Trade-Offs and Causal Relationships Between Cues to Grammatical Subject and Object in English and Creole Languages

Master's Thesis

Anna Pommersbach
1074775

The work submitted here is the sole responsibility of the undersigned, who has neither committed plagiarism nor colluded in its production.

Signed 

Name of student: Anna Pommersbach
Student number: 1074775

Abstract
Languages have different cues to express "who did what to whom" (case marking, word order, animacy, agreement), which are usually correlated (Sinnemäki, 2010; Levshina, 2021) and are influenced by processing factors. For example, Hawkins (2018) explains that richer, morphologically complex words allow for quicker and error-free online processing. However, it is longer than in more analytical languages like English. As claimed by Levshina (2021), language contact also plays an important role in the distribution of these cues. However, actual data from contact languages have not been studied in that regard. The present thesis aims to address this gap. The study investigates the cues to subject and object in two creoles: Tok Pisin and Bislama, and English as their lexifier. Data was extracted from the ANNIS and BNC corpora. The predictions are that the cues in creoles will be similar to the English one, because grammars of creoles investigated in this paper evolved from inter alia English grammar. Hence, the latter had major influence on the former. Second predication is that the analysis of creoles will replicate results of Levshina (2021) study–a strong negative correlation between word order and case marking. Moreover, the study is expected to give evidence for the hypothesis that language contact can result in semantic 'looseness' (Hawkins, 1986).

Table of Contents

1. Introduction

1.1. Background

1.1.1. Trade-offs and grammatical cues

It has been many decades since linguists dedicated their work to extracting and discovering dependencies and correlations of syntactic as well as semantic phenomena both within a language and cross-linguistically (Sapir, 1921; Greenberg, 1963; Hawkins, 1986). Findings in that field paved the way for establishing language universals and constructing a comprehensive theory on language evolution (for instance, Haspelmath (2016)). Language universals can take two different forms. Regularities like: "the two arguments of a transitive clause should be distinguishable" (de Hoop et al., 2008) or implications: if language *l* possesses a feature *f*, then it also has a feature *f'*. For instance, Haspelmath (2016) proposes a universal: if a language has an analytic and a synthetic causative, then the analytic causative tends to be used with transitive/unergative base meanings, and the synthetic causative with intransitive/unaccusative verb meanings, respectively

One of widely discussed topics of language universals are so-called trade-offs. These are inverse correlations between different features of grammar. The concept of trade-offs in syntax was proposed as early as the first half of 20th century. Sapir (1921) drew the attention to the importance of the word order and the case marking in interpreting the utterance and its parts. Moreover, he hinted at their (case's and word order's) mutual exclusivity. To illustrate that phenomenon by an example, in Polish with flexible word order and rich case system (7 different cases) it is the case that shows one who is the subject in the sentence. To illustrate that consider (1) and (2):

(1) English sentence
The mother gave her the apple.

(2) Polish translation(s):
(2a) Mama dała jej jabłko.

Mama	dała	jej	jabłko
mother.NOM (zero)	give.PAST.3PER.SING.FEM	she.DAT	apple.NOM-ACC

(2b) Jej dała mama jabłko.

(2c) Jabłko mama jej dała.

Although (2a) is the literal translation, (2b) and (2c) are completely grammatical and convey the same message. Of course, (2a) is slightly more natural, so choosing (2b) and (2c) would imply a desire to stress the component that is placed in the front, so that it was given to her and that it was the apple that was given, respectively. However, the factual meaning of the three is the same.

As Sapir (1921) comments it, "there is here no question of functional poverty, but of formal economy" (p. 33). In the 1960s, Joseph Greenberg (1963) proposed a set of language universals by presenting a correlation between various linguistic variables, *inter alia*:

- If the verb has categories of person-number or if it has categories of gender, it always has tense-mode categories;
- If in a language the verb follows both the nominal subject and nominal object as the dominant order, the language almost always has a case system.

Sapir and Greenberg paved the way for more complex theories on linguistic trade-offs whose validity was enhanced by modern scientific tools such as the corpus-based analysis. Some of the so-called Greenberg universals have been confirmed in a corpus-based study conducted by Hahn et al. (2020). They created a dataset of sentences from 51 languages and looked for correlations between different grammatical features and the word order. The study showed among others that in a group of languages that have SVO word order the proportion of prepositions to postpositions is 456 to 14. Analogously, in the SOV group in proportion is 472 to 42. Hahn et al. (2020) conclude that there is a correlation between the word order and the place of the adposition: SVO co-occurs with prepositions and SOV with postpositions. That in turn confirms Greenberg's theory.

Another example of developments in corpus linguistics is the study by Fenk-Oczlon and Fenk (2008). They argue that there is a trade-off between semantic complexity and morphological complexity (frequent multiplicity of meanings of words correlates with small number of syllables in words). Moreover, in some languages if word order conveys much information, then word structure does not have that function. We could interpret this correlation as an efficient trade-off. Morphological marking is not necessary when word order provides sufficient information about the utterance (Levshina, 2021). Going further, different rigid grammatical features in different languages can help the user (addressee) disambiguate the utterance and its parts. In other words, they can provide the user with cues how to interpret given utterance. In 2010, a study conducted by Sinnemäki showed that there is a universal corelation between SVO word order and zero-marking (no case system). One example is English, it has a rigid word order and a poor case marking, consequently it is the SVO rule that hints to the user which word (i.e. noun) is the subject and which the object. Studies conducted by MacWhinney et al. (1984) have shown that English speakers rely primarily on the word order to determine the subject of the sentence, whereas Italian speakers decide on the agency based on the agreement of the verb. When it is not decisive, they use animacy and the combination of stress and word order to interpret the sentence correctly. Moreover, German has strict word order rules and a case system of four cases. It is the latter that serves as a primary cue. However, in ambiguous cases, German speakers resort to animacy and word order.

This trade-off between two (or more) features of a language can be explained by communicative efficiency. Wilson & Sperber (2004) suggest that the foundation of communication is relevance, that is, an attempt to maximize the communicative effect with simultaneous minimalization of cognitive effort. Subsequent studies proved the existence of trade-offs between grammatical units as well as in the semantic domain (Kemp et al., 2018; de Coupe 2019; Berdicevskis et al., 2020). However, those claims did not take into account other

factors that contribute to the decision of inserting a cue or not. There are at least two more facets that are vital during communication. First, the accessibility of the message the speaker wants to convey. In other words, whether the information conveyed is new, hard to process or it requires highlighting. Second, the context of the utterance: the speaker's knowledge of the listener's knowledge and their language proficiency, their surroundings, to name a few (Wilson & Sperber, 2004). As Sinnemäki (2014) and Levshina (2019) show, in the majority of languages, rules concerning grammatical structures such as word order or case marking are not strict and therefore allow for variations. So, if a Polish speaker wants to stress the patient of the sentence, they would mark it via word order as well as case.

Also, a considerable number of studies showed that multiple cues facilitate perception and processing. Holler et al. (2018) make a case that people respond faster to questions when they are reinforced by a gesture. Holler & Levinson (2019) also suggest that information conveyed via various modalities is easier to process than one conveyed by only one means. It is due to the synergy effects and creation of Gestalts.

To conclude, many scientists suggest that there are trade-offs between various grammatical as well as semantic features of given language. A plethora of studies and theories refute the suggestion that the existence of communicative efficiency effects necessitates trade-offs. They are modulated and influenced by other factors such as the accessibility and the context (understood in a broader sense).

1.1.2. Looseness of a language and its grammatical complexity

Another aspect that requires an explanation is the distinction between loose-fit and tight-fit languages. The concept was first proposed by Hawkins (1986). A language exhibiting semantic looseness would have weak associations between grammatical roles and semantic classes of nominals. For instance, in English the usage of verbs is wider and the rules behind the creation of verbs are more flexible than in German. For example,

(3a) The book sold 10 000 copies.
(3b) *Das Buch verkaufte 10 000 Exemplare.
(3c) Das Buch verkaufte sich 10 000 mal. / 10 000 Exemplare von dem Buch wurden verkauft. [1]

As Mueller-Gotama (1994) describes Hawkins' observations, "the case system can…be held responsible for the fact that the basic grammatical relations of German are semantically more focused than those of English" (p. 4). He goes on by arguing that in English, as opposed to German, there is a wider semantic range of noun phrases that can potentially act as a subject and simultaneously maintain grammaticality. The tightness or looseness of a given language concerns not only semantics but also its grammatical aspects. For instance, raising or WH-movement are more widely spread in loose languages, according to Hawkins (1986).

Hawkins (1986) also observes that the tightness of languages can evolve. English epitomizes such a language–it lost case system. Also, its zero-marked NPs became more dependent on the verb for theta-role assignment. As a consequence, the rigid SVO order emerged and serves as a hint to the interpretation of "who did what to whom" (Sapir, 1921).

There is one more concept that should mentioned and it is strongly linked to the tightness of the language. It is the grammatical complexity, understood as *inter alia* richness of inflectional and declensional system as well as the number of grammatical units a language has. For example, English has articles, which are not present in Polish. As Siegel (2008) points out, grammatical complexity covers so many different aspects of a language, that it is hard to

[1] Examples taken from Mueller-Gotama (1994).

credibly judge without distortions whether one language is more grammatically complex than the other. Even though English has more articles than Polish, the latter has far more complex case system. Consequently, ranking the complexity of one higher than of the other would have to involve disregarding some aspects of grammar.

1.1.3. Complexity and language contact

Many researchers believe that grammatical complexity is shaped by sociolinguistic factors. For example, Trudgill (2011) proposed that brief language contact that mainly involves language learning by adults triggers the grammatical simplification (regularization and increased morphological transparency). On the other hand, long-term contact that is also co-territorial creates perfect conditions for bilingualism and first language acquisition. That, in turn, fosters complexification of syntax. This tendency is the effect of commonly known phenomenon–there is a sensitive period during which a child has the ability to learn a language quickly and naturally (for details see: (Granena & Long, 2012)). When one crosses that "threshold," one loses that ability and when using the second language resorts to simplifications to communicate effectively, but not necessarily correctly.

This intuition is reflected in the study by Sinnemäki and DiGarbo (2018). They investigated the relation between population size of first language (L1) and second language (L2) speakers of a given language with (1) its verbal inflectional synthesis (the number of morphemes or morphological categories that are realized in a word) and (2) a number of grammatical genders in those languages. The results showed that indeed there is a negative correlation between population sizes and the verbal inflectional synthesis. The bigger the L2 and L1 populations, the smaller the verbal inflectional synthesis. However, the authors did not find a correlation between number of genders and the population sizes. Another feature of grammar whose complexity is tested and compared cross-linguistically is the case system. Bentz and Winter (2013) conducted a study which investigated the relation between the case system and the population of L2 speakers. They found the statistical association between the binary variable (lack of or existence of case system) and the proportion of L2 speakers. Moreover, their study showed an inverse association between the number of case markers and the proportion of L2 speakers.

