https://doi.org/10.22409/gragoata.v27i58.50708

Article

A probabilistic approach to the distribution of subject and anacoluthon NPs in Topics in spontaneous speech

Luis Filipe Lima e Silva[a]

Heliana Mello[a]

ABSTRACT

The definition of Topic as well as that of information structure in the literature is very broad (cf. BARBOSA, 2005; MELLO; SILVA, 2015). Here we assume the definition as proposed by the Language into Act Theory (CRESTI, 2000), which says that Topic is the textual unit that is performed by an intonational profile of the prefix type ('t HART et al. *1990), and that has the function of constituting the domain over which the illocutionary force applies. An NP in Topic either can be the subject of the following verb in Comment or an anacoluthon. Anacolutha NPs are phrases that bear no syntactic relations with the predication in Comment. In this paper, we show how NPs are distributed probabilistically between these two conditions when they are performed as Topics in spontaneous speech. For this purpose, we collected data from available spontaneous speech*

Recebido em: 02/07/2021
Aceito em: 10/10/2021

[a] Universidade Federal de Minas Gerais, Faculdade de Letras, Belo Horizonte, MG, Brasil
E-mail: luisf.1397@gmail.com; hmello@ufmg.br

How to cite:
Lima e Silva, L.F.; MELLO, H. A probabilistic approach to the distribution of subject and anacoluthon nps in topics in spontaneous speech. *Gragoatá*, Niterói, v.27, n.58, p. 86-117, mai.-ago 2022. <https://doi.org/10.22409/gragoata.v27i58.50708>

corpora informationally labeled – including the Topic unit as defined above – from three languages: European Spanish (NICOLÁS MARTÍNEZ; LOMBÁN SOMACARRERA, 2018), American English (CAVALCANTE; RAMOS, 2016), and Brazilian Portuguese (PANUNZI; GREGORI; MITTMANN, 2014). The statistical method used to calculate the probability was a mixed-effects logistic regression with crossed random effects conducted with the aid of R (R CORE TEAM, 2018). Three variables were chosen: accessibility of referent, animacy, and definiteness. The model showed that there are about five times more chances for an NP performed in Topic to be the subject of the verb in Comment if it is animate, definite, and given.

KEYWORDS: *Topic. Subject. NP. Spoken syntax. Probabilistic grammar.*

Introduction

Although syntax is a field comprising a range of theoretical frameworks covering both formalist and functionalist traditions, spoken syntax, according to Crystal (1980), has been misunderstood by most well-meaning linguists. This is due to the fact that there is a strong writing bias underlying the study of syntax. Firstly, in order to effectively study spoken language, syntax included, it is necessary to take into account the acoustic signal of recordings, and not just their transcriptions. Additionally, we need to consider that the structures performed in speech can be very different from those found in writing, and in some cases we don't find a realistic correlate of the structures of the two diamesia. The difficulty to study spoken syntax comes in part from the aforementioned bias as well as from theoretical assumptions regarding data handling, since different researchers carry particular beliefs toward the way they are distributed and organized in the continuum of speech. Consequently, one and the same phenomenon could be treated within diametrically opposite viewpoints leading to very different conclusions (cf. SILVA, 2020).

In this study, we deal with the data from a probabilistic standpoint, taking into account theoretical input from different traditions to explain the results found. Thus, we are not concerned with a particular theory, because we believe that a theory of spoken syntax should be built taking into account sound methodological aspects, and an important step toward it involves, firstly, treating the data probabilistically. Our study concerns the realization of NPs in Topic (TOP) and the chances for such NPs to be either the subject of the verb in the following Comment (COM) or constitute independent elements in the utterance, thus being anacolutha. The underlying information structure theory that informed this analysis is the Language into Act Theory (L-AcT) applicable to spoken data, as proposed by Cresti (2000).

Within an utterance, TOP here is understood to be the linguistic element that is performed through a specific intonational profile called prefix in the prosodic typology proposed by 't Hart et al. (1990) and whose function is to constitute the domain over which the illocutionary force (CRESTI, 2000) or, in other approaches, the predication (CHAFE, 1976) applies. The illocutionary force is carried through the COM, the major information unit distributionally following the TOP, which guarantees the TOP-COM pragmatic-informational articulation. The problem tackled in this paper is illustrated through the following two examples of spontaneous spoken Spanish. Example (1) shows a subject NP of the verb form *explicó* 'explained', while (2) shows an anacoluthon NP, an independent item, which holds no syntactic relationship with the sentence in COM. Both NPs are hosted in TOP.[1]

(1) epubcv02[2]
RMA: porque *Dolores* /^{TOP} nos explicó también en el Consejo que / es que a estas sustancias no vale con ponerles límites //^{COM}
'because Dolores also explained to us in the council that it is the case that these substances are not worth setting limits to'

[1] Approximate translations are provided for examples and glosses are inserted only when relevant to the discussion.

[2] Examples used in this paper were taken from informationally annotated minicorpora and are identified through their file id codes and portray original annotations featured in them. The three capital letters indicate the participant. Single slashes indicate non-terminal prosodic breaks, while double slashes indicate terminal prosodic breaks. The symbol [/n] refers to linguistic material that has been retracted. Here n indicates the number of words that were retracted. Angular brackets indicate overlapping speech. The & symbol refers to interrupted words.

(2) efamcv01c[3]

PIZ: *las e' matemáticas de COU* /[TOP] *que son todas facilísimas y de entender* /[APT] no conseguimos jamás enterarnos de nada //[COM]

'the mathematics of COU, that are all very easy and to understand, we never manage to find out anything'

[3] In this example, APT stands for Appendix of Topic, which means a separated prosodic unit that gives a delayed integration of the content of Topic, in this case it is a relative clause (for more details about this unit, cf. CRESTI, 2000).

In order to carry our analysis and to verify the chances for the NP in TOP to be either a subject or an anacoluthon, selected variables integrated a statistical method we used – a mixed-effects model of logistic regression with crossed random effects. The data analyzed were extracted from informationally annotated spontaneous speech minicorpora of American English, European Spanish, and Brazilian Portuguese.

This paper is organized as follows. Firstly, we introduce the subcorpora that served as data sources for the research reported. Next, we explain each variable selected for the analysis. After that, we show the data analysis, as well as the results obtained by the logistic regression method. Lastly, we present some final remarks.

The corpora

In order to carry out the analysis hinted at in the previous section, the following three subcorpora were used: European Spanish (NICOLÁS MARTÍNEZ; LOMBÁN SOMACARRERA, 2018), American English (a sample from DU BOIS et al. 2000-2005 as published by CAVALCANTE; RAMOS, 2016) and Brazilian Portuguese (PANUNZI; GREGORI; MITTMANN, 2014). Such subcorpora are informationally annotated samples of their parent corpora. The informational annotation follows that adopted by the C-ORAL corpora (CRESTI; MONEGLIA, 2005) through which prosodically identified tone units pair with informational units. Each minicorpus portrays three interaction types: conversations (more than two participants), dialogues, and monologues. The extraction of data from European Spanish and Brazilian Portuguese was done through queries in the DB-IPIC platform (PANUNZI; GREGORI, 2011). The American English data was manually extracted from the original subcorpus texts.

The variables

In this section, the variables that made up the set of independent variables in the study will be explained. Justification for the variables is provided in each of the sections in which they are presented, respectively.

Accessibility of referent

The notion of accessibility of a referent has a lot of variation in the literature, and there is no real consensus on how to define and analyze it (cf. PRINCE, 1981; GIVÓN, 1983; CHAFE, 1987; ARIEL, 1990). The terminology of the two main categories that make up this notion reflects this fact: there are, on the one hand, terms such as given, presupposed, familiar, accessible, inferable information, and on the other hand, new, not presupposed, not accessible, etc. Despite this, numerous studies show how this notion interferes in the grammatical expression of some linguistic components (cf. VIRTANEN, 1992; LAMBRECHT, 1994; MEUERMAN-SOLIN et al. 2012).

Some approaches, such as that of Prince (1981), consider a scalar series of the manifestation of referent accessibility, dividing it into the sources that the listener has to apprehend: his own mind and the discursive context. Thus, content can be new from the point of view of the discourse, but given from the point of view of the listener, for example. Despite being a more analytically refined approach, in the processing dimension, it seems to be reduced, as pointed out by Givón (1987). According to the author,

> while the psycho-cognitive dimensions which underlie semantics and pragmatics may indeed be scalar and non-discrete, the imperatives of processing within finite time require discretization and reduction along any functional-cognitive continuum (GIVÓN, 1987, p. 185).

