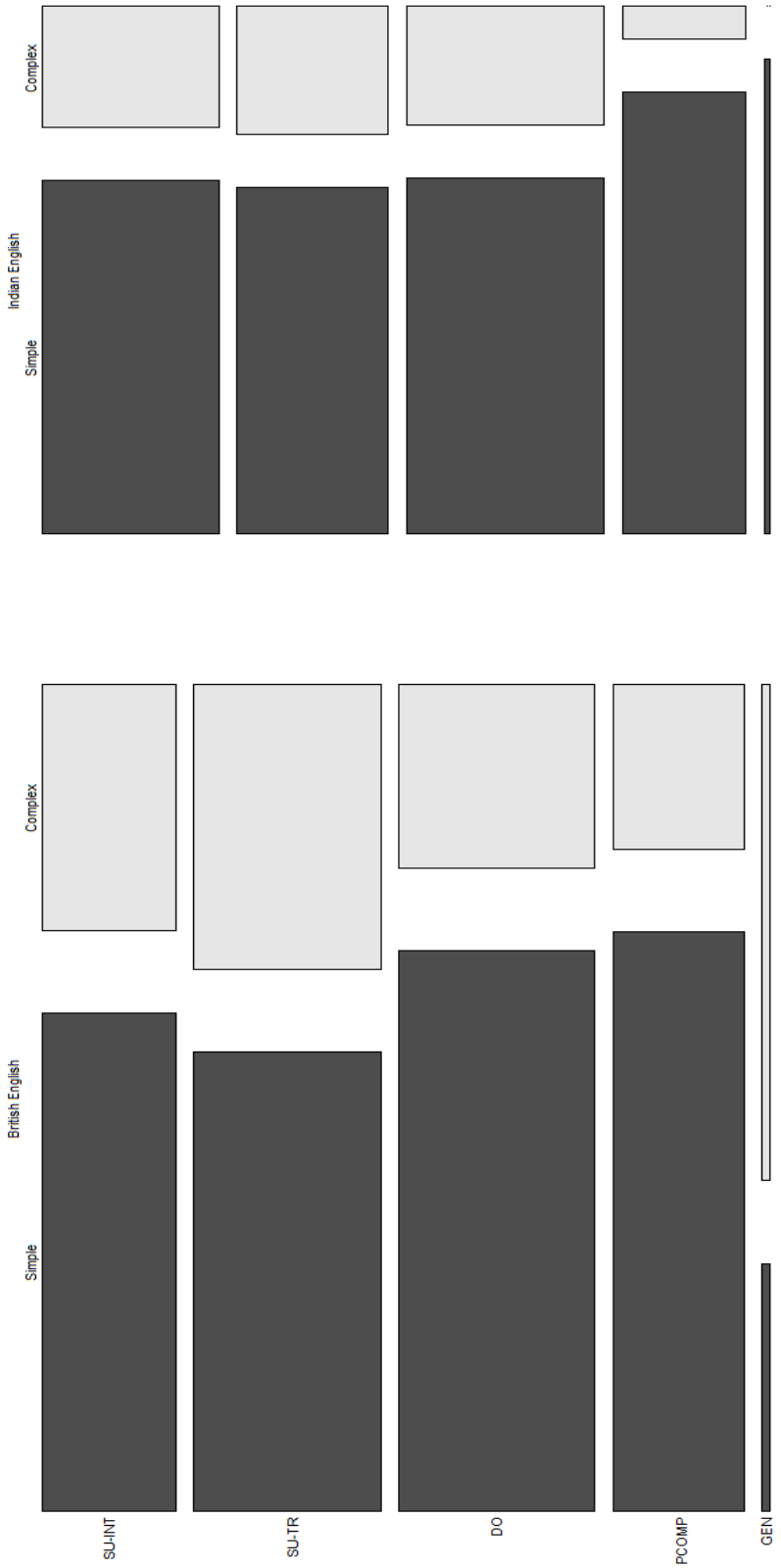


Figure 4: Conjoined effect of relativized position and relative clause complexity in BrE and IndE

There are two interesting statistically significant configurations: complex relatives are a type in SU-TR position in BrE, and an antitype in PCOMP position in IndE. No significant results were found in GEN position, probably due to the low number of instances, but the distribution is suggestive: there are no cases of GEN complex relatives in IndE, while in BrE they are more frequent than simple ones. Both varieties comply with the Accessibility Hierarchy, with fewer instances of simple and complex relatives in the lower positions. However, we do find differences between them: complex relatives are more strongly disfavoured in PCOMP and GEN positions in IndE, i.e., in those in which the process of gap identification is more difficult, and they are more strongly favoured in SU-TR in BrE, i.e., in the SU position in which forming relatives is more complicated. This distribution points to the conclusion that relative clauses are simpler in IndE than in BrE, an issue that is further discussed in section 6.2.

