Corela
Cognition, représentation, langage

12-1 | 2014
Vol. 12, n° 1

Magnitude Estimation: can it do something for your pragmatics?

Rudy Loock and Cyril Auran

Electronic version
URL: http://journals.openedition.org/corela/3406
DOI: 10.4000/corela.3406
ISSN: 1638-573X

Publisher
Cercle linguistique du Centre et de l'Ouest - CerLICO

Electronic reference
Rudy Loock and Cyril Auran, « Magnitude Estimation: can it do something for your pragmatics? », *Corela* [Online], 12-1 | 2014, Online since 30 June 2014, connection on 19 April 2019. URL : http://journals.openedition.org/corela/3406 ; DOI : 10.4000/corela.3406

Magnitude Estimation: can it do something for your pragmatics?

Rudy Loock and Cyril Auran

This material was presented at the 5th AFLiCo conference in Lille, France, in May 2013; we would like to thank the audience for useful comments and suggestions. In addition, we would like to thank all our informants, as well as colleagues who helped us get in contact with them: Kathleen O'Connor, Joseph Tyler, and above all Doron Cohen from the University of Manchester. Special thanks also go to Erin Myers and Aaron Hutson for their help with the recording of the stimuli. Finally, we thank the anonymous reviewers of the article for their insightful comments, and Kathleen O'Connor for her proof-reading.

1 The aim of this article, whose title echoes that of Featherston (2005), "Magnitude estimation and what it can do for your syntax: some wh-constraints in German", is to determine whether the Magnitude Estimation protocol can be implemented in the field of discourse pragmatics to provide scientifically robust, finer-grained evaluations of data, in particular data considered "negative" or pragmatically unacceptable. Our pilot study comes at a time when researchers feel the need for more reliable grammaticality/acceptability judgments to avoid debatable, biased introspective data (see e.g. Schütze 1996, Bard et al. 1996) and are searching for more scientific, objective means to evaluate them (speeded grammaticality judgments, Likert n-point scales). This article intends to find out whether the Magnitude Estimation protocol, which has its origins in psychophysics (Stevens 1957) but has also been used in phonology (e.g. Grosjean and Lass 1977, Green 1987, Toner and Emanuel 1989) and syntax, (e.g. Keller and Alexopoulou 2005, Hoffman 2007b, Bard et al. 1996, Keller and Asudeh 2001), can be used for acceptability judgments in discourse pragmatics, which relies heavily on negative data (labeled '#') as evidence for definitions of discourse functions and constraints. The case studies retained for this pilot experiment are based on Birner and Ward's (1998) data on non-canonical word order (topicalization phenomena specifically) and on Loock's (2010b) definition of the "fame effect" as a constraint governing the choice between a nominal appositive and an appositive relative clause (*Barack Obama, president of the US* vs. *Barack Obama, who is*

president of the US). Through a comparison of the introspective judgments from these studies with judgments obtained using a Magnitude Estimation experiment, we discuss the relevance of the protocol for pragmatic acceptability judgments, which are often more subtle and thus more subject to debate (Ariel 2008).

2 The article is organized as follows. In the first section, we provide some background context on the need for more reliable data in general, and more reliable judgments in particular, and how this need has been dealt with in recent literature. We then show in section 2 that the Magnitude Estimation protocol has interesting potential for exploring the data from which conclusions are drawn. Thus after a discussion of the origins of the technique, we examine how it has been used quite successfully in phonology and syntax, but has remained marginal in pragmatics. In sections 3 and 4, we report on the results of a pilot study carried out on data extracted from two pragmatics-based studies; one of them focuses on the constraints governing the felicitousness of non-canonical word order in English sentences (Birner and Ward 1998) while the other is on a familiarity constraint (called the "fame effect") governing the alternation between nominal appositives and appositive relative clauses (Loock 2010b). We finally discuss the results and try to determine to what extent the technique can be successfully used for studies within the discourse pragmatics field. Note however that this is an exploratory study with its specific limitations.

1. The need for more reliable acceptability judgments

1.1. Acceptability vs. grammaticality judgments

3 One of the aims of linguistic studies is to determine what belongs or not to a linguistic system. A description of the data in terms of grammaticality and semantic or pragmatic felicity is essential whether the approach is formal or functional. Researchers thus often rely on so-called 'negative data' to draw conclusions and define the constraints that govern the system of a language, whether the data are ungrammatical (labeled '*') or semantically/pragmatically infelicitous (labeled '#') because they do not fit into the linguistic co-text and/or context or because they are difficult to process. Consider the following:

(1) *Sheldon gave flowers Amy.
(2) #Sheldon gave the flowers that he had picked up in one of the neighbors' front garden before getting back to his apartment on the fourth floor in the morning to Amy.

4 Over the past twenty years, researchers have come to realize that such judgments are actually problematic. The legitimacy of grammaticality judgments has been called into question, since the judges, whether researchers or not, can only evaluate the acceptability of data, and not its grammaticality, meaning that the term 'grammaticality judgments' is a "misnomer" (Schütze 1996: 6). In this view, only acceptability judgments exist, and these can then be correlated with grammaticality (well-formedness) or felicity (suitability to situational context and linguistic co-text). We shall not delve into the debate of whether grammaticality judgments exist alongside acceptability judgments. As our pilot experiment addresses the evaluation of pragmatic data, we will refer to judgments on data as 'acceptability judgments', but like Schütze (1996), we use the same

cover term for judgments meant to account for both the grammaticality and acceptability of some data.

1.2. The need for more reliable judgments

5 Judgments, whether obtained introspectively or through questionnaires with informants, are known to be problematic because they can be influenced by many external parameters such as prescriptivism, frequency of the linguistic phenomenon (Bader and Häusler (2010: 15-316)), extra-linguistic plausibility (cf. Hill 1960), but also the researcher's bias. For this reason, it is not rare for acceptability judgments in the literature to be criticized and called into question, with differences in acceptability judgments leading to differences in theoretical conclusions (see Schütze 1996 for a series of examples). Furthermore, the reliability of judgments is problematic, since researchers have shown that acceptability is a matter of gradience rather than a binary opposition between acceptable and non-acceptable (e.g. Sorace and Keller 2005). Researchers then resort to different labels and include '?' or '??' alongside '*' for ungrammatical and '#' for pragmatically unacceptable, but they have proved to be inconsistent within the same field or even in their own work. Bard et al. (1996) mention the example of Haegeman (1991), who uses 4 different labels (Ø/?/*/**), while Müller (1995) uses 5 (Ø/ ?/ ??/* ?/*), Lakoff (1973) uses 6 (Ø/ ?/ ??/ ?*/*/**) and Wurmbrand (2001) and Belletti and Rizzi (1988) use up to 7 (Ø/#/%/ ?/ ?/ ??/* and Ø/(?)/?/??/*?/(*)/* respectively). These examples are not exceptional and have led to great inconsistency and endless debates about theoretical conclusions due to a lack of irrefutable acceptability judgments that take gradience into account.

6 Data evaluation is particularly problematic for discourse pragmatics. Even though the grammaticality of some data can be difficult to determine, pragmatic felicity is actually even more difficult to evaluate because it is generally a matter of preference (Ariel 2008), with judgments being more context-sensitive, subtler and consequently more open to debate. In turn, the conclusions drawn from pragmatically unacceptable data can be fragile and open to debate.

7 The fragility of acceptability judgments has been widely acknowledged for more than thirty years now[1] (see e.g. Labov (1996), entitled "When intuitions fail", or Cowart (1997)) and different methods have recently been proposed to remedy the problem: elaborate questionnaires constructed with semantic tests (e.g. Erteschik-Shir and Lappin (1979)'s lie-test), speeded grammaticality judgments (SGJs), Likert n-point scales, and experiments with the Magnitude Estimation protocol used in psychophysics. The latter has the advantage of taking real gradience into account. Overall, it seems that there is a general consensus now on the fact that introspection by the researcher is not the ideal way of evaluating linguistic data, as linguists thus produce both the data and the analysis at the same time (Labov (1972: 99); Hoffmann (2007b: 1)). However, no consensus has been reached on the best method of evaluating linguistic data; hence, we aim to contribute to the debate with our pilot study.

2. Magnitude Estimation: Definition and Uses

2.1. Magnitude Estimation in psychophysics

8 The Magnitude Estimation protocol (ME) was first used in the field of psychophysics, where the goal is to determine the link between the magnitude of a physical stimulus and the way it is perceived by human subjects. For more than fifty years now (cf. Stevens 1957), the protocol has been applied to different kinds of stimuli: length of lines, light intensity, volume of sound and heat to name a few. What the different studies have shown is that there is a correlation between the different physical stimuli and perceptions of them, although the correlation is not always a direct one. For example, while the doubling of the initial length of a line is perceived as such, doubling the intensity of light is only perceived as an increase of 1.5 times the initial stimulus. Psychophysical relationships have thus been analyzed as corresponding to a set of mathematical functions (Bard et al. 1996: 1).

9 The protocol itself is simple to implement. Each subject is first asked to evaluate an initial stimulus, called the modulus, to which a numerical value is assigned. This numerical value must be superior to zero (e.g. 0.5; 1; 2; 10; 34; 89; 100,000). This initial stimulus becomes the reference stimulus against which all of the following stimuli are evaluated through the assignment of numerical values. For example, depending on the protocol, if the subject perceives a stimulus, for instance a line, as being twice as long as the modulus, then s/he should multiply the numerical value attributed to the modulus by two; if the subject perceives the second line as being half as long, then s/he should divide the numerical value of the modulus by two. To allow for inter-subject comparisons, results are then normalized, first by dividing each numerical value by that of the modulus – each modulus is thus assigned the value of 1 – and second by applying the decimal logarithm (log10) to the normalized results, a standard though not systematic practice within the ME protocol (see Sprouse (2007) for a discussion on the necessity of log transformations).