All in all, there is an agreement among a plethora of scientists that extensive (in time and scale) language contact influences language complexity. Languages tend to accumulate grammatical complexity, but it can be kept and transmitted only by native speakers of given language. L2 speakers (whose number coincide with brief language contact) tend to simplify the grammar and disregard contextual grammatical nuances. As Maitz and Nemeth (2014) put it, "language contact can also lead to broad typological changes and in particular, to structural simplification. The cause of such contact-induced simplification is adult second language acquisition, which in most cases does not lead to native proficiency attained through first language acquisition" (p. 3). However, as Sinnemäki's (2020) results showed that the decrease of probability of case is influenced by the increase of L2 speakers only in languages that have certain word order (SVO, VSO, VOS). Thus suggesting the conditional influence of L2 speakers on language complexity (or at least some of its aspects).

1.2. Aim of the Study

Results of Levshina's (2021) study showed a strong negative correlation between case marking and rigid word order, as well as a positive correlation between case marking and tight semantics, and an association between verb-final position and case marking. One of the explanations presented was that language looseness, like grammatical complexity, can be influenced by language contact (via the increased number of L2 speakers). The claim was also that extensive

language contact can increase rigidity of word order and decrease complexity of morphological system. One could expect a trade-off between word order and case marking (cf. Sinnemäki 2008, 2010), but how these and other correlations manifest themselves in high-contact languages has not been explored based on corpus data.

This thesis aims at testing Levshina (2021)'s and Sinnemäki's (2010) results showing correlations between different cues to subject and object in contact languages, thus filling a gap in previous research. Also, the study investigates the relation between semantic looseness and language contact. Creoles as languages that evolve as a result of intensive language contact provide perfect data for testing this hypothesis. This thesis' goals can be summed up by three questions:

- Which semantic and formal properties serve as cues to subject and object in English and creoles if we take naturally occurring speech?
- What are the relationships between these cues across the languages? Are there trade-offs?
- Is semantic tightness lower in high-contact languages like creoles than in English?

Consequently, in this thesis there are three hypotheses that are tested. First, following Levshina (2021) it is predicated that word order and case-marking are reliable cues to subject and object. Second, a trade-off (negative strong correlation) between casing marking and word order is expected. Lastly, there is a positive effect of language contact on sematic looseness.

1.3. Methods and Procedure

The above-mentioned hypotheses on (1) word order and case-marking as cues to subject and object, (2) a trade-off (negative strong correlation) between casing marking and word order and (3) positive effect of language contact on sematic looseness were checked by means of the analysis of the grammatical structure "who did what to whom." The reason for choosing such a structure is twofold. First, it is a simple structure, which is common in language use. This makes it easy to find, extract and analyse. Secondly, its potential ambiguity resulting from the fact that it contains at least two participants that are bound by only a verb (no adpositions) can only be solved by grammatical means. Since grammatical cues are necessary in this structure, they are "visible" and easy to detect.

The analysis of two creole languages was performed. These are Tok Pisin and Bislama. All data was extracted from ANNIS, an online corpus[2]. Moreover, the analysis of informal spoken as well as written English was added as a point of reference and acknowledgement of the origin of the said creoles. For that purpose, the British National Corpus (BNC)[3] was used. I extracted 100 examples of transitive clauses per language, with a finite verb, an overt subject and overt object. These clauses were annotated for the following variables:

- Case marking of subject and object.
- Word order (SO or OS; verb-medial or not).
- Semantic class of subject and object. The sematic class had three values: Human, Animate and Inanimate. Words describing groups of people (government, counsel, party, etc.) were assigned as human.
- Part of speech (common noun, pronoun, proper name).

Those variables were used to operationalize three concepts: case marking as a cue (relation between grammatical role and case), semantic tightness (relation between part of speech and

[2] http://www.corpus.dynamicsoflanguage.edu.au/
[3] https://www.english-corpora.org/bnc/

grammatical role as well as relation between semantic class and grammatical role) and word order rigidity (relation between grammatical role and the place in the sentence (first, second)).

The analysis included description of grammatical cues used in the structure and frequency of the relevant grammatical features. First, basic descriptive statistical measures were calculated. I also compared the strength of the correlations in the creole languages and in English. Finally, multivariate analyses were carried out (using conditional inference trees and random forests), which allow for the evaluation of which cues allow one to predict the identity of a nominal argument the best, and which are less relevant.

1.4. How the Creole Languages Emerged and Evolved

To fully grasp the relations within grammatical systems of creole languages one has to first analyse and understand how they emerged. The first roots of such languages can be traced down to the Europeans' travels to remote locations in Latin America, Africa and Oceania. They forced indigenous people out of their homes and hired them at, inter alia, plantations. This situation created a need for a common language or communication system between workers as well as between workers and plantation governors.

Melanesian Pidgin epitomizes an effect of such a process. In early 1800s Europeans started to trade with Melanesians, mostly in sandalwood. Both parties got to know and acquired phrases and expressions from each other's language. They started to spontaneously create a way of communication based on simple structures and words from the other language that they already knew. It is important to note that at that stage every user of this system had their own individualized way of communicating. That is how a pre-pidgin emerged. When a few decades later Europeans involved, both voluntarily and involuntarily, Melanesians in their plantations in Queensland Samoa and Fiji, this jargon had perfect conditions for a stabilization: a big group of speakers of different languages that had to communicate with each other and stayed together for a longer period of time. With stable contact and "continuous use, new features were added, norms began to emerge and a stable pidgin language began to develop-early Melanesian Pidgin" (Siegel, 2008, p. 84). The third stage of the development is when the workers come back home and bring a stabilized but still developing pidgin. Since the laborers came from different countries, after their coming back to the homeland the pidgin was under the influence of different indigenous languages. Thus, three current dialects developed: Tok Pisin, Bislama and Pijin. The pidgin starts to be used in increasing number of aspects of life of a community and people start to mingle between villages. Consequently, there are more and more mixed marriages and families in which children hear the pidgin from the beginning and it is their first language. Making the language official (a language of newspapers, parliament proceedings, etc.) and the emergence of its native speakers are the points where a pidgin becomes a creole, a fully-fledged language from syntactic, semantic as well as sociological perspective.

All aforementioned stages (pre-pidgin, pidgin and creole) are effects of socio- and psycholinguistic processes and thus possess specific linguistic characteristics. The pre-pidgin and restricted pidgin (so not yet stabilized) are stages governed by second language acquisition (SLA) processes. A few scientists[4] compared how people use language in the early stages of its acquisition (both a first and second language) with the grammar of (not stabilized) pidgin languages. The general conclusion was that both restricted pidgins and the early stages of SLA can be characterized by preference of lexical items over grammatical ones. For instance, expressing tense and aspect by lexical means–adverbs and adjectives. Moreover, the results showed striking number of similarities in very specific domains[5]:

- no inflections,

[4] See, for example, Schumann (1978), Givón (1979), Kotsinas (2001),

[5] I only name a few of them.

- multifunctionality of lexical items,
- few quantifiers and prepositions,
- no complementizers and expletive elements,
- no tense, mood and aspect (TMA) markers.

The poverty of the grammatical system is explained by perceptual prominence of content words. In other words, lexical items are those that are stressed and, with regard to English, contain strong syllables. Consequently, they are more distinguishable, easier to process and acquire. On the other hand, grammatical items are "squeezed in" and often are reduced to consonant(s) and a vowel shwa. Bates and Goodman (1999) argue that grammar is a function of vocabulary, the latter's size has a direct impact on the former's development. Additionally, overgeneralization (disregarding irregularities) and regularization, which are very characteristic symptoms of early stages of LA, are also present in many restricted pidgins and creoles. For example, in Hawai'i Creole a first-person singular possessive is "mainz" (mines) suggesting the overgeneralization of "s" (yours, hers, ours, etc.). Also, the plural marking "s" is used more often: junks, furnitures, baggages, etc. The restricted pidgin, a lexically-dominant, grammatically-absent, highly individualized communication system, is then stabilized. In other words, various ways of communicating a sentence are reduced to only one (or a few) that are deemed correct. "The choice of features is affected by a combination of environmental factors, (such as frequency) and linguistic factors (such as semantic transparency) …The remaining features make up the stable contact variety" (Siegel, 2008, p. 40).

The next step is the direct reaction to the language's poor ability to answer communicative demands, which is twofold. Firstly, the speaker is not proficient in that language and second the language itself is not developed enough. Under such circumstances, the morphological expansion appears, which is a main factor that constitutes the creolization of a pidgin. The process involves the development and emergence of new grammatical structures and units. One source of changes are language-internal developments. They are usually the effects of grammaticalization, that is, a process of a lexical item becoming a grammatical one. For instance, in Tok Pisin a future marker "baimbai" (and later "bai") is derived from English "by and by" and a possessive preposition "bilong" is derived from an English verb "belong to." The second source of change are the influences from other languages: the lexifier and substrate (indigenous) languages. The examples of direct targeting from the lexifier have been already mentioned, Hawai'i Creole took both the form ("s" as a suffix) as well as the function (possessive marker) from English grammar. There are also instances of full substrate influence, however "the mixture" is the most common. Frequently, the form is taken from the lexifier and the function from the substrate. Notably, Keesing (1988) identified seven syntactic features that are present both in Central Eastern Oceanic languages (substrates) and the ancestors of Melanesian Pidgin. These are:

- subject-referencing pronoun in the verb phrase,
- transitive suffix on the verb,
- adjectives functioning as stative verbs,
- preverbal causative marker,
- post-nominal possessive marker,
- third-person plural pronoun used as a plural marker,
- exclusiveness and dual number marker in the pronoun system.

Importantly, these features are not present in English grammar. This widespread phenomenon is explained by so-called imposition transfer. It is a process characteristic of SLA "in which the linguistic features of one language are used in learning or using another language" (Siegel, 2008, p. 106). The "imposition" implies that the user is only fluent in one language, that is the

source language. Transfer, in turn, has two types: syntactic (word order) and functional (using L2 forms with L1 grammatical features). As Siegel (2008) argues, the former is characteristic of SLA. To illustrate it by an example, it is more than plausible that an English native speaker would construct a sentence in French (L2) in such a way:

(6) *Louise toujours mange du pain.
(7a) Louise always eats bread.