6. Discussion

6.1. Relativizer choice and preposition placement

As seen in section 5.1, BrE shows a preference for covert relativizers (*that* and *zero*) in restrictive relative clauses: *that* is favoured in speech (both informal and formal) and SU/DO positions, and *zero* in written informal texts and DO/PCOMP positions. *Wh*-pronouns are favoured in written formal texts and in SU/GEN positions in restrictive relatives, and they are the only relativizers used in nonrestrictive ones. This distribution agrees with previous descriptions of relativizer choice in English, with *wh*-pronouns being preferred in SU position in written formal contexts and in nonrestrictive relatives, and *that* and *zero* in speech and informal texts. In IndE, on the other hand, overt relativizers, i.e., *wh*-pronouns, are the preferred choice: in restrictive relatives, they are the most frequent option in spoken informal and written formal texts, and they are in competition with *that* and *zero* in spoken formal and written informal ones respectively. They are again favoured in SU and GEN positions, and they are not infrequent in PCOMP relatives in written formal contexts. *That* is only frequent in spoken formal texts and DO position, while *zero* is favoured only in written informal texts and DO/PCOMP positions. As in BrE, *wh*-pronouns are the default option in nonrestrictive relative clauses, with only a few cases of marginally non-restrictive *that* relatives. Overall, these results agree with Suárez-Gómez (2014), with a preference for *wh*-pronouns in SU position and competition between *that*, *zero*, and *wh*-pronouns in non-SU positions, with *wh*-pronouns being the most common relativizers overall.[18] These findings, as argued by Suárez-Gómez (2014: 259), characterize IndE as being more formal and similar to educated BrE in the 1950s. Additionally, the IndE preference for *wh*-pronouns over other relativizers may be the result of substrate influence (see § 3.1).

There are two findings in the present study that are unexpected and require further explanation: the high frequencies of *zero* relatives in written informal texts (in both IndE and BrE) and *that* relatives in spoken formal ones (in IndE). With respect to the former, *zero* relativizers, as has been mentioned before (§ 3), do tend to occur in

informal registers, but why should they be mostly restricted to written language? The *zero* relativizer is the variant that provides the least information: there is no relative marker at the beginning of the embedded clause, contrary to *wh*-pronoun and *that* relatives, which do have an explicit marker, and there is no information about the function of the gap in the clause, which is provided in *wh*-pronoun relatives in most cases. As a consequence, identifying the relative clause and the gap within it is harder for the addressee in *zero* relatives. In written language, however, the temporal constraints on spoken communication can be ignored, since the reader has access to the previous discourse and does not have the same pressure of keeping information in short-term memory (Mass, 2009: 166). Therefore, the lack of an explicit marker in *zero* relatives does not cause as many problems for the addressee in written language as it does in speech. As regards the frequency of *that* relatives in spoken formal language in IndE, this is indeed a very unexpected finding, taking into account what has been found in previous studies (§ 3.1), and is still in need of further research.

With respect to preposition placement in PCOMP relatives, BrE favours pied-piping in written formal texts in restrictive clauses, and there is competition between stranding and deletion in informal registers (both spoken and written). In nonrestrictive relatives, pied-piping is preferred in writing, and stranding in speech. These findings are only in partial agreement with previous research: pied-piping is not disfavoured in nonrestrictive relatives, although it is indeed more frequent in written formal contexts while stranding and deletion dominate in informal registers. In IndE, there is a higher frequency of pied-piping overall than in BrE: it is favoured in writing in restrictive relatives, and it is the only structure in nonrestrictive ones. Deletion is the preferred option in speech, and stranding is only frequent in written informal texts, where it is in competition with pied-piping. These results are again not completely aligned with those of previous studies, since deletion is not infrequent in speech (although it is rare overall) and stranding is only common in written informal texts. Pied-piping is preferred only in written formal texts and nonrestrictive relative clauses.