2.2. Magnitude Estimation in linguistics

10 The ME technique has been used in linguistics, first by phonologists and then by syntacticians, and has been encouraged by some researchers for the last twenty years (e.g. Schütze 1996; Bard et al. 1996; Myers 2009). As Bard et al. (1996: 42) note, it has been used in phonology to investigate the perception of speech rate (Grosjean 1977, Grosjean and Lass 1977, Green 1987), vowel roughness (Toner and Emanuel 1989), similarity of syllables from different languages (Takefuta et al. 1986) and quality of synthesized speech (Pavlovic et al. 1990). In syntax, in spite of criticism (see below), the ME technique has also been successfully used to investigate several problematic phenomena: word order and clitic doubling in Greek (Keller and Alexopoulou 2005), the choice between pied piping and preposition stranding (Hoffman 2007b), auxiliary selection (Bard et al. 1996) and the influence of resumptive pronouns on linguistic acceptability (Alexopoulou and Keller 2002). Generally speaking, the aim of these studies was to put an end to long-lasting debates on the (non-)acceptability of some data. For instance, Featherston (2005) showed that the superiority condition and discourse linking phenomena do exist in German, in spite of what had been claimed in the literature. Keller and Asudeh (2001)

purport to have solved the data conflict on so-called 'picture NPs' and binding phenomena (3): while the literature on the topic is divided between an explanation based on structural constraints and an explanation based on pragmatic factors, Keller and Asudeh (2001) show that it is the structural factors that govern the binding possibilities in picture NPs.

(3) a. Hanna found a picture of her_i/$herself_i$.
b. $Peter_i$ found $Hanna_j$'s picture of *her_j/$herself_j$.
c. $Hanna_i$ took a picture of *her_i/$herself_i$.

11 Alexopoulou and Keller (2002) have demonstrated that, in opposition to the literature on the subject, it is not necessarily correct to assume that the use of a resumptive pronoun can "save" an utterance in the case of a violation of a wh-island constraint when the absence of the resumptive pronoun would result in unacceptability. Their study shows that examples in (4) and (5) are in fact evaluated in the same way by informants (initial judgments from Alexopoulou and Keller):

(4) a. ?*Who did Mary wonder whether they will fire *t*?
b.*Who did John meet the girl who will marry *t*?
(5) a. Who did Mary wonder whether they will fire him?
b. Who did John meet the girl who will marry him?

12 To our knowledge, setting aside Keller and Asudeh's (2001) study on the influence of structural and pragmatic factors on picture NPs, the only example of a pragmatics study using the ME technique is Davies (2011). The aim of her research is to evaluate the perception of over-informativeness as a violation of Grice's Maxim of Quantity by providing too much information, and to compare these perceptions with those of under-informativeness. The prototypical example of over-informativeness is a situation in which a subject is asked to give someone "the big apple" in a case where only one apple is available, while under-informativeness is characterized by a situation in which a subject is asked to give someone "the apple" in a situational context where more than one apple is available. Davies (2011) compares results obtained using three different methods: (i) binary choice, (ii) a Likert 5-point scale, and (iii) ME. What is of particular interest for us here is the fact that her study concludes that, for both adults and children, the ME technique provides relevant results that are comparable to results obtained using the other techniques.

13 In our pilot study, we examine whether the ME technique can be used more generally for studies conducted within the discourse pragmatics field, where acceptability judgments can sometimes be very subtle but are usually obtained introspectively, as emphasized by Noveck and Sperber (2007). These authors emphasize the fact that experimental methods are greatly needed to test pragmatic hypotheses, as pragmatic intuitions are for the moment generally nothing but educated guesses about hypothetical utterances.

2.3. Criticisms and objections

14 Before we move on to our ME experiment, it is important to note that the use of the Magnitude Estimation technique in linguistics is not without criticism. Critics include Bader and Häusler (2010), Sprouse (2008) and Fukuda et al. (2012). These authors present two main arguments. First, no objective data is available for comparison with informants' judgments, contrary to other types of stimuli. Although it is possible to compare the evaluation of light intensity with the actual, physical stimuli submitted to informants, no

such objective reality exists for linguistic data. Second, some researchers have shown that the results obtained through ME are similar to results obtained with other, simpler techniques such as questionnaires or binary evaluations. This has led some of them to wonder whether the use of ME in linguistics is "worth the trouble", to quote the title of Fukuda et al.'s (2012) paper.

15 These criticisms reveal that the use of ME for the evaluation of data in linguistics remains open to debate, and our pilot study will thus provide new elements for the discussion on the topic. We now turn to the description of our experiment.

3. Experiment: data and methodology

3.1. Description of the data

16 We have retained two different sets of data for this pilot experiment, both of which rely on acceptability judgments that do not concern grammaticality issues but rather pragmatic acceptability/felicity in context. However, the two data sets differ in that their unacceptability does not stem from pragmatic violations that have the same strength. We first use data with non-standard word order (topicalization) in contexts where this marked word order clearly leads to acceptability problems. One such example is provided in (6). The data have been adapted from Birner and Ward (1998) and are described in the next sub-section. The second set of data is more difficult to evaluate and includes sentences with appositive/non-restrictive relative clauses[2] modifying different types of antecedents composed of proper nouns. An example is provided in (7). According to the fame effect hypothesis developed in Loock (2010a, 2010b), some of these examples are infelicitous in context as a function of the informational status (old or new information) of the identificational relationship between the antecedent and the predicate in the relative clause (here, *Barack Obama is the president of the US*). The acceptability of such data is more problematic, and one of our goals is to see whether informants can provide a relevant evaluation with the ME technique – confirming or invalidating Loock's own acceptability judgments.

(6) A: What kind of sports do you like?
B: #Baseball I like. (Birner and Ward 1998)
(7) #Barack Obama, who is the president of the US, has arrived today. (Loock 2010b)

17 For the first set of data, acceptability judgments are straightforward and simple to formulate, as bad topicalizations generally provide clear judgments. Our goal is to verify that the ME technique provides the same results as the judgments based on Birner and Ward's own evaluation of their data. The objective here is thus an evaluation of the ME technique. Essentially, we aim to check to what extent there is a match between the results obtained with the ME technique and the results provided in Birner and Ward (1998). Regarding the second set of data, we will determine whether the ME technique provides results that match Loock's (2010b) data on the fame effect, where judgments are less straightforward. In this case, it is the data itself that we seek to evaluate: are the results obtained with the ME protocol in line with Loock's own judgments? At the same time, this is also an evaluation of the technique with the observation of finer-grained data.

18 At this stage, we wish to remind the reader that this article presents the results of a pilot study, which as such has certain limitations insofar as the selection of data is concerned

(e.g. limited number of stimuli, only one ungrammatical stimulus for control, no systematic pairing of examples), but we are confident that the results shed interesting light on the use of the ME technique for discourse pragmatics.

3.1.1. Birner and Ward's (1998) data on non-standard word order

19 Birner and Ward (1998) investigate the constraints that govern the (in)felicity of sentences with a non-canonical word order: topicalization, focus preposing, passive sentences, right/left dislocation, existential *there*-sentences. The constraints that they provide are related to the informational status of the different elements in the sentence (discourse new/old, hearer new/old information, following Prince's (1981, 1992) taxonomy of given/new information).[3] To define their constraints, Birner and Ward took into account negative data (labeled '#'), using introspective judgments. With such data, the judgments seem fairly consensual as any violation of the constraints governing the choice of non-canonical word orders generally provides very bad results in context. This is why we have selected and adapted some of Birner and Ward's data for our ME experiment: we have selected a number of short dialogues involving felicitous (8) or infelicitous (9) topicalizations, and have also adapted some of them for the purpose of our experiment, that is, we modified some of their examples to make them shorter or to make up new dialogues similar to theirs from a structural point of view (see e.g. examples (23)-(25) below, based on (9)).

(8) Customer: Can I get a bagel?
Waitress: No, sorry. We're out of bagels. A bran muffin I can give you.
(9) Customer: What kind of breakfast baked goods do you have?
Waitress: #A bran muffin I can give you.

3.1.2. Loock (2010a, 2010b)'s fame effect hypothesis

20 In his examination of the discourse functions of appositive relative clauses (henceforth ARCs), Loock (2010a) provides both a positive and a differential definition of this structure. While the positive definition provides a taxonomy of the discourse strategies underlying speakers' use of ARCs, the differential definition provides constraints that govern the choice between ARCs and other structures that fulfill the same kinds of discourse functions. Among these structures are sentential parentheticals, independent clauses, non-restrictive premodifiers and nominal appositives. Among other pragmatic constraints, it is suggested that a familiarity constraint, called "fame effect", can explain the choice between an ARC and a nominal appositive. The choice between these competing structures, or allostructures, is dependent on the hearer new/old status of the information conveyed by their insertion. The fame effect hypothesis can be summarized as follows: The more familiar the relation of the kind A is B, the less felicitous the use of an ARC, as it makes the relation explicit (anaphoric pronoun + *be*), while the use of a nominal appositive keeps the identificational relation implicit.