When grammatically correct sentence would have the verb on the second position:

(7b) Louise mange toujours du pain.[6]

The second type of transfer, functional, mostly occurs during the usage of L2, but also there are some studies arguing its presence in bilingual first language acquisition (see: Mueller, 2006). One of the examples is Singapore English. Word "already" is used there as a perfect aspect marker which corresponds to the word "liau" in Sinitic (a substrate). For instance:

(8a) Gun thauke trig chhu l**iau**.
(8b) Our boss return home **already**. / Our boss has returned home.[7]

The question remains why, from the socio- and psycholinguistic point of view, such a process occurs. Why are functions rather than forms of grammatical units more likely to be transferred? As mentioned above, the first cause of transfer is the communicative need which increases with the spread of language use from smaller communities (plantations) to whole communities.

To conclude, the processes that make up the emergence and development of creoles are not unique and specific to that kind. Similar processes (and their results) can be observed during second (and first) language acquisition, such as omission of grammatical units, overgeneralization and functional transfer. Although pidgins and creoles are succumbed to the same language change processes as other "normal" languages, linguists[8] acknowledge their uniqueness in how rapidly, not in a gradient fashion, the changes occurred. Also, DeGraffs argues that there is a "sociohistorical difference from other situations involving language change in that a larger number of language groups were in contact in the creole situation than in other situation" in (Siegel, 2008, p. 52). The conclusion is that "the conditions that bring about ordinary language change and those that lead to creole features have fundamental differences, including intense language contact" (Siegel, 2008, p. 52).

1.5. Summary and hypotheses

This thesis focuses on grammatical cues to subject and object in transitive clauses. The analysis concerns two creole languages as they present qualitative uniqueness, which was argued in the section above. Also, it includes English as a lexifier and consequently, the point of reference for the creoles. In the sections below, I discuss the evolution of said languages, present grammars and show results of the multivariate analysis. In Section 2, English is discussed: its evolution in terms of grammar as well as vocabulary triggered by, inter alia, language contact and the present state of syntax. The next part discusses the methods used and the results of statistical analysis (both bivariate and multivariate). Section 3 concerns Tok Pisin and Section 4 Bislama. Their structures are analogous to that of English: I describe basic, and relevant to the analysis, features of grammar, present results and discuss their implications. The last Section is devoted to the comparison of the results of different languages and the general conclusion.

[6] Example from Siegel (2008).
[7] Example from Platt and Weber (1980) in Siegel (2008).
[8] DeGraffs (2001, 2003).

2. English Grammar

2.1. Evolution of English Grammar

English, and therefore English grammar, has roots in Germanic languages. The period between 450 and 1150[9] marks the development stage of English called Old English (OE). Its characteristics were of Germanic sort with constant influence of Latin (via Romans and Christian missionaries). One of the features shared by the Germanic languages is the stress on the first syllable. Consequently, non-initial vowels were mostly weak and hence subject to reduction. As Baugh and Cable (2002) comment it, this is the "feature of great importance in all the Germanic languages because it is chiefly responsible for the progressive decay of inflections in these languages" (p. 46). One of the key features of verbs in those languages was the distinction between strong and weak verbs. In the former the preterite was indicated by the change of the vowel, which remnants we can observe in irregular English verbs (drink - drank), whereas in the latter the dental suffix was added, whose remnants are visible in regular verbs (walk - walked). As for nouns, Germanic had three genders and two (out of three used in Indo-European) numbers. In Germanic languages, adjectives gained another declension–a weak one. Its evolved variation is still visible in German: "(strong adjective) guter Mann (OE god mann): (weak adjective) der gute Mann (OE se goda mann)" (Bammesberger, 2005, p. 54). The eight cases that the Indo-European had had have been reduced to two (nominative and possessive) throughout the centuries. Rich case system allowed for flexible word order; Germanic languages as well as the Indo-European used both SVO and SOV, the latter being regular one and the former used for the sake of emphasis.

English from the period between roughly 1150 and 1500 is called Middle English and it marks French influence and its fossilisation in the language (and culture, but it is not the subject of this paper). The language of invaders, French, became the language of upper-class and politics. Hence, very soon after the conquest English people started to learn and speak French. During that time English lost its inflectional character. Endings of nouns, adjectives and verbs were gradually reduced (a phonetic processes) making case, grammatical gender and number disappear. Since the "s" sound was subject to neither reduction nor change, the masculine declension of plural form remained (-es and -s) later becoming universal plural marker, which is still today[10]. Analogically, genitive singular marker "es" spread onto all forms. One of the grammatical aspects that maintained in a similar form to the Old English ones were personal pronouns. As Baugh and Cable (2002) explain it, "most of the distinctions that existed in Old English were retained…However the forms of the d[a]tive and accusative cases were early combined, generally under that of the dative (him, her, [t]hem). In the neuter the form of the accusative (h)it became the general objective case…One other general simplification is to be noted: the loss of the dual number" (p. 150).

Verbs, on the other hand, were changed twofold. First, weak declension became dominant (which effects we still see in English). The explanation is that there is the tendency to follow and choose option that is familiar and exhibits comprehensible rule–"the weak conjugation offered a fairly consistent pattern for the past tense and the past participle, whereas there was much variety in the different classes of the strong verb" (Baugh & Cable, 2002, p. 151). Second, English verbs were often replaced by French equivalents.

Also, it is the period of change in the word order. As explained above, Old English was a SOV-dominant language, but allowed for variations (SVO). In subordinate clauses, verb was

---

[9] I leave out Celtic influence on English. "Outside of place-names, however, the influence of Celtic upon the English language is almost negligible" (Baugh & Cable, 2002).

[10] With some remnants of weak masculine declension, like in "oxen."

at the final position (as it is often the case in today's German). Contrastingly, Middle English became an analytic language with more and more rigid word order–SVO, which was proceeded by the loss of inflection.

The period from 16th to early 17th century marks the early Modern English. Democratization of books (by printing press) and advancement of science fostered vocabulary development (which relied on Latin, French and Italian words) on the one hand and stressed the importance of preservation and stabilization of grammar on the other. Hence, there are no dramatic changes in syntax, rather a unification and fossilization of rules (a shift from -eth to -s as a third person singular or dominance of s as a plural marker). When it comes to phonology, 15th century was the time of Great Vowel Shift–a revolutionary change in pronunciation that covered the whole system of English vowels.

After the upheavals of the 17th century and emerging of rationalistic thought, English society felt the need of systematization and introducing order also into language. "Now for the first time, attention was turned to the grammar, and it was discovered that English had no grammar" (Baugh & Cable, 2002, p. 241).

During the mid-eighteenth century, English has reached the status which, with some changes, has today. As Denison (1999) comments, there has not been much syntactic change for the last two centuries. Shifts were rather statistical, not qualitative. In other words, some constructions became more popular and some obsolete. Denison (1999) remarks that "the overall, rather elusive effect can seem more a matter of stylistic than of syntactic change" (p. 93; Woolford, 1979).

2.2. Basic Features of English Grammar

Today's English is an effect of 1500 years of processes constituting language evolution and the influence from language contact. The latter was constantly present–Celtic, Scandinavian languages, Latin, French have all imprinted on English language. In today's form, English is an analytic language being on a "grammaticizing" side of the Semantic Typology[11] spectrum. That implies much semantic diversity of grammatical relations, no overt role markers for grammatical relations, fixed intraclausal word order and frequent extraction and adposition stranding (Mueller-Gotama, 1994).

The first characteristics is the flexibility of words to acquire or assume once a grammatical function of a different part of sentence. Naturally, it also involves assuming different meaning. The symptoms of the first characteristic are very frequent in English. Consider this:

(9) "You don't act very grandmotherly."[12]
(10) I water my plants every day.
(11) "Don't Donna me."[13]

(9) and (11) use words that do not exist (are not included in the dictionary). The adverb "grandmotherly" is formed by adding "-ly" suffix, an adverb marker, whereas the verb "Donna" acquires the meaning and new grammatical function via its position–between the subject and the object. Analogically, in (10) the word "water" did not change its form (no morphological units added) in becoming a verb.

The second feature, fixed intraclausal word order, is an extension of the general SVO word order. It means that the clauses constituting one sentence does not influence grammatically one another (this rule is also visible in so-called reported speech where tense is

[11] (Mueller-Gotama, 1994)
[12] From *Young Sheldon* series.
[13] From *Suits* series.

treated universally, not subjectively). The rigidity of intraclausal word order indicates to what extent is that grammatical aspect dominant in English grammar. In German (as well as Old English) the subordinate clause's word order is determined by the main clause. For instance,

(12a) Mattis kommt später zur Arbeit, weil er noch sein Fahrrad reparieren **muss**.
(12b) *Mattis comes later to work, because he his bike repair has to.
(12c) Mattis comes later to work, because he has to repair his bike.

Even though there is a strict rule in German that a verb comes in the second place, in the subordinate clause the verb "muss" is after both the subject and the rest of verb phrase.

The third feature, frequent extraction (topicalization, wh-movement), is strongly linked to the word order. Its frequency can be easily explained by the rigidity of word order. If SVO is a prevailing feature of English grammar, it leaves no room for variability. Movements within word order are common means for stressing and highlighting parts of sentences (see: Section 1.1). Consequently, English speakers resort to extractions to express salience of given part of information (and sentence).

The last feature mentioned is adposition stranding:

(8a) What are you looking **at**?
(8b) * At what are you looking?
(7b) **Na** co patrzysz?

In English, the preposition remains adjacent to the verb and thus is separated from the noun phrase. Conversely, in Polish sentence the preposition "na" is attached to the noun phrase, which is a more common way of forming questions.

The development or in some cases even existence of grammatical features mentioned above is the effect of rigid word order. Its dominance renders it a reliable and efficient grammatical cue to subject and object. For that reason, it was excluded from the main analysis.

2.3. Analysis and Results

2.3.1. Data

2.3.1.1. *Corpus*

The British National Corpus was used to extract examples for this analysis. The corpus contains both spoken and written English, from various sources: newspapers, literature, legal documents, political speeches, etc. The utterances were collected in the 20th century.

2.3.1.2. *Dataset*

The dataset consisted of 100 examples of sentences with a transitive verb. They were chosen twofold. The first half were the result of running a formula that looked for structures: Pronoun / Noun Verb Pronoun / Noun. Despite being efficient, this method excluded sentences like: "We want cast-iron guarantees that…" Hence, for the second part of the dataset a different formula was used. I run an analysis of most frequent verbs, picked those that were transitive and chose sentences with them.

