Naive v. expert intuitions: An empirical study of acceptability judgments

EWA DĄBROWSKA

Abstract

Judgments about the grammaticality/acceptability of sentences are the most widely used data source in the syntactic literature. Typically, syntacticians rely on their own judgments, or those of a small number of colleagues. Although a number of researchers have argued that this is problematic, there is little research which systematically compares professional linguists' intuitions with those of linguistically naive speakers.

This article examines linguists' and nonlinguists' judgments about one particular structure: questions with long distance dependencies. Linguists' judgments are shown to diverge from those of nonlinguists. These differences could be due to theoretical commitments (the conviction that linguistic processes apply 'across the board', and hence all sentences with the same syntactic structure should be equally grammatical) or to differences in exposure (the constructed examples of this structure found in the syntactic literature are very unrepresentative of ordinary usage). Whichever of these explanations turns out to be correct, it is clear that linguists' judgments are not representative of the population as a whole, and hence syntacticians should not rely on their own intuitions when testing their theories.

1. Introduction

The most widely used data source in syntactic research is speakers' intuitions about the well-formedness of sentences. Traditionally linguists have relied on their own intuitions, or those of a few colleagues: indeed, some linguists have argued that this is the most reliable data source available (see, for example, Newmeyer 1983: 50ff). This, however, is problematic, since individual judgments are often unreliable (cf. Cowart 1997; Schütze 1996); thus, to obtain

The Linguistic Review 27 (2010), 1–23
DOI 10.1515/tlir.2010.001

0167–6318/10/027-01

stable measures of grammaticality, it is necessary to average over responses provided by a number of informants.

Another problem with linguists' reliance on their own intuitions is observer bias: the possibility that judgments can be influenced by the observer's beliefs and expectations. Observer bias has been found even in observations involving objective phenomena such as contractions and head turns made by flatworms (Planaria) in response to light: observers who had been told to expect such movements recorded more instances than those who did not expect the reaction (Cordaro and Ison 1963). Obviously, the danger of beliefs and expectations contaminating observation is compounded when the observer is reporting on internal states. It is well known that in clinical trials, patients often report improvement in their condition even if they are given a placebo rather than a drug. More surprisingly, clinicians who know that the patient is getting a drug rather than a placebo are more likely to record clinical improvement in their patient's condition than blinded clinicians (Noseworthy et al. 1994) – which is why all serious medical trials are double-blind. A syntactician reporting on his or her own intuitions is like an unblinded patient and clinician rolled into one: not only are they observing their own internal states, but also interpreting them.

It is also possible that linguists' and nonlinguists' intuitions differ as a result of differences in experience. Repeated exposure to some types of ungrammatical or borderline structures can make them sound more acceptable – a phenomenon known as syntactic satiation and demonstrated experimentally by Hiramatsu (1999) and Snyder (2000). As Snyder observes, "... many linguists admit that they can no longer perceive the (presumed) ungrammaticality of certain syntactic violations and that they have simply memorized the judgments that are standard in the linguistics literature" (2000: 575). The possibility that judgments could be learned from the literature or in the course of linguistic education is also mentioned by Schütze (1996: 47) and Cowart (1997: 60).

Although many researchers have expressed concerns about the practice of collecting grammaticality judgments 'in house', there is surprisingly little research which systematically compares professional linguists' judgments with those made by linguistically naive informants, and the few existing studies have produced contradictory results. Spencer (1973) collected naive speakers' judgments about 150 exemplar sentences from the literature, and found that naive judges agreed among themselves about over 80 % of the sentences, but agreed with published judgments of the linguists for only half of the sentences. However, it is not clear whether this reflects a genuine difference between the two populations or is simply due to the fact the published judgments were made by individuals and hence were less reliable. A more systematic study by Bradac et al. (1980) also found significant differences; Snow and Meijer (1977), in contrast, report a very high correlation (Spearman's $\rho = 0.89$) between linguists' and nonlinguists' judgments. The discrepancy may be due to the fact that the two studies used different types of stimuli. Snow and Meijer elicited judgments about Dutch sentences involving various non-canonical word order patterns. Although word order variation is clearly of interest to syntacticians, no major theoretical controversies hinge on the relative well-formedness of the sentences used in their experiment, and hence there is no obvious reason for the linguists' judgments to be different from those of nonlinguists. The Bradac et al. study used sentences containing various kinds of 'errors': 'theoretical errors' (i.e., ungrammatical sentences of the kind that one frequently finds in the syntactic literature, for instance complex NP violations), 'foreign errors' (sentences containing errors typical of non-native speakers), 'native errors' (sentences involving non-standard structures such as split infinitives and stranded prepositions), and sentences which were grammatical but unacceptable (e.g., triple center embedding). This stimulus set did reveal differences between linguists and nonlinguists; unfortunately, the authors do not state which sentences were rated differently, so it is difficult to draw any conclusions about the reasons for the discrepancies. Group differences in the rating of sentences containing native errors, for example, are of limited interest, since they are likely to reflect prescriptive attitudes. Different intuitions about sentences with theoretical errors, on the other hand, would be much more revealing.