In addition, in empirically-based research such as ours, the limitation of the number of data often causes subcategories to be grouped within a common macro-category (cf. HUNDT et al. 2018, for example). The category of inferable that Prince (1981) lists in his analysis, although somewhat intuitive, is often difficult to apply to empirical data. With created data, it is quite straightforward to conclude that the accessibility of the noun

motor is inferable in a discursive context in which one is talking about cars, for example. However, attribution of inferable status is not always so intuitive, especially in spontaneous speech corpora data, in which the recorded situations have a very high level of actional dynamism – something that is certainly reflected in speech. This makes defining what would be inferable for the interlocutor in a given context an even more subjective task.

In order to distance ourselves from subjective judgments about the accessibility status of a referent and considering the nature of the data analyzed, as well as the limitations that this type of data imposes, we sought to classify NPs as either given or new only. The NP was classified as given if it had already been formally, previously mentioned in the speech file. If the NP had not been previously mentioned anywhere in the text, it would be considered new. Below are two examples to illustrate how our classificatory decisions were taken.[4] In (3) we consider that in the last utterance of the excerpt *Durepox* is given because it had been previously introduced in the discourse; however, in (4) *gente velha* 'old folks' is treated as a new referent:

[4] Note that the possible objections about our bipartite classification would be limited to the analysis of referents considered as new information by the adopted criterion, since referents considered given would receive this status in any other type of analytical criterion choice.

(3) bfamcv01 (friends talk about soccer teams): given referent
LUI: <com certeza es nũ vão participar / uai> //
LEO: <eles são piores do que *o*> *Durepox* //
EVN: é / pois <é> //
LUI: <agora> manda uma barrinha <minha> //
EVN: <porque *o Durepox*> /[TOP] pelo menos jogava bola //[COM]
LUI: for sure they won't participate //
LEO: they are worse than *Durepox* //
EVN: yeah / that's right //
LUI: now give me a health bar one of mine //
EVN: because *Durepox* /[TOP] at least played some ball //[COM]

(4) bfamcv02 (Family members discuss wedding arrangements): new referent
TER: é que ea ganhou tudo / né //
TER: ganhou tudo / dos lado do Anderson //
RUT: oh / <que maravilha> //
JAE: <ganhou não> //

RUT: <chama só o lado do> Anderson / pa ser <padrim da Dani> //
TER: <o'> //
TER: <escuta> //
JAE: <ganhou / não> //
TER: <não> //
JAE: vai ganhar / <né> //
TER: <vai ganhar / mas> +
JAE: <ea nũ tem nada> na mão //
TER: ô Jael //
TER: mas / *gente velha* /[TOP] já prometeu o [/1] os presente /[TOP] <já / pode> garantir que ganhou //[COM]

TER: she got everything as gifts / right //
TER: she got everything / from Anderson's side of the Family //
RUT: oh / that's great //
JAE: she did not //
RUT: just invite Anderson's Family / to be bridesmaids and groomsmen //
TER: look //
TER: listen //
JAE: she did no [get everything] //
TER: no //
JAE: she will get / right //
TER: she will / but //
JAE: they haven't got anything / on their hands [yet] //
TER: hey Jael //
TER: but / *old folks* /[TOP] if they have promised the gifts /[TOP] it's guaranteed //[COM]

According to the given-new strategy (CLARK; CLARK; HAVILAND, 1974; CLARK; CLARK, 1977), which postulates that there is a tendency in the discourse for the given information to precede the new information, the subject category would tend to constitute given, shared information by the interlocutors (cf. GIVÓN, 1979; ITAGAKI; PRIDEAUX, 1985). In this sense, it is expected, as a hypothesis, that anacoluthon NPs are more likely to occur as new referents in the discourse[5]. As subject and anacoluthonconstructs from

[5] An interesting point raised by a reviewer must be clarified in this part of the text. We are assuming as a hypothesis that for the anacoluthon NPs to occur as new referents in the discourse (as well as indefinite and inanimate referents, as will be mentioned in the following sections) would be a hypothesis. The literature shows that there is a tendency for the subject NP to be given, definite and animate. An NP that is a subject displays a syntactic integration towards the predicate that is in Comment, considering the data we are analyzing. An NP that is an anacoluthon doesn't display this syntactic integration. So it is somehow natural to assume that the opposite features of the subject NP could be associated probabilistically to the performance of the anacoluthon NP in the discourse. Languages should exhibit a way to differentiate these two categories uttered in the same context, i.e. in the Topic information unit. The anacoluthon relation in TOP is exclusevely informational in relation to the predicate in COM, therefore it requires activation, which is a cognitive principle related to the "(...) amount of attention a concept receives" (DEANE, 1992, p. 34), differently from the subject, which is a syntactic relation strongly related to the verb argument stracuture properties. It is also important to say that we are not defining anacoluthon by these features. We believe that a possible characterization of anacolutha could have some of these features as defining properties associated

to probabilistically tendencies (as will be clear in the results of the data we analyzed), but at this moment we are taking these characterizating features just as a hypothesis to be tested.

given/new information, the analysis of this variable is more exploratory, although it is not at all incompatible to test it with the aforementioned hypothesis.

Definiteness

Definiteness involves not only semantic notions but also has strong pragmatic correlates, which make it a very complex category. Even a morphological mark that indicates a definite phrase will not be enough to classify it as such, considering that its status is subject to change depending on the linguistic-pragmatic context in which it is inserted (KRÁMSKY, 1972; LYONS, 1999). There are several explanatory hypotheses that try to pinpoint what definiteness is by employing other concepts such as familiarity, identifiability, uniqueness, and inclusiveness.

As the name implies, familiarity indicates that the NP is familiar to the interlocutors, whether situationally, anaphorically, associatively, or through general knowledge. Lyons' example (1999, p. 3) below illustrates these notions:

(5) Put these clean towels in *the bathroom* please.
(At the moment of speech, the interlocutors share the same space, so the NP is familiar to both)

Identifiability, roughly speaking, is related to the possibility of retrieval of the referent by the interlocutor via the inference brought about by a definite article. The listener, in this case, is able to associate the referent to an entity in the world that he can see, hear or at least infer its existence. In the words of Lyons "(...) while on the familiarity account *the* tells the hearer that he knows which, on the identifiability account it tells him that he knows or can work out which" (LYONS, 1999, p. 6). The author presents a situation in which Ann is trying to hang a painting on the wall and, without looking back, says to Joe (LYONS, 1999, p. 6):

(6) Pass me *the hammer*, will you?

The author says that Joe looks around and sees a hammer in a chair. It indicates that although the referent is not familiar,

it can be identified not only by the presence of the object but also by the verb *pass*, which already presupposes an action that is within the reach of the interlocutor to perform.
Uniqueness, in turn, refers to the fact that the definite article signals that, in a given context, there is only one entity that fills the description that is being used. Thus, uniqueness is not always absolute, it is subject to a given context. Lyons' example (1999, p. 7) illustrates this notion:

(7) I've just been to a wedding. *The bride* wore blue.

In the situation described, the speaker informs that he was present at a wedding and that the bride wore blue. World knowledge tells us that there is usually only one bride at a wedding, so the NP is definite by indicating a unique referent in that given situation.

The notion of inclusiveness refers to the possibility that the reference indicates the totality of objects in a context that satisfies the description (HAWKINS, 1978). The NP is understood, then, as a set, and not as an individual. Consider the example from Lyons (1999, p. 10) below:

(8) [Nurse about to enter operating theatre]
I wonder who *the anaesthetists* are.

The reference of the NP in the example above indicates the anesthesiologists as a set, therefore all those who can participate in the operation.

