THE LEARNABILITY OF THE RESULTATIVE CONSTRUCTION IN ENGLISH L2: A COMPARATIVE STUDY OF TWO FORMS OF THE ACCEPTABILITY JUDGMENT TASK

Ricardo Augusto de SOUZA
Universidade Federal de Minas Gerais (UFMG)

Cândido Samuel Fonseca de OLIVEIRA
Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG)

RESUMO

A tarefa de julgamento de aceitabilidade é um método crucial em sintaxe experimental. Este estudo compara duas formas da tarefa – escalas Likert e estimativas de magnitude – na investigação da aprendizibilidade da construção resultativa do inglês por bilíngues do português do Brasil e do inglês. A resultativa é difícil para essa população, uma vez que se faz necessário aprender mapeamentos entre sintaxe-semântica inexistentes na L1 e restrições de estrutura de eventos que licenciam a construção na L2. Os resultados indicam que ambas as tarefas controladas atestam a aprendizibilidade da construção, mas a alegada superioridade das estimativas de magnitude não foi verificada.

ABSTRACT

The acceptability judgment task is a crucial method in experimental syntax. This study compares two forms of the task –Likert-scale and magnitude estimations – in the investigation of the learnability of the English resultative construction for bilinguals of Brazilian Portuguese and English. The resultative poses a challenge for this population, as not only must they learn the syntax-semantics mapping unlicensed in their L1, but they must learn event structure constraints that govern such construction in the L2. The results indicate that while both controlled acceptability judgment tasks attest the learnability of the construction, the alleged superiority of magnitude estimations was not verified.

PALAVRAS-CHAVE

Escala Likert, estimativa de magnitude, construção resultativa, aprendizibilidade em L2.

KEYWORDS

Likert-scale, magnitude estimation, resultative construction, L2 learnability.

Introduction

The acceptability judgment task is a method of eliciting participants' data for research in syntax, as well as research on other areas of linguistic organization. The method allows the researcher to observe reactions of speakers to linguistic constructions that may not exist in common usage of a given language, thus allowing for the inquiry of hypotheses about restrictions imposed by the knowledge systems underlying specific languages, irrespective of whether such systems are treated as representations accessed either implicitly or explicitly. However, there is controversy about the validity of acceptability judgment data. This controversy is mostly motivated by recognition that such method of data elicitation is potentially liable to be affected by a number confounding variables that are not easily integrated into models of strictly linguistic knowledge.

More recently the rigorous implementation of guidelines of experimental designs to the acceptability judgment task gave rise to an emerging tradition referred to as experimental syntax. Through experimental syntax, approaches that seek to mitigate possible threats to the validity of acceptability judgments are pursued. In such approaches, measurement scales are a vital concern. Two examples of the acceptability judgment task that differ in terms of their measurement rationale are "scalar acceptability judgments" and "magnitude estimations." In the first type of task, judgment responses are usually recorded on categorical

Likert-type scales. In magnitude estimation the primary goal is the establishment of individualized scales, typically without manipulation of the temporal ceiling for judgment manifestation. The first approach, based on categorical scales, involves a task that is usually easy to learn by participants, but the ensuing analysis typically involves adaptation of statistical tests originally designed for continuous measures to categorical data. On the other hand, magnitude estimation in theory offers the advantage of ready adequacy to established statistical procedures for hypothesis testing based on continuous measures. However, it can be a more demanding task from participants' viewpoint.

The present study aimed primarily at comparing the results yielded by these two types of acceptability judgment tasks in an investigation of the learnability of the English resultative constructions by bilinguals of Brazilian Portuguese and English. The English resultative construction is a potentially complex structure for bilinguals of this particular linguistic profile. Not only does its licensing depend on semantic composition of subtle configurations of event structure, but also the surface syntactic structure it maps to is ambiguous with respect to the Portuguese language, where the same order of constituents link to a different meaning.

A secondary goal of the present study is to examine the viability of implementation of experimental syntax techniques to the psychometric exploration of second language knowledge. There has been some controversy as to whether the acceptability judgment task fits the investigation of L2 knowledge, with doubts mainly related to whether such tasks would tap into implicit grammatical knowledge rather than explicit knowledge (MANDELL, 1999). More recently, though, there has been a revival of interest in the viability of use of this type of task as a reliable indicator of both implicit and explicit knowledge of L2 users (GUTIÉRREZ, 2013).

The next section outlines details of the acceptability judgment task, with a special focus on the "magnitude estimation paradigm". A description of the resultative construction of English and a comparison and contrast with similar patterns in Brazilian Portuguese follows, leading to some considerations about the significance of evidence of learnability of this construction by native speakers of Brazilian Portuguese to current debates in second language acquisition. Then the methods of the present study are described, and the results analyzed and discussed. We finish with some concluding remarks about our interpretation of our findings.

1 Acceptability judgments in experimental syntax

The cognitive revolution of the second half of the twentieth century, and especially the emergence of generative linguistics, has sparked interest in the cognitive processes related to the faculty of language. To investigate these processes, the primary source of evidence in various sub-fields of linguistics has been the acceptability of certain strings of linguistic units, namely sentences in the context of syntactic research (SORACE & KELLER, 2005). Acceptability – how good and/or acceptable a sentence sounds – is regarded as property of overt linguistic material to which speakers have reasonable access through introspection. Sentence acceptability is a construct consisting of several factors (SCHUTZE, 1996; SPROUSE et al, 2013). We understand the acceptability judgment (AJ) task as a psychometric operationalization of this construct, and as such it should ideally be mostly sensitive to three factors: (1) grammaticality – or well-formedness according to some governing linguistic principle; (2) representation – or availability of implicit or explicit knowledge of that governing principle; and (3) processing capacity – or access to such representation over the course of performance of the AJ task. Thus, the systematic observation of the behavior – acceptance or rejection – of a speaker in relation to certain constructions or strings of linguistic units can be considered a method to reveal properties of the grammar of a language as well as the functioning of the human mind regarding this linguistic knowledge. It follows that it is also a method for experimental exploration of hypotheses about both such properties and their psychological reality.

Data from speakers' intuition have great value since there is no correspondence between linguistic knowledge and language use (SORACE & KELLER, 2005; SORACE, 2010). In other words, AJ data provide evidence of linguistic phenomena that are not readily observable in samples of spontaneous usage, but which nonetheless provide important insights into the functioning of language. This is the reason why AJ tasks and improvement of the methodological procedures that support them have been gaining ground among linguists, as attested by this very issue of the ABRALIN journal.