2.3.1.3. *The annotation procedure and variables*

Then, the sentences were annotated manually in the excel file. Various features were described: word order, subject's part of speech, how (and whether) the subject is case-marked and subject's animacy (henceforth: semantics). Analogous annotation was done for the object of each entry. Features and their examples are shown in the Table 1:

Variable	Meaning	Values	Examples
A_POS	Part of speech of Agent	Noun, Pronoun, Proper name	The girl likes ice-cream. I like ice-cream. Sue likes ice-cream.
A_Semantics	Semantics of Agent	Human Animate Inanimate	Sue likes ice-cream. My cat likes ice-cream. The hammer hit the nail.
A_Case	Whether Agent is case-marked	Zero Nom	Sue likes ice-cream. She likes ice-cream.
P_POS	Part of speech of Patient	Noun Pronoun Proper name	I like ice-cream. I like it. I like Cracow.
P_Semantics	Semantics of Patient	Human Animate Inanimate	I like my mum. I like his cat. I like ice-cream.
P_Case	Whether Patient is case-marked	Zero Oblique	I like it. I like her.

Table 1: Table with variables used in the analysis.

2.3.2. Methods

2.3.2.1. *Bivariate statistics*

Bivariate statistics, as the name indicates, measures and tests relations between two variables. It shows correlations (or their lack) as well as causal relations. In this study, bivariate statistics helped in the separate analysis of relation of pairs of variables (for instance, grammatical role-POS). That is, regardless of the other variables and their potential effect on each variable of the pair as well as the pair's relation.

A few bivariate statistics tests were included to investigate the relationship between the grammatical roles (subject and object) and the other variables. First, the Fisher's Exact Test for Count Data was run. It is used to check for associations between categorical variables. In this analysis, the Fisher's test was run to ascertain that the potential association between variables is not due to chance.

The second statistic was residuals, which compares the observed frequencies to those expected. As Levshina (2015) explains it, "expected frequencies are the frequencies that one can expect if the variables are independent, that is, if the null hypothesis were true and there were no differences in the proportions" (pp. 210-211). This allows me to determine to what extent the distribution of variables diverges from the expectances and thus, what "the scale" and direction of the association is. The straightforward interpretation of residuals is that "the greater the absolute value of a residual, the greater the discrepancy between the observed and expected frequencies, and the more it contributes to the test statistic" (Levshina, 2015, p. 218).

The last descriptive statistics test used in this analysis was Mutual Information (MI), a measure from information theory. It shows how much we know (can predict) about one variable

when having the information on the other. It is based on the individual probabilities of the roles and the values of the other variables, as well as the joint probability of every combination of their values. In other words, it is the means for operationalizing three concepts mentioned above: word order rigidity, casing marking as a cue and semantic tightness. Moreover, MI allows for the comparison of results cross-linguistically.

2.3.2.2. *Multivariate statistics*

The multivariate statistics is used to test and analyze relation between multiple variables. In other words, what is the relation between x and y when we take into account z? Also, multivariate statistics "ranks" variables according to their ability to explain (and predict) investigated phenomenon (a variable). In this particular study, multivariate statistics tests helped determine which grammatical features (case, POS, semantics) predict the grammatical role of a given word. Going further, if we know which variable predicts grammatical role most effectively, we know which grammatical feature is the best cue in determining what is the subject and what is the object. That, in turn, is the main aim of this study.

The multivariate statistics in this analysis investigated if we can predict the grammatical role (subject or object), which is called the response variable, based on the other variables (predictors). The main difference from the bivariate analyses is that we can test the association between every predictor and the response variable while controlling for all other predictors. The multivariate analyses had two elements: conditional inference trees and conditional random forests, which returned the conditional importance of all predictors of the roles. Conditional inference trees have two aspects: recursive splitting and permutation to obtain p-value. When it comes to the former, the algorithm repeatedly splits consecutive variables into two (into values) to find an optimal scenario, in which the most important predictor is the strongest of all possible. For example, when predicting the feature of having polished nails, the gender is the strongest predictor, and within the "female" group we can add the age factor and split into {Child, Teenager, Adult}. Contrastingly, the age variable is weaker most important predictor of having polished nails. In other words, age variable would not split (predict) polished nails as accurately as gender[14]. The second aspect is obtaining the p-value by permutation. It "means that the labels on the observed data points are rearranged many times, and for each rearrangement the relevant test statistic is computed…Next, one determines the proportion of the permutations that provide a test statistic greater than or equal to the one observed in the original data" (Levshina, 2015, p. 292).

The random forest is essentially a collection (summary) of hundreds or thousands of conditional inference trees, when a certain number of predictors and observations is sampled from the entire data. They return more reliable results than individual trees.

As Levshina (2015) shows, there are several advantages of conditional inference trees and random forests. For instance, in the case of "data sparseness ('small n large p', where n is the number of observations and p is the number of predictors)" (p. 292). Moreover, the reason for using conditional random forests is their ability to deal with correlated predictors. As Strobl et al. (2008) point out, "correlations between predictor variables…severely affect the original random forest variable importance measures, because they can be considered as measures of marginal importance, even though what is of interest in most applications is the conditional effect of each variable" (p. 2).

The last element of the multivariate statistics was the conditional variable importance test. The conditional importance of a given variable for the entire forest is computed as an average over all trees. Thus, it also shows the strongest predicator of a given response variable. However, it also calculates the importance of other variables.

[14] This example is not based on any data, just on author's experience and assumptions.

2.3.3. Results

2.3.3.1. *Descriptive results*

The co-occurrence frequencies of the roles (A and P) and of values of POS, case, semantics and word order were calculated (for results see Table 2). The analysis of POS showed that two out of three parts of speech (POS) had a balanced subject-object distribution: noun (56 to 62), pronoun (34 to 37). Proper names, on the other hand, appeared in 10 out of 11 times as the subject.

The residuals of chi-square test (df = 2) showed the highest value in proper name class: 1.919 (for the subject). Noun and pronoun classes had -0.3906 and -0.252, respectively. The interpretation is that in proper name class the frequency of subjects is greater than expected, whereas in noun and pronoun classes it is the opposite, the frequency of subjects is less than expected (and objects appear more often).

	Noun	**Pronoun**	**Proper Name**
A	**56** [-0.3905667]	**34** [-0.2517544]	**10** [1.9188064]
P	**62** [0.3905667]	**37** [0.2517544]	**1** [-1.9188064]

Table 2: Frequencies and residuals of POS in English

As Table 3 shows, the distribution of case was not as balanced as in the case of POS. There were three values: the value "Nom" was assigned to the basic form of words that have a second form that is a remnant of accusative. Namely, these are pronouns: I, she, he, we, they. The "Oblique" value was assigned to the second form: me, her, him, us, them. The third value represents the rest which does not have any variety in form and marking. The analysis showed the highest residual value on "Nom", 3.937 (for subjects). Close to that value was "Oblique", with the value of -3.316. The most balanced value was "Zero", -0.525. In other words, "Nom" subjects and "Zero" and "Oblique" objects were more frequent than expected.

	Nom	**Zero**	**Oblique**
A	**31** [3.9370039]	**69** [-0.5248907]	**0** [-3.3166248]
P	**0** [-3.9370039]	**78** [0.5248907]	**22** [3.3166248]

Table 3: Frequencies and residuals of Case in English

As Table 4 shows, the frequencies of semantics values (Human, Animate, Inanimate) were not balanced. For example, 22 out of 82 instances of inanimate class were subjects. When it comes to absolute values, human and inanimate class had similar residuals of chi-square test (df = 2): 2.364 and -2.967, respectively. It means that human subjects and inanimate objects are more frequent than expected. Also, animate subjects were more frequent, value: 1. The situation with the objects is the opposite–they appeared more often as inanimate and less often as animate and human.

	Animate	**Human**	**Inanimate**

A	**2** [1]	**76** [2.363516]	**22** [-2.967301]
P	**0** [-1]	**40** [-2.363516]	**60** [2.967301]

Table 4: Frequencies and residuals of Case in English

Since there is a rigid word order in English (SVO), 100% of subjects were in the first position and 100% of objects were in the second position. For results, see Table 5:

	First	**Second**
A	**100** [7.071068]	**0** [-7.071068]
P	**0** [-7.071068]	**100** [7.071068]

Table 5: Frequencies and residuals of Word Order in English

2.3.3.2. *Fisher's test*

The Fisher's test was calculated for all pairs of role and other variables. The results are as follows:

	Role-Case	**Role-Semantics**	**Role-POS**	**Role-Word Order**
p-value	< 0.0001	< 0.0001	0.02102	< 0.0001

Table 6: Fisher's test for all variables

As the Table 6 clearly shows, all p-values for Fisher's test are smaller than 0.05. Hence, the association between the role and given variable is not due to chance. In other words, the association is significant.

2.3.3.3. *Mutual Information*

The next step was computing Mutual Information to see how much one can infer about the role from the information about the POS / case / semantics / word order. Its results are presented in the Table 7:

Case	**Semantics**	**POS**	**Word Order**
0.2669886	0.1169774	0.03238583	1

Table 7: Mutual Information for role in English.

2.3.3.4. *Conditional inference trees*

The second part of the study was multivariate analysis (conditional inference trees and random forests). Because word order separated perfectly between subject (A) and object (P), it was

excluded from the analysis. I only tested how much POS, semantics and case contribute to distinguishing between A and P, when all predictors were controlled for.

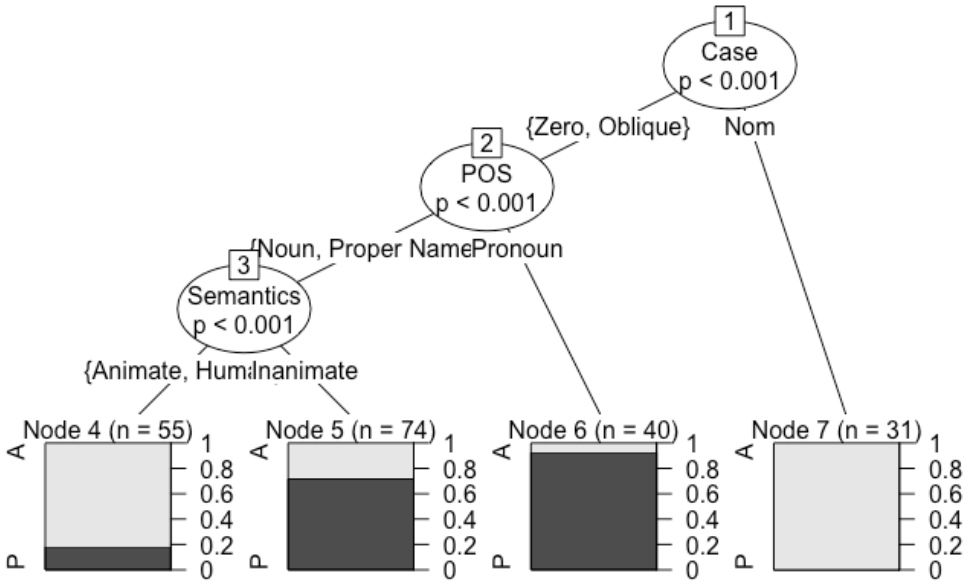

Table 8: Conditional inference tree, English

As it is clearly visible in the graph (Table 8), the case variable is the most important predictor in subject-object determination. The distinction is Nom vs {Zero, Oblique}. 100% of "Nom" instances were subjects. The second most important predicator, within {Zero, Oblique}, was POS. The distinction was pronouns vs nouns and proper names. Within the former category 90% of its instances were objects. The third, and last, significant predicator was semantics. The animacy served as the demarcation line. In other words, the distinction was {Human, Animate} vs Inanimate. Within the former group less than 20% of instances were objects and within the latter it was 70%.