The discussion of these four notions points to the difficulty of identifying a definite NP. Some probabilistic studies prefer to amalgamate the categories of definite/given and indefinite/new, assuming that if the speaker is able to identify the referent by means of a definite NP, then that referent is shared, therefore it would constitute given information (cf. VOGELS; VAN BERGEN, 2017, for example). If an NP is indefinite, it is considered that it is not accessible, thus constituting new information in the discourse. To individualize whether the NP is definite or indefinite, these studies generally consider the lexical factor (presence or absence of an article), and relate the referent's status to a definiteness scale: personal pronoun > proper name > definite NP > specific indefinite NP > Non-

specific NP (cf. GIVÓN, 1976; CROFT, 1988; GUNDEL et al. 1993; AISSEN, 2003). The leftmost elements would mark topical information – with highly accessible referents – in a continuum that would develop until the last two elements, which would indicate non-topical information – with inaccessible referents. These studies emphasize that theirs is a theoretical choice that carries its particularities. In this paper, we tried to separate the notions of definiteness and accessibility of the referent, not only due to the fact that another criterion was used to analyze the second notion, but also due to the complexity that each category exhibits. It is not known to what extent the pairing of the two variables would prove reliable for the study of spontaneous speech data in different interactive situations. Consider the excerpt below:

(9) bfammn06
JOR: se o brasileiro nũ lê os manuais / hhh no mercado de reposição / &auto [/1] de autopeça / eles acham que abrir *uma empresa* é comprar um produto por um real / na base cem / e vender por dois / acha que tá ganhando o &do [/2] o dobro //
JOR: na verdade nũ é assim que isso funciona //
JOR: *uma empresa* /[TOP] tem a sua despesa administrativa tributária fiscal / é lucro bruto pa poder projetar o lucro líquido //[COM]

JOR: if Brazilians don't even read manuals / hhh in the reposition market / &auto [/1] auto parts / they think that opening *a business* is to buy a product for one real / say buying one hundred / and selling it for two / they think they are profiting the &do [/2] double //
JOR: in reality it doesn't work like that //
JOR: *a business* /[TOP] has its administrative tributary fiscal expenses / it is gross profit to project net profit //[COM]

It is possible to observe through example (9) that the NP *uma empresa* 'a business' is indefinite even though it clearly has a given referent status. The same participant had already introduced the NP *uma empresa* 'a business', as can be seen through the inspection of the spoken excerpt. Thus, equating the category of definiteness with the accessibility of the

referent is not always a reliable strategy, although it would be more practical to do so from the point of view of research that deals with a large set of data. Either way, the analysis of definiteness is inherently more interpretive. Thus, the pragmatic-discursive context must be taken into account. For the research reported here, NPs that exhibited the notions of uniqueness, identifiability, familiarity, or inclusiveness in the referred contexts were considered definite, on the other hand, generic or non-specific NPs were considered indefinite. The NP in (9) above presents the notion of non-specificity, that is, the speaker does not refer to a specific business.

In the example below, two friends are grocery shopping. At a given moment, REN asks FLA if they needed to buy disinfectant. The NP exhibits a generic reading, so it was included in the category of indefinite.

(10) bfamdl01[6]
REN: *desinfetante* /[TOP] a gente precisa //[COM]
FLA: precisa //
FLA: desinfetante / não //

REN: *disinfectant* /[TOP] do we need //[COM]
FLA: we do //
FLA: disinfectant / no //

[6] An important point made by a reviewer about this example needs clarification. We are assuming that the analyzed structures are the actual ones really uttered in the discourse flow. We don't assume any underlying structures. Therefore the NP "desinfetante" in this example isn't analyzed as a result from a derivation involving phrase movement with subsequent preposition deletion (cf. "A gente precisa de desinfetante"). We are, therefore, following on the lines of the analysis proposed by Pontes (1987) for these cases.

The literature points out that there is a tendency for subject NPs to be definite (KEENAN, 1976). In some languages, such as Samoan, it is even a necessary condition that all NPs in subject position are definite (HYMAN, 1984). Therefore, it would be a hypothesis to be tested probabilistically that anacoluthon NPs are more likely to be indefinite.

Animacy

Animacy is an inherent property of nouns and would encompass a hierarchical scale represented as human > animal > inanimate (COMRIE, 1989). In some languages, such as Asmat, Igbo, Kobon, Marind Chukchi, Hindi, Finnish, Yidiny, Russian, Even, and Nepali, this category is related to others, such as number and case, for example (CORBETT, 2000; MALCHUKOV, 2008). It is worth mentioning that this scale

may be subjected to some variation according to the language under analysis. According to Hall (1990), in Kashaya, the *–yac* 'Agentive' clitic would indicate a subcategory called by the author as superanimate. The adapted scale that the author elaborates for Kashaya is: superanimate > human > animal > inanimate. The way in which a language conceptualizes a referent would then focus on the level of animacy culturally dictated. This is clearly evidenced in Kashaya.

A more complex gradience scale is proposed by Yamamoto (1999) and can be seen in Figure 1. Human beings are the center of the scale, thus representing a prototypical exemplar. Additionally, there is a gradient in relation to the human category itself, with the speaker placing himself as more empathic/animate than the listener or third parties on an empathy scale that develops into speaker > listener > human > animal > physical object (SILVERSTEIN, 1976; LANGACKER, 1991). This is due to the fact that an interaction takes place between interlocutors who are present in a given situation. Therefore, the direct reference, formalized with the use of first and second-person pronouns, is related to the common goals between speaker and listener, thus encoding the perception of intentionality and sensitivity – features that presuppose a high level of animacy.

The impact that animacy has on grammatical phenomena is quite broad, which makes this variable widely explored in probabilistic studies (VOGELS; VAN BERGEN, 2017; SZMRECSANYI et al. 2017; HUNDT et al. 2018). Such studies report that subject NPs are more likely to be animate. Therefore, it is hypothesized that anacoluthon NPs in TOP would be more likely to be inanimate.

Evidently, for our purposes, just as it is necessary to taper the number of subcategories for accessibility of the referent and definiteness, it is also necessary to do the same with animacy. Subcategories other than animate or inanimate were not considered. Therefore, both animals and humans were considered animate. No level of animacy was distinguished for gradience, so first and second-person pronouns were allocated within the same category of animate. Abstract entities or concepts (maths, this travel, etc.) and lifeless entities (restroom, window, etc.) were classified as inanimate.