In fields such as syntax, much of the data are extracted from methods that are considered informal by traditional cognitive science standards (DRABOWSKA, 2010). However, especially with the emergence of "experimental syntax" (COWART, 1997), formal procedures have come to be more consistently used and have drawn ever growing interest, opening a debate about which methods have greater empirical validity (BARD ET AL, 1996; SCHUTZE, 1996; FEATHERSTON, 2005, 2009; FERREIRA, 2005; CULBERTSON & GROSS, 2009; MYERS, 2009a, b; SORACE, 2010; PHILIPS, 2010; WESKOTT & FANSELOW, 2011; SPROUSE, 2011; FUKUDA et al, 2012; GIBSON & FEDORENKO, 2013; SPROUSE et al., 2013). Thus, studies seeking to illuminate the differences between two or more procedures (formal vs. informal; formal-x vs formal-y) are important in both validating proposals once defended, as well as in directing future studies.

A practice widely used especially in much of generative linguistic research is for linguists themselves to judge the acceptability of sentences in order to probe theoretical arguments. Such practice is what has been commonly referred to as informal AJ, and it has generated distrust among many authors (FERREIRA, 2005; MYERS, 2009a, b; GIBSON & FEDORENKO, 2013). Manifestations of distrust are due to a number of issues, among which Weskott & Fanselow (2011) highlight (1) the possibility of theoretical bias, (2) the lack of generalizability across other lexicalizations of the structure under scrutiny and (3) the inherent non-replicability.

Another characteristic of informal AJ that inspires distrust is the fact that this subjectivist introspectionism is likely to be performed by a person who deals with linguistic analysis far more often than ordinary people. Certain studies suggest this fact originating from the linguists' professional lives may bias judgments (SCHUTZE, 1996; BARILE & MAIA, 2008; CULBERTSON & GROSS, 2009; DRABOWSKA, 2010). Some authors report that it is common to witness a linguist admitting their lack of sensitivity in relation to certain ungrammatical structures with which they work (SNYDER, 2000; MAIA 2013).

A formal approach to the AJ task may be regarded as attempted responses to all such key criticisms. So we now pass over to a description of the main task characteristics and measurement guidelines of the formal AJ.

1.1 Formal Acceptability Judgments:

Within a formal approach, the AJ task can be viewed as an experimental paradigm designed to measure intuitive assessments on the formation of different strings of linguistic units (KELLER, 1998). As opposed to the informal procedure, the formal AJ needs a representative sample of a given group of speakers, as well as a set of sentences that represent distinct lexicalizations of the construction under scrutiny. Basically, participants are exposed to a set of sentences – target-sentences, control-sentences and distractors – and after viewing/ listening to each of them, they use some kind of scale to indicate how acceptable each sentence is. This methodology allows the application of statistical tests for the verification of hypotheses. Despite their suitability from the point of view of experimental methodology, the formal judgments of acceptability also present methodological aspects that may still need to be improved, and which are still open to debate. Among such aspects, the choice of measurement scale for trial conduction plays a leading role.

There are at least four types of scales that can be used to perform the judgment of acceptability tests: nominal scale, ordinal scale, interval scale and ratio scale (BARD et al, 1996.). As argued by KELLER (1998), the scale used in the elicitation of judgments has crucial importance, since it determines what kind of data will be obtained and what kind of mathematical operations will best suit the ensuing statistical inference procedure.

The nominal scale is used to label different items. The objective is that participants determine whether different items are the same or different with respect to certain properties. It would be possible to use such a scale, for example, to sort words in relation to which language they belong. Needless to say, the items labeled in this test cannot be ordered and, moreover, the study is limited in regard to mathematical operations. It is not uncommon to find nominal scales to describe the acceptability of linguistic constructions (acceptable X unacceptable, ✔x *, etc.), but such an operation hinders the perception of gradient differences between items, as these can only be revealed if a large number of participants are recruited and the statistical frequencies are computed.

The ordinal scale focuses on two properties: order and equivalence. This scale is used to sort different items according to the amount or intensity of a particular property. Despite the fact that it is adequate to test scalar properties, this kind of scale does not assume that there is equality in the interval between the groups, as in the classification adopted by some linguists – “√” “?” “??” “?*” “*” “**” – in introspective judgments. Thus limitations regarding the application of statistical tests must be taken into consideration.

The interval scale, dissimilarly, seems appropriate for judging acceptability tests, precisely because it allows the use of several analytical tools in data treatment. This scale not only assumes that the groups are ordered, but it also assumes that the intervals between them are equal. Thus, it is possible to compare the differences between pairs of items in relation to a given property. Most studies that utilize this scale to carry out an AJ task do so with a Likert Scale (LS) of 5 or 7 points. Usually, the minimum value of these scales represents total unacceptability, while the maximum value is the total acceptability. The midpoint is neutral while the other points express partial acceptability or partial unacceptability. According to FUKUDA et al. (2012), AJ tasks with the LS are user-friendly and generate refined results, which can receive statistical treatment. However, the author also points out that the LS cannot express all distinctions in acceptability as perceived by the participants.

The ratio scale, though, is argued by BARD et al. (1996) to be the most informative scale. While this scale has the same benefits as the interval scale, it also has more freedom and flexibility in the implementation of judgments and, consequently, more fine-grained data. Whereas points on the interval scale – usually represented by numbers – have fixed boundaries, on a ratio scale there are no pre-established lower limit values across integers, making it possible to express very subtle differences between items. Therefore, it is argued that the ratio scale does not place any restrictions on the representation of the acceptability, allowing participants to express all personally perceived differences irrespective of how minute they are.

However, the use of such a scale to measure the acceptability of different constructions is not easy to implement. In response to this challenge, Bard et al. (op. cit.) propose a method originating from psychophysics research called the "magnitude estimation" (ME), which aims to take, to the interval scale, the fine-grained data of the proportional scale (KELLER, 1998). We now turn to a description of both the details and the controversy surrounding this proposal.

1.2 Magnitude Estimation (ME):

The ME was proposed by Stevens (1956[1], apud BARD et al., 1996) in order to measure more efficiently people's perception of different continuous physical stimuli, such as brightness and sound. These perceptions are traditionally measured by a numerical judgment, but can also be represented by lines whose lengths reflect an individual's perception. With the ME, participants compare stimuli proportionally, instead of classifying, ordering, or labeling them as in most other scales. The judgments are made based on a comparison with a standard stimulus, i.e., the standard stimulus has the function of being the basis for estimating the magnitude of all other stimuli.