On the whole, there is more variation in restrictive relatives than in nonrestrictive ones. With respect to the choice of relativizer, there is more competition between the different alternatives in restrictive clauses, while in nonrestrictive ones *wh*-pronouns seem to be the default option. As regards the position of the preposition in PCOMP relative clauses, pied-piping is the only structural option available in nonrestrictive relatives in IndE, and a very common one in BrE. The situation again is more complex in restrictive relatives, with more competition between pied-piping, stranding, and deletion in both varieties.

6.2. Complexity effects

Both varieties follow the Accessibility Hierarchy proposed by Keenan and Comrie (1977): we find fewer instances of relative clauses as we go down the hierarchy due to the increased difficulty of relativizing a NP in the lower positions. With respect to the complexity of relative clauses, there is also a preference for simple relatives in both BrE and IndE, rather than for more complex ones with coordination or further embedding.

The most interesting findings, however, emerge from the interaction between the two variables relativized position and relative clause complexity: complex relatives decrease in frequency in the lower positions in the Accessibility Hierarchy, and this tendency is stronger in IndE than in BrE, with almost no cases of complex relatives in PCOMP and GEN positions in the former (4 in PCOMP and 0 in GEN). Additionally, in BrE there are many cases of complex relatives in SU-TR position, i.e., the SU position with which children had more problems in Diessel and Tomasello's (2005) study. It seems that BrE speakers opt for more complicated SU relatives (complex SU-TR) more often than expected by chance, a tendency that is not found in IndE.

The differences in relative clause formation between BrE and IndE can be characterized as differences in complexity. IndE speakers tend to produce simpler relative clauses than BrE speakers, with simpler domains for the processing of head NP-gap dependencies: they disfavour complex relatives in PCOMP and GEN positions, i.e., relative clauses with coordination and further embedding, constructions which, as argued in section 4, increase the number of nodes in the way from the head NP to the gap. PCOMP and GEN are the most difficult to relativize, and, therefore, IndE speakers tend to produce syntactically simpler and shorter relatives in these positions. The added complexity of relativizing a PCOMP or GEN NP and producing a complex clause seems to be too costly for these speakers, who, as mentioned in section 3.1, are mostly L2 or L3 speakers of English with non-native proficiency. The findings of the present study thus reinforce what has been found before in the literature: language contact that is characterized by short-term adult acquisition has a simplifying influence on languages/dialects.

7. Conclusions

This study has provided interesting results with respect to, on the one hand, relativizer choice and preposition placement, and, on the other, relative clause complexity in BrE and IndE. As concerns the selection of the relativizer and the position of the preposition in PCOMP relatives, the inclusion of different text types and nonrestrictive relative clauses in the analysis yielded a clearer picture of the variation found in the data. In line with previous research (cf., for instance, Gut and Coronel, 2012), text type turned out to be very important, since some relativizers are more characteristic of certain text types and not others: *zero* relativizers were mostly found in written informal texts in both BrE and IndE, a finding that was explained in terms of register and processing considerations (§ 6.1). Other distributions encountered in the present study, such as the unexpected high frequency of *that* relatives found in spoken formal texts in IndE, are still in need of further research. With regard to restrictiveness, more variation and competition between the different structural options were found in restrictive relatives. Nonrestrictive clauses, on the contrary, seem to be more homogeneous.

The analysis of relativizer choice and preposition placement is also significant for another reason: it can be used to test the reliability of the results of the present study against the background of previous research on relative clause formation in varieties of

English, which has so far been mostly focused on these two issues. The data for the present article has been extracted from a relatively small corpus (fewer than 90,000 words per variety), which casts doubt on the generalizability of the results. However, the distribution of relativizers and structural options regarding preposition placement in PCOMP relatives in the present data is very similar to that found in previous research based on larger corpora (§ 3), which suggests that the results of the present study are indeed worthy of consideration.