2.3.3.5. *Conditional random forests and variable importance*

The analysis also included calculation of conditional importance of the variables. Its results are presented in the Table 9:

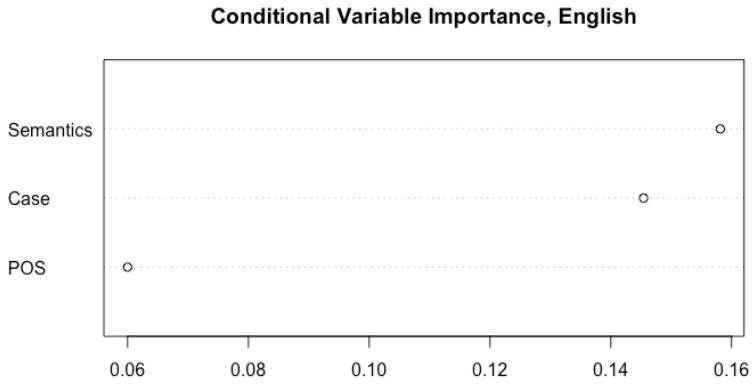

Table 9: Conditional importance of the variables, English.

The graph clearly shows the primacy of semantics as a contributor to subject-object determination which is in the opposition of what the conditional interference tree analysis showed. Case was the second-best predictor. POS has a non-zero value, but it is the weakest.

2.3.3.6. *Summary of results*

The analysis showed that semantics is the strongest predictor of the role in the sentence, when other variables are taken into account. If analysed separately, the case is the most important one, on its own. In other words, semantics is the most reliable grammatical cue to subject and object in English. The strong subject-nominative co-occurrence does not influence the distribution and relations of other values.

Also, what is crucial, it is the exclusion of the word order from the analysis. This feature is the ultimate and unambiguous grammatical cue to subject and object. Consequently, one should not interpret the findings of the analysis without taking into account this grammatical feature.

3. Tok Pisin

Since much was written about the emergence and evolution of Tok Pisin in the Section 1, this chapter will focus on grammar.

3.1. Basic Features Grammatical Features of Tok Pisin

As mentioned above, Tok Pisin has acquired features from the lexifier (English) as well as the substrates. The influence of the latter can be seen in grammatical aspects listed below.

Subject-referencing pronoun in the verb phrase.
Mi yet **i** no traim ol dispela marasin bilong ol tumbuna.
I have not myself tried these kinds of medicine of the ancestors.

As Verhaar (1995) comments it, "*i* is triggered by a remote subject pronoun, that is, a subject followed immediately by a constituent not the predicate. Hence the essential pick-up function of *i*, picking up a subject not preceding immediately" (p. 75).

transitive suffix on the verb
Em bai inap wok**im** haus.
3PER.SG FUT ABIL make.TRANS house.
He will be able to make a house.

adjectives functioning as stative verbs
In other words, in sentences without predicate, sometimes adjective assumes its function. "While *raunim* and *was* may function as genuine verbs, a modifier like *amamas* is different: it may mean '(to be) happy' (which is stative) or 'to enjoy oneself (something clearly indicating a process)" (Verhaar, 1995, p. 144).

preverbal causative marker
In Tok Pisin, we can observe "the use of periphrastic causatives using *mek-* and grammaticalised causatives using the transitive suffix -im" (Tryon, 1991, p. 511).

post-nominal possessive marker
"*Bilong* (followed by a noun) may be called possessive in some rather general sense. What precedes *bilong* is what is possessed (or the possessee); what follows *bilong*, the possessor. Thus, in *kantri bilong ol* 'their country *kantri* is the possessee, and *ol* is the possessor, which is marked by *bilong*" (Verhaar, 1995, p. 190).

third-person plural pronoun used as a plural marker
"*Ol* is a plural marker for nouns. It is not an article, *ol* is not necessarily, and in fact is often not, "the same" as English (plural) *the*. English *the* is typically definite, but *ol* need not be" (Verhaar, 1995, p. 346) However, it can also be used in sentence like this:

Tasol Moses i no i go waniaim **ol**.
'But Moses did not go with them.'

exclusiveness and dual number marker in the pronoun system
Yumi amamas.
We [incl.] are happy.

Mipela i amamas.
We [excl.] are happy.

What is important, Tok Pisin possesses also grammatical features characteristic of English. There is no declension of either number or case–it is applicable to both nouns and adjectives. Tok Pisin does not recognize grammatical genders and the tense is realized by separate markers, not morphological units:

(9) Dispela man bai i go long taun.
This man FUT PM go PREP town.
This man will go to town.

(10) Yutupela i bin lokim ka?
2PL PM PST lock car?
Did you (two) lock the car?

Notably, one crucial grammatical feature, in the context of this analysis, is the rigidity of the word order–SVO.

Many of the features mentioned above push Tok Pisin toward the "grammaticizing" side of the Semantic Typology spectrum. One can observe semantic diversity of grammatical relations in the fact that adjectives function also as stative verbs. Also, Tok Pisin is characterized by a rigid word order that holds also in subordinate clauses:

Dedi	**blo**	**mi**	**bin**	**lukim**	**wanpla**	**krokodail**
Father	POSS	1SG	PST	see.TRANS	one.MOD	crocodile

we	**em**	**i**	**traim**	**lo**	**atekim**	**mipla.**
REL	1SG	PM	try	to	attack.TRANS	1PL.EXCL

My father saw a crocodile which tried to attack us.[15]

3.2. The Analysis and Results

3.2.1. Data

3.2.1.1. *Corpus*

The ANNIS corpus was used for the analysis of Tok Pisin. It is an open-source browser that allows corpora analysis. It contains databases of various languages. Among others, a database of more than 500 entries of Tok Pisin (spoken and written).

3.2.1.2. *Dataset*

Again, 100 examples of sentences with transitive verbs were extracted. For every verb in the list of transitive verbs[16] in Tok Pisin a search in ANNIS was run. First 2-3 sentences that fulfilled requirements (no modal verb before the target verb, target verb functioning as a predicate, etc.) were included into the dataset. The examples were from *inter alia* sermons, news and speeches of politicians.

3.2.1.3. *The annotation procedure and variables*

Similarly to the English analysis, the sentences were annotated manually in the excel file. I do not have a good command of Tok Pisin, as it is the case with English, the annotations involved using online dictionaries and thesauruses. Various features were described: word order, subject's part of speech, how (and whether) subject is case-marked and subject's semantics. Analogous annotation was done for the object of each entry. For instance, for the sentence: "Mi lukim planti ol man na meri i save lusim lotu":

Mi	lukim	planti	ol	man	na	meri	i	save	lusim	lotu
1.PR.SG	See / look at	plenty	Plural marker	man	and	woman	subject-referencing pronoun	Used to / know	lose	faith

[15] https://apics-online.info/valuesets/22-7
[16] https://en.wiktionary.org/wiki/Category:Tok_Pisin_transitive_verbs

Verb	A_POS	A_Semantics	A_Case	P_POS	A_Semantics	A_Case
lukim	Pronoun	Human	No	Noun	Human	No

Figure 10: Example of an annotated Tok Pisin sentence.

3.2.2. Methods

The same procedure and methods were used as in the English grammar analysis (see: Section 2.3.2.)

3.2.3. Results

3.2.3.1. *Descriptive results*

The distribution, conditioned on the role, of values of POS, case, semantics and word order was calculated. Then, the chi-square test was run for all variables. Test's results are presented in the Table 11.

The analysis of POS showed that the most balanced POS was proper name class, 10 instances out of 19 were subjects. The highest residual value was in pronoun class: 1.967 (for the subject). Noun and proper name classes had -1.397 and 0.162, respectively. The interpretation is that noun objects, and pronoun and proper name subjects are more frequent than expected. Also, noun subjects and pronoun and proper name objects appeared less often.

	Noun	Pronoun	Proper Name
A	**51** [-1.3970014]	**39** [1.9668302]	**10** [0.1622214]
P	**73** [1.3970014]	**18 [-**1.9668302]	**9** [-0.1622214]

Table 11: Frequencies and residuals of POS in Tok Pisin.

Since there is no variation in case in Tok Pisin, that is, it has no case system, 100% of subjects and 100% of objects had no casing marking. The results are presented in the Table 12.

	Zero
A	**100** [0]
P	**100** [0]

Table 12: Frequencies and residuals of Case in Tok Pisin.

The frequencies of semantics values were not balanced (except animate class). For example, 40 out of 132 instances of human class were objects. The analysis of residual values shows that human subjects (3.2) and inanimate objects (4.648) are more frequent than expected. Also, animate subjects were more frequent, value: 0.408. Notably, Tok Pisin exhibits the same pattern as English–frequent human subjects and inanimate objects. Moreover, the most balanced class in the Animate one. For details, see Table 13:

	Animate	Human	Inanimate
A	**2** [0.4082483]	**92** [3.2003788]	**6** [-4.6484075]
P	**1** [-0.4082483]	**40** [-3.2003788]	**59** [4.6484075]

Table 13: Frequencies and residuals of Case in Tok Pisin.

As Table 14 shows, similarly to English, there is a rigid word order in Tok Pisin (SVO), so 100% of subjects were in the first position and 100% of objects were in the second position.

	First	Second
A	**100** [7.071068]	**0** [-7.071068]
P	**0** [-7.071068]	**100** [7.071068]

Table 14: Frequencies and residuals of Word Order in Tok Pisin.

3.2.3.2. *Fisher's test*

The Fisher's test was performed for all pairs of role and other variables. The results are as follows (see Table 15):

	Role-Case	Role-Semantics	Role-POS	Role-Word Order
p-value	1	< 0.0001	0.002629	< 0.0001

Table 15: Fisher's test for all variables

As the Table 15 clearly shows, all p-values for Fisher's test, except case, are smaller than 0.05. Hence, the association between the role and given variable is not due to chance. In other words, the association is significant. The p-value of Fisher's test of role-case pair is due to the absence of case forms.