There are two variants of the ME with respect to presentation of the standard stimulus, or modulus, against which comparisons are to be made (FEITOSA, 1996; SORACE, 2010). In the first, the participants receive the standard stimulus with a fixed value – the modulus – and should estimate the value of the other items based on a comparison with the modulus. The modulus should preferably have an intermediate value and may or may not be visible throughout the task. In the second variant, participants attribute the value of the standard stimulus, whose availability throughout the task is also optional. Imagine, for example, that the items to be compared are different intensities of the same sound. First, each participant assigns a numerical value to the first item, which will be used as the experiment modulus. If the participant assigns the value 50 for the standard stimulus and he/ she perceives the second sound as being 10 times more intense, he/ she should assign the second sound the value 500. Similarly, if the third stimulus seems

[1] STEVENS, Stanley S. The direct estimation of sensory magnitudes: Loudness. *American Journal of Psychology*. 1956. v.69, p.1-25.

to have half the intensity of the module, then it must be assigned the value 25. Instead of classifying the items, the participants compare them in relation to the standard stimulus. Since there is no restriction on the numbers that can be used to express this comparison, participants in the linguistic experiment as explained earlier could not only tell whether a sentence is better or worse than another, but could also point out the numerical degree to which it is better or worse.

It should be noted that one of the peculiarities of the ME is also one of its advantages: the judgments made with this method are relative, i.e., the judgments are made through comparisons. Most other methods require an absolute judgment or, in other words, the judges must use their own points of reference to judge the intensity of a certain property – such as acceptability – in the items being tested. According to Sorace (2010), from a psychometric point of view, the participants tend to be better at relative judgments than at absolute ones.

According WESKOTT & FANSELOW (2011), the extension of this methodology to areas other than perceptual psychology began in the 1980's with studies in the social sciences. Despite being a debated methodology in this area, it remains a methodological option for researches on subjective attitudes. BARD et al. (1996) states that a considerable number of studies have demonstrated that social opinion can also be analyzed by methods and quantitative analysis. SORACE (2010) goes on to state that ME can be applied to validate social scales, which yields a powerful quantitative measure of social opinions.

The ME has also been extended to linguistic studies. More specifically, it has been used to conduct AJ tests. This method has several aspects that make some linguists see it as gold standard for conducting this kind of task. According to SORACE & KELLER (2005), the ME allows linguistic acceptability to be treated as a continuous property and is able to reveal subtle differences in judgments. KELLER (1998) agrees, stating that this method has been applied successfully by different linguists to investigate various linguistic phenomena. Moreover, BARD et al. (1996) and Sorace (2010) argue that normal adults are able to perform this kind of task reliably. They also claim that if the instructions are clear, even less experienced participants are able to provide consistent estimation of magnitude of sentences' relative acceptability. According to these authors, even though the test may seem non-traditional at first sight, participants have the ability to express the proportionality between the sentences' difference in acceptability just as they can express differences in sound intensity.

SORACE (2010) argues that high sensitivity; no restriction of judgment values; relative and gradient judgments; and liability to powerful statistical tests are some advantages of the ME over competitor forms of the AJ task. FURTHERMORE, SORACE & KELLER (2005) advocate the use of the ME for its ability to generate fine-grained data, which enables the investigation of important linguistic aspects such as the differences between soft constraints and hard constraints.

The superiority of the ME compared to other methods, however, is still a questionable point. As noted by FUKUDA et al. (2012), problems that may be present in other scales, such as non-uniform distances between distinct points and limited representation of perceived differences do not seem to be present in the ME method. However, the learnability of a ME task may be severely jeopardized due to its counter-intuitiveness and the mathematical reasoning requirements it imposes on participants.

Therefore, studies aimed at contrasting the data from the ME with others – such as the LS – are of great importance for the validation of AJ methods. BARD et al. (1996) argue that the ME provides more informative data than those extracted from experiments with 7-point scales. WESKOTT & FANSELOW (2011) put this claim to test with a series of three experiments, which aimed to compare the degree of informativeness of a 2-point scale, a 7-point scale, and the ME. In none of the tests in the authors' study was the ME significantly more informative than the other scales. In fact, there were data that pointed to the greater informativeness of the 7-point scale. Thus, the authors consider the claim that the ME is the best method for AJ tasks doubtful.

FUKUDA et al. (2012) conducted a study divided into 3 experiments that aimed at contrasting three AJ task methods: a binary yes/no task, a multiple-point Likert scale and the ME. In general, all methods appear to be sensitive enough to capture the different degrees of acceptability with respect to specific contexts, such as: inversion and non-inversion structure in wh-questions; the *that*-trace effect; sub-extraction from objects, subjects and wh-subjects; and the effect of subject type on wh-questions without inversion. There was data, however, that indicated that only the ME was not able to represent the difference in acceptability between WH-object sentences with and without "that". In this case, either the ME was less sensitive than other methods or the other methods generated an inferential error.

Some authors have pointed out to aspects of the ME that may cause it not to be as superior as other methods as previously thought (FEATHERSTON, 2009; WESKOTT & FANSELOW, 2011; SPROUSE 2008, 2011). SPROUSE (2008) argues that the constant repetition of the module can affect the processing mechanisms of the participants due to a priming effect. Moreover, SPROUSE (2011) suggests that participants do not seem to use the default sentence as a unit of measure. Such influence can cause changes in the patterns of acceptability assigned by the participants.

WESKOTT & FANSELOW (2011) and SPROUSE (2011) present psychophysical studies showing that people do not seem to provide the levels of accuracy advocated by the proponents of the ME method regarding the mental arithmetic involved in this task. These studies have primarily tested if particpants' magnitude estimation of sound intensity presented commutativity and multiplicability, which are the basic assumptions of the ME. Commutativity is the property that makes the order in which successive adjustments are made irrelevant – p ★ (q ★ X) ≈ q ★ (p ★ X) – while multiplicability is the property that makes the result of two successive adjustments equal to one adjustment that is numerically equivalent to the product of the two mentioned adjustments – p ★ (q ★ X) ≈ r ★ X, when p.q = r (SPROUSE, 2011). The results of these studies indicate that the production of the magnitude of sound intensity has commutativity, but not multiplicability. These findings are interpreted by WESKOTT & FANSELOW (2011) as causing several implications to the use of the ME in linguistic studies, especially because according to the authors the scale underlying sound intensity is less complex than the one involved in sentence AJ tasks.

SPROUSE (2011) conducted a study consisting of two experiments to investigate whether ME judgments presented commutativity. The results indicate that less than 20% of participants presented this property in their judgments, i.e., over 80% of the participants performed the task differently from the manner they were instructed. Among the justification presented by the author, there is the absence of a 0-point for the acceptability scale, which is required in this type of judgment. The author suggests that there is no reason to believe that the ME provides more significant data than other methods of AJ.

The present study brings a contribution to this debate by exploring the behavior of two AJ task paradigms in the context of L2 knowledge investigations. We now pass over to the description of the resultative construction of English, and to an exploration of the reasons why it may pose a learnability challenge to speakers of Brazilian Portuguese.