The main focus of the article lay, however, on complexity effects in the domain of relative clause formation in BrE and IndE. The results replicate a pattern that has been previously found in the literature: contact in which short-term adult L2 acquisition dominates tends to simplify languages. In this case, IndE, a high-contact L2 variety, shows a preference for simpler relative clauses in comparison with BrE. In order to properly understand this finding, it is crucial to consider the manner in which complexity was measured. The metric used here was a structural one, i.e., it measured the complexity of individual structures. Therefore, we cannot claim that the syntax of IndE is simpler than that of BrE in the domain of relative clauses, but it can be stated that IndE speakers tend to produce relative clauses that are easier to process, and that this is a consequence of their condition of L2/L3speakers of English. The metric used in the present study was also a relative one, i.e., it was rooted on the preferences of language users. Contrary to absolute metrics which, by definition, are independent from considerations of language use, relative metrics can locate the source of the complexity differences between BrE and IndE, which in this case emerge from the cost of processing relative constructions. IndE speakers, being non-native users of English, have more problems than BrE speakers when it comes to producing and comprehending relative clauses that relativize a PCOMP or GEN NP and that contain coordinate clauses and/or further embedding, due to the added processing cost. This, in turn, suggests that at least part of the decrease in complexity that takes place in contact situations is not due to imperfect L2 acquisition, but a result of structural simplifications that originate in language use. Thus, in future research, performance effects should be distinguished from simplifications that stem from the process of adult language learning.

Acknowledgements

Some of the material in this paper was originally presented at the *ChangE 2015* conference held at the University of Helsinki, 8-10 June 2015; I would like to thank the participants at the conference for helpful discussion. For generous financial support, I am grateful to the European Regional Development Fund and the following institutions: Regional Government of Galicia (Directorate General for Scientific and Technological Promotion, grant GPC2014/004); Spanish Ministry of Economy and Competitiveness (grants FFI2014-52188-P and BES-2015-071233). Finally, I would like to thank two anonymous reviewers for their useful comments and suggestions.

I declare that the work presented in this paper is original, and that it does not copy or reproduce the work of any other person, except as acknowledged in the references.

Notes

1. This does not mean that all the differences found in the present study between BrE and IndE with respect to the domain of relative clauses can be attributed to contact. While contact does in fact exert a strong influence in language variation, it is not its only cause; substrate effects, for instance, are another important motivation for change. I would like to thank an anonymous reviewer for drawing attention to this issue.

2. Syntactic complexity has also been operationalized in terms of the length in words or syllables of a syntactic unit (Wasow, 1997; Arnold et al., 2000; Wasow and Arnold, 2003). However, as demonstrated by Szmrecsanyi (2004), this metric is highly correlated with the number of nodes that a syntactic unit contains.

3. As an anonymous reviewer points out, relativizers differ in how complex/simple they are. For example, in the English relativization system, *that* relatives are considered to be simpler than *zero* relatives because the former are more transparent than the latter: in *zero* relatives the relativizer is not explicit. Furthermore, *that* is simpler than *wh*-pronouns, since these must agree with the animacy of the antecedent (*who(m)* with animate and *which* with inanimate antecedents) while *that* is invariable in this respect. Agreement adds redundancy to a structure because it implies that one meaning is expressed by means of two or more forms and, therefore, it increases complexity (cf. Suárez-Gómez, 2017). While these issues are very important to account for the complexity of relative clauses, the present paper focuses on complexity from a syntactic perspective.

4. Indirect objects and prepositional complements are not different with respect to their complexity in the metric proposed by Hawkins (1999).

5. Objects of comparison are excluded from Hawkins' (1999: 253) metric because "the coding of this position is highly variable across languages".

6. Many studies found support for the Accessibility Hierarchy in both L1 and L2 acquisition (cf. Izumi, 2003, and references therein).

7. Relative adverbs (*where*, *when*, and *why*) also occur in English when the gap is an adjunct, but these forms lie out of the scope of the present study.

8. *Zero* does occur in subject position in a restricted set of constructions, although very infrequently (Quirk et al., 1985: 1250; Biber et at, 1999: 619; Huddleston and Pullum et al., 2002: 1055).

9. Adnominal relative clauses are those "which depend on an explicitly mentioned nominal antecedent" (Suárez-Gómez, 2014: 246).

10 For example, in *I met the man to whom you gave the book*, the relative pronoun functions as the complement of the preposition *to*. No cases of actual IO relatives (e.g., *I met the man whom you gave the book*) were found.

11. As an anonymous reviewer observes, the complexity of the relative clause is not the only factor that influences the processing cost of relative structures. Another important issue to take into account is the position of the relative in the main clause (Izumi, 2003, and references therein; Diessel and Tomasello, 2005). Thus, relative clauses embedded in head NPs functioning as subjects of the main clause (as in example (6) above) are harder to process than those embedded in a nonsubject NP (as in (5)), because they hinder the parsing of the sentence by adding extra material between the subject and the main verb. This factor was explored in an initial stage of the investigation but was then abandoned since no significant differences between BrE and IndE were found.

