CrossMark

Language processing, acceptability, and statistical distribution: A study of null and overt subjects in Brazilian Portuguese

Amit Almor [a,b,*], Jefferson de Carvalho Maia [b], Maria Luiza Cunha Lima [c], Mirta Vernice [d], Carlos Gelormini-Lezama [e]

[a] Psychology Department, University of South Carolina, Columbia, SC, USA
[b] Linguistics Program, University of South Carolina, Columbia, SC, USA
[c] Universidade Federal de Minas Gerais, Brazil
[d] University of Milano-Bicocca, Italy
[e] Departamento de Ciencias Sociales, Universidad de San Andrés, Buenos Aires, Argentina

ARTICLE INFO

ABSTRACT

Two experiments and a corpus study tested whether Brazilian Portuguese (BP), which has been argued to be shifting from null subjects toward overt subjects indeed shows a comprehension preference for reduced over fuller anaphors for salient antecedents, and whether comprehension is better explained as an imbalance between processing cost and discourse function (pragmatics account), or simply the frequency of different constructions (usage-based account). Sentences with overt pronouns were read slower (Experiment 1) but rated more acceptable (Experiment 2) than sentences with null pronouns when the antecedents were salient. Sentences with repeated names showed weaker effects than overt pronouns. The corpus analysis confirmed that BP is changing, but that null subjects are not yet infrequent, especially in academic writings. We argue that these results reflect a "pronoun avoidance strategy" in BP related to its transitory state, and propose a new view that integrates elements from both the pragmatics and usage-based accounts.

© 2016 Elsevier Inc. All rights reserved.

Null pronouns, overt pronouns and repeated names in Brazilian Portuguese

Most discourses and texts span over multiple informational units such as sentences and utterances, and maintaining coherence between these units is necessary for effective linguistic communication (Halliday & Hasan, 1976). An important linguistic device for establishing and maintaining discourse coherence is repeated reference by means of anaphoric[1] expressions (Garrod & Sanford,

1982a; Grosz, Joshi, & Weinstein, 1983, 1995; Kintsch & Van Dijk, 1978). Because this is true of all human languages, it is not surprising that they all display anaphors. More surprising is the observation that, despite differences in the inventory of referential forms in different languages, less informative expressions, such as the English unstressed pronoun, are commonly preferred when the antecedent is readily accessible to comprehenders (Gundel, Hedberg, & Zacharski, 1993).

This preference suggests the existence of some underlying constraints that are not language-specific. Indeed, many of the factors that are known to affect the processing of anaphors appear to be universal in nature and not specific to only some languages. For example, reference resolution is known to be affected by the relative syntactic configuration of the anaphor and its antecedent (Chambers & Smyth, 1998; Crawley, Stevenson, &

---

* Corresponding author at: Department of Psychology, 1512 Pendleton St., Columbia, SC 29208, USA.
E-mail address: almor@sc.edu (A. Almor).

[1] In this paper we use the term "anaphor" to denote any expression that refers to an already established discourse entity (the "antecedent"). This is to be distinguished from the use of this term in certain work in linguistics, where it is used to denote only certain forms of referential expressions.

Kleinman, 1990; Frederiksen, 1981; Gordon, Grosz, & Gilliom, 1993), by discourse pragmatic principles (Almor, 1999; Ariel, 1990; Prince, 1978), and by memory constraints (Almor, 1999; Gernsbacher, 1989; Sanford & Garrod, 1981), all of which are quite likely involved in the processing of any natural language.

Despite the arguably universal nature of many of the factors that have been shown to affect anaphor processing, most of the related empirical research has been carried out in the framework of a single language, most often English. This could limit our understanding of the extent to which general principles can account for anaphor processing in different languages with different referential inventories. Moreover, research on reference processing in bilinguals, which has recently become a focus of considerable effort (e.g., Cho, 2010; Filiaci, Sorace, & Carreiras, 2013; Roberts, Gullberg, & Indefrey, 2008; Sorace & Filiaci, 2006), can clearly benefit from the better understanding of the differences and similarities in coreference processing in languages other than English.

The research we report here was undertaken to fill this gap in the literature by examining and comparing the processing of different types of anaphors (null pronouns, overt pronouns, and repeated names) in Brazilian Portuguese, a language that offers a window into the factors underlying reference choice and processing because it is currently undergoing a change in its referential preferences.