3.2.3.3. *Mutual Information*

The next step was computing Mutual Information to see how much one can infer about the role from the information about the POS / word order / semantics. The results are presented in the Table 16:

Word Order	Semantics	POS
1	0.2578086	0.04291566

Table 16: Mutual Information for role in Tok Pisin.

3.2.3.4. *Conditional inference trees*

The second part of the study was multivariate analysis (conditional inference trees and random forests). Test's results are presented in Table 17:

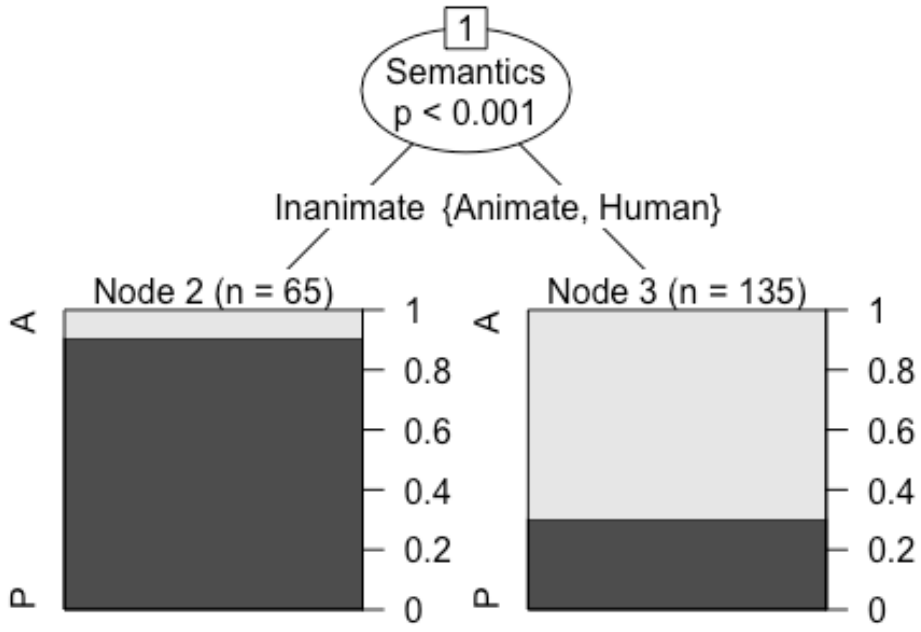

Table 17: Conditional inference tree, Tok Pisin

The results (presented on the Figure 17) show that the strongest and only predictor of subject and object in Tok Pisin is semantics, that is, whether the participant of the event described by the sentence is alive (human or animate) or not (inanimate). Around 90% of inanimate instances were objects. On the other hand, in the {Animate, Human} class only around 30% of instances were objects.

3.2.3.5. *Conditional random forests and variable importance*

The analysis also included calculation of conditional importance of the variables. Its results are presented in the Table 18:

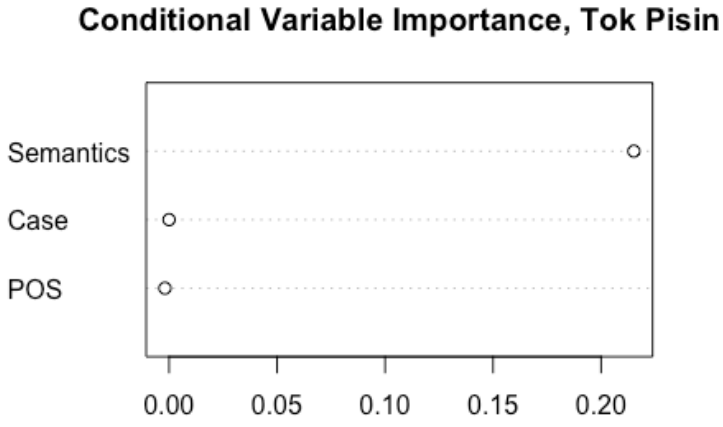

Table 18: Conditional importance of the variables, Tok Pisin.

It unambiguously shows that the only important grammatical feature in distinguishing subject and object is semantics (human, animate, inanimate). Both case and POS contribute nothing. That results in line with the conditional interference tree.

3.2.3.6. *Summary of results*

The analysis showed that semantics is the strongest, and the only one, predictor of the role in the sentence. In other words, it is the most reliable grammatical cue to subject and object in Tok Pisin. Notably, the primacy of semantics stands only when one excludes word order, which, as in English, is the ultimate grammatical cue.

4. Bislama

4.1. History and evolution of Bislama

Bislama is a sister of Tok Pisin threefold. They emerged in the same circumstances, they are linguistically similar (when it comes to both grammar and vocabulary), and they are spoken in the same region. Since the history of emergence of Melanesian Pidgin was described in the section above and as Troyn (1991) points out, "the concordance of the archival materials of the last century and the morpho-syntactic and lexical regionalisms of Vanuatu Bislama today suggests that the other Melanesian pidgins, Tok Pisin and Solomons Pidgin, indeed shared a lengthy period of common development with Bislama. The evidence further suggests that they separated only after a considerable degree of stability had been achieved, differentiation and individual stabilization being achieved after separation late in the nineteenth century" (p. 518). That, in turn, limits the "field of fire" of processes that would characteristic only of Bislama making the chapter on the history of emergence of Melanesian Pidgin sufficient for the purposes of this paper.

4.2. Basic Features Grammatical Features of Bislama

As argued above, all descendants of Melanesian Pidgin possess certain grammatical characteristics. Consequently, they can be observed in Bislama grammar as well. For instance:

subject-referencing pronoun in the verb phrase
In Bislama there two such pronouns: *i* (for singular subjects) *and oli* (for plural ones).

transitive suffix on the verb
In Bislama, it has three variations: *-em, -im* and *-um*. However, there a few exceptions, like: *save* (to know), *lukluk* (to look at) or *singaot* (to call) (Camden, 1996).

adjectives functioning as stative verbs
As Meyerhoff (2013) explains, "adjectives such as gud 'good', red 'red' and laki 'lucky, fortunate' may be used adjectively modifying nouns, or as predicates, e.g. pepa we yu wantem i red 'the paper you want is red'. Predicative adjectives may also be suffixed with -wan, e.g. redwan 'red'."

preverbal causative marker

For instance, “mekem i slip,” which means “to put to sleep.”[17]

post-nominal possessive marker
Mama blong woman ya nao.
mother POSS woman that DEIC
This is the mother of that woman.

third-person plural pronoun used as a plural marker
The word “olgeta” functions as the third person pronoun, an adverb “completely” and a pronoun “all”, which individually marks plurality:

Hem i katemdoan **olgeta wud** ya long jenso.
3SG PM cut.down.TR all tree this PREP chainsaw
He cut down all these trees using a chainsaw.

exclusiveness and dual number marker in the pronoun system
For example, there are seven variations of the first person: mi (singular), mitufala (exclusive) and yumitu (inclusive) for dual, mitrifala (excl) and yumitri (incl) for trial and mifala (excl) and yumi (incl) for plural.

Moreover, like Tok Pisin, Bislama has rigid SVO and no gender system. It marks plurality by separate marker (not affixes). However, as Meyerhoff (2013) notices, Bislama “marks plurality with the determiner *ol* before the head noun. A minority of speakers, mainly in urban centres, and perhaps only those with high English proficiency, may irregularly use a noun with plural suffix -s, with or without a Bislama determiner.” Regarding tense, aspect and mood, the rules in Bislama are analogous to those in Tok Pisin–both those features are realized by separate markers. They were derived from English words. For examples, see Table 19:

form	**lexifier etymon**	**meaning**
bae; bambae	*by and by*	irrealis
bin	*been*	anterior
finis	*finish*	completive
save	*savvy* (< Portuguese *saber*)	abilitive; habitual
stap	*stop*	imperfective; progressive; habitual
Ø		all, except progressive

Table 19: TAM markers in Bislama.

Moreover, it is common in Bislama to dislocate the subject. In other words, the noun phrase constituting the subject consists of both noun/proper name and the pronoun. The literal translation to English would go as follows: “My mum she made the dinner.”[18]

[17] From Siegel (2005).
[18] This kind of dislocation is also found in informal English, see: Prince (1997).

4.3. The Analysis and Results

4.3.1. Data

4.3.1.1. *Corpus*

The ANNIS corpus was used for the analysis of Bislama. It is an open-source browser that allows corpora analysis. It contains databases of various languages. Among others, a database of more than 4000 texts in Bislama (spoken and written).

4.3.1.2. *Dataset*

Again, 100 examples of sentences with transitive verbs were extracted. For every transitive verb mentioned in Bislama Handbook[19] a search in ANNIS was run. First 1-3 sentences (the exception: verb *givim*, 7 entires) that fulfilled requirements (no modal verb before the target verb, target verb functioning as a predicate, not object, etc.) were included into the dataset. The examples were from, inter alia, Bible, news and speeches of politicians.

4.3.1.3. *The annotation procedure and variables*

Similarly to the English analysis, the sentences were annotated manually in the excel file. Since the author does not have a good command of Tok Pisin, as it is the case with English, the annotations involved using online dictionaries and thesauruses. Various features were described: word order, subject's part of speech, how (and whether) subject is case-marked and subject's semantics. Analogous annotation was done for the object of each entry. For instance, for the sentence: "Ebram i sevem Lot":

Ebram	i	sevem	Lot
Abram	subject-referencing pronoun	Save	Lot

Verb	A_POS	A_Semantics	A_Case	P_POS	A_Semantics	A_Case
sevem	Proper name	Human	No	Proper name	Human	No

Table 20: Example of an annotated Bislama sentence.

4.3.2. Methods

The same procedure and methods were used as in the English grammar analysis (see: Section 2.3.2.)

[19] https://moet.gov.vu/docs/textbooks/Peace%20Corps%20Bislama%20Handbook_2014.pdf

4.3.3. Results

4.3.3.1. *Descriptive results*

The distribution, conditioned on the role, of values of POS, case, semantics and word order was calculated. Then, the chi-square test was run for all variables–its results are presented in the Table 21.

The analysis of POS showed the highest value in pronouns: 3.674 (for the subject). Noun and proper name classes had -2.648 and 1.512, respectively. The interpretation is that in noun objects, and pronoun and proper name subjects are more frequent than expected.

	Noun	**Pronoun**	**Proper Name**
A	**47** [-2.648489]	**42** [3.674235]	**11** [1.511858]
P	**91** [2.648489]	**6** [-3.674235]	**3** [-1.511858]

Table 21: Frequencies and residuals of POS in Bislama

Since there is no variation in case in Bislama, that is, it has no case system, 100% of subjects and 100% of objects had no casing marking (see: Table 22).