2 The resultative construction:

English and Brazilian Portuguese (BP) present differences and similarities in the semantics assigned to a sentence formed by a noun phrase (NP), a transitive verb, a second noun phrase and an adjectival phrase (AP). This sequence is illustrated by the following sentences:

(1)

a. Lucy ate the salmon raw

b. Lucia comeu o salmão cru

(2)

a. John arrived at the meeting late

b. *João chegou à reunião atrasado*

In both English and BP, the AP can be interpreted as a modifier of either the first or the second NP. In (1), for instance, the AP – raw/ *cru* – modifies the second NP – salmon/ *salmão* – whereas in (2) the AP – late/ *atrasado* – modifies the first NP – John/ *João*. This construction is denominated "depictive" (cf. WECHSLER, 2001; PYLKKÄNNEN & McELREE, 2006) and has as its main characteristic the fact that the AP can be interpreted only as a modifier of one of the arguments of the verb.

Sentences in (1) and in (2) behave similarly in both English and BP. Nevertheless, there are sentences with the same pattern – NP-VP-NP-AP – that map to different interpretations in each of these languages. In English, in a sentence such as (3a) the AP – open – refers the state achieved by the second NP – the package – as a result of the action described by the verb, i.e., George tore the package until it became open. This construction in which the AP describes the result of an action is denominated "resultative." In BP, however, a sentence such as (3b) triggers a different interpretation: it only allows for a depictive reading. In other words, the AP – *aberto* – refers to the state of the second NP – *pacote* – during the action described by the verb.

(3)

a. George tore the package open

b. *Jorge rasgou o pacote aberto*

Both the depictive and the resultative constructions are licensed in English (WECHSLER, 2001; GOLDBERG & JACKENDOFF, 2004), but only the former seems to be licit in BP (MARCELINO, 2014), as discussed. This difference between the grammar of English and BP has been hypothesized to reflect contrastive settings of the compounding parameter (SNYDER, 1995[2], apud MARCELINO, 2014). Such parameter is proposed as clustering the licensing of a number free-standing open class morpheme compounds in languages that allow for its positive setting. Such languages are languages that permit complex predicates like the resultiative construction.

2.2 The Resultative Construction in English

According to Goldberg and Jackendoff (2004), the English resultative construction has been a major focus in research related to the interface between syntax and semantics. According to the authors, in this type of construction there are two sub-events. The verb phrase determines a sub-event, whereas the resultative construction as a whole determines the other. In order for both sub-events to occur in harmony, it is necessary that the arguments licensed by the verbs and the arguments licensed by construction share some syntactic positions. In (4), for instance, the sub-event determined by the construction is "BILL CAUSE [TULIPS BECOME FLAT]", in which "Bill" occupies the subject position and "flat" occupies one of the internal argument positions. Such sub-event occurs through the second sub-event "BILL WATER TULIPS" whose subject coincides with the one in the first sub-event and therefore the subject position is shared. The NP "tulip" occupies the other internal argument. Thus, all thematic roles are performed harmoniously.

[2] SNYDER, W. *Language Acquisition and Language Variation: The Role of Morphology*. Ph.D. Dissertation. Cambridge, MA: The Massachusetts Institute of Technology, 1995.

(4) Bill watered the tulips flat

As presented by GOLDBERG and JACKENDOFF (2004), both the semantics and the syntax of the resultative construction can vary. In turn, the authors propose that the resultatives should be considered a family, which can be divided into different sub-constructions. In this study, our focus will be only in the subconstruction instantiated by (4), which is formed by a subject (NP), a direct transitive verb (VP), a direct object independently selected by the verb (NP), and a resultative predicate expressing property (AP).

This resultative sub-construction must be telic in order to be grammatical, but not all APs are able to provide a sentence with this property, as maintained by WECHSLER (2001). According to the author, a sentence with a durative verb, such as (5), is atelic. Adding an AP to the sentence may make it telic (6) or keep it atelic (7). Wechsler goes on to contend that a telic event needs to have three aspects: an affected theme, a property scale, and a bound, related as follows: "Some property of the affected theme argument changes by degrees along a scale due to the action described by the verb, until it reaches a bound." (WECHSLER, 2001, p.7). The resultative construction, therefore, requires APs that are capable of constituting a bound.

(5) He wiped the table

(6) He wiped the table dry / clean

(7) *He wiped the table wet / dirty

The only differences between (6) and (7) are the APs that constitute the resultative predicate. Both sentences have a theme – table, which is transformed by an action – wipe – and acquire a scalar property. In

(6) the APs can be considered a definite bound because they are the maximum point of their own scale since "dry" and "clean" represent respectively 0% dirt and 0% liquid. The APs in (7) are different because neither "wet" or "dirty" represent the maximum value of their own scale. They basically represent any value above 0% and, in turn, cannot impose a bound for the action. In sum, in order for a sentence instantiating the resultative construction to be telic/ grammatical, it needs an AP that represents the maximum point of a scale property.

2. 3 The Resultative Construction in Brazilian Portuguese

Some studies have investigated the presence of the resultative construction in BP. For FOLTRAN (1999), for instance, the BP resultative construction is exemplified by sentences like (8) and (9). MARCELINO (2000), differently, argues that in BP the resultative predicate is formed by the adverb "*bem*" (well) followed by an AP that represents the change that the object NP has undergone. Moreover, this AP should contain the diminutive suffix "*-inho*", as illustrated in (10).

(8) Ele fez o chá fraco.
He made the tea weak

(9) Ele construiu a casa muito grande.
He built the house very big

(10) Joana picou o papel bem picadinho.
Joana cut the paper well cut(DIMINUTIVE)

LOBATO (2004) also argues for the existence of the resultative construction in BP. Besides the structure shown in (10), according to the author, there are at least four major groups within the resultative construction in BP, which are illustrated below:

(11) Deus criou os homens fracos.
God created the men weak.

(12) A manteiga congelou torta.
The butter froze crooked.

(13) João pintou a casa torta.
John painted the house crooked

(14) Ele cortou o cabelo curto.
He cut the hair short.

MARCELINO (2014) compares these potential resultative structures in order to verify if they have the same properties of the English resultative sentences. The author proposes that these two languages differ from each other in two main aspects.

First, the English resultative sentences are formed by an activity verb, but the sentence has an accomplishment reading due to the telicity generated by the resultative predicate. In turn, telic modifiers such as "in an hour" can be added to these sentences (15). However, if the resultative predicate is removed from these resultative sentences, they become atelic and it is no longer possible to add telic modifiers such as "in an hour" (16). In BP, however, the sentences that have been presented as instances of the resultative construction do not have an activity verb. Instead, they are formed by an accomplishment verb which makes the sentence telic regardless of the presence of the APs. Thus, the presence of the APs does not have an impact on the grammaticality status of these so-called BP resultatives, as illustrated in (17) and (18).

(15) He hammered the metal flat in an hour

(16) *He hammered the nail in an hour.