21 We understand from this hypothesis that not all nominal appositives can be rephrased as ARCs, and while some examples are clearly infelicitous (#), others are acceptable, or have a problematic status (? or ??):

(10) #Some expected Barack Obama, who is the president of the United States, to appoint a completely new economic team so as to implement another New Deal.
(11) #/?? Bill Clinton, who is the former president of the United States, will attend an international seminar on AIDS and SARS (Severe Acute Respiratory Syndrome)

and deliver a lecture on global AIDS prevention and control efforts, a seminar official said Friday.
(12) #/?? Angela Merkel, who is the German chancellor, on Friday described Barack Obama's presidency as a "unique opportunity" to revive the Middle East peace process as the US leader continued his international tour with a stop in the historic eastern city of Dresden.
(13) ?Nancy Pelosi, who is the Speaker of the House, is among those on the Left now seeking to find common ground with the conservative populism that is sweeping across the United States.
(14) ?According to Arne Duncan, who is the secretary of education, the president will discuss the importance of hard work, educational goals and other topics.
(15) Martin Townsend, who is the editor of the Sunday Express, has made a personal appeal for her safe return, and said the paper had given its full support to her decision to enter the country illegally.
(16) Edgar Griffin, who is the father of the BNP leader Nick Griffin, was sacked as a vice-president of the Duncan Smith campaign in Wales after he admitted answering a BNP telephone inquiry line.

22 According to the fame effect hypothesis, this is directly related to the hearer new/old status of the A is B relationship:[4]

(10a) Barack Obama is the president of the United States.
(11a) Bill Clinton is the former president of the United States.
(12a) Angela Merkel is the German Chancellor.
(13a) Nancy Pelosi is the speaker of the House (in Congress).
(14a) Arne Duncan is the Secretary of Education.
(15a) Martin Townsend is the editor of the Sunday Express.
(16a) Edgar Griffin is the father of the BNP leader Nick Griffin.

23 Since the hypothesis is only valid if the evaluation of the data is reliable and since the evaluation of the data by native speakers of English is questionable, Loock (2010b) used data collected from the internet to establish a parallel between the judgments and the number of results for queries involving an ARC (e.g. *Barack Obama, who is (the) President of the United States/America/the US*). The results reveal a correlation between the number of hits as retrieved through a search engine and the use of an ARC (examples of *Barack Obama* followed by an ARC are extremely rare while examples with *Angela Merkel* are more frequent – see Loock (2010b) for exact figures). Our paper aims to go further by evaluating the data used in Loock (2010b) in an experiment based on the Magnitude Estimation protocol and to compare the two types of evaluation. As mentioned above, such data are particularly problematic, as acceptability judgments are more subtle than those concerning word-order phenomena. The difference between a nominal appositive and an ARC is clearly a matter of preference (Ariel 2008), and not a question of grammaticality. So our goal, here, is to investigate the kind of information that evaluation through the ME protocol can provide, and to see whether the judgments in Loock (2010b) can be confirmed.

3.2. Description of the experiment

3.2.1. The stimuli

24 The experiment was divided into 4 sub-experiments, corresponding to 4 tests which the informants – to be described in the next sub-section – were asked to take. Using a slideshow designed with Microsoft Office PowerPoint, the informants were asked to numerically evaluate 4 types of stimuli: (i) length of lines (LINES TEST), (ii) size of circles (

CIRCLES TEST), (iii) naturalness of short dialogues (DIALOGUES TEST), and (iv) naturalness of sentences (SENTENCES TEST). The informants were shown a first stimulus that served as the modulus to which the subsequent stimuli were compared. The modulus was repeated systematically from one slide to another to allow for easier comparisons and to avoid any memory-related problems. Thanks to clear instructions and examples, the informants were guided through the experiment and invited to write down their numerical evaluation, obligatorily superior to 0, on a separate sheet, to be sent back by e-mail or to be filled in and returned to the researchers in person. The total number of stimuli amounted to 80, divided as follows: 10 lines, 10 circles, 20 dialogues, 40 sentences. No time limit was given to the informants, though they were asked to provide spontaneous evaluations and never to go back and modify their previous answers. On average, filling in the evaluation sheet took between 10 and 15 minutes.

25 The first two series of stimuli (LINES TEST and CIRCLES TEST) were used as training, as the evaluation of line lengths and circle sizes is considered standard stimuli for ME experiments (Bard et al. 1996; Alexopoulou and Keller 2005). In psychophysics studies, it is generally acknowledged that ME permits an evaluation of such stimuli to be directly correlated with objective reality. The evaluation of line lengths and circle sizes is thus regularly used as training to allow informants to be at ease with the procedure. These first two tests allowed us to ensure that the informants had understood the instructions correctly and were able to provide reliable answers.

26 The third series ('DIALOGUES TEST') consisted in the evaluation of the naturalness of 20 short dialogues extracted or adapted from Birner and Ward (1998) with pragmatically felicitous and infelicitous topicalizations. The instructions clearly stated that the informants were meant to evaluate the natural-sounding character of the exchanges, without taking into account any stylistic, plausibility or politeness considerations. Below we provide a sample of the data; the acceptability judgment based on Birner and Ward's constraints is provided in square brackets ('OK' for acceptable, '#' for pragmatically infelicitous). One ungrammatical example was inserted for control (26).

(17) Customer: Can I get a bagel?
Waitress: No, sorry. We're out of bagels. I can give you a bran muffin. [OK]
(18) Customer: Can I get a bagel?
Waitress: No, sorry. We're out of bagels. A bran muffin I can give you. [OK]
(19) Customer: What kind of breakfast baked goods do you have?
Waitress: A bran muffin I can give you. [#]
(20) Customer: What kind of breakfast baked goods do you have?
Waitress: I can give you a bran muffin. [OK]

(21) A: Do you like football?
B: Yeah. Baseball I like a lot better. [OK]
(22) A: Do you like football?
B: Yeah. I like baseball a lot better. [OK]
(23) A: What kind of sports do you like?
B: I like baseball. [OK]
(24) A: What kind of sports do you like?
B: Baseball I like. [#]
(25) A: What kind of sports do you like?
B: Baseball. [OK]

(26) A: What did Sheldon give to Amy?
B: He gave flowers her. [*]

27 The fourth series ('SENTENCES TEST') consisted of sentences (provided out of context) with structures of the type <proper name, ARC> (e.g. *Barack Obama, who is the president of the US*) and <proper name, nominal apposition> (e.g. *Barack Obama, the president of the US*), to which a number of distractors/fillers were added (felicitous and infelicitous examples of existential and presentational constructions taken from Birner and Ward (1998) or constructed for the purpose of the experiment), for a total of 40 stimuli, 24 of which correspond to the ARC/nominal apposition alternation after a proper name taken from Loock (2010a, 2010b). As for the dialogues, the informants were asked to evaluate the naturalness of sentences, starting with a modulus. They were given clear instructions and examples. A selection from among the 24 stimuli is provided in (27), together with the acceptability judgment as provided in Loock (2010b) ('OK' for acceptable, '#' for infelicitous, '?' for questionable, '??' for very questionable). The 40 stimuli from the sentences test are provided in Appendix A.[5]

(27) a. A. Raja, the Indian environment minister, said his country would accept help to reduce emissions but would not be forced into cuts. [OK]
b. A. Raja, who is the Indian environment minister, said his country would accept help to reduce emissions but would not be forced into cuts. [OK]
c. Angela Merkel, the German chancellor, described Barack Obama's presidency as a "unique opportunity" to revive the Middle East peace process. [OK]
d. Angela Merkel, who is the German chancellor, said her country would accept help to reduce emissions but would not be forced into cuts. [??]
e. Some expected Barack Obama, the president of the United States, to reduce emissions. [OK]
f. Some expected Barack Obama, who is the president of the United States, to appoint a completely new economic team so as to implement another New Deal. [#]
g. According to Arne Duncan, the secretary of education, the president will discuss the importance of hard work, educational goals and other topics. [OK]
h. According to Arne Duncan, who is the secretary of education, the president will accept help to reduce emissions but will not be forced into cuts. [?]

3.2.2. The informants

28 Forty informants were recruited to participate in the experiment. All of them were adult native speakers of English and none of them were linguists. They were contacted either via e-mail or in person at the Languages and Psychology Departments of the University of Manchester. There were no restrictions based on gender, age, social or geographical origin; their nationality (British or American) was noted (24 British, 16 American) but was not used as a criterion for the distribution of the informants into different groups. All of them were volunteers and did not receive any compensation for their participation.

29 The 40 informants were randomly divided into 4 groups (4x10), with each of the informants taking part in the experiment only once. A distinction was made between Experiment 1, in which the linguistic data (dialogues and sentences) were submitted to the informants in their written form only, and Experiment 2, in which the written stimuli were accompanied by audio stimuli, consisting of recordings of native speakers reading the dialogues and sentences. In this condition, both types of stimuli (spoken and written) were submitted at the same time. Thus, informants in Experiment 1 evaluated written stimuli only, while informants in Experiment 2 evaluated written stimuli accompanied by their audio counterpart. For both experiments, the order of the stimuli was randomized to avoid any influence of the order of presentation on the results. Experiments 1 and 2 were thus divided into Experiments 1a/1b and 2a/2b, each of the sub-experiments being

assigned to 10 informants, for a total number of 40 participants. The results of 5 informants were not taken into account, as they had not followed the instructions, e.g. they assigned values that were negative or equal to zero to some of the stimuli, contrary to standard procedure for ME experiments, or they did not evaluate some of the stimuli. Therefore, only the results for 35 informants are presented; the number of informants for each sub-experiment is provided in Table 1 (note that the number of informants in each group is not always identical, but this imbalance is neutralized with the statistical analysis).