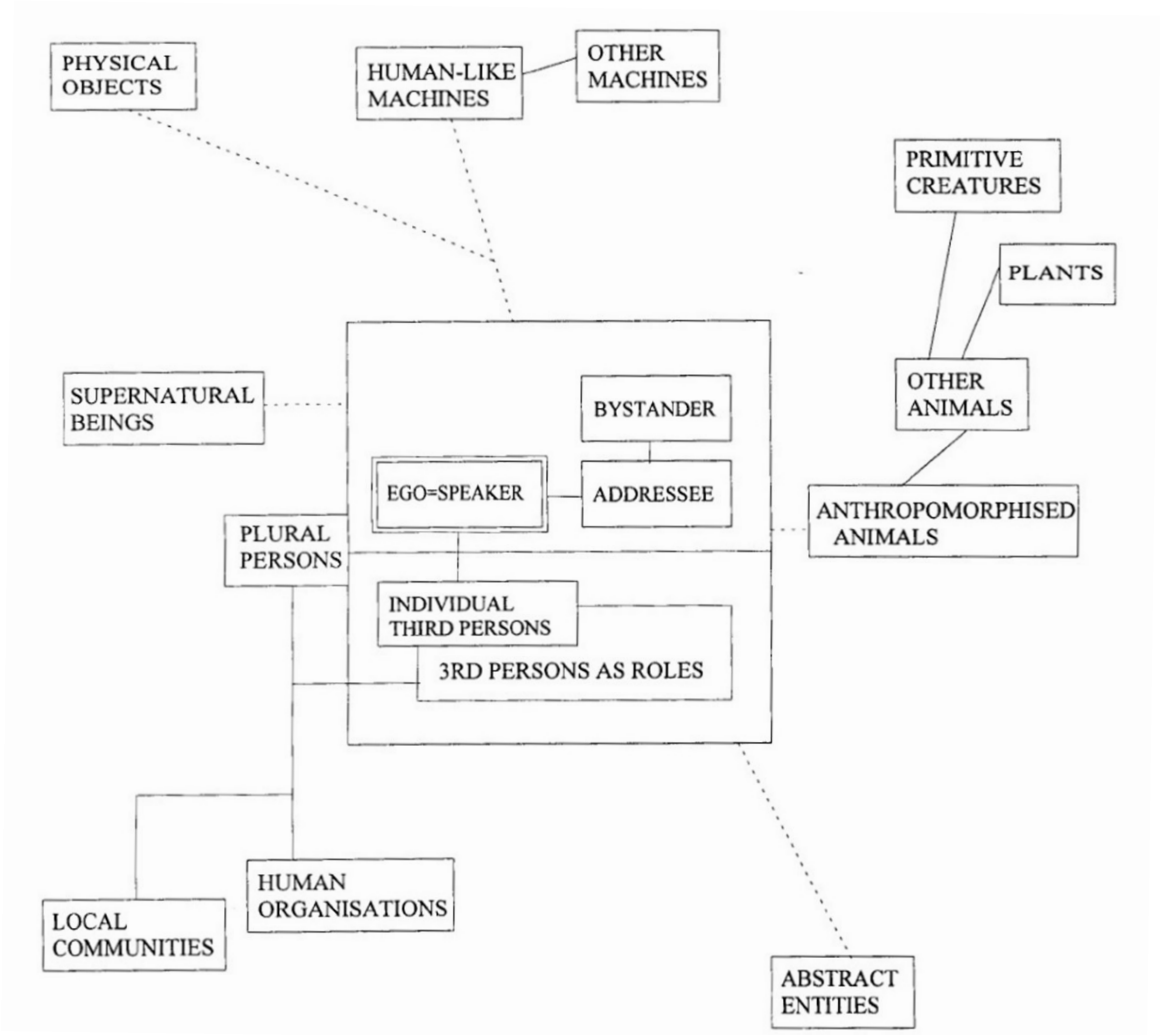


Figure 1 – Animacy: Radial gradience with human subcategorization
Source: Yamamoto (1999)

Results and data analysis

In this section, the results of the research carried out in three different languages (American English, Brazilian Portuguese, and European Spanish) are presented. Some descriptive aspects will be considered before proceeding to the regression statistical analysis.

Descriptive statistics

The results shown in this subsection relate to descriptive factors only. Any trend shown must be verified with an

augmentation in the dataset, since only samples from the selected corpora were analyzed, thus, the study requires future validation through the application of statistical tests. The parameters shown refer to the realization of NPs in TOP, that is, NPs that function as subjects or NPs that do not hold a syntactic relationship with content that follows them in the subsequent information COM unit.

Subject NP vs. anacoluthon NP

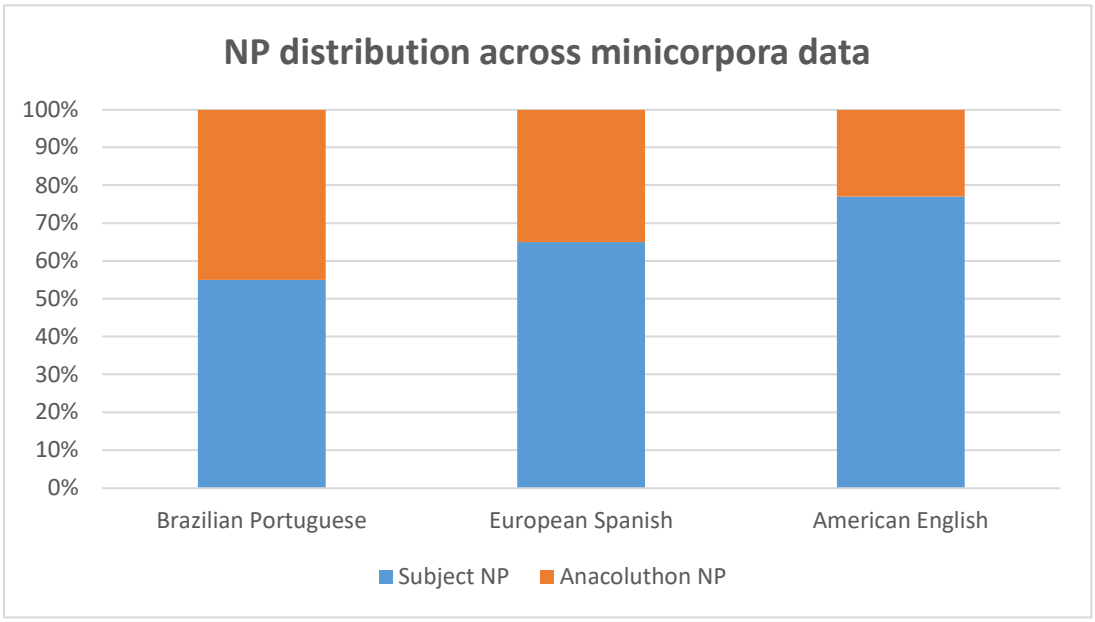
Firstly, it is important to note that the number of subject NPs is always greater than that of anacoluthon NPs. In comparison with AE and ES, BP presents a certain balance regarding this distribution. Considering only our dataset, such a result would support what the literature says about BP, that is, it would be both a Topic and a subject-prominent language (PONTES, 1987). The number of subject NPs in the three languages covers 64% of the data, as can be seen in Table 1 and Graph 1.

Table 1 – NP distribution in TOP across three languages

		Subject NP	Anacoluthon NP	Total
Brazilian Portuguese		84	68	152
	%	(55)	(45)	
European Spanish		166	91	257
	%	(65)	(35)	
American English		82	25	107
	%	(77)	(23)	
Total		332	184	516
	%	(64)	(36)	

Source: Authors

In comparison to what takes place in BP, the proportion of subject NPs increases in Spanish and reaches a peak in English (77% of the data). The results found in the descriptive analysis empirically reflect the assumption that English is a subject-prominent language, as the literature points out (LI; THOMPSON, 1976). The distribution of NP subjects in AE is so prominent that no significant results would be achieved if regression analysis were applied for this type of data. If this



Graph 1 – NP distribution in TOP across three languages
Source: Authors

trend does reflect how NPs are distributed in the language as a whole, not even a much larger sample would be able to get around this problem. Thus, regression analysis will only be applied to BP and ES.

Accessibility of the referent

Regarding the parameter of accessibility of the referent, it is possible to observe that in the three languages, subject NPs occur more often as given referents in the speech (see Table 2). In BP, 38% of subject NPs are given, while 18% are new. In ES, 45% are given and 19% are new. In AE, 48% are given and 29% are new. This is true even for the few occurrences of this category in English (see Graph 2).

Such results would show that, in principle, our data converge with what has been postulated for the category of subject NPs in relation to the accessibility of the referent, that is, they tend to be elements given in the discourse (GIVÓN, 1979).

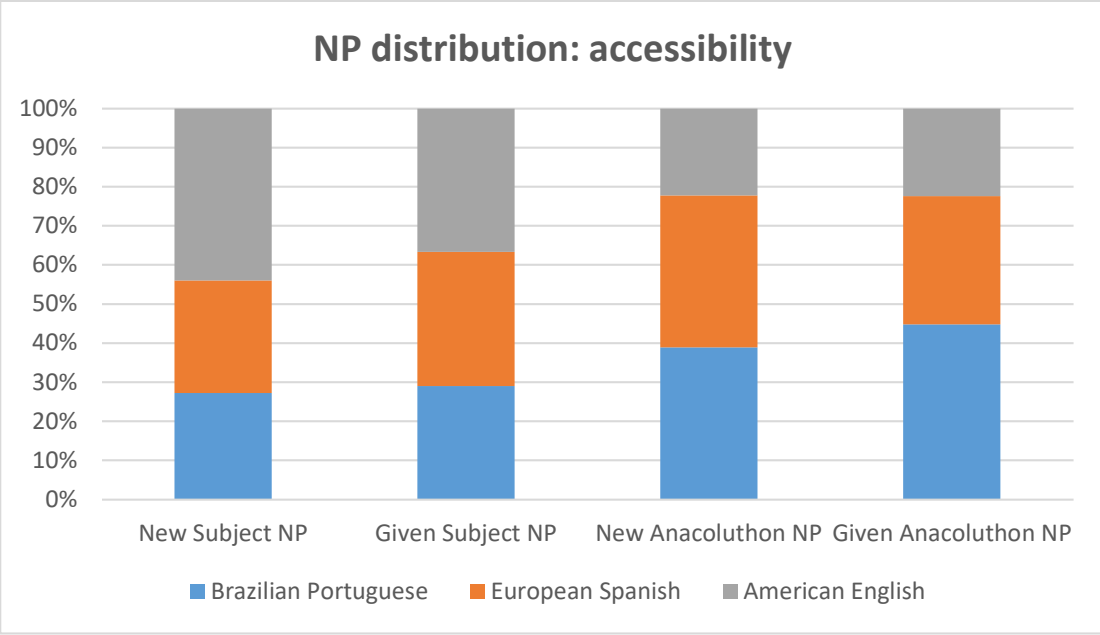
Definiteness

Regarding the definiteness parameter, it is possible to note that both in the set of subject NPs and in the set of anacoluthon

Table 2 – NP distribution in TOP across three languages: accessibility of referent

	New subject NP	Given subject NP	New anacoluthon NP	Given anacoluthon NP	Total
Brazilian Portuguese	27	57	22	46	152
%	(18)	(38)	(14)	(30)	
European Spanish	50	116	35	56	257
%	(19)	(45)	(14)	(22)	
American English	31	51	9	16	107
%	(29)	(48)	(8)	(15)	
Total	**108**	**224**	**66**	**118**	**516**
%	**(21)**	**(43)**	**(13)**	**(23)**	

Source: Authors.