(17) Ela costurou a saia bem justinha em uma hora.
She sewed the skirt well tight(DIMINUTIVE) in an hour

(18) Ela costurou a saia em uma hora.
She sewed the skirt in an hour
(MARCELINO, 2014)

Second, MARCELINO (2014) shows that BP and English respond differently to the how-test. Notice that in English the resultative predicate cannot be utilized as an answer to a how-question, i.e., a question about the manner the action was performed (19). In BP, dissimilarly, the APs can be utilized as answer to the how-question (20).

(19) How did he hammer the nail? (*flat) (slowly/ rapidly)

(20) Como ela custurou a saia? (bem justinha)
How did she sew the skirt? (well tight(DIMINUTIVE))
(MARCELINO, 2014)

MARCELINO (2014) argues that in this case we are dealing with two different constructions: the "adverbial resultatives" and "true resultatives". The sentences that have been proposed as instances of the resultative construction in BP are, in reality, "adverbial resultatives" because instead of denoting a resultant state, their AP modifies the resulting state, which is indicated by the verb itself. "Adverbial resultatives" can also be found in English in sentences such as (21). Sentences such as "he hammered the metal flat" are instances of the "true resultatives". They have an activity verb and their telicity depends on the presence of the resultative predicate. "True resultatives" are licensed and productive in English, but not in BP.

(21) How did he cut the meat? (thick)
He cut the meat thick

2.4 Learnability of the resultative construction for BP-English bilinguals

From a second language acquisition perspective, verification of the development of representations that support the interpretation of the English resultative by L2 users of English whose L1 is Brazilian Portuguese will be of theoretical interest. The contrast between the resultative construction in English and in Brazilian Portuguese, as well as the subtle semantic requirements for the licensing of such construction in English, imposes non-trivial cognitive operations to the second language learner.

Let us assume, together with MARCELINO (2007, 2014), that the "true" English resultative emerges from a positive setting of the compounding parameter. Such assumption entails that evidence of L2 learning of the English resultative by native speakers of Brazilian Portuguese would constitute evidence suggestive of behavior that resembles parameter re-setting in L2 acquisition. Whether or not second language learners do re-set L1 parameter values has been a core empirical question within generative studies of second language acquisition. Such question has been theoretically framed as related to the ultimate question of whether second language acquisition is guided or not by access to principles of Universal Grammar (for a review see WHITE, 2003).

In the case of the English resultative construction the task the second language learner would face outreaches the level of morphemic compositionality, however. As discussed above, there are complex semantic configurations related to event structure that need to be computed for a grammatical resultative to ensue. Therefore, not only would the learner need to figure out the right mapping of syntactic structure to semantic reading, but also he/she would need to figure out the right event structure that supports the construction.

Another empirical question in second language acquisition research is whether or not fine semantic configurations are prone to be learnable. SLABAKOVA (2006, 2008) defends that there is no maturational constraint that would impede L2 learners from acquiring syntax and semantics of a non-primary language. To the researcher, the bottleneck of second language acquisition is functional morphology proper. So, according to his perspective, the learnability of the resultative construction is predicted.

Based on these facts about the resultative construction in English and BP, as well as based on the challenges it imposes for second language acquisition, the present study sought to test the behavior of the two AJ tasks as psychometric instruments to probe the following two learnabiliy questions:

(i) Do bilinguals of BP and English with high proficiency in their L2 accept the English resultative construction?

(ii) Do they learn the resultative predicate rules, therefore rejecting unlicensed APs?

3 Method

In order to investigate the acquisition of the resultative construction by high-proficient BP-English bilinguals, we carried out two acceptability judgment tasks. As observed above, this experimental procedure for eliciting responses to verbal stimuli is a practical way to verify the existence of mental representation of aspects of the grammar since it does not depend on the observation of spontaneous speech. The task in Experiment 1 was conducted in person with a 7-point Likert scale, whereas the one in Experiment 2 was conducted online with the Magnitude Estimation

3.1 Experiment 1

3.1.1 Participants

In total, there were 25 participants in Experiment 1, who were English majors at the School of Letters at the Federal University of Minas Gerais. They were bilinguals whose English language learning process had occurred in contexts of formal education in a society that does not have English as the dominant language for social interactions.

All the selected participants had scores close to the maximum in a test of English vocabulary knowledge – Vocabulary Levels Test, or VLT (NATION, 1990). In this test, participants perform associations between lexical items and meanings, and their lexical competence is ranked in a five-band scale that reflects access to lexical items of progressively decreasing frequency in corpora of general English. In other words, test takers scoring in the highest bands will have shown capacity to figure out meanings of lexemes of relatively low frequency, a behavioral trait that is assumed to tap into the size of the mental English L2 lexicon. In this study, the VLT was administered with a time limit of ten minutes, procedure adopted in order to increase the discriminatory effect of the test. The selected participants were in levels 4 or 5 or the VLT. The underlying assumption of this screening criterion is that high levels of competence in L2 lexical access are associated with high levels of proficiency in this language.

3.1.2 Materials

The AJ task contained 56 sentences to be judged in relation to their acceptability. They were balanced in terms of grammaticality, so that 50% of the sentences were grammatical and 50% were ungrammatical. 16 of the sentences were the target stimuli (sentences instantiating the resultative construction). 8 of these sentences were grammatical

– as illustrated in (22) – and 8 were ungrammatical according to the rules proposed by WECHSLER (2001) – as illustrated in (23). All the sentences had the pattern NP-VP-NP-AP and were formed mostly by words that are among the most 2000 frequent ones according to the Corpus of Contemporary American English[3].

(22) The driver loaded his car full.

(23) *The farmer burned the wood dark.

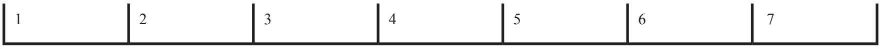
3.1.3 Procedures:

Participants were tested in groups at the School of Letters at the Federal University of Minas Gerais. After obtaining the participants consent and applying the VLT, the experimenter explained the participants how the task should be performed and conducted a training session.

The AJ task sentences were displayed on a white screen by a data projector. Sentences were displayed one-by-one through a Microsoft Powerpoint slideshow presentation. Slides moved along automatically after a 9-second display. Therefore, participants had a 9-second ceiling of exposure to each sentence, and were instructed to provide their judgments within this ceiling. As soon as the training session was over, participants started the acceptability judgment task, which lasted for about 8 minutes. Items were numbered, and participants recorded their judgments on correspondingly numbered 7-point Likert scales (Fig. 1) on a printout.

[3] Available at http://www.americancorpus.org/. The 2000 most frequent lexeme threshold was based on the first band of the VLT, which covers precisely such frequency plateau. Therefore, for the design of the present study the participant screening criterion combined with the lexical control of stimuli as a measure to maximally reduce the possibility that performance on the AJ tasks was modulated by L2 vocabulary difficulty.