Table 1. Distribution of informants

Sub-experiment	**Data**	**Number of informants**
1a	Written stimuli, order 1	8
1b	Written stimuli, order 2	9
2a	Written and spoken stimuli, order 1	10
2b	Written and spoken stimuli, order 1	8
Total		**35**

3.2.3. Methodology

30 The individual results were compiled and listed in a single table. The stimuli in the 4 sub-experiments were ordered following the order in experiments 1a/2a (so results from experiments 1b/2b were reordered). As is the case with many experiments based on the ME protocol, results were normalized twice: (i) each numerical value assigned by the subjects was divided by the value that they had assigned to the modulus in each of the four sub-experiments. This means that each modulus received the score of 1, while the other stimuli received a new value based on that given to the modulus. Thus, any stimulus perceived as being twice as long/big/natural now receives the value of 2, irrespective of the value chosen for the modulus. A second normalization was performed using the decimal logarithm (log10), a standard though unsystematic practice in ME experiments (Sprouse 2007) which aims to provide finer normalization. It is this twice-normalized value that was taken into account here. Table 2 provides an example of this double normalization process.

Table 2. Example of the double normalization process (Informant 1, First 5 line length stimuli)

LINES TEST	**Informant 1**		
	Raw value	Raw value/modulus	Log_{10}
Line 1 MODULUS	2	1	0
Line 2	5	2,5	0,3979

Line 3	0,33	0,165	-0,783
Line 4	3	1,5	0,1761
Line 5	0,5	0,25	-0,602

31 The influence of different parameters was then investigated. First of all, we examined parameters that could potentially lead to variation between the informants due to the presentation mode of the stimuli (written vs. written/spoken), the order of the stimuli (order 1 vs. order 2), nationality (GB vs. US). If we can show that these parameters had no influence on the results and that there is no statistically significant inter-subject variation between the 35 informants, then the results for all informants can be grouped together to check for the influence of the relevant parameters for Birner and Ward's (1998) data (word order) and Loock's (2010b) data (nominal appositive vs. appositive relative clause). Recall that the parameters investigated were the choice of structure (topicalizations vs. SVO word order for the dialogues test; nominal appositives vs. ARCs for the sentences) but also the hearer new/old status of the referent in the sentences test.

32 All statistical analyses were carried out within the R environment (R Core Team, 2013) and mainly relied on classical parametric and non-parametric tests (ANOVAs and corrected *t test* and *Tukey Honest Significant Difference* post hoc tests).[6]

3.3. Preliminary examination of the data

33 Before investigating the results of our experiment through a comparison with the data from Birner and Ward (1998) and Loock (2010b), we checked whether informants - properly performed on the ME protocol by checking their results for the lines and circles tests. Linear regressions for both circles and lines and for all sub-groups indicate a strong correlation (corrected R^2>.78) between subjective evaluations and the objective reality of the different stimuli, thus confirming that our informants had no particular difficulty with the technique (Figures 1a/1b show the results for the estimation of line length and circle diameter respectively).[7] We thus conclude that our informants understood the technique and that, after a training session with 20 stimuli, mastered it sufficiently to apply it to the linguistic stimuli that followed.

Figure 1a. Results for the lines tests (4 sub-groups of informants)

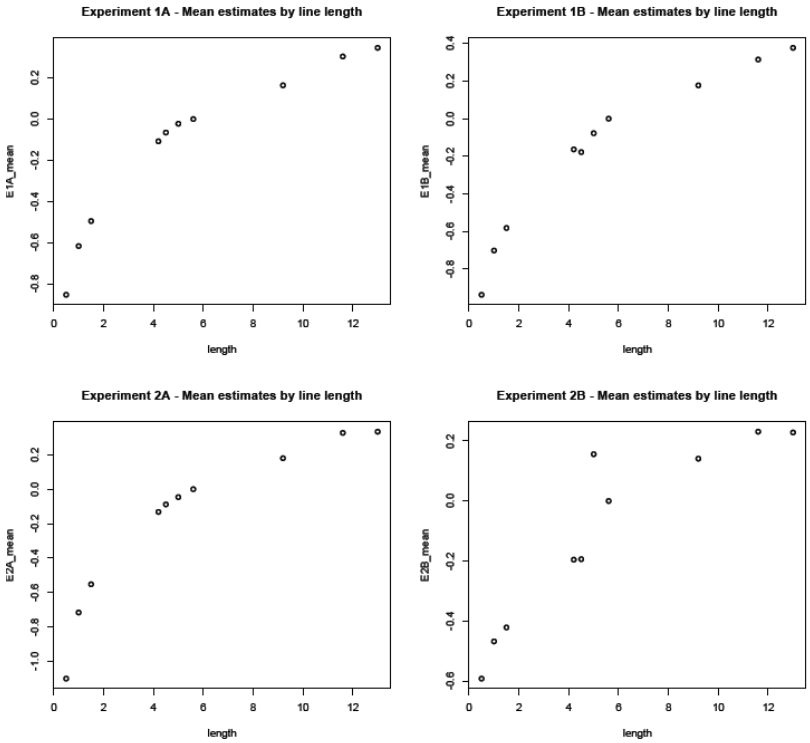
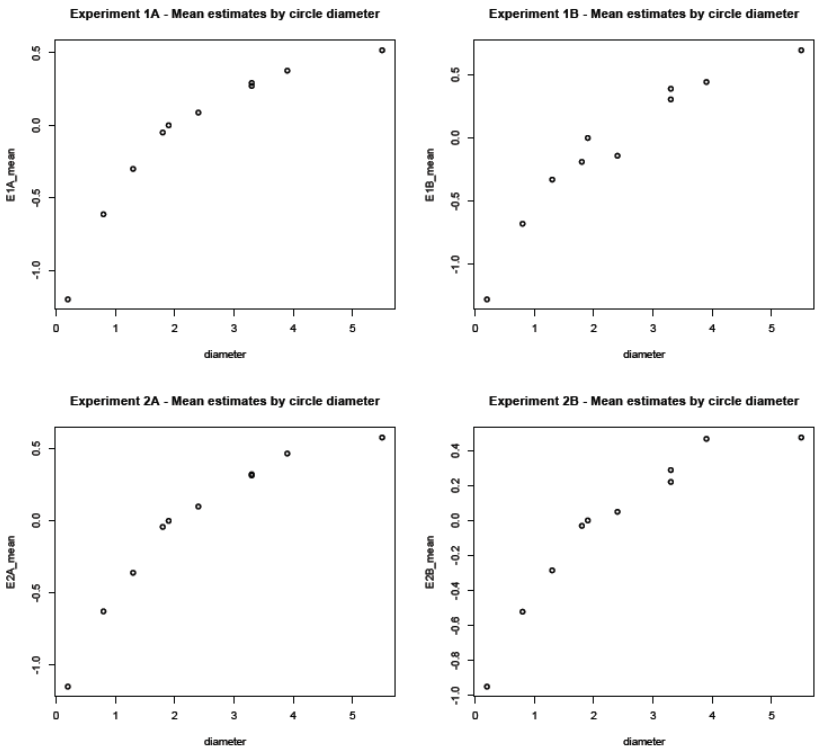


Figure 1b. Results for the circles tests (4 sub-groups of informants)



4. Results for dialogues and sentences[8]

4.1. Preliminary results

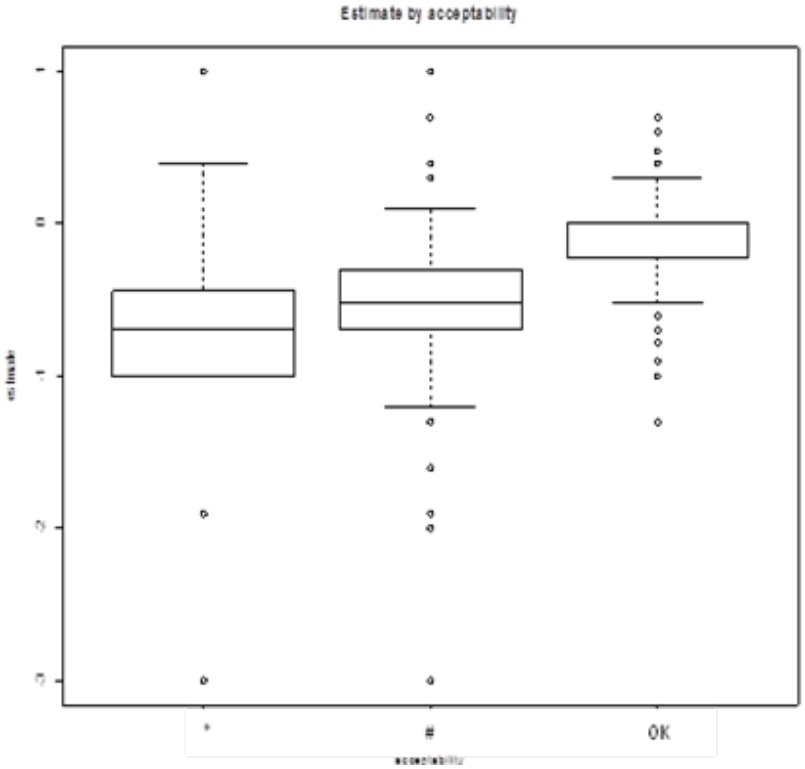
34 Before investigating the results of the linguistic part of our experiment (dialogues and sentences tests), we first investigated whether the results differed as a function of the mode of delivery of the stimuli (written vs. written/spoken), the order of the stimuli (order 1 vs. order 2) and the informants' nationality (GB vs. US). We also looked at whether the level of inter-subject variation was great enough to invalidate the experiment. These preliminary results show that none of the three parameters has any influence on the results (ANOVA: $p>0.05$). Although there is a certain amount of inter-subject variation, it is not statistically significant (ANOVA: $F=12.9263$, $p<0.01$; Tukey HSD post hoc test shows that no subject significantly differed from all the others). Consequently, the results for the 4 sub-groups of informants (1A, 1B, 2A, 2B) have been collapsed and the results will be treated as a whole. We can now turn to a comparison of the ME results and the data evaluations from Birner and Ward (1998) and Loock (2010b).