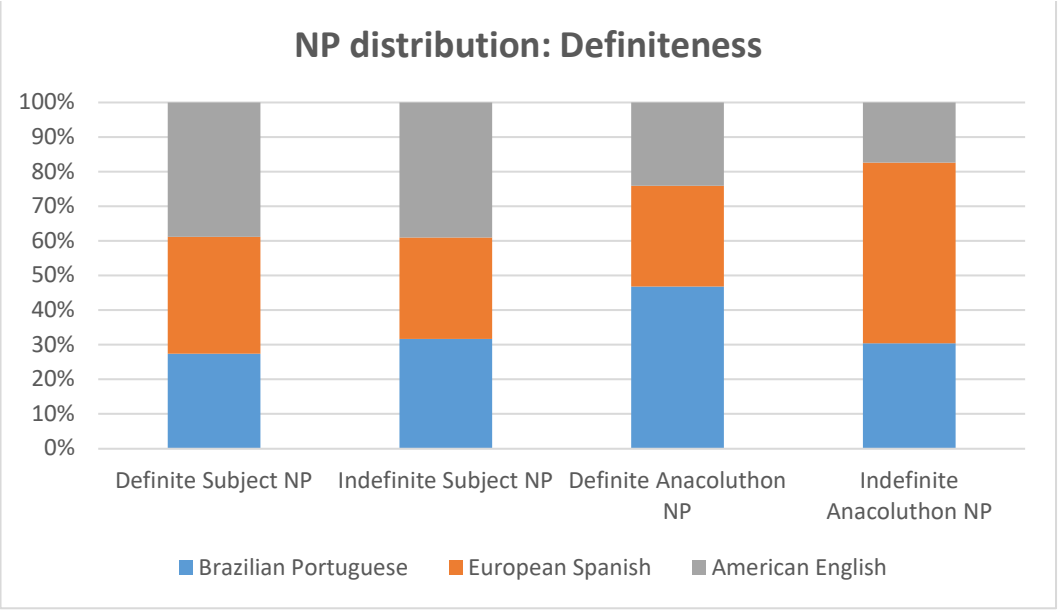
Graph 2 – NP distribution in TOP across three languages: accessibility of referent
Source: Authors

NPs the presence of definite referents is predominant in the three languages, and this is more accentuated in NPs with subject function (see Table 3). In BP, 43% of these NPs are definite, while only 13% are indefinite. In ES, 53% are definite and only 12% are indefinite. Finally, in AE 61% are definite and 16% are indefinite (see Graph 3).

Table 3 – NP distribution in TOP across three languages: Definiteness

	Definite subject NP	Indefinite subject NP	Definite anacoluthon NP	Indefinite anacoluthon NP	Total
Brazilian Portuguese	65	19	58	10	152
%	(43)	(13)	(37)	(7)	
European Spanish	136	30	61	30	257
%	(53)	(12)	(23)	(12)	
American English	65	17	20	5	107
%	(61)	(16)	(19)	(4)	
Total	266	66	139	45	516
%	(52)	(13)	(26)	(9)	

Source: Authors



Graph 3 – NP distribution in TOP across three languages: Definiteness
Source: Authors

We believe that these results reflect a general tendency in language use that favors definite NPs. Probably, the use of indefinites would be more expected in monological situations, in which there is the construction of long narratives that allow greater textual articulation (MITTMANN, 2013). This, in turn, could contribute to the expansion of entities introduced

in the discourse. In conversations and dialogues, in which the dynamics of situations lead to less textual articulation, there would probably be a smaller number of entities in the discourse and most of them would already be within reach of the interlocutors in the spatial environment in which they find themselves. Nevertheless, all the considerations made above would require further investigation through a larger body of data, so that NPs in other syntactic positions could also be considered.

Animacy

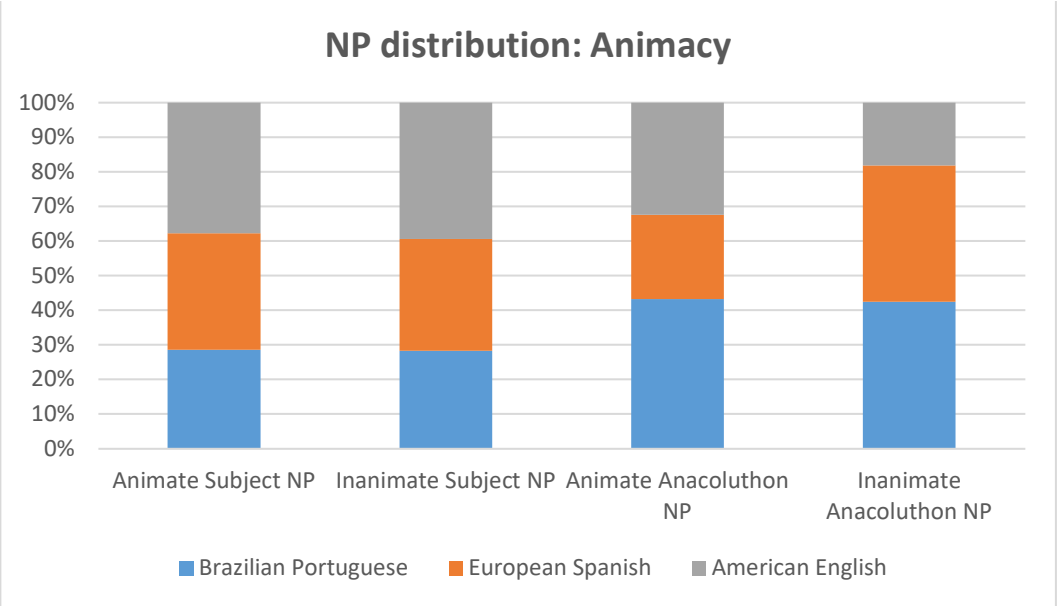
The distribution of animate subject NPs compared to inanimate ones was very close in the three languages. Such distribution is quite balanced: in BP 28% of subject NPs are animate and another 28% are inanimate. In ES, 33% are animate and 32% are inanimate. Finally, in AE 37% are animate and 39% are inanimate (see Table 4 and Graph 4). For anacoluthon NPs, this balanced distribution does not occur. Both in BP and ES there is a predominance of inanimate anacoluthon NPs. In AE, anacoluthon NPs have an identical distribution between animate (12%) and inanimate (12%). However, as previously mentioned, the occurrence of anacoluthon NPs in AE is very limited, which may have contributed to the distribution found.

Table 4 – NP distribution in TOP across three languages: Animacy

	Animate subject NP	**Inanimate subject NP**	**Animate anacoluthon NP**	**Inanimate anacoluthon NP**	**Total**
Brazilian Portuguese	42	42	25	43	152
%	(28)	(28)	(16)	(28)	
European Spanish	84	82	24	67	257
%	(33)	(32)	(9)	(26)	
American English	40	42	12	13	107
%	(37)	(39)	(12)	(12)	
Total	**166**	**166**	**61**	**123**	**516**
%	**(32)**	**(32)**	**(12)**	**(24)**	

Source: Authors

It can be noted that there is a balanced distribution concerning animate and inanimate subject NPs in the three languages, although it would be expected to find more animate



Graph 4 – NP distribution in TOP across three languages: Animacy
Source: Authors

subject NPs, since the literature points toward a prevalence of this feature in the category of subject.