The present study will compare linguists' and nonlinguists' judgments about sentences instantiating a particular construction – questions with long distance dependencies – which has played a central role in the development of modern syntactic theory, and which, consequently, features quite prominently in examples cited in the literature. As I will show in the following section, the constructed examples found in the literature differ in a number of ways from naturally occurring instances of the construction. Thus, linguists' experience of questions with long-distance dependencies is different from that of ordinary language users, and this could be reflected in their judgments.

Unlike most earlier research, which investigated intuitions about sentences involving violations of various constraints, the main focus is on sentences which are traditionally regarded as fully grammatical. This will enable us to examine how lexical and structural factors interact and the extent to which they affect acceptability judgments of the two types of informants.

1.1. The status of acceptability judgments

Before going into the details of the present study, we must address an important methodolgical issue, namely, what exactly does an acceptability judgment test measure?

Ever since Chomsky (1965), most linguists have distinguished between grammaticality (whether or not a sentence conforms to the rules of grammar) and acceptability (the degree to which a sentence is judged by native speakers to be permissible in their language). Acceptability, Chomsky argues,

> is a concept that belongs to the study of performance, whereas grammaticalness belongs to the study of competence. … Grammaticalness is only one of the many factors that interact to determine acceptability (1965: 11).

Thus, speakers may judge perfectly grammatical sentences as unacceptable because they violate some prescriptive notion (e.g., *This is something I will not put up with*), because they are difficult to process (*The horse raced past the barn fell*) or because they are semantically anomalous (*Colorless green ideas sleep furiously*); conversely, a sentence could be acceptable but ungrammatical (e.g., *Watched some TV, then went to bed,* produced in response to *What did you do last night?;* see also Newmeyer 1983 for an extensive discussion of these issues).

However, the distinction is not without problems, and some linguists have suggested that it should be abandoned. Featherston (2005), for instance, justifies this proposal as follows:

> We may advance two reasons for this: first, studies such as this one demonstrate that the dividing line is being drawn in the wrong place, and second, it is not obvious on what grounds we might decide where the right place is. For a construct such [as] the Grammaticality/Acceptability distinction to be of any use, it must be possible to judge where the dividing line is located. But this criterion is lacking: in practice linguists tend to assume traditional assignments in the literature, and in new cases apply the criterion of categoricity; a few seem to use it indiscriminately as a weapon (if data supports my theory it must be Grammatical, if it supports your theory it is just markedness) without offering any evidence to support the assignment. (2005: 701–702)

Others agree with the distinction in principle, but point out that grammaticality can only be operationalized as acceptability. Thus, Riemer (2009) argues that

> The only way predictions of grammaticality can be checked is by assuming that, other things being equal, the sets of grammatical and acceptable sentences coincide: in effect, then, the grammatical strings are those which are predicted to be acceptable in conditions unaffected by interference from performance factors.

In other words, if we want a scientific theory – i.e., one that could be falsified by empricial data – we must rely on acceptability judgments, because grammaticality is not directly accessible to intuitions (Newmeyer 1983: 51; Schütze

1996: 26). We must, of course, acknowledge that judgments can be, and often are, influenced by extragrammatical factors, and therefore researchers must take care to either neutralize them (by balancing stimuli for length, lexical content, processing difficulty, plausibility, etc., whenever possible) or to control for them (by setting up control conditions which will allow them to assess the extent to which the confounding factors affect speakers' judgments; see Schütze 1996; Cowart 1997; and Featherstone 2005 for further discussion).[1]