	Zero
A	**100** [0]
P	**100** [0]

Table 22: Frequencies and residuals of Case in Tok Pisin

Like in English, the frequencies of semantics values were not balanced. For example, 31 out of 125 instances of human class were objects. When it comes to absolute values, human and inanimate class had high residual absolute values: 3.984 and -5.119, respectively. It means that human subjects and inanimate objects are more frequent than expected. Also, animate subjects were less frequent, value: -0.707. Notably, Bislama exhibits the same pattern as English and Tok Pisin–frequent human subjects and inanimate objects. Moreover, the most balanced class in the Animate one. For details, see Table 23:

	Animate	**Human**	**Inanimate**
A	**1** [-0.7071068]	**94** [3.9844699]	**5** [-5.1190063]
P	**3** [0.7071068]	**31** [-3.9844699]	**66** [5.1190063]

Table 23: Frequencies and residuals of case in Bislama

As Table 24 clearly shows, similarly to English, there is a rigid word order in Bislama (SVO), so 100% of subjects were in the first position and 100% of objects were in the second position.

	First	**Second**
A	**100** [7.211103]	**0** [-7.211103]
P	**0** [-7.211103]	**100** [7.211103]

Table 24: Frequencies and residuals of Word Order in English

4.3.3.2. *Fisher's test*

The Fisher's test was calculated for all pairs of role and other variables. The results are as follows (see: Table 25):

	Role-Case	Role-Semantics	Role-POS	Role-Word Order
p-value	1	< 0.0001	< 0.0001	< 0.0001

Table 25: Fisher's test for all variables

As the Table 6 clearly shows, all p-values for Fisher's test, except case, are smaller than 0.05. Hence, the association between the role and given variable is not due to chance. In other words, the association is significant. The p-value of Fisher's test of role-case pair is due to case system having only one variation.

4.3.3.3. *Mutual Information*

The next step was computing Mutual Information to see how much one can infer about the role from the information about the POS, word order and semantics. The test's results are presented in the Table 26:

Word Order	Semantics	POS
1	0.3482545	0.1785659

Table 26: Mutual Information for role in Bislama.

4.3.3.4. *Conditional inference trees*

The second part of the study was multivariate analysis (conditional inference trees and random forests).

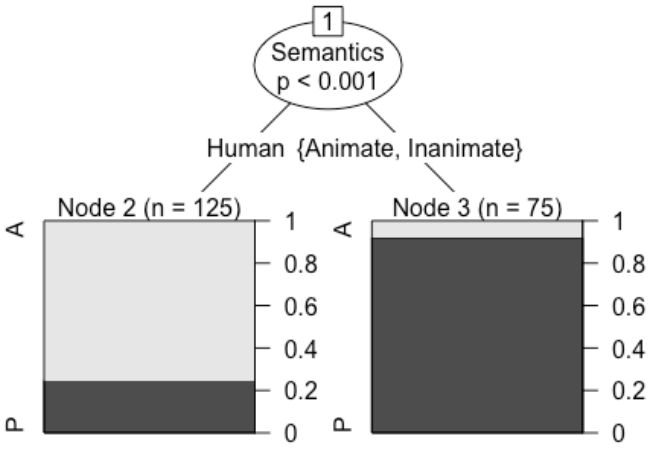

Table 27: Conditional inference tree, Bislama

As the graph in the Table 27 shows, in Bislama there is only one grammatical cue to subject and object that was statistically significant-being human. As graph shows, in {Animate, Inanimate} group 90% of instances were objects, whereas in human group around 80% of instances were subjects.

4.3.3.5. *Conditional random forests and variable importance*

The analysis also included calculation of conditional importance of the variables. Its results are presented in the Table 28:

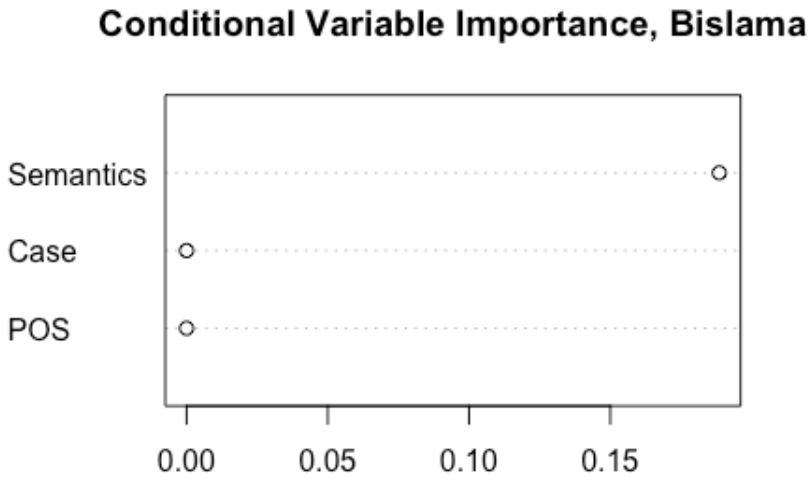

Table 28: Conditional importance of the variables, Bislama.

It unambiguously shows that the only important grammatical feature in distinguishing subject and object is semantics (human, animate, inanimate). Both Case and POS contribute nothing. Notably, only the analysis of Bislama yielded results that were consistent, both random forest and conditional variable importance showed one significant predictor–semantics.

4.3.3.6. *Summary of results*

The analysis showed that semantics is the strongest, and the only one, predictor of the role in the sentence. In other words, it is the most reliable grammatical cue to subject and object in Bislama. Notably, the primacy of semantics stands only when one excludes word order, which, as in English, is the ultimate grammatical cue.

5. Conclusions

5.1. Cues to subject and object in creoles and English

One similarity that is crucial in the discussion on the grammatical cues to subject and object is word order. In all three languages (English, Tok Pisin and Bislama), we can observe rigid SVO in the "who did what to whom" structures. It is unquestionably the most important and most reliable grammatical cue to subject and object in those languages. Moreover, all three languages do not have case system, that is, words are not case-marked in any form. The exception are

several English pronouns which kept accusative forms. This peculiarity significantly influenced results: the association between grammatical role and case is significant (Fisher's test).

The comparison of chi-square test results shows cross-linguistic tendencies. In all three languages noun objects and proper name subjects are more frequent than expected. Assuming a probabilistic approach towards interpretation of heard or read sentences, the addressee hearing a proper name will be more likely to assign it as a subject of that sentence. Moreover, in all three languages words describing human(s) (one person "a mother", a group of people "party" and institutions "court") are more often subject than objects. Also, words describing inanimate entities are more often objects.

There are some aspects that are shared by creoles and are different in English. For instance, pronouns in Tok Pisin and Bislama appear more often as objects of a sentence, whereas the analysis showed that in English the case is the opposite–more often as subject (the difference is quite small with a residual value: 0.25).

5.2. Relationships between these cues across languages

When it comes to case, in English it is a significant predicator of the grammatical role, as conditional importance of variables test showed. In other words, even if we take into account other factors (grammatical features), the case's contribution to determining subject and object is valid.

As explained above, there are cross-linguistic tendencies between the semantics and the grammatical role. They are mirrored in Mutual Information and Conditional Importance of Variables results. Except for case and word order, semantics is the strongest predictor of grammatical role in all three languages. However, semantics plays greater role in discriminating the subject in creoles than it does in English. In the latter, case as well as POS are also significant predictors.

Moreover, in both creole languages POS ceases to be significant predictor when we control for semantics. This can be explained by a classic chocolate-Noble Prize winners situation. There is a famous positive correlation between amount of chocolate eaten per capita and the number of Noble Prize winners per capita. However, if we control for the GDP for a given country, then the correlation disappears. Both amount of chocolate eaten, and number of Noble Prize winners are dependent of the richness of the country. Hence, there is no correlation between those two, rather a shared cause. Analogously, correlation between POS and grammatical role stands only if we don't control for semantics. That would suggest that semantics is a predictor of both POS and grammatical role creating a fictional correlation between them.

When it comes to differences between Tok Pisin and Bislama, even though semantics proved to be the only grammatical cue to subject and object, the demarcation line went differently. In Tok Pisin, the inanimate value predicted the object and {animate, human} values predicated the subject. In contrast, in Bislama only human value predicted the subject. However, it must be taken into account that the sample size of animate class ("swing" group) was quite small. Thus, making the results less certain.

5.3. Semantic tightness in high-contact languages

The comparison of Mutual Information results allows us to conclude on semantic tightness of given languages and the trade-offs. As the Table 29 shows, both semantics and POS are weaker predicators of grammatical roles in English than they are in creoles. It means that English is the least semantically tight language of those three.

One explanation is linked to the present status of creoles. Even though they are a result of language contact since their emergence they had time to become more complex and semantically tighter. Moreover, at present creoles are not influenced by language contact processes to a significant extent. Of course, both of those languages function as one of a few official languages. However, it is a language contact between languages and communities, in which most of the members are bilinguals (native speakers of both languages). Hence, they are able and tend to maintain the grammatical complexity and pass it on to the next generation. In English case, the situation is quite different. With the globalization and establishing English as lingua franca the percentage of its non-native speakers dramatically increased. The influence of language contact on English is not strong (as it was with French invaders imposing French), but worldwide. Thus, making the influence significant. Consequently, one could argue that population size and number of second language (L2) speakers (like in English) matters more for semantic looseness, rather than emerging in a contact situation (creoles' case).

Moreover, as described above (see: Sections 1.4 and 2.5) today's English is a direct descendent of Old English which had rich case system, whereas at no point in history of Tok Pisin or Bislama did they possess case system. Remnants of accusative case in English help determining subject and object, which is not the case in Tok Pisin or Bislama. Consequently, in those language semantics could compensate for the lack of case marking as a cue.

The cross-linguistic analysis shows a trade-off between semantic tightness and case-marking. English, with some form of case marking, is semantically looser than analyzed creoles, which do not have case system. Moreover, the analysis did not confirm a trade-off between word order and any other variable. Going further, there is no trade-off between any two variables. However, there is a cross-linguistic positive correlation between semantics and POS–the higher MI score for semantics, the higher the MI score for POS, and vice versa.