Mixed-effects logistic regression model

As mentioned in the previous section, regression analysis was not conducted for English data, since this language presented 77% of subject NPs, an incompatible number for the analysis proposed in this section. A very low occurrence of anacolutha prevents any level of statistical significance when the methods chosen in this study are applied. Therefore, we cannot know if this is due to the distribution of data in the sample or to the very nature of the phenomenon in that language. Regression analysis was conducted with the aid of the R programming environment (R CORE TEAM, 2018) using the lme4 package (BATES et al. 2014). Initially, a model was run for Spanish and another for Portuguese, taking into account the same effects. As fixed effects, accessibility of referent, animacy, and definiteness were listed. As there was more than

one exemplar within the same text, as well as there could be more than one occurrence by the same speaker, random effects selected were text and participant. However, when the models were run, the text effect did not show any level of explanation (variance 0), so it was excluded. Thus, the only random effect considered in the final models was the speaker.

Table 5, below, shows the results obtained with the application of the mixed model for the Spanish data. It is important to note that only one value is shown for each fixed effect. This happens because a factor from each effect is compared with the factors that are at the reference level. In our case, the reference level is anacoluthon NP, given referent, animate and definite. The factors subject NP, new referent, inanimate and indefinite, thus, will be compared with their respective reference level factor. In the R package used, the reference level is chosen based on the alphabetical order of the factors for each effect. For didactic purposes, there is the possibility of changing factors at the reference level; however, statistically, this does not change the results. The values in the "Estimate" column indicate the coefficients related to the chance that each effect will occur comparing it to the reference level. These values are transformed into a logarithmic scale (known by the terms log-odds or logits). They are centered at 0 because the natural logarithm of 1 (that is, when the chance of occurrence of the two factors of the dependent variable – subject NP and anacoluthon NP – is equal) is 0. The values can therefore vary from - ∞ to + ∞. When the coefficient is negative, the chance of the reference level factor occurring will increase – in this case, it would be the anacoluthon NP factor. On the other hand, when the coefficient is positive, the chance of the second factor occurring, in this case, subject NP, will increase. However, this coefficient will only be considered if it is statistically significant. This can be verified using the values in the last column. If the value is lower than 0.05, the null hypothesis is rejected and the coefficient can be interpreted. Table 5 below shows the results obtained with the application of the model.

Table 5 – Mixed-effects model: European Spanish

Fixed-effects				
	Estimate	Standard error	z value	Pr (>|z|)
Intercept	1,6656	0,3705	4,495	6,95e-06 ***
New referent	-0,2914	0,3448	-0,845	0,39800
Inanimate NP	-1,1817	0,3655	-3,233	0,00122 **
Indefinite NP	-0,8220	0,3792	-2,168	0,03016 *
Random-effects				
Variable	Variance	Standard dev.		
Speaker (intercept)	1,396	1,181		
AIC	BIC	logLik	deviation	df.resid
302,6	320,4	-146,3	292,6	252

Source: Authors

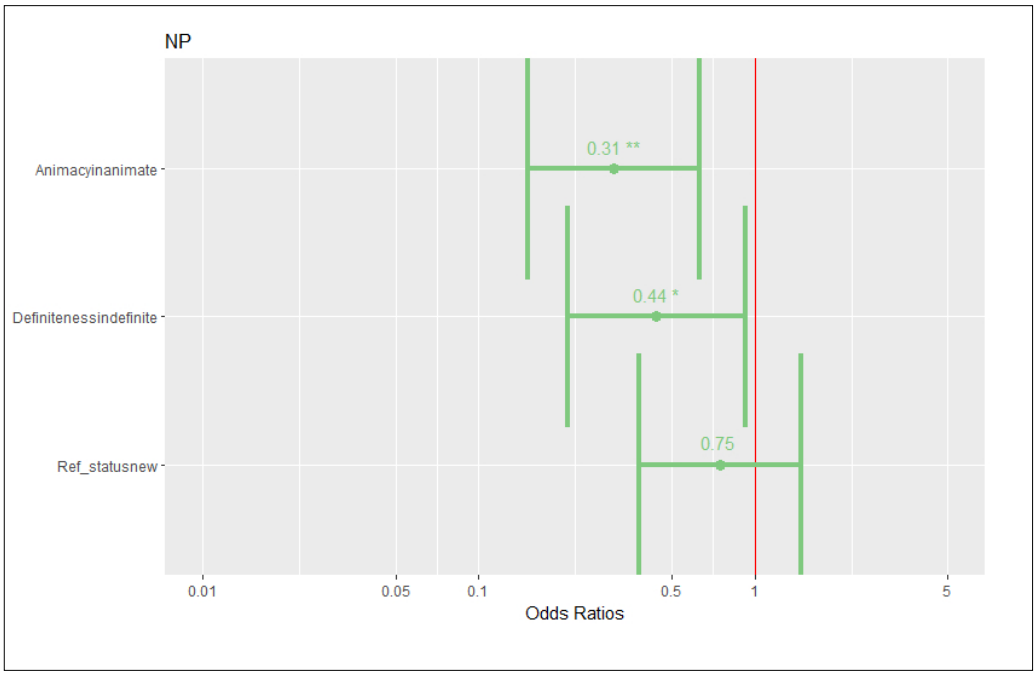
The model shows that when an NP is inanimate, it is more likely that it is an anacoluthon NP, which is in line with the hypothesis listed in section "Animacy". In addition, the model also indicates that an NP is more likely to be an anacoluthon NP if it is indefinite. This result is also in line with the hypothesis launched in section "Definiteness". Regarding the accessibility of the referent, it cannot be explored, since there was no statistically significant result for this purpose (0.39). If this result were significant and maintained a value close to what was achieved with this model, the interpretation would be that there is more chance of an NP being anacoluthon if it bears new information.

The first reported value, that is, the intercept must be interpreted as follows: the intercept coefficient indicates a greater chance of occurrence of one or another factor of the dependent variable when all fixed effects are at the reference level. In other words, the intercept coefficient will indicate whether there is a greater chance of occurrence of a subject NP or an anacoluthon NP when such NP is animate, definite, and is given information. Therefore, by observing the value of the intercept, it can be noted that there is more chance for a subject to occur when the NP is animate, definite, and given. More precisely, when the NP has these properties, the chance of occurrence (simple odds) of a subject is 5.28. The simple odds are the ratio of the probability of one event to the probability

of another event or the frequency of X over the frequency of non-X. If the values are between 0 and 1, the chance of an anacoluthon NP occurring is lower. If the value is greater than 1, the chance of a subject NP occurring is higher. Therefore, when the NP is aligned with the effects at the reference level, the chance of a subject NP occurring is 5.28 higher than that of an anacoluthon NP.

[7] This graph was elaborated with the sjPlot package (LÜDECKE, 2018).

Graph 5 below shows the results of the mixed model[7]. Note that the values are below 1, so this indicates that the fixed effect factors shown on the left favor the occurrence of an anacoluthon NP. The red line drawn in value 1 indicates that the chances of occurrence of a subject NP and an anacoluthon NP are equal. Note that no effect has reached this value and that the value of the accessibility of the referent effect is not statistically significant.



Graph 5 – Mixed-effects model for European Spanish: Fixed-effects in odds ratio
Source: Authors

The mixed logistic regression model was also applied to BP data. However, no fixed effect showed a statistically significant result. We believe that this is due to the fact that there is less BP data compared to the ES data set. In this way, statistically significant results could only be obtained if the sample were increased, making it at least comparable to that of ES.

The fact that an animate NP is more likely to be a subject may be due to aspects of human cognition. According to Bock (1982), animate nominal concepts are lexicalized more easily and quickly than inanimate nominal concepts, since speakers process information from an egocentric point of view and are more attentive to the former than to the latter. Animate nouns tend to be more easily remembered than inanimate ones (cf. ROHRMAN, 1970; GLANZER; KOPPENAAL, 1977). Because they are, therefore, more cognitively favored, phrases that contain animate nouns would tend to occur at the beginning of the sentence. Bock (1982) also highlights the role of egocentric processing of information by humans as a relevant factor for animate beings to occur at the beginning of the sentence. According to the author,

> (...) several of the theories of constituent ordering (...) assume that humans process information egocentrically and are, therefore, predisposed to attend to personally relevant stimuli. Among such personally relevant stimuli are other animate beings, particularly human animate beings. Thus, animate entities should tend to occur early in sentences more often than inanimate entities (BOCK, 1982, p. 15).