12. The fact that we find fewer cases of relative clauses in IndE than in BrE in a similar number of words may already be an indication of the simplification processes at play in the former variety in this grammatical domain. Relative clauses are hard to process (see § 2.1) and, therefore, it is not surprising that we find fewer instances of this structure in the L2 variety.

13. There is one case of a *zero* relativizer in SU position in BrE. It functions as the subject of an embedded clause, a context in which relativizer omission in SU is allowed in English (cf. Huddleston and Pullum et al., 2002: 1047): *She played her "Minstrel Showboat" the one [ø you said sounded Chinese] for about 100 people […]* (ICE-GB: W1B-007 #114:3).

14. The global chi-square results are the same for BrE and IndE because variety was another variable in the HCFA test, i.e., the data from both varieties was included in the test in order to compare them.

15. There is also one case of a *zero* relativizer in SU in IndE. It occurs in the same type of context as the example found in BrE (see endnote 13): *The other point [ø I think is <,> important] Korean going to do a lot of interceptions* (ICE-IND:S2A-004#96:1:D).

16. There are five cases of *that* relatives in IndE that are not prototypical examples of restrictive clauses and were classified as nonrestrictive, as in *The whole misunderstanding about Hume's philosophical position is the outcome of his treatment of causation [that is often misunderstood]* (ICE-IND: W2A-001#58:1). However, as suggested by Denison and Hundt (2013: 162), a binary distinction between restrictive and nonrestrictive relatives might not be the best way to classify the data. As the focus of the present study is not on how to categorize relative clauses with respect to the relation between the head NP and the relative clause, these examples are not discussed further.

17. Only restrictive clauses are taken into account in the analysis of complexity effects. The syntactic relation between the head NP and the relative clause is different in restrictive and nonrestrictive relatives (cf. Huddleston and Pullum et al., 2002: 1058), and, therefore, processing the dependency between the head NP and the gap may also be different.

18. Suárez-Gómez's (2014) study focuses on restrictive relative clauses in the private spoken component of ICE-IND.

Primary sources

ICE-GB = International Corpus of English - the British Component (1990). Project Coordinated by Prof. Gerald Nelson at the Chinese University of Hong Kong. URL: <http://www.ice-corpora.net/ice/download.htm>.

ICE-India = International Corpus of English - the Indian Component (2002). Project Coordinated by Prof. S. V. Shastri at Shivaji University and Prof. Dr. Gerhard Leitner at FreieUniversität Berlin. URL: <http://www.ice-corpora.net/ice/download.htm>.

References

Arnold, Jennifer, Thomas Wasow, Anthony Losongco and Ryan Ginstrom (2000): "Heaviness vs. newness: The effects of structural complexity and discourse status on constituent ordering". *Language*, 76(1): 28-55.

Biber, Douglas; Stig Johansson, Geoffrey Leech, Susan Conrad and Edward Finegan (1999): *Longman Grammar of Spoken and Written English*. Essex: Longman.

Dahl, Östen (2004): *The Growth and Maintenance of Linguistic Complexity*. Amsterdam: John Benjamins.

Dahl, Östen (2009): "Testing the assumptions of complexity invariance: The case of Elfdalian and Swedish". In G. Sampson, D. Gil and P. Trudgill, ed., *Language Complexity as an Evolving Variable.* Oxford: Oxford University Press, 50-63.

Denison, David and Marianne Hundt (2013): "Defining relatives". *Journal of English Linguistics*, 41(2): 135-167.

Diessel, Holger (2009): "On the role of frequency and similarity in the acquisition of subject and non-subject relative clauses". In T. Givón and M. Shibatani, ed., *Syntactic Complexity: Diachrony, Acquisition, Neuro-cognition, Evolution*. Amsterdam: John Benjamins, 251-276.

Diessel, Holger and Michael Tomasello (2005): "A new look at the acquisition of relative clauses". *Language*, 81(1): 1-25.

Givón, Talmy (2009): *The Genesis of Syntactic Complexity: Diachrony, Ontogeny, Neuro-Cognition, Evolution*. Amsterdam: John Benjamins.

Gries, Stefan Th. (2004): *HCFA 3.2. A program for R*. Available at: <http://www.linguistics.ucsb.edu/faculty/stgries/>.