Language \ Feature	Case	Word Order	Semantics	POS
English	0.2669886	1	0.1169774	0.03238583
Tok Pisin	-	1	0.2578086	0.04291566
Bislama	-	1	**0.3482545**	**0.1785659**

Table 29: Mutual Information results, all three languages

5.4. General Discussion

The study partly confirmed Sinnemäki's (2010) and Levshina's (2021) results, supporting the hypothesis of a trade-off between case-marking and rigid word order. The rigidity of word order is associated with a lack of case marking in the creoles, or its relative poverty in English. In case of creoles, the correlation was as high as 1. Notably, there is no correlation at the level of languages, that is, it is not the case that English, which has some case marking, has more flexibility in word order. This aspect flies in the face of Levshina's (2021) results (although the trade-off she found was probabilistic, rather than categorical). One explanation for that is that the number of languages included (only three) and their properties could not have allowed for testing the hypothesis quantitatively and in a more fine-grained way. For instance, we could test a potential difference between word orders of languages that have two cases and more than ten cases. For instance, the analysis performed by Sinnemäki (2020) included languages that have from 0 up to 10 cases.

The question remains whether the results are in line with Bentz and Winter's (2013) conclusions. They argued that the larger the L2 proportion the poorer the case system. Here we see that the creoles, which develop as L2 languages in the situation of high

contact, have no cases, whereas English has distinct case forms of some personal pronouns. We can explain this difference by the role of language contact. On the other hand, what is important is the situation now. That is, what is the L2 speakers' proportion now, not in the times of the language's evolution? The statistics varies when it comes to the L1 and L2 speakers count. Despite the inconsistency, the numbers (see: Table 30) show us some tendencies.

Language	All speakers [in thousands]	L1 speakers [in thousands]	Percentage of L2 speakers	Source
Tok Pisin				
	3 000[20]	500	83,3333333	Ethnologue.com
	5 000	500	90	Ethnologue.com
	4 000	120	97	Omniglot.com
Bislama				
	210	10	95,2380952	Ethnologue
	206,6	6,2	96,9990319	Omniglot.com
English				
	1200000	350000	70,8333333	Omniglot.com
	1360000	360000	73,5294118	Ethnologue.com
	1121000	378000	66,280107	Ethnologue, 21st edition

Table 30: Speakers counts of Tok Pisin, Bislama and English

Clearly, the creoles have much higher L2 proportion than English. Combining those statistics with this paper's results, we can conclude that creoles with higher L2 speakers' proportion have poorer case system than English with richer case system and a lower L2 speakers proportion. That, in turn, is in line with Bentz and Winter's (2013) conclusions.

In all three languages semantics (animacy) was proved to be strong grammatical cue to subject and object. In creoles, it was also the only significant cue. Those findings are in line with the results from MacWhinney's et al. (1984) paper. It clearly showed that even though speakers of all three languages (English, German, and Italian) use primarily different grammatical cues to subject and object, the animacy had a significant effect and became crucial in difficult, linguistically ambiguous situations.

When it comes to the third hypothesis (semantic looseness as a result of language contact), the study did not fully confirm it. Results showed that English is a semantically looser language than Tok Pisin and Bislama. That however does not reject the hypothesis completely assuming the explanation given above (see: Section 5.2). Also, although Levshina (2021) found a positive correlation between semantic tightness and case marking, this analysis showed a negative one: English with case marking is semantically looser, whereas the creoles without case marking are semantically tighter. This was unexpected. Further analyses, which would control for register and text type in a systematic way, are necessary. It is also possible that the absolute number of L2 users, or the total number of users, who can introduce innovations, are relevant for making a language tighter or looser.

Surely, the presented results suggest a redefinition of our approach towards the creoles as they do not exhibit some features that could be expected assuming presented universals. For instance, Hawkins' (1984) theory (case system positively influences semantic tightness) does not hold when applied both to nouns and pronouns. English with more developed case system

[20] Ethnologue estimates the number of Tok Pisin between 3 and 5 million.

is semantically looser than Bislama or Tok Pisin, which have no case marking. In other words, either creoles do constitute a qualitatively different type of language, or some language universals are not valid globally.

5.5. Limitations and further research

The data used in the study was limited twofold. First, it only included two creole languages. Secondly, for each language there were only 100 examples. Further research could investigate more languages and include bigger samples. Also, there are cases when the object can come first in English in conversations. That makes controlling for text type and register a reasonable next step. We do not know yet how semantic tightness, for example, varies across different registers and modalities.

As stated above, the study did not reject the third hypothesis (semantic looseness as a result of language contact) completely. We need more data to confirm or refute the hypothesis, performing a comparison between creoles and English and more isolated languages, which are subject to less intense language contact and have fewer L2 speakers, like Polish.

Since creoles have much poorer vocabulary, and consequently the usage of metaphors is frequent, for instance, hair in Tok Pisin in literal translation is the grass of head, future studies could focus on encoding both the source and target domain for metaphoric expressions.

Bibliography

Bammesberger, A. (2005). The Place of English in Germanic and Indo-European. In R. Hogg (ed.), *The Cambridge History of the English Language: Volume I The Beginnings to 1066* (pp. 26-66). Cambridge: Cambridge University Press.

Bates, E., & Goodman, J. (1999). On the Emergence of Grammar From the Lexicon. In B. MacWhinney, *The Emergence of Language* (pp. 29-80). London: Psychology Press.

Baugh, A., & Cable, T. (2002). *History of English Language.* London: Routledge.

Bentz, C., & Winter, B. (2013). Languages with More Second Language Learners Tend to Lose Nominal Case. *Language Dynamics and Change, 3*, 1-25.

Camden, B. (1996). The Transitive Verbs Using Long in Bislama. *Oceanic Studies: Proceedings of the First International Conference on Oceanic Linguistics* (pp. 319-351). Canberra: The Australian National University.

Denison, D. (1999). Syntax. In S. Romaine, *The Cambridge History of the English Language* (Vol. 4, pp. 92-329). Cambridge: Cambridge University Press.

Fenk-Oczlon, G., & Fenk, A. (2008). Complexity trade-offs between the subsystems of language. In M. Miestamo, K. Sinnemaki, & F. Karlsson, *Language Complexity. Typology, contact, change* (pp. 43-66). Amsterdam: John Benjamins.

Granena, G., & Long, M. H. (2012). Age of onset, length of residence, language aptitude, and ultimate L2 attainment in three linguistic domains. *Second Language Research, 29*(3), 311-343.

Greenberg, J. (1963). Some Universals of Grammar with Particular Reference to the Order of Meaningful Elements. In J. Greenberg, *Universals of Language* (pp. 73-113). London: MIT Press.

Greenberg, J. (1966). Some Universals of Grammar with Particular Reference to the Order of Meaningful Elements. In J. Greenberg, *Universals of Grammar* (pp. 73-113). Cambridge: MIT Press.

Hahn, M., Jurafsky, D., & Futrell, R. (2020). Universals of Word Order Reflect Optimization of Grammars for Efficient Communication. *Proceedings of the National Academy of Sciences, 117*(5), 2347-2353.

Haspelmath, M. (2016). Universals of Causative and Anticausative Verb Formation and the Spontaneity Scale. *Lingua Posnaniensis, 58*(2), 33-63.

Hawkins, J. (1986). *A Comparative Typology of English and German: Unifying Contrasts.* Austin: University of Texas Press.

Hawkins, J. (2018). Word-external Properties in a Typology of Modern English: a Comparison with German. *English Language and Linguistics, 23*(3), 701-727.

Holler, J., & Levinson, S. C. (2019). Multimodal Language Processing in Human Communication. *Trends in Cognitive Science, 23*, 639-652.

Holler, J., Kendrick, K. H., & Levinson, S. C. (2018). Processing language in faceto- face conversation: questions with gestures get faster responses. *Psychonom. Bull. Rev., 25*, 1900-1908.

Hoop, H. d., & Malchukov, A. (2008). Case-Marking Strategies. *Linguistic Inquiry, 39*(4), 565-587.

Levshina, N. (2015). *How to Do Linguistics with R.* Amsterdam: John Benjamins.

Levshina, N. (2019). Token-based typology and word order entropy: A study based on Universal Dependencies. *Linguistic Typology, 23*(3), 533-572.

Levshina, N. (2020). How Tight Is Your Language? A Semantic Typology Based on Mutual Information. *Proceedings of the 19th International Workshop on Treebanks and Linguistic Theories*, (pp. 70-78).

Levshina, N. (2021). Cross-Linguistic Trade-Offs and Causal Relationships Between Cues to Grammatical Subject and Object, and the Problem of Efficiency-Related Explanations. *Frontiers in Psychcology, 12*(648200), 1-20.

MacWhinney, B., Bates, E., & Kliegl, R. (1984). Cue Validity and Sentence Interpretation in English, German, and Italian. *Journal of Verbal Learning and Verbal Behaviour, 23*, pp. 127-150.

Maitz, P., & Németh, A. (2014). Language Contact and Morphosyntactic Complexity: Evidence from German. *Journal of Germanic Linguistics, 26*, 1-29.

Meyerhoff, M. (2013). *The Atlas of Pidgin and Creole Language Structures Online*. Retrieved from Bislama: https://apics-online.info/surveys/23

Mueller-Gotama, F. (1994). *Grammatical Relations. A Cross-Linguistic Perspectiveon their Syntax and Semantics.* Berlin: Mouton de Gruyter.

Sapir, E. (1921). *Introduction of the Study of Speech.* New York: Harcourt.

Siegel, J. (2005). Possession in South Pacific Contact Languages. *Monash University Linguistics Papers, 4*(1).

Siegel, J. (2008). *The Emergence of Pidgin and Creole Languages.* Oxford: Oxford University Press.

Sinnemäki, K. (2010). Word Order in Zero-marking Languages. *Studies in Language, 34*(4), 869-912.

Sinnemäki, K. (2020). Linguistic System and Sociolinguistic Environment as Competing Factors in Linguistic Variation: A Typological Approach. *Journal of Historical Sociolinguistics, 6*(2), 1-39.

Sinnemäki, K., & Di Garbo, F. (2018, August). Language Structures May Adapt to the Sociolinguistic Environment, but It Matters What and How You Count: A Typological Study of Verbal and Nominal Complexity. *Frontiers in Psychology, 9*.

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional Variable Importance for Random Forests. *BMC Bioinformatics, 9*(307), 1-11.

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

Tryon, D. T. (1991). Wanem Bislama? In R. Blust, *Currents in Pacific linguistics: Papers on Austronesian linguistics and ethnolinguistics in honour of George W. Grace* (pp. 509-519). Canberra: The Australian National University.

Verhaar, J. W. (1995). *Toward a Reference Grammar of Tok Pisin: An Experiment in Corpus Linguistics* (Vol. 26). Honolulu: University of Hawaii Press.

Wilson, D., & Sperber, D. (2004). Relevance Theory. In L. Horn (ed.), *The Handbook of Pragmatics* (pp. 606-632). Hoboken: Blackwell.

Woolford, E. B. (1979). *Aspects of Tok Pisin.* Canberra: Australian National University.