FIGURE 1: 7-point Likert scale



3.2 Experiment 2

3.2.1 Participants

28 subjects comprised the group of participants in Experiment 2. The profile of these participants was similar to those who were in Experiment 1. They were also students and high proficient BP-English speakers whose dominant language was BP. These subjects also performed an online version of the VLT and were rated at levels 4 or 5.

3.2.2- Materials:

80 sentences were presented to the participants in Experiment 2. As in Experiment 1, the items were controlled in terms of grammaticality and frequency. 16 of the sentences were the target stimuli, 8 of which were grammatical – as illustrated in (24) – whereas the other 8 had an unlicensed AP – as exemplified in (25).

(24) One of the classrooms was very dirty, so Desiree swept it clean

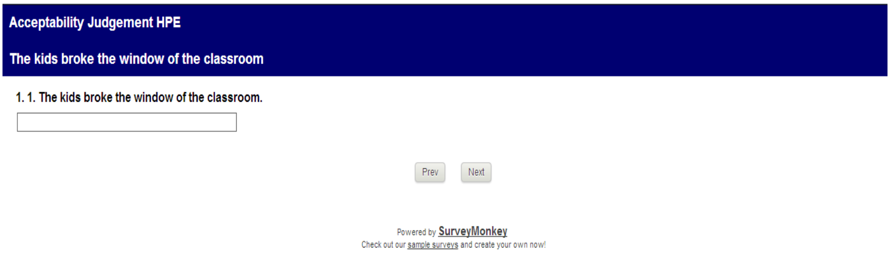
(25) *Chelsea had straightened her hair, but her little brother watered it curly.

3.2.3 Procedure

Experiment 2 was conducted online on the website Survey Monkey (www.surveymonkey.com), which offers enough tools for experiments using the magnitude estimation. 8 pages were created to present the experiment instructions in a succinct, direct and exemplified manner.

After this, the stimuli were presented one-be-one without a time limit. Below them there was a space for the participants to type the number representing their judgment, as illustrated in Figure 2.

FIGURE 2: Presentation of Experiment 2 first item.



Acceptability Judgement HPE

The kids broke the window of the classroom

1. 1. The kids broke the window of the classroom.

The procedure was relatively simple. First, the participants assigned a number to the first sentence, which remained visible throughout the whole task. The following sentences were judged in comparison to the module – the number that represented the judgment of the first sentence. Thus, the participants performed proportional judgments, as not only did they judge whether a sentence was more or less acceptable than the standard sentence, but they also expressed through their numbers how many times more or less acceptable the sentence was.

4 Analysis and discussion.

For comparison of the measures o AJ elicited by the two forms of the AJ task described above, both the individual scores obtained through the Likert scale and the individual scores obtained in the magnitude estimation mode were converted into scores ranging from 0 to 1. Two subjects with incomplete sets of judgments from the data set obtained with the Likert scale AJ task were excluded from the analysis. After such conversion, valid subjects' means across all critical items and items' means across all valid subjects were compiled from the raw data of both variants of the AJ task. The data is described in the table below.

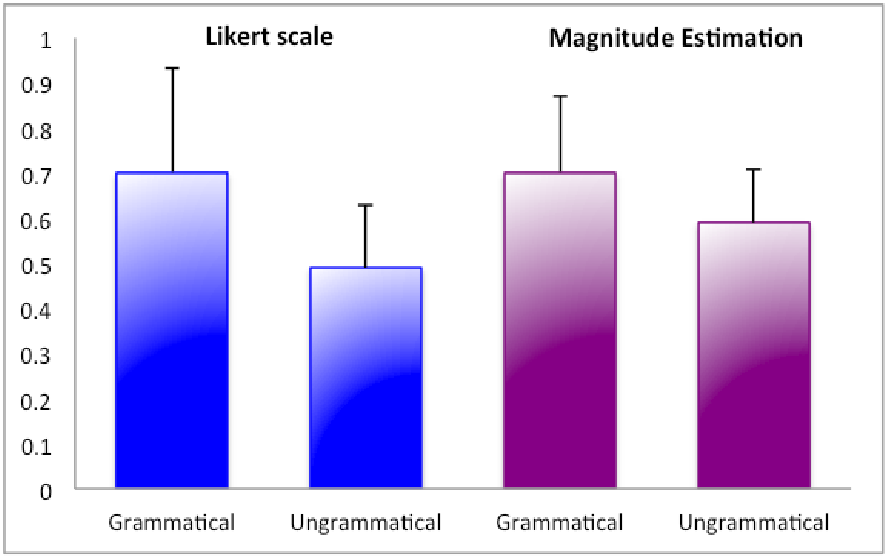
TABLE 1: Means and standard deviation of the judgment data sets from the two tasks.

Judgment data	Means	SD
Grammatical resultatives – Likert scale	0.70	0.23
Grammatical resultatives – Magnitude Estimation	0.70	0.14
Ungrammatical resultatives – Likert Scale	0.49	0.17
Ungrammatical resultatives – Magnitude Estimation	0.59	0.12

The compiled means were tested for normality with the Kolmogorov-Smirnov test. The Likert scale set of means did not differ significantly from the normal distribution for either the grammatical sentences (K-S=0.15; p=0.17) or the ungrammatical sentences (K-S= 0.10; p=0.63). Likewise, the magnitude estimation set of means did not differ significantly from the normal distribution for the grammatical sentences (K-S=0.13; p=0.21), and only differed marginally for the ungrammatical sentences (K-S=0.16; p=0.07).

As stated above, the primary goal of the present study was a comparison the sensitivity of AJ task conducted according to the magnitude estimation paradigm with the sensitivity of a speeded version of the Likert scale AJ task to capture L2 knowledge. To achieve such aim, we investigated the learnability of constraints on the English resultative construction by bilinguals of Brazilian Portuguese and English. We wanted to investigate if such bilinguals were not only capable of categorizing instances of such construction as well-formed sentences in English, but also capable of telling the truly well-formed instances from others that violated the subtle semantic constraints that modulate the well-formedness of the construction. A constrasting view of the resulting subject-based means obtained for both the grammatical and the ungrammatical sentences through the two forms of the AJ task applied in this study are displayed in the following graph.

GRAPH 1: Grammatical and ungrammatical resultaive sentences



Both the AJs collected through the 7-point Likert scale task and the AJs collected through the magnitude estimation task reveal a clear differentiation between the grammatical and the ungrammatical resultative sentences. Pairwise T-tests were performed with the data sets from both tasks. The 7-point Likert scale task yielded a significant difference in the judgments of the grammatical and ungrammatical sentences by subjects (t1=4.96 (df=22), p<.001), and marginally significant difference by items (t2=2.2 (df=7), p=.06). The magnitude estimation task yielded a similarly significant difference in the judgments of the grammatical and ungrammatical sentences by subjects (t1=5.68 (df=27), p<.001), but no significant difference by items (t2=1.64 (df=7), p=.14).