Additionally, a possible explanation for the fact that there is a greater chance that indefinite NPs in TOP do not present a syntactic relationship with the subsequent content of the utterance may perhaps reside in the tendency that indefinites generally introduce a new element in a distinct mental space (cf. FAUCONNIER, 1985, for more information on definiteness and mental spaces)[8]. This mental space would be housed in the TOP unit precisely to safeguard the discursive function without losing the interpretative condition of the utterance. It would be for this reason that a structure without a syntactic relation like this would be unlikely to occur, without a prosodic break, in the information unit of COM, in which predicates

[8] In a nutshell, mental spaces are "very partial assemblies constructed as we think and talk for purposes of local understanding and action" (FAUCONNIER, 2007, p. 351). For the relation between definiteness and mental spaces, cf. Fauconnier (1985), Ohashi (1995), Epstein (2002), among others.

are usually lodged. In order to be individualized as a distinct mental space and to safeguard the interpretability of the utterance, the NP would be carried out in a separate unit that still fulfills a discursive function (TOP), but not necessarily a syntactic one. This would result in a different access from that in which the NP maintains a subject relationship with the COM content.

Final remarks

This paper tried to show an application of how syntax can be studied from a probabilistic point of view. The syntactic aspect discussed in this regard was the possibility that NPs fulfilling the informational function of Topic are either subject or an independent item in the utterance called anacoluthon. The variables selected for the analysis were accessibility (given or new), animacy (animate or inanimate), and definiteness (definite or indefinite). These variables are at the core of much debate in the literature and, in most cases, there is no real consensus on their representation. However, in this study, they were individualized according to certain assumptions established through a discussion of the literature. The statistical method employed to carry out the data analysis was a mixed-effects model of logistic regression with crossed random effects, which showed that there is more chance that an NP in a TOP unit will not present a syntactic relationship with the subsequent content in COM if it is inanimate and indefinite. The intercept showed that there is more chance that an NP will perform the function of a subject if it is animate, definite, and brings information given in the discourse. Future studies need to consider further qualitative aspects of the results portrayed here, especially the cognitive structure of the utterances, in order to fully tackle the complexity of the phenomenon explored in this paper.

Acknowledgments

We would like to thank the two anonymous reviewers for their careful reading of our manuscript, as well as for their valuable and constructive suggestions that helped improve this paper. Any remaining errors are our sole responsibility.

Financial Disclosure

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

REFERENCES

AISSEN, Judith. Differential object marking: iconicity vs. economy. *Natural Language and Linguistic Theory*, v. 21, n. 3, p. 435-483, aug. 2003.

ARIEL, Mira. *Accessing noun phrase antecedents*. London: Routledge, 1990.

BARBOSA, Joaquim. Foco e tópico: algumas questões terminológicas. *In:* RIO-TORTO, Graça; FIGUEIREDO, Olívia; SILVA, Fátima (ed.). *Estudos em homenagem de Mário Vilela*. Porto: Faculdade de Letras da Universidade do Porto, 2005. p. 339-351.

BATES, Douglas et al. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, v. 67, n. 1, p. 1-48, jun. 2014.

BOCK, Kathryn. Toward a cognitive psychology of syntax: Information processing contribution to sentence formulation. *Psychological Review*, v. 89, n. 1, p. 1-47, 1982.

CAVALCANTE, Frederico; RAMOS, Adriana. The American English spontaneous speech minicorpus. Architecture and comparability. *CHIMERA: Romance Corpora and Linguistic Studies*, v. 3, n. 2, p. 99-124, 2016.

CHAFE, Wallace. Givenness, contrastiveness, definiteness, subjects, topics and point of view. *In:* LI, Charles (ed.). *Subject and topic*. New York: Academic Press, 1976. p. 27-55.

CHAFE, Wallace. Cognitive constraints on information flow. *In:* TOMLIN, Russell (ed.). *Givenness, contrastiveness, definiteness, subjects, topics, and points of view*. Amsterdam: Benjamins, 1987. p. 21-51.

CLARK, Herbert; CLARK, Eve. *Psychology and language*: an introduction to psycholinguistics. New York: Harcourt Brace Jovanovich, 1977.

CLARK, Herbert; CLARK, Eve; HAVILAND, Susan. Psychological processes as linguistic explanation. *In:* COHEN, David (ed.). *Explaining linguistic phenomena*. Washington: Hemisphere, 1974. p. 91-124.

COMRIE, Bernard. *Language universals and linguistic typology*. 2. ed. Oxford: Basil Blackwell, 1989.

CORBETT, Greville. *Number*. Cambridge: Cambridge University Press, 2000.

CRESTI, Emanuela. *Corpus di Italiano parlato*. Firenze: Accademia Della Crusca, 2000.

CRESTI, Emanuela; MONEGLIA, Massimo (ed.). *C-ORAL-ROM*: integrated reference corpora for spoken romance languages. Amsterdam; Philadelphia: John Benjamins Publishing Company, 2005.

CROFT, William. Agreement vs. case marking in direct objects. *In:* BARLOW, Michael; FERGUSON, Charles (ed.). *Agreement in natural language*: approaches, theories, descriptions. Stanford: CSLI, 1988. p. 159-180.

CRYSTAL, David. Neglected grammatical factors in conversational English. *In:* GREENBAUM, Sidney; LEECH, Geoffrey; SVARTVIK, Jan (ed.). *Studies in english linguistics:* for Randolph Quirk. London: Longman, 1980. p. 153-166.

DEANE, Paul. *Grammar in mind and brain:* explorations in cognitive syntax. Berlin; New York: Mouton de Gruyter, 1992.

DU BOIS, John et al. *Santa Barbara corpus of spoken American English parts 1-4*. Philadelphia: Linguistic Data Consortium, 2000-2005.

EPSTEIN, Richard. The definite article, accessibility, and the construction of discourse referents. *Cognitive Linguistics*, v. 12, n. 4, p. 333-378, jan. 2002.

FAUCONNIER, Gilles. *Mental spaces:* aspects of meaning construction in natural language. Cambridge: The MIT Press, 1985.

FAUCONNIER, Gilles. Mental spaces. *In:* GEERAERTS, Dirk; CUYKENS, Hubert (ed.). *The oxford handbook of cognitive linguistics*. New York: Oxford University Press, 2007.

GIVÓN, Talmy. Topic, pronoun and grammatical agreement. *In:* LI, Charles (ed.). *Subject and topic*. New York: Academic Press, 1976. p. 149-188.

GIVÓN, Talmy. *On understanding grammar*. New York: Academic Press, 1979.

GIVÓN, Talmy (ed.). *Topic continuity in discourse:* a quantitative cross-language study. Amsterdam: John Benjamins, 1983.

GIVÓN, Talmy. Beyond foreground and background. *In:* TOMLIN, Russell (ed.). *Coherence and grounding in discourse*. Amsterdam: Benjamins, 1987. p. 175-188.

GLANZER, Murray; KOPPENAAL, Lois. The effect of encoding tasks on free recall: stages and levels. *Journal of Verbal Learning and Verbal Behavior*, v. 16, n. 1, p. 21-28, feb. 1977.

GUNDEL, Jeanette; HEDBERG, Nancy; ZACHARSKI, Ron. Cognitive status and the form of referring expressions in discourse. *Language*, v. 69, p. 274-307, jun. 1993.

HALL, Kira. Agency and the Animacy Hierarchy in Kashaya. Paper presented at the *Hokan-Penutian Languages Workshop*, University of California, San Diego, 22-23 June 1990.

HAWKINS, John. *Definiteness and indefiniteness*: a study in reference and grammaticality prediction. London: Croom Helm, 1978.

HUNDT, Marianne, RÖTHLISBERGER, Melanie; SEOANE, Elena. Predicting voice alternation across academic Englishes. *Corpus Linguistics and Linguistic Theory*, v. 17, n. 1, apr. 2018.

HYMAN, Larry. Form and substance in language universals. In: BUTTERWORTH, Brian; COMRIE, Bernard; DAHL, Östen (ed.). *Explanations for language universals*. New York; Amsterdam: Mouton Publishers, 1984. p. 67-86.

ITAGAKI, Nobuya; PRIDEAUX, Gary. Nominal properties as determinants of subject selection. *Lingua*, v. 66, p. 135-149, 1985.

KEENAN, Edward. Towards a universal definition of subject. *In:* LI, Charles (ed.). *Subject and topic*. New York: Academic Press, 1976. p. 303-333.