4.2. Results based on Birner and Ward's non-canonical word order data

35 Recall that our dialogues test consisted of 20 stimuli extracted or adapted from Birner and Ward's (1998) work on non-standard word order. Our goal was to compare Birner and Ward's acceptability judgments of felicitous and infelicitous topicalizations with the evaluation of the data used as stimuli in our ME experiment. What we find is that the acceptability categorization (*, #, OK) based on Birner and Ward is echoed in the evaluation provided by our informants following the ME protocol (ANOVA: $F=113.816$, $p<2e^{-16}$).

36 A post hoc Tukey HSD test confirms that informants clearly make a distinction between acceptable data on the one hand (OK) and ungrammatical/infelicitous data (*/# respectively) on the other ($p<2e^{-16}$ in both cases). These results are illustrated in Figure 2 and can be schematized as the gradient provided in (28): the evaluation of ungrammatical (*) and unacceptable data (#) is significantly inferior to that of acceptable data (OK). They show that the informants distinguish between felicitous and grammatically/ pragmatically unacceptable data, but fail to discriminate between grammatically and pragmatically unacceptable data (Tukey HSD: $p>.05$). Another way of looking at the results is to note that infelicitous topicalizations (as defined by Birner and Ward) result in sentences that sound so bad that they are treated as if they violate a rule of grammar. However, we must be very careful here, as we have used only one ungrammatical stimulus for a control. Of course, this is not sufficient for a reliable comparison between ungrammatical and unacceptable data, but was intended, in the pilot study, to provide information on how our informants evaluated data that are unequivocally ungrammatical. Needless to say, this is an area that will require further development in future studies.

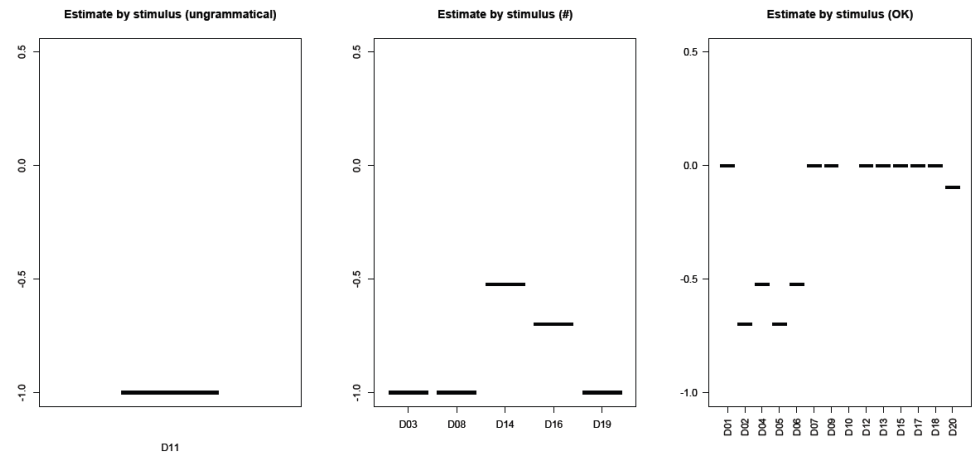
Figure 2. Overall results based on Birner and Ward's acceptability categorization[9]



(28) */# < OK

37 If we now consider the examples in detail, that is stimulus by stimulus, and if we compare the results of our experiment with the categorization predicted by Birner and Ward's (1998) analysis, we see that, although the overall results provided above are quite clear, they do not line up completely. Figure 3 below provides the results (only the median result is provided for each stimulus) for the 3 categories (*, #, OK).

Figure 3. Overall results based on Birner and Ward's acceptability categorization



38 Figure 3 thus reveals that the results for some stimuli do not match: Dialogues 14 and 16, which were expected to be infelicitous, are not rated as being as bad as the other infelicitous stimuli. Dialogues 2, 4, 5 and 6, which were expected to be felicitous, are

actually evaluated as less felicitous than the other stimuli in the acceptable category. These problematic cases are provided in Appendix B.

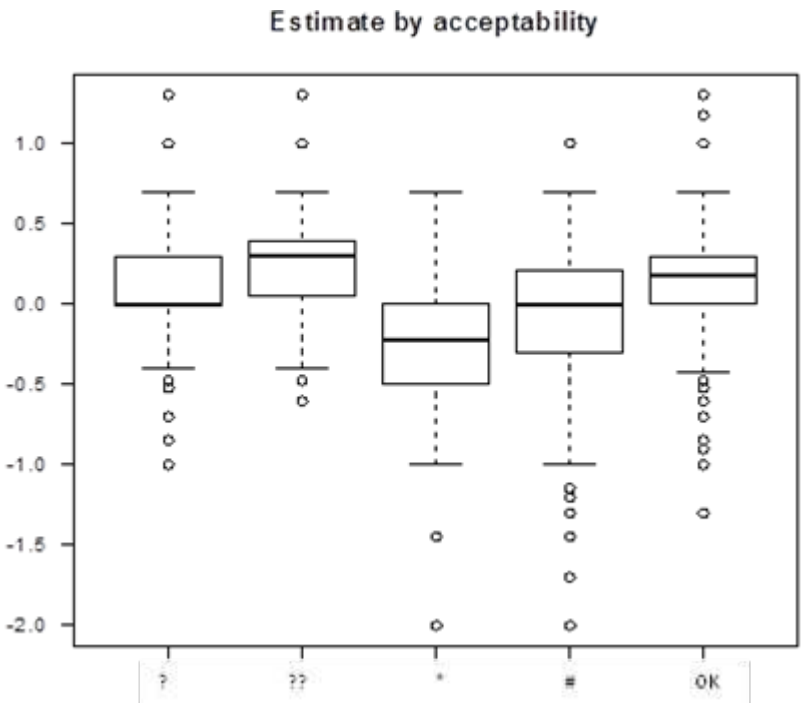
39 Apart from these exceptions, which remain to be accounted for as they represent 6 examples out of 20, the ME results globally match Birner and Ward's evaluation of the data. Given that their data are fairly straightforward (bad topicalizations are clearly infelicitous), we can (almost) safely conclude that the ME technique is reliable for the evaluation of pragmatic data. We now turn to Loock's (2010b) more subtle data.

4.3. Results based on Loock's fame effect data

4.3.1. Acceptability categorization

40 If we consider the 40 stimuli used in the sentences test (examples of ARCs, nominal appositives, and distractors), we find a general correlation between our informant's judgments and the ones predicted by Loock's fame effect hypothesis and Birner and Ward's predictions (ANOVA: $F=43.327$, $p<2.2e^{-16}$). These results are shown in Figure 4. However, it is important to note that there is no statistically significant difference (Tukey HSD: $p>.05$) between acceptable (OK) and (very) questionable data (?/??). However there is a significant difference between pragmatically infelicitous (#) and ungrammatical (*) data (Tukey HSD: $p<.01$; note that, once again, we had only one ungrammatical stimulus for control, as our focus here is on pragmatic acceptability). The results can also be schematized with the gradient in (29).

Figure 4. Overall results for acceptability categorization in the sentences test (all 40 stimuli)

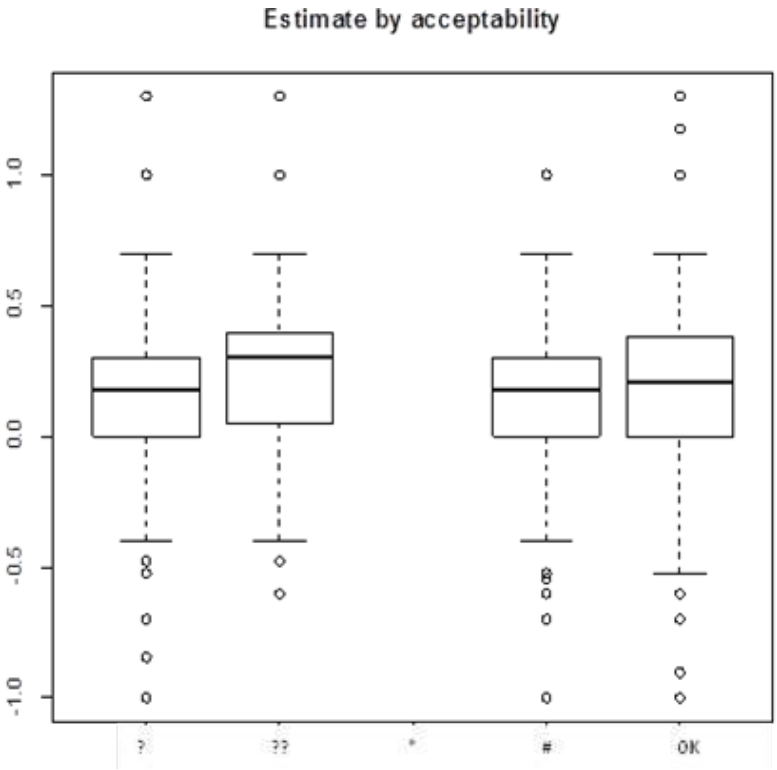


(29) *< # < (?/??/OK)

41 If we remove the distractors and focus only on the "fame effect" data, that is, stimuli with ARCs or nominal appositives, the results are similar. There is a correlation between our informants' evaluations and the results predicted by the fame effect hypothesis (ANOVA:

F=5.325, p<.01). Yet, at the same time, the informants fail to discriminate between acceptable (OK) and (very) questionable data (?/??) (Tukey HSD: p>.05). These results are illustrated in Figure 5 and schematized as a gradient in (30).