In order to further scrutinize the comparability of the two forms of the AJ tasks examined in this study, we compared the recorded mean judgments for both grammatical and ungrammatical stimuli. As can be seen in the graph, the mean of AJs obtained grammatical sentences was practically identical in both tasks. Therefore, they did not yield statistically significant differences either by subjects (t1=-.15 (df=25), p=.88) or by items (t1=-.27 (df=7), p=.79). When the AJs elicited by the ungrammatical sentences are compared, the data set produced by the magnitude estimation task achieved a subject-based mean judgment that was 0.1 point higher than the mean judgment obtained by way of the Likert scale task. This difference was enough to yield a significant difference by subjects (t1=-22.46 (df=25), p<.05), but not by items (t2=-1.34 (df=7), p=.22).

The comparison of the two forms of the acceptability judgment task conducted in this study revealed that they were equally successful in demonstrating the learnability of the resultative construction in English as an L2 by bilinguals whose L1 is Portuguese. Even though the results of the magnitude estimation version of the task showed a significantly less pronounced rejection of the ungrammatical set of resultative sentences, such rejections gathered as a set of mean responses significantly different from the level of acceptability granted to the licensed counterparts. We understand this difference to be negligible, as the learnability of the construction would have been testified even if only the magnitude estimation data had been considered. The two data sets clearly tell a very similar case of learnability of the resultative construction in the context of English as an L2.

The relative loss of power to discriminate between the grammatical and the ungrammatical resultatives that was observed in the magnitude estimation data could be explained by a number of plausible factors. First of all, it can be the case that the subject pool that participated in Experiment 2 had a higher proportion of individuals who have not fully acquired the relevant distinction. Also, it can be the case that the online delivery of the task played a role in decreasing the level of attention to task that was fostered in the face-to-face experiment. Finally, at the present stage we cannot rule out the possibility that the task requirements of the magnitude estimation task had an effect on the participants' judgments that may not have been detected had the task been a simpler, more straightforward one.

The general picture that emerges from the data is that the bilinguals of Brazilian Portuguese and English in our two samples are capable of categorizing the resultative construction of English as a grammatical construction of their L2. This shows that this particular population of bilinguals, who demonstrate a high level of proficiency in the second language in an independent measure of L2 lexical knowledge, is able to depart from the restriction of their native language grammar. In other words, they exhibit parameter resetting-like behavior with respect to at least this particular construction, as they seem to acquire a construction that can be linked to a parameter absent from their L1.

Furthermore, the data sets analyzed in the present study indicate that the bilingual participants are also capable of systematically perceiving a distinction between the type of resultative construction that is actually licensed in English and the type that is not grammatical. This is evidenced by the fact that the two AJ tasks yielded mean judgments that were significantly different for the two types of resultative sentences. Such capacity to differentiate the two types of sentences suggests that the bilinguals are sensitive to the semantic distinctions that modulate the licensing of grammatical resultatives in English. L2 acquisition of this kind of fine-grained semantic configuration is consistent with SLABAKOVA'S (2006, 2008) proposal that L2 semantic specifications are ultimately learnable by adult L2 learners, as according the researcher there is no evidence of a critical period for the L2 acquisition of semantic constraints.

5 Concluding remarks:

For any long-term scientific enterprise, comparability of results obtained through different protocols for data gathering and analysis is an extremely important issue. Such comparability is after all a prerequisite for cumulative advancement of knowledge. The present study sought to contribute to the establishment of comparability of methods in experimental syntax by examining their application to a particular linguistic profile: L2 speakers.

Our data demonstrated that both a traditional Likert-scale speeded judgment task and a magnitude estimation task were equally successful at revealing the learnability of the resultative construction in English L2 by native speakers of Brazilian Portuguese. These findings attest to the psychometric potential of these techniques for the context of bilingualism studies. Specifically, they proved to be trustworthy in the exploration of what is likely to be implicit linguistic knowledge, as the details that govern the licensing of grammatical resultatives as opposed to ungrammatical resultatives in English are too subtle to have been acquired through explicit training following the typical English as a Foreign Language program syllabus.

Notwithstanding the interest of these observations concerning the learnability of the resultative construction by bilinguals of Brazilian Portuguese and English, important questions remain to be answered. Specifically, more detailed explorations of whether or not such knowledge is accessed automatically must be conducted. The time frame for trials in the two forms of the AJ task employed in the present study was too broad for any claims about processing to be made: 9 seconds in the Likert-scale task, and an indefinite temporal ceiling in the magnitude estimation task. Therefore, investigations of the resultative construction that employ a speeded version of AJ task set at the minimum temporal ceiling for judgments, and studies that look at online processing are due. After all, as argued in CARNEIRO & SOUZA (2012), information revealing the mechanism of online processing and its cost is critical for experimental studies of L2 syntax, since it is the speed and fluidity of access to and manipulation of linguistic representations during language processing that will in great part determine the degree of skill in second language use.

The magnitude estimation paradigm has been proposed as a technique that qualitatively surpasses Likert-scale AJ tasks, especially because it may be both more powerful to capture acceptability gradience and more readily adequate to parametric statistic tests. However, the results of the present study actually suggest that the Likert-scale AJ task provided a finer distinction of the acceptability difference in question. SORACE (2010) points out some negative points of the ME such as low face validity, the restriction on the applicability in some contexts – because of the possible need for longer training sessions, restricting applicability to certain linguistic profiles and data that may show individual variation. This last point is also discussed by WESKOTT & FANSELOW (2011), who argue that the freedom which is given to judges with the ME can cause a within-subject variance that does not contribute to the interpretation of the variance between the means of experimental conditions in the statistical analysis. The present study is yet another one that fails to depict the alleged promises of the magnitude estimation paradigm for acceptability judgment tasks as truly justifiable.

References:

BARD, Ellen G.; ROBERTSON, Dan; SORACE, Antonella. **Magnitude estimation of linguistic acceptability**. Language. 1996. v.27, n.1, p.32-68.

BARILE, Wendy; MAIA, Marcus. **Aspectos prosódicos do Qu in-situ no português brasileiro**. ReVEL. 2008. v.6, n.11, p.1-21.

CARNEIRO, Marisa; SOUZA, Ricardo. **Observação do processamento online: uma direção necessária para o estudo experimental da sintaxe bilíngue**. Revista Virtual de Estudos da Linguagem. 2012. v. 10, n. 18, p. 107-127.

COWART, Wayne. **Experimental Syntax:** Applying Objective Methods to Sentence Judgments. Thousand Oaks: Sage Publications, 1997.

CULBERTSON, Jennifer; GROSS, Steven. **Are Linguists Better Subjects?** British Journal for the Philosophy of Science. 2009. v.60, p.721-736.