Gut, Ulrike and Lilian Coronel (2012): "Relatives worldwide". In M. Hundt and U. Gut, ed., *Mapping Unity and Diversity World-Wide: Corpus-based Studies of New Englishes*. Amsterdam: John Benjamins, 215-241.

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

Hawkins, John A. (1999): "Processing complexity and filler-gap dependencies across grammars". *Language*, 75(2): 244-285.

Hawkins, John A. (2004): *Efficiency and Complexity in Grammars*. Oxford: Oxford University Press.

Hilpert, Martin (2013): *Constructional Change in English: Developments in Allomorphy, Word Formation, and Syntax*. Cambridge: Cambridge University Press.

Hofmann, Thomas (2005): "Variable vs. categorical effects: Preposition pied piping and stranding in British English relative clauses". *Journal of English Linguistics*, 33(3): 257-297.

Huddleston, Rodney, Geoffrey K. Pullum, Laurie Bauer, Betty Birner, Ted Briscoe, Peter Collins, David Denison, David Lee, Anita Mittwoch, Geoffrey Nunberg, Frank Palmer, John Payne, Peter Peterson, Lesley Stirling and Gregory Ward (2002): *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.

Izumi, Shinichi (2003): "Processing difficulty in comprehension and production of relative Clauses by learners of English as a second language". *Language Learning*, 53(2): 285-323.

Karlsson, Fred (2009): "Origin and maintenance of clausal embedding complexity". In G. Sampson, D. Gil and P. Trudgill, ed., *Language Complexity as an Evolving Variable.* Oxford: Oxford University Press, 192-202.

Keenan, Edward and Bernard Comrie (1977): "Noun phrase accessibility and universal grammar". *Linguistic Inquiry*, 8(1): 63-99.

Keenan, Edward and Bernard Comrie (1979): "Data on the noun phrase accessibility hierarchy". *Language*, 55(2): 333-351.

Kortmann, Bernd and Benedikt Szmrecsanyi (2009): "World Englishes between simplification and complexification". In L. Siebers and T. Hoffmann, ed., *World Englishes – Problems, Properties and Prospects: Selected Papers from the 13th IAWE Conference.* Amsterdam: John Benjamins, 265-285.

Kortmann, Bernd and Benedikt Szmrecsanyi (2011): "Parameters of morphosyntactic variation in World Englishes: Prospects and limitations of searching for universals". In P. Siemund, ed., *Linguistic Universals and Language Variation*. Berlin/New York: Mouton de Gruyter, 264-290.

Kusters, Wouter (2003): *Linguistic Complexity: The Influence of Social Change on Verbal Inflection*. Utrecht: LOT.

Kusters, Wouter (2008): "Complexity in linguistic theory, language learning and language change". In M. Miestamo, K. Sinnemäki and F. Karlsson, ed., *Language Complexity: Typology, Contact, Change.* Amstredam: John Benjamins, 3-22.

Maas, Utz (2009): "Orality versus literacy as a dimension of complexity". In G. Sampson, D. Gil and P. Trudgill, ed., *Language Complexity as an Evolving Variable.* Oxford: Oxford University Press, 164-177.

Matras, Yaron (2009): *Language Contact*. Cambridge: Cambridge University Press.

McWhorter, John H. (2001): "The world's simplest grammars are creole grammars". *Linguistic Typology*, 5(2-3): 125-166.

McWhorter, John H. (2007): *Language Interrupted: Signs of Non-native Acquisition in Standard Language Grammar*. Oxford: Oxford University Press.

Miestamo, Matti (2008): "Grammatical complexity in a cross-linguistic perspective". In M. Miestamo, K. Sinnemäki and F. Karlsson, ed., *Language Complexity: Typology, Contact, Change.* Amstredam: John Benjamins, 23-41.

Miestamo, Matti; Kaius Sinnemäki, and Fred Karlsson (2008): *Language Complexity: Typology, Contact, Change*. Amsterdam: John Benjamins.

Mukherjee, Joybrato (2007): "Steady states in the evolution of New Englishes: Present-day Indian English as an equilibrium". *Journal of English Linguistics*, 35(2): 157-187.

Mukherjee, Joybrato and Stefan Gries (2009): "Collostructional nativization in World Englishes: Verb-construction associations in the International Corpus of English". *English World-Wide*, 30(1): 27-51.