Figure 5. Overall results for acceptability categorization in the sentences test (Loock's (2010b) data only based on his judgments)

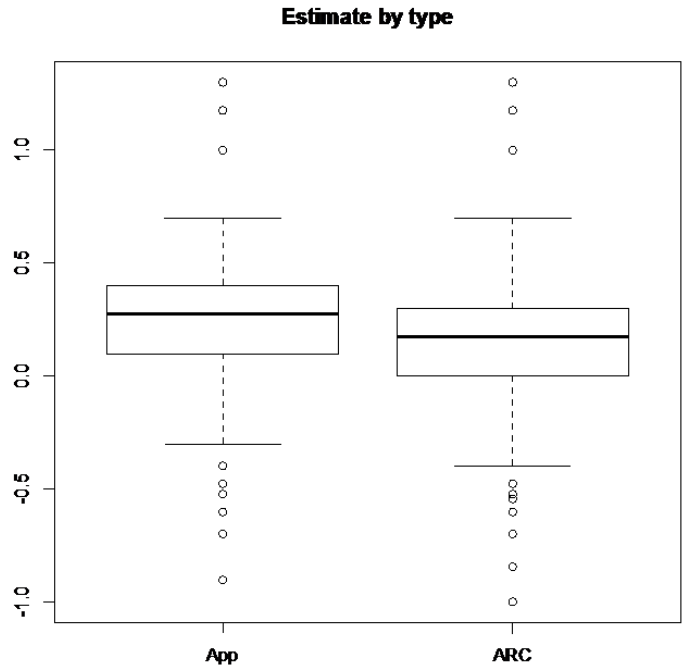


(30) # < (?/??/OK)

4.3.2. Nominal appositives vs. Appositive Relative Clauses

42 According to the fame effect hypothesis, nominal appositives are systematically felicitous after a proper noun, whereas ARCs are only felicitous if the identificational relationship 'A is B' is hearer new (as in *Martin Townsend is editor of the Sunday Express*, where the relationship M. Townsend/be editor of the Sunday Express represents hearer new information for most, if not all addressees). As a result, we predicted that our stimuli with nominal appositives would receive a better evaluation than our stimuli with ARCs. This is indeed the case (corrected t-test: t=3.102; p<.01), with respective mean estimates of 0.248 and 0.181, as illustrated by Figure 6 and schematized by the gradient in (31).

Figure 6. Results for nominal appositives (App) vs. Appositive Relative Clauses (ARCs) in the sentences test

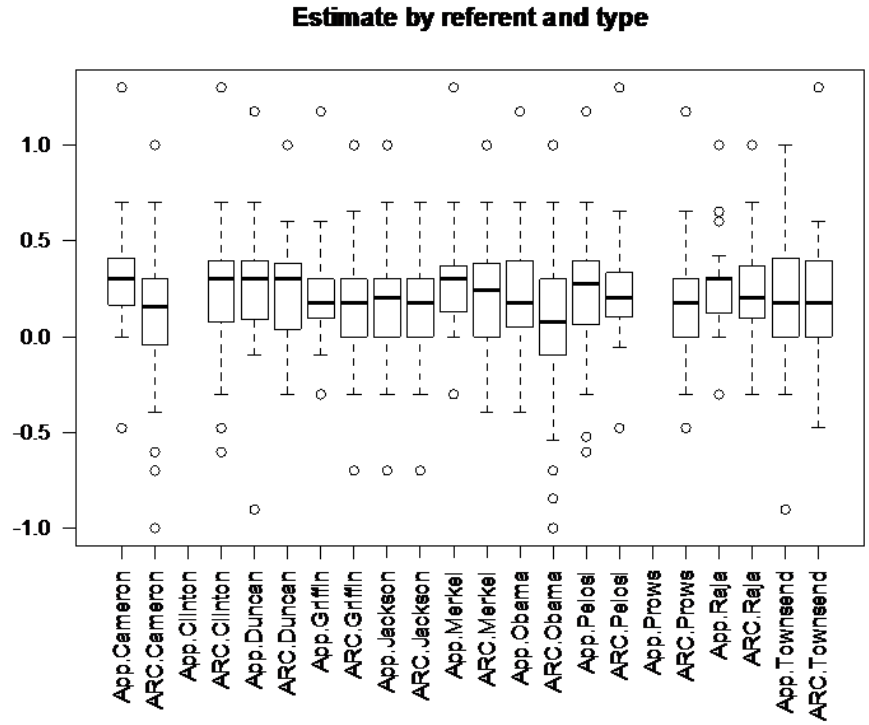


(31) App > ARC

4.3.3. The fame effect hypothesis

43 According to the fame effect hypothesis, the more familiar the identificational relation 'A is B', the less felicitous the use of an ARC. The hypothesis thus predicts a lower evaluation for ARCs than for nominal appositives when the identificational relationship is hearer old (e.g. *Barack Obama is president of the US*) and an absence of difference between the two structures when the identificational relationship is hearer new (e.g. *Martin Townsend is editor of the Sunday Express*). The results of our ME experiment are, however, not so straightforward. While we do notice *some* interesting tendencies, the results are not systematic and not always confirmed statistically. Figure 7 provides the detailed results for a series of referents (Cameron, Clinton, Duncan, Griffin, Jackson, Merkel, Obama, Pelosi, Prows, Raja, Townsend) for both the use of an ARC and a nominal appositive.

Figure 7. Results for nominal appositives (App) vs. Appositive Relative Clauses (ARCs) for each referent in the sentences test



44 These results show an absence of difference between the judgments when an ARC is employed as compared to a nominal appositive for a series of 'non-famous' referents: this is the case for Duncan, Griffin, Pelosi, Raja, Townsend (note we have no stimulus with an ARC for Prows) (Tukey HSD: p>.05). By non-famous, we mean that the identificational relationship 'A is B' is hearer new for most addressees. This is in line with the one-way fame effect hypothesis: the use of both structures is felicitous. However, the hypothesis that the use of an ARC with 'famous' referents (those in which the identificational relationship 'A is B' is hearer old for most addressees at the time of utterance), that is Cameron, Jackson, Merkel and Obama (note that we have no stimulus with a nominal appositive for Clinton), is not confirmed: the results seem to indicate a positive though not statistically significant trend for Cameron and Obama (with lower Tukey HSD p values, below .4), but the results for Jackson and Merkel are not confirmed statistically (Tukey HSD: p=.99). It is therefore impossible to say that the results provided by our ME experiments line up completely with the judgments based on Loock's (2010b) fame effect hypothesis. We see two possible explanations for this: either this type of judgment, which is not as straightforward as the evaluation of (in)felicitous topicalizations and has been produced here by informants with minimal contexts, is too subtle for informants, or the data evaluation predicted by the fame effect hypothesis is not correct in the first place and needs to be revised in the light of the ME results.

5. Conclusions and future research

45 If we return to the question in our title, "Can Magnitude Estimation do something for your pragmatics?", the answer based on our pilot study appears to be yes, but only to

some extent. We have seen that there is a general correlation between introspective judgments and ME judgments for cases of (in)felicitous topicalizations. In spite of specific cases where the two judgments provided different results, most cases showed that the results obtained with our ME experiment did in fact confirm those in Birner and Ward's (1998) analysis. These results seem to be confirming Birner and Ward's analysis and its predictions, although the 6 cases where the results do not match need to be accounted for, by determining whether it is Birner and Ward's evaluation or the ME technique that is problematic (this is left open for further research). However, the results are not clear for subtler acceptability judgments: we have not been able to confirm that the acceptability judgments based on the fame effect hypothesis developed in Loock (2010a, 2010b) are unequivocally correct. While examples with nominal appositives receive a better evaluation than examples with ARCs, the core of the hypothesis is not confirmed. The predicted difference between hearer new and hearer old identificational relationships is not confirmed by our ME results, although some tendencies do exist. We can account for this lack of correspondence in the results with two possible explanations: the limitations of the ME protocol in dealing with subtle acceptability judgments or problems in Loock's (2010b) analysis and evaluation of the data. Moreover, we have seen that with the ME technique, informants may fail to discriminate (i) between pragmatically infelicitous (#) and ungrammatical data (*), and (ii) between acceptable and questionable data (? or ??). In other words, finer-grained judgments need to be teased apart, as acceptability has been shown to be a matter of gradience. However, it is not clear whether the ME protocol can take such gradience into account.

46 Further research is therefore needed to determine whether it is the ME technique or the fame effect hypothesis that is problematic in our case. This will enable us to distinguish protocol-related vs. data-related issues. In addition, further research should be carried out in order to find a more rigorous and relevant way of determining which referents are famous and which are not: we have relied here on common sense by claiming that *Barack Obama is the president of the US* or *David Cameron is the Prime Minister of Great Britain* represents hearer old information while *Martin Townsend is the editor of the Sunday Express* represents hearer new information, but cases like *Nancy Pelosi is the Speaker of the House*, which we assumed would represent hearer new information, lie perhaps in a grey area. Although relying on common sense would do for a pilot study and the referents that we selected, a much more rigorous approach will be required in future to investigate the problem under study here. Finally, further research is also necessary to determine how gradient acceptability in discourse pragmatics can be dealt with in ME experiments.