DRABOWSKA, Ewa. **Naive v. expert intuitions**: An empirical study of acceptability judgments. The Linguistic Review. 2010. v.27, n.1, p.1-23.

FEATHERSTON, Sam. **Relax, lean back, and be a linguist**. Zeitschrift Für Prachwissenschaft. 2009. v.28, p.127-132.
______. **Magnitude estimation and what it can do for your syntax**: Some WH-Constraints in German. Lingua. 2005. v.115, p.1525-1550.

FEITOSA, Maria Ângela Guimarães. **Teoria e métodos em psicofísica**. In: PASQUALI, Luiz. Teoria e métodos de medida em ciências do comportamento. Brasília: Editora da Universidade de Brasília. 1996. p.43-71.

FERREIRA, Fernanda. **Psycholinguistics, formal grammars, and cognitive science**. The Linguistic Review. 2005. v.22, p.365-380.

FOLTRAN, Maria José Gnatta Dalcuche. **As construções de predicação secundária no português do Brasil: aspectos sintáticos e semânticos**. São Paulo: USP. 1999. 205f. Tese de Doutorado em Linguística, Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo.

FUKUDA, Shin; GOODALL, Grant; MICHEL, Dan; BEECHER, Henry. **Is magnitude estimation worth the trouble**? In: CHOI, Jaehoon; HOGUE, Alan; PUNSKE, Jeffrey; TAT, Deniz; SCHERTZ, Jessamyn; TRUMAN, Alex. Proceedings of the 29th West Coast Conference on Formal Linguistics, Somerville: Cascadilla Proceedings Project, 2012. p.328- 336.

GIBSON, Edward; FEDORENKO, Evelina. **The need for quantitative methods in syntax and semantics research**. Language and Cognitive Processes. 2013. v.28, n.1/2, p.88-124.

GOLDBERG, Adele; JACKENDOFF, Ray. **The English Resultative as a family of constructions**. Language. 2004. v.80, p.523-567.

GUTIÉRREZ, Xavier. **The construct validity of grammaticality judgment tests as measures of implicit and explicit knowledge**. Studies in Second Language Acquisition. 2013. v.35, p.423-449 .

KELLER, Frank. **Grammaticality judgments and linguistic methodology**. Centre for Cognitive Science University of Edinburgh. Research Paper EUCCS-RP. 1998. p.1-15

LOBATO, L. **Afinal, existe a construção resultativa em português?** In: NEGRI, Lígia; FOLTRAN, Maria José; OLIVEIRA, Roberta Pires. Sentido e Significação. São Paulo: Contexto. 2004. p.142-181

MAIA, Marcus. **Sintaxe experimental**: uma entrevista com Marcus Maia. Revista Virtual de Estudos da Linguagem. 2012. v.10, n.18, p.184-193.

MANDELL, Paul. **On the reliablity of grammaticality judgment tests in second language acquisition research**. Second Language Research. 1999. v.15, n. 1, p.73-99.

MARCELINO, Marcello. **Resultativas em português brasileiro**. Veredas. 2014. v.18, n.1, p.121-137.
______. **O Parâmetro de Composição e a Aquisição/Aprendizagem de L2**. Campinas: UNICAMP. 2007. 211f. Tese de Doutorado em Linguística, Institudo de Estudos da Linguagem, Universidade Estadual de Campinas.
______. **Construções Resultativas em Português e em Inglês**: Uma Nova Análise. São Paulo: PUC-SP. 2000. 97f. Dissertação de Mestrado em Linguística Aplicada e Estudos da Linguagem. Faculdade de Filosofia, Comunicação, Letras e Artes, Pontifícia Universidade Católica de São Paulo.

MYERS, James. **Syntactic judgment experiments**. Language and Linguistics Compass. 2009a. v.3, n.1, p.406-423.
______. **The design and analysis of small-scale syntactic judgment experiments**. Lingua. 2009b. v.119, p.425-444.

NATION, Paul. **Teaching and Learning Vocabulary.** Boston: Heinle & Heinle. 1990.

PHILLIPS, Colin. **Should we impeach armchair linguists?** In: IWASAKI, Shoichi; HOJI, Hajime; CLANCY, Patricia M.; SOHN, Sung-Ock. Japanese-Korean Linguistics 17. Stanford: CSLI Publications. 2010. p.49-64.

PYLKKÄNNEN, Liina.; McELREE, Brian. **The syntax-semantics interface**: On-line composition of sentence meaning. In: TRAXLER, Matthew; GRENSBACHER, Marton. The Handbook of Psycholinguistics – 2nd Edition. London/Burlington: Academic Press. 2006. p.1-69

SCHÜTZE, Carson. **The Empirical Base of Linguistics:** Grammaticality Judgments and Linguistic Methodology. Chicago: University of Chicago Press. 1996.

SLABAKOVA, Roumyana. **Meaning in the Second Language**. Berlin: Mouton de Gruyter, 2008.
____. **Is there a critical period for the acquisition semantics?** Second Language Research. 2006. v. 22, n. 3, p. 302-338.

SNYDER, William. **An experimental investigation of syntactic satiation effects**. Linguistic Inquiry. 2000. v.31, n.3, p.575-582.

SORACE, Antonella. **Using Magnitude Estimation in developmental linguistics**. IN: BLOM, Elma; UNSWORTH, Sharon. Experimental Methods in Language Acquisition Research. Amsterdam/Philadelphia: John Benjamins. 2010. p.57-72

SORACE, Antonella; KELLER, Frank. **Gradience in linguistic data**. Lingua. 2005. v.115 p.1497-1524.

SPROUSE, Jon. **A test of the cognitive assumptions of magnitude estimation**: Commutativity does not hold for acceptability judgments. Language. 2011. v.87, n.2, p.274-288.
______. **Magnitude Estimation and the Non-Linearity of Acceptability Judgments**. In: ABNER, Natasha; BISHOP, Jasoon. Proceedings of the 27th West Coast Conference on Formal Linguistics. Somerville, MA: Cascadilla Press. 2008. p.397-403.

SPROUSE, Jon; SCHÜTZE Carson T.; ALMEIDA Diogo. **A comparison of informal and formal acceptability judgments using a random sample from Linguistic Inquiry 2001-2010**. Lingua. 2013. v.134, p.219-248.

WECHSLER, Stephen. **An analysis of English resultatives under the event-argument homomorphism model of telicity**. Proceedings of the 3rd Workshop on Text Structure. University of Texas, Austin. 2001. p.1-15

WESKOTT, Thomas; FANSELOW, Gisbert. **On the informativity of different measures of linguistic acceptability**. Language. 2011. v.87, p.249-73.

WHITE, Lydia. Second Language Acquisition and Universal Grammar. Cambridge: Cambridge University Press, 2003.