1.2. Questions with long distance dependencies

Dąbrowska (2004, 2008, in preparation) and Verhagen (2005) point out that 'real life' questions with long distance dependencies (henceforth LDDs) are extremely stereotypical. In the vast majority of spontaneously produced LDD questions, the main clause auxiliary is *do*, the main clause subject *you* or another pronoun, and the main verb *think* or *say*; moreover, there are generally no additional elements in the main clause, and no complementizer. This is illustrated by the examples in (1), all taken from the spoken part of the British National Corpus:

(1) a. *What do you think you're doing?*
b. *Who do you think you are?*
c. *What do you think it means?*
d. *Where do you think that goes?*
e. *What did you say the score is?*

Dąbrowska and Verhagen both suggest that speakers have lexically specific templates (*WH do you think S-GAP?, WH did you say S-GAP?*) which enable them to produce and understand new LDD questions by inserting appropriate items in the WH and S-GAP slots. According to usage-based theories of language (Bybee 2006; Langacker 1988, 2000; Barlow and Kemmer 2000; Dąbrowska 2010), such lexically specific units are psychologically more basic than more abstract constructions. Of course speakers are also able to produce and understand questions which do not fit the templates, which suggests that they either have more general schemas in addition to the more specific ones or that they resort to analogy when processing non-prototypical LDD questions (see Dąbrowska 2008 for some suggestions about how this could work). Either way, the usage-based accounts proposed by Dąbrowska and Verhagen predict that processing such questions would involve more effort (since, according to

[1] It is worth noting that acceptability judgments are routinely used in the L2 literature, and are assumed to reflect L2 linguistic knowledge.

usage-based theories, low-level, 'local' schemas are psychologically more basic and hence easier to access than more general representations: see Langacker 2000; Dąbrowska 2010). Hence, one would expect prototypical LDD questions (i.e., those which match one of the templates) to be produced more fluently, remembered better, and judged to be more acceptable than non-prototypical ones. All three of these predictions appear to be correct (see Dąbrowska 2008, in preparation; Dąbrowska, Rowland and Theakston 2009).

As suggested earlier, linguists' experience of LDD questions is different from that of native speakers, due to the central role that this and other related constructions have played in the development of syntactic theory. In addition to hearing naturally occurring exemplars of this construction in their daily lives, linguists are also exposed to a fair number of constructed examples in the literature; and these constructed examples are very different from those attested in spontaneous speech, as illustrated by the sentences in (2).

(2) a. *What are you expecting that he will say to her?* (Radford 2004)
b. *Who did Mary hope that Tom would tell Bill that he should visit?* (Chomsky 1977)
c. *Who do you think Hobbs said he imagined that he saw?*
(Borsley 1999)
d. *What might she think that they will do?* (Radford 2004)

A comparison of constructed examples from the syntactic literature[2] with corpus sentences reveals that they show much more variation in all main clause positions, are three times more likely to contain additional elements in the main clause (e.g., an adverb, a direct object or prepositional phrase, or a negative particle), and more than ten times more likely to contain a complementizer (see Table 1). Most strikingly, a sizable proportion (about 9 %) involve a dependency over more than one clause boundary, as in (2b) and (2c). Such structures are extremely rare, perhaps nonexistent, in naturally occurring spoken language: the sample of 423 LDD questions with finite complement clauses analyzed by Dąbrowska (in preparation) does not contain a single instance of such a construction.[3]

Linguists' experience of LDD construction differs from that of ordinary speakers in other ways, too. In the syntactic literature, the canonical position

[2] Examples from the following sources were used in the analysis of linguistic texts: Roberts 1997; Borsley 1999; Radford 2004; Levine and Hukari 2006; Chomsky 1977; Carnie 2002; Wekker and Haegeman 1985.

[3] An anonymous referee pointed out that examples from written texts may be a more appropriate standard of comparison for the constructed examples in linguistics texts. However, LDD questions in written texts occur almost exclusively in dialogue, and they are very similar to spoken LDD questions (cf. Verhagen 2005).

Table 1: Comparison of spontaneously produced and constructed LDD questions (%)

	Spoken BNC (N=423)	Linguistics texts (N=87)
main subject = *you*	90	62
main auxiliary = *do/does/did*	94	85
main verb = *think* or *say*	86	67
overt complementizer	5	52
another element in main clause	2	6
dependency over 2+ clause boundaries	0	9

of the displaced constituent is often marked by some symbol (*t, e*, or __), and the relationship between the filler and the gap may be explicitly indicated by subscripts or connecting lines. This