BIBLIOGRAPHY

Alexopoulou, Theodora, Keller, Frank. (2002). Resumption and locality: a crosslinguistic experimental study. In Andronis, M., Debenport, E., Pycha, A., Yoshimura, K. (eds), *Papers from the 38th Meeting of the Chicago Linguistic Society Vol. 1.* Chicago, University of Chicago Press. pp. 1-14.

Ariel, Mira. (2008). *Pragmatics and Grammar*. Cambridge, Cambridge University Press.

Bader, Markus, Häussler, Jana. (2010). Toward a model of grammaticality judgments. *Journal of Linguistics* 46 (2). 273–330.

Bard, Ellen G., Robertson, Dan, Sorace, Antonella. (1996). Magnitude estimation of linguistic acceptability. *Language* 72 (1), 32-68.

Belletti, Adriana, Rizzi, Luigi. (1988). Psych verbs and O-Theory. *Natural Languages and Linguistic Theory* 6, 291-352.

Birner, Betty J. (2006). Inferential relations and noncanonical word order. In Birner, B., Ward, G. (eds), Drawing the Boundaries of Meaning: Neo-Gricean Studies in Pragmatics and Semantics in Honor of Laurence R. Horn. Amsterdam/Philadelphia, John Benjamins. Pp. 31-51.

Birner, Betty J. (2004). Discourse functions at the periphery: Noncanonical word order in English. In Shaer, B., Frey, W., Maienborn, C. (eds.), *Proceedings of the Dislocated Elements Workshop, Zentrum für Allgemeine Sprachwissenschaft,* Berlin, November 2003, Vol. 1 (ZAS Papers in Linguistics 35). Berlin, Zentrum für Allgemeine Sprachwissenschaft. pp. 41–62.

Birner, Betty. J., Ward, Gregory. (1998). Information Status and Noncanonical Word Order in English. Amsterdam/Philadelphia, John Benjamins.

Chomsky, Noam. (1962). Explanatory models in linguistics. In Nagel, E., Suppes, P., Tarski, A. (eds), *Logic, Methodology and Philosophy of Science.* Stanford, Stanford University Press. pp. 528-550.

Cowart, Wayne. (1997). Experimental Syntax: Applying Objective Methods to Sentence Judgments. Thousand Oaks (CA), Sage Publications.

Davies, Catherine (2011). Over-Informativeness in Referential Communication. Ph.D. dissertation, Magdalene College, University of Cambridge.

Erteschik-Shir, Nomi, Lappin, Shalom. (1979). Dominance and the functional explanation of island phenomena. *Theoretical Linguistics* 6. 41–85.

Featherston, Sam. (2005). Magnitude estimation and what it can do for your syntax: Some wh-constraints in German. *Lingua* 115. 1525–1550.

Fukuda Shin, Michel, Dan, Goodall, Grant, Beecher, Henry. (2012). Is Magnitude Estimation worth the trouble? In Choi, J., Hogue, E. A., Punske, J., Tat, D., Schertz, J., Truman, A. (eds), *29th West Coast Conference on Formal Linguistics, Univ. of Arizona, April 22-24 2011*. Somerville, MA, Cascadilla Proceedings Project. pp. 328-336.

Green, Kerry P. (1987). The perception of speaking rate using visual information from a talker's face. *Perception and Psychophysics* 42. 587-593.

Grosjean, François. (1977). The perception of rate in spoken and sign languages. *Perception and Psychophysics* 22. 408-413.

Grosjean, François, Lass, Norman J. (1977). Some factors affecting the listener's perception of reading rate in English and French. *Language and Speech* 20. 198-208.

Haegeman, Liliane. (1991). Introduction to Government and Binding Theory. Oxford, Blackwell.

Hill, Archibald A. (1960). Grammaticality. *Word* 17. 1–10.

Hoffmann, Thomas. (2007a). 'Good is good and bad is bad': but how do we know which one we had? *Corpus Linguistics and Linguistic Theory* 3(1). 87-98.

Hoffmann, Thomas. (2007b). I need data which I can rely on: Corroborating empirical evidence on preposition placement in English relative clauses. In Featherston, S., Sternefeld,W. (eds), *Roots: Linguistics in Search of its Evidential Base*. Berlin, Mouton de Gruyter. pp. 161-183.

Keller, Frank, Alexopoulou, Theodora (2005). A crosslinguistic, experimental study of resumptive pronouns and that-trace effects. In Bara, B. G., Barsalou, L., Bucciarelli, M. (eds), *Proceedings of the 27th Annual Conference of the Cognitive Science Society*. Mahwah, New Jersey, Lawrence Erlbaum. pp. 1120-1125.

Keller, Frank, Asudeh, Ash. (2001). Constraints on linguistic coreference: Structural vs. pragmatic factors. In *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*. Mahwah, New Jersey, Lawrence Erlbaum. pp. 483-488.

Labov, William. (1996). When intuitions fail. In McNair, L., Singer, K., Dolbrin, L., Aucon, M. (eds), *Papers from the Parasession on Theory and Data in Linguistics Chicago Linguistic Society 32*. pp. 77-106.

Labov, William. (1972). Sociolinguistic Patterns. Philadelphia, University of Pennsylvania Press.

Lakoff George. (1973). Fuzzy grammar and the performance/competence terminology game. *Chicago Linguistics Society* 9. 271-291.

Loock, Rudy. (2013). Extending Further and Refining Prince's Taxonomy of Given/New Information: a case study of non-restrictive, relevance-oriented structures. *Pragmatics* 23(1), 69-91.

Loock, Rudy. (2010a). Appositive Relative Clauses in English: Discourse Functions and Competing Structures. Amsterdam/Philadelphia, John Benjamins.

Loock, Rudy. (2010b). The "Fame Effect" or How the syntactic choices of writers can be explained by their assumptions about their addressees' states of knowledge: the case of relevance-oriented, non-restrictive noun modifiers. *Discours 7*. http: /discours.revues.org/8027 (accessed 07.08.2013)

Müller Gereon. (1995). A-bar syntax: A study in movement types. Studies in Generative Grammar 42. Berlin/New York, De Gruyter.

Myers, James. (2009). Syntactic judgment experiments. *Language and Linguistics Compass* 3, 406-423.

Noveck, Ira, Sperber, Dan. (2007). The why and how of experimental pragmatics: The case of 'scalar inferences'. In Burton-Roberts, N. (ed.), *Advances in Pragmatics*. Houndmills, Basingstoke, Hampshire, Palgrave/MacMillan. pp. 184-212.

Pavlovic, Chaslav V., Rossi, Mario, Espesser, Robert. (1990). Use of the magnitude estimation technique for assessing the performance of text-to-speech synthesis systems. *Journal of the Acoustical Society of America* 87 (1). 373-382.

Prince, Ellen F. (1992). The ZPG Letter: Subjects, definiteness, and information-status. In: Thompson, S., Mann, W. (Eds). *Discourse Description: Diverse Analyses of a Fund Raising Text*. Amsterdam/Philadelphia, John Benjamins. pp. 295-325.

Prince, Ellen F. (1981). Toward a taxonomy of given/new information. In Cole, P. (ed.). *Radical Pragmatics*. New York, Academic Press. pp. 223-254.

R Core Team. (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, url = http: /www.R-project.org (accessed 07.08.2013)

Schütze, Carson T. (1996). The Empirical Base of Linguistics. Chicago, University of Chicago Press.

Sorace, Antonella, Keller, Frank. (2005). Gradience in linguistic data. *Lingua* 115(11): 1497-1524.

Sprouse, Jon. (2008). Magnitude estimate and the non-linearity of acceptability judgments. *Proceedings of the West Coast Conference on Formal Linguistics* 27.

Sprouse, Jon. (2007). A program for experimental syntax: Finding the relationship between acceptability and grammatical knowledge. PhD dissertation, University of Maryland, College Park.

Stevens, Stanley S. (1957). On the psychophysical law. *Psychological Review* 64 (3), 153-181.

Takefuta, Yukio, Guberina, Peter, Pizzamigio, Luigi, Black, John W. (1986). Cross-lingual measurements of interconsonantal differences. *Journal of Psycholinguistic Research* 15, 489-507.

Toner, Mary Ann, Emanuel, Floyd W. (1989). Direct magnitude estimation and equal appearing interval scaling of vowel roughness. *Journal of Speech and Hearing Research* 32 (1), 78-82.

Wurmbrand Susanne. (2001). Infinitives: Restructuring and Clause Structure. Berlin/New York, Mouton de Gruyter.

APPENDIXES

Appendix A: Sentences

1. "I think the hyper-cars are more hype than anything," Larry Carlat, editor of Toy And Hobby World, which is the leading industry trade journal, said.

2. A. Raja, the Indian environment minister, said his country would accept help to reduce emissions but would not be forced into cuts.

3. Prince Charles was "amazed" and "shocked" by Bolland's comments yesterday to a Sunday newspaper -- he is the prince's former deputy private secretary and press adviser.

4. In addition to interest-rate risk, there is the added risk that when interest rates fall, mortgages will be prepaid, thereby reducing the Portfolio's income stream.

5. Martin Townsend, who is the editor of the Sunday Express, was "amazed" and "shocked" by Bolland's comments yesterday to a Sunday newspaper.

6. The notable alumni of Columbia include one of the most powerful men in the world: Barack Obama, who is the President of the United States, was a part of Columbia College Class of 1983.

7. Arrested were Nathan Thomas, 23, of New York, and his brother, WO Victor Thomas, 32, a 13-year Army veteran.

8. A. Raja, who is the Indian environment minister, said his country would accept help to reduce emissions but would not be forced into cuts.