Parkvall, Mikael (2008): "The simplicity of creoles in a cross-linguistic perspective". In M. Miestamo, K. Sinnemäki and F. Karlsson, ed., *Language Complexity: Typology, Contact, Change.* Amstredam: John Benjamins, 265-285.

R Core Development Team (2015): *R: A Language and Environment for Statistical Computing*. Viena: The R Foundation for Statistical Computing. URL: <http://www.R-project.org>.

Quirk, Randolph (1957): "Relative clauses in educated spoken English". *English Studies*, 38(1-6): 97-109.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech and Jan Svartvik (1985): *A Comprehensive Grammar of the English Language*. Essex: Longman.

Sharma, Devyani (2010): "Indian English". In B. Kortmann and K. Lunkenheimer, eds., *The Mouton World Atlas of Variation in English*. Berlin/New York: Mouton de Gruyter, 523-530.

Schneider, Edgar (2007): *Postcolonial English: Varieties around the World*. Cambridge: Cambridge University Press.

Schröter, Verena and Bernd Kortmann (2016): "Pronoun deletion in Hong Kong English and colloquial Singaporean English". *World Englishes*, 35(2): 221-241.

Sinnemäki, Kaius (2009): "Complexity in core argument marking and population size". In G. Sampson, D. Gil and P. Trudgill, eds., *Language Complexity as an Evolving Variable.* Oxford: Oxford University Press, 126-140.

Suárez-Gómez, Cristina (2014): "Relative clauses in Asian Englishes". *Journal of English Linguistics*, 42(3): 245-268.

Suárez-Gómez, Cristina (2017): "Transparency and language contact in the nativization of relative clauses in World English". *English World-Wide*, 38(2): 211-237.

Szmrecsanyi, Benedikt (2004): "On operationalizing syntactic complexity". In G. Purnelle, C. Fairon and A. Dister, ed., *Le poids des mots. Proceedings of the 7th International Conference on Textual Data Statistical Analysis*. *Louvain-la-Neuve, March 10-12, 2004*. Louvain-la-Neuve: Presses Universitaires de Louvain, 1032-1039.

Szmrecsanyi, Benedikt and Bernd Kortmann (2009a): "The morphosyntax of varieties of English worldwide: A quantitative perspective". *Lingua*, 119(11): 1643-1663.

Szmrecsanyi, Benedikt and Bernd Kortmann (2009b): "Between simplification and complexification: Non-standard varieties of English around the world". In G. Sampson, D. Gil and P. Trudgill, ed., *Language Complexity as an Evolving Variable.* Oxford: Oxford University Press, 64-79.

Szmrecsanyi, Benedikt and Bernd Kortmann (2012): "Introduction: linguistic complexity – Second language acquisition, indigenization, contact". In B. Kortmann and B. Szmrecsanyi, ed., *Linguistic Complexity: Second Language Acquisition, Indigenization, Contact*. Berlin/Boston: Walter de Gruyter, 6-34.

Thomason, Sarah and Terrence Kaufmann (1988): *Language Contact, Creolization, and Genetic Linguistics*. Berkeley: University of California Press.

Tottie, Gunnel (1997): "Relatively speaking: Relative marker usage in the British National Corpus". In T. Nevalainen and L. Kahlas-Tarkka, ed., *To Explain the Present: Studies in the Changing English Language in Honour of Matti Rissanen*. Helsinki: Société Néophilologique, 465-481.

Trudgill, Peter (2009): "Sociolinguistic yypology and complexification". In G. Sampson, D. Gil and P. Trudgill, ed., *Language Complexity as an Evolving Variable.* Oxford: Oxford University Press, 98-109.

Trudgill, Peter (2011): *Sociolinguistic Typology: Social Determinants of Linguistic Complexity*, Oxford: Oxford University Press.

Valdés, Guadalupe (2012): "Multilingualism". *Linguistic Society of America.* Available at: <http://www.linguisticsociety.org/resource/multilingualism> [accessed 24 November 2017].

Wasow, Thomas (1997): "Remarks on grammatical weight". *Language Variation and Change*, 9(1): 81-105.

Wasow, Thomas and Jennifer Arnold (2003): "Post-verbal constituent ordering in English". In G. Rohdenburg and B. Mondorf, ed., *Determinants of Grammatical Variation in English.* Berlin/New York: Mouton de Gruyter, 119-154.