Co-referential processing

As already mentioned, the choice of an anaphoric expression is highly dependent on the discourse salience of the entity introduced by the antecedent (Almor & Nair, 2007; Ariel, 1990; Garrod & Sanford, 1982b; Gernsbacher, 1989; Givón, 1987; Gordon et al., 1993; Grosz et al., 1983; Gundel et al., 1993; van-Dijk & Kintsch, 1983). Salient discourse entities are usually referred to by means of the least informative expression available in the language (e.g., null pronouns in a null-subject language), whereas non-salient entities are usually referred to by means of heavier anaphors such as definite expressions or proper names (Gordon & Scearce, 1995; Grosz et al., 1983; van-Dijk & Kintsch, 1983). Thus, there appears to be an inverse relationship between the salience of the antecedent and the amount of information contained in the anaphor: the more salient the antecedent is, the less informative the anaphor is likely to be.

Almor's Informational Load Hypothesis (ILH; Almor, 1999; Almor & Nair, 2007) is a pragmatics-based theory that explains this inverse relationship between the salience of the antecedent and the information load of the anaphor as the outcome of an optimization process reflecting a balance between the amount of information activated by the anaphor in working memory and its discourse function. When salient discourse entities are referred to by means of a highly informative anaphor, a larger amount of information is activated in working memory without serving any discourse function, causing a processing delay. However, when the same highly informative anaphor refers to a non-salient antecedent, the reduced salience of the latter results in less overall working memory

activation and, at the same time, the information activated by the anaphor now serves the function of helping to correctly identify the antecedent and reactivate its representation. In this case, the cost associated with activating the information conveyed by the anaphor is balanced with its discourse function; therefore, no overall processing delay occurs.

The slower reading of repeated names relative to pronouns when the antecedent is salient is one notable example of imbalance between cost and function: Gordon et al. (1993) showed that, in English, sentences containing a repeated name cause a processing delay relative to sentences with pronouns when the antecedent is salient, but not when the antecedent is not salient. This "repeated-name penalty" (RNP) has also been extended to Mandarin Chinese (Yang, Gordon, Hendrick, & Wu, 1999) and Spanish (Gelormini-Lezama & Almor, 2011, 2013), suggesting that this might be a universal phenomenon.

Another instance of imbalance between cost and function is the "overt pronoun penalty" (OPP), which is the slower reading of overt pronouns referring to salient antecedents relative to null pronouns. The OPP was first reported in Spanish (Gelormini-Lezama & Almor, 2011) in which overt pronouns can serve the function of either retrieving a non-salient antecedent, or marking a contrast or emphasis. In non-contrastive contexts, an overt pronoun referring back to a salient antecedent represents a case where the richer semantic content of the overt pronoun in comparison to a null pronoun is unnecessary. According to Gelormini-Lezama and Almor, the use of a more complex anaphoric form than would be required to retrieve an already accessible antecedent is penalized with longer reading times.

While there has also been other research on null and overt pronouns in Spanish and Italian, which are typologically close to BP, much of this research has focused on intra- rather than inter- sentential anaphora (e.g., Carminati, 2002; Filiaci, 2011; Filiaci et al., 2013). For example, contrasting Italian and Iberian Spanish, Filiaci (2011) and Filiaci et al. (2013) found a stronger preference for null over overt pronouns subject anaphors in Italian than in Spanish. The relevance of their results to the present discussion is somewhat limited however, because they only examined intra-sentential anaphora. As argued and shown by Carminati (2002), these two types of anaphora are likely subject to different types of constraints and should therefore be studied separately. Nevertheless, the body of studies mentioned above, together with other research showing regional variability in preferences for null pronouns in different regional variants of Spanish (Cameron, 1992; Flores-Ferrán, 2004), suggest that there may exist differences between languages in the likelihood of observing an OPP.

Yang et al. (1999) reported a study that did compare the processing of null and overt pronouns as inter-sentential anaphors with subject antecedents in Chinese, which is a typologically different language from the Romance languages discussed thus far. This study found that although Chinese utilizes null pronouns, the OPP was not elicited in this language, in that no significant differences between reading times of sentences containing null and overt