9. Angela Merkel, the German chancellor, described Barack Obama's presidency as a "unique opportunity" to revive the Middle East peace process.

10. There were the neighbors at the City Council meeting yesterday.

11. Prince Charles was amazed and shocked by Bolland, who is the prince's former deputy private secretary,'s comments.

12. David Cameron, the British Prime Minister, said his country would accept help to reduce emissions but would not be forced into cuts.

13. Daniel told me that shortly after Grumman arrived at Wideview Chalet there arrived also a man named Sleeman.

14. Edgar Griffin, who is the father of the BNP leader Nick Griffin, described Barack Obama's presidency as a "unique opportunity" to revive the Middle East peace process.

15. The death photo of Michael Jackson, who was The King of Pop, has been released.

16. I have some interesting news for you. At today's press conference there was President Clinton.

17. Angela Merkel, who is the German chancellor, said her country would accept help to reduce emissions but would not be forced into cuts.

18. I'm pleased that dogs eat cheese, but eat cheese they do.

19. Some expected Barack Obama, the president of the United States, to reduce emissions.

20. Bill Clinton, who is the former president of the United States, will attend an international seminar on AIDS and SARS.

21. Mr Miliband is scheduled to hold bilateral talks with Mr A. Raja, who is the Indian Environment Minister.

22. Nancy Pelosi, the Speaker of the House, is among those on the Left now seeking to find common ground with conservative populism.

23. No one gets to take their original birth certificate home with them, idiot: you can get a certified copy for identification from the county, but you don't get to go anywhere with the original, you, me or Barack Obama, who is the President of the United States.

24. There was the wedding picture of the Clintons on his table.

25. According to Arne Duncan, who is the secretary of education, the president will accept help to reduce emissions but will not be forced into cuts.

26. There are all sorts of variations on the term insurance: policies structures to pay off your mortgage debt, term riders tacked on to permanent insurance, and many others.

27. Nancy Pelosi, who is the Speaker of the House, said her country would accept help to reduce emissions but would not be forced into cuts.

28. There was the usual crowd at the beach today.

29. Martin Townsend, editor of the Sunday Express, said his country would accept help to reduce emissions but would not be forced into cuts.

30. Edgar Griffin, the father of the BNP leader Nick Griffin, was sacked as a vice-president of the Duncan Smith campaign in Wales after he admitted answering a BNP telephone inquiry line.

31. The death photo of Michael Jackson, the King of Pop, has been released.

32. After the flood were a number of devastated families.

33. Some expected Barack Obama, who is the president of the United States, to appoint a completely new economic team so as to implement another New Deal.

34. I'm pleased that dogs eat cheese, if eat cheese they do.

35. David Cameron, who is the British Prime Minister, yesterday that the Europeans had until the end of July to come up with a more concrete offer.

36. In addition, as the review continues, there isn't the chance that we'll uncover something additional that is significant.

37. According to Arne Duncan, the secretary of education, the president will discuss the importance of hard work, educational goals and other topics.

38. Bill's parents visited last month; an entire week they were here.

39. "I think people were scared," said Peter Prows, who is a politics student form Oberlin College, Ohio.

40. Bill Clinton, who is the former president of the United States, described Barack Obama's presidency as a "unique opportunity" to revive the Middle East peace process.

Appendix B: 'problematic' cases in data from or adapted from Birner and Ward's (1998)

Dialogue 14:
A: Someone broke into the garage last night.
B: Your father you need to tell. [#]

Dialogue 16:
A: I'm really tired tonight.
B: Maybe a movie you should rent. [#]

Dialogue 2:
Customer: Can I get a bagel?
Waitress: No, sorry. We're out of bagels. A bran muffin I can give you. [OK]

Dialogue 4:
Customer: What kind of breakfast baked goods do you have?
Waitress: I can give you a bran muffin. [OK]

Dialogue 5:
A: Do you like football?
B: Yeah. Baseball I like a lot better. [OK]

Dialogue 6:
A: Do you like football?
B: Yeah. I like baseball a lot better. [OK]

NOTES

1. As early as in the 1960s, Chomsky himself, who is often blamed for the use of introspection in data evaluation, acknowledged that introspection does not suffice: "I dislike reliance on intuition as much as anyone (...) We should substitute rigorous criteria just as soon as possible, instead of clinging to intuition" (Chomsky (1962), cited in Schütze (1996: 9)).
2. Appositive (also called non-restrictive) relative clauses do not contribute to the identification of the referent of their antecedent and thus contrast with determinative/restrictive relative

clauses. This traditional opposition between the two types of relative clauses is to be found in most, if not all, standard grammar books, and is illustrated by a series of formal differences, including the presence/absence of punctuation between antecedent and relative pronoun, and the distribution of the latter (*wh-* pronouns, *that*, or zero pronouns) (see Loock (2010a: 35) for a critical review of the distinctive criteria).

3. Prince's (1981, 1992) taxonomy of given/new information distinguishes between givenness/newness in the discourse and givenness/newness in the hearer's knowledge store as assumed by the speaker, to which the case of inferable information needs to be added. The taxonomy has been complemented and refined by several researchers (e.g. Birner 2004, 2006 for the inferrables category; Loock 2013 for the creation of an extra-category, the indeterminables).

4. Note that in Loock (2010a, 2010b), and therefore in this article, the familiarity of the relation A is B is based on common sense. This issue, which could be problematic for the reliability of our results because of the subjectivity that is involved, is discussed in the conclusion.

5. Note that, although examples in (27) show a systematic pairing of examples containing a nominal appositive with examples containing an ARC, this has not been systematized (see Appendix A and section 4.3.3).

6. A t-test is a standard statistical test which checks whether the means of two groups are statistically different from each other or not. The ANOVA (or ANalysis Of Variance) is a statistical method which can be seen as a generalization of t-tests to more than two samples and whose aim is to determine whether the means of the groups are similar or not. As for the Tukey Honest Significant Difference test, it allows for multiple comparisons between means; used in association with the ANOVA, it allowed us to determine whether some of the means in the different samples were significantly different from each other. All these tests allowed us to determine the (non-)influence of a series of parameters (both linguistic and non-linguistic) on the informants' evaluation of the data and measure the differences, if any, between our different sub-groups.

7. Figures 1a/b should be read as follows: each figure shows the results for our 4 sub-groups of informants (Experiments 1A, 1B, 2A and 2B). In each graph, the horizontal axis shows the actual length of the lines and of the circle diameters (in centimeters); the vertical axis shows the twice-normalized evaluation by the informants.

8. In this section, the value of F corresponds to the result of the statistic used in the ANOVA; p (or p-value) corresponds to the probability according to which the result is significant or not (the significance level used here is 0.05: any p-value higher than 0.05 means that the observed differences are not significant).

9. Each boxplot provides the median value (black line inside the rectangle), the upper and lower quartiles (that is the 25% of the results that are greater and lower than the median value, the rectangle thus representing the middle 50% of results), the greatest and lowest values (so-called 'whiskers' outside the rectangle), as well as the outliers (small circles).

ABSTRACTS

We report the results of a pilot study investigating whether the technique of Magnitude Estimation, widely used in psychophysics but also in syntax and phonology, can be exploited in discourse pragmatics. In this domain, unacceptable data (i.e. data labeled '#') play an important role but acceptability judgments can be subtle, context-dependent, and thus a topic of debate, as shown by endless disputes in the literature. Using two sets of data, (in)felicitous topicalizations

from Birner and Ward (1998) and the (in)felicitous use of appositive relative clauses as discussed in Loock (2010a, 2010b) on the fame effect, we aim to determine whether the Magnitude Estimation technique can provide reliable results that would overcome the difficulties with intuitive acceptability judgments. We discuss whether this technique provides acceptability judgments that are less problematic than those obtained introspectively or via regular questionnaires.

Avec cet article, nous souhaitons vérifier, dans le cadre d'une étude pilote, si le protocole expérimental nommé « Magnitude Estimation », largement utilisé en psychophysique mais également en syntaxe et en phonologie, peut l'être en pragmatique du discours, où les données inacceptables (étiquetées '#') jouent un rôle important mais où les jugements d'acceptabilité peuvent être subtils, influencés par le contexte, et donc prêter à débat, comme le montre la littérature. A partir de deux types de données, des topicalizations (a)pragmatiques extraites de Birner et Ward (1998), et l'utilisation (a)pragmatique de relatives appositives par opposition à des appositions nominales d'après la définition du *fame effect* de Loock (2010a, 2010b), nous souhaitons voir ici si la Magnitude Estimation peut fournir des résultats fiables qui viendraient pallier les difficultés liées aux jugements d'acceptabilité obtenus par introspection ou par le biais des questionnaires traditionnels.

INDEX

Mots-clés: magnitude estimation, évaluation des données, acceptabilité, topicalisation, relatives appositives, apposition
Keywords: data evaluation, acceptability, topicalization, appositive relative clauses

AUTHORS

RUDY LOOCK

Rudy Loock is Professor of English Linguistics and Translation Studies in the Applied Languages department of the University of Lille 3 and affiliated with the CNRS laboratory Savoirs, Textes, Langage (UMR 8163). His research interests include discourse pragmatics, information packaging, the discourse-prosody interface, corpus-based translation studies, and translation quality.

CYRIL AURAN

Cyril Auran was until 2013 a Lecturer in Linguistics and Phonetics at the University of Lille 3 and affiliated with the CNRS laboratory Savoirs, Textes, Langage (UMR 8163). His doctoral thesis and subsequent corpus-based and corpus-oriented research in both English and French explore the roles of prosody in discourse structure, cohesion and coherence.