KRÁMSKÝ, Jirí. *The article and the concept of definiteness in language*. The Hague; Paris: Mouton, 1972.

LAMBRECHT, Knud. *Information structure and sentence form*: topic, focus and the mental representations of discourse referents. Cambridge: Cambridge University Press, 1994.

LANGACKER, Ronald. *Foundations of cognitive grammar:* descriptive application. Stanford: Stanford University Press, 1991. 2 v.

LI, Charles; THOMPSON, Sandra. Subject and topic: a new typology of language. *In:* LI, Charles (ed.). *Subject and topic*. Santa Barbara: University of California, 1976. p. 457-489.

LÜDECKE, Daniel. *sjPlot*: Data Visualization for Statistics in Social Science. R package version 2.6.0, 2018. Available from: https://CRAN.R-project.org/package=sjPlot. Accessed in: 13 oct. 2018.

LYONS, Christopher. *Definiteness*. Cambridge: Cambridge University Press, 1999.

MALCHUKOV, Andrej. Animacy and assymetries in differential case marking. *Lingua*, v. 118, n. 2, p. 203-221, feb. 2008.

MELLO, H.; SILVA, L. F. L. Focus and information patterning: refining terminology and distinguishing categories in a spoken corpus. *Revista Virtual de Estudos da Linguagem*, n. 10, p. 138-169, 2015.

MEURMAN-SOLIN, Anneli; LOPEZ-COUSO, Maria Jose; LOS, Bettelou (eds.). *Information structure and syntactic change in the history of english*. Oxford: Oxford University Press, 2012.

MITTMANN, Maryualê. Análise da estruturação de diálogos e monólogos na fala informal: quantificando as diferenças. *Domínios de Lingu@gem*, v. 7, n. 2, p. 338-372, jul./dec. 2013.

NICOLÁS MARTÍNEZ, Carlota; LOMBÁN SOMACARRERA, Marina. Mini-Corpus del español para DB-IPIC. *CHIMERA - Romance Corpora and Linguistic Studies*, v. 5, n. 2, p. 95-113, oct. 2018.

OHASHI, Hiroshi. Discourse referents in mental spaces: a case of anaphoric reference to indefinite noun phrases with nonspecific interpretation. *Journal of UOEH*, v. 17, n. 4, p. 287-298, 1995.

PANUNZI, Alessandro; GREGORI, Lorenzo. DB-IPIC: an XML database for the representation of information structure in spoken language. *In:* MELLO, Heliana; PANUNZI, Alessandro; RASO, Tommaso (ed.). *Pragmatics and prosody*: illocution, modality, attitude, information patterning and speech annotation. Firenze: Firenze University Press, 2011. p. 133-150.

PANUNZI, Alessandro; GREGORI, Lorenzo; MITTMANN, Maryualê. The IPIC resource and a cross-linguistic analysis of information structure in Italian and Brazilian Portuguese. *In:* RASO, Tommaso; MELLO, Heliana (ed.). *Spoken corpora and linguistic studies*. Amsterdam; Philadelphia: John Benjamins, 2014. p. 129-151.

PONTES, Eunice. *O tópico no português do Brasil*. Campinas: Pontes, 1987.

PRINCE, Ellen. Toward a taxonomy of given-new information. *In:* COLE, Peter (ed.). *Radical pragmatics*. New York: Academic Press, 1981. p. 223-255.

R CORE TEAM. *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing, 2018.

ROHRMAN, Nicholas. More on the recall of nominalizations. *Journal of Verbal Learning and Verbal Behavior*, v. 9, n. 5, p. 534-536, 1970.

SILVA, L. F. L. O estatuto da sintaxe na fala: considerações acerca da proposta da Language into Act Theory. *Revista de Estudos da Linguagem*, v. 28, n. 1, p. 271-330, 2020.

SILVERSTEIN, Michael. Hierarchy of features and ergativity. *In:* DIXON, Robert (ed.). *Grammatical categories in Australian languages*. Canberra: Australian Institute of Aboriginal Studies, 1976. p. 112-171.

SZMRECSANYI, Benedikt et al. Spoken syntax in a comparative perspective: the dative and genitive alternation in varieties of English. *Glossa:* a Journal of General Linguistics, v. 2, n. 1, p. 1-27, 2017.

't HART, Johan; COLLIER, René; COHEN, Antonie. *A perceptual study of intonation:* an experimental-phonetic approach to speech melody. Cambridge: Cambridge University Press, 1990.

VIRTANEN, Tuija. Given and new information in adverbials: clause-initial adverbials of time and place. *Journal of Pragmatics*, v. 17, n. 2, p. 99-115, feb. 1992.

VOGELS, Jorrig; VAN BERGEN, Geertje. Where to place inaccessible subjects in Dutch: The role of definiteness and animacy. *Corpus Linguistics and Linguistic Theory*, v. 13, n. 2, p. 369-398, sep. 2017.

YAMAMOTO, Mutsumi. *Animacy and reference*: a cognitive approach to corpus linguistics. Amsterdam; Philadelphia: John Benjamins, 1999.

Uma abordagem probabilística para a distribuição de nps sujeito e anacoluto em tópicos na fala espontânea

RESUMO

A literatura apresenta muitas definições acerca da noção de Tópico e de estrutura informacional (cf. BARBOSA, 2005; MELLO; SILVA, 2015). Neste trabalho, assume-se a definição da Language into Act Theory (CRESTI, 2000) que diz que o Tópico é a porção textual que se realiza por meio de um padrão entoacional do tipo prefix ('t HART et al. 1990) e que tem por função constituir o domínio sobre o qual a força ilocucionária se aplica. Um SN em Tópico pode ser o sujeito do verbo no Comentário ou um anacoluto. SNs anacolutos são sintagmas que não possuem relação sintática com o conteúdo presente no Comentário. Neste trabalho, mostra-se como SNs estão distribuídos probabilisticamente entre essas duas condições quando são realizados como Tópico na fala espontânea. Para isso, foram coletados dados de corpora de fala espontânea etiquetados informacionalmente – incluindo a unidade de Tópico conforme definida acima – de três línguas: espanhol europeu (NICOLÁS MARTÍNEZ; LOMBÁN SOMACARRERA, 2018), inglês americano (CAVALCANTE; RAMOS, 2016) e português brasileiro (PANUNZI; MITTMANN, 2014). O método estatístico utilizado para o cálculo da probabilidade foi um modelo misto de regressão logística com efeitos aleatórios cruzados, conduzido com auxílio do R (R CORE TEAM, 2018). Três variáveis foram selecionadas para o cálculo: acessibilidade do referente, animacidade e definitude. O modelo mostrou que há cerca de cinco vezes mais chances de que o SN realizado em Tópico seja sujeito do verbo no Comentário caso ele seja animado, definido e seja classificado como informação dada no discurso.

PALAVRAS-CHAVE: Tópico. Sujeito. SN. Sintaxe da fala. Gramática probabilística.

Luis Filipe Lima e Silva. Possui graduação em Letras (habilitação em Linguística) pela Universidade Federal de Minas Gerais (2013), mestrado (2016) e doutorado (2020) em Estudos Lingüísticos pela mesma instituição. Tem experiência na área de Linguística, com ênfase em Teoria e Análise Linguística, atuando principalmente nos seguintes temas: sintaxe, pragmática, gramaticalização e corpus de fala espontânea. É integrante do grupo de pesquisa InCognito - Interfaces Linguagem, Cognição e Cultura.

Heliana Mello. Graduação em Letras (1987) e Mestrado em Linguística (1990) pela Universidade Federal de Minas Gerais; Master of Arts in Linguistics (1992), Master of Philosophy (1993), PhD (1997) e pós-doutorado (1998) em Linguística pela City University of New York; pós-doutorado pela Escola de Matemática Aplicada da Fundação Getúlio Vargas (2012-2013). Professora Titular da Universidade Federal de Minas Gerais. Atua na área de Estudos Linguísticos na graduação e na pós-graduação, com ênfase em análises baseadas em corpora, relacionadas a mudanças de sistemas gramaticais e contato linguístico, semântica, sintaxe e pragmática da fala espontânea. Seu foco em Linguística de Corpus e Linguística Computacional abarca a compilação de corpora e metodologias quantitativas de análise linguística baseadas em corpora. Dentre seus projetos de pesquisa destaca-se o C-ORAL-BRASIL (www.c-oral-brasil.org) e os seus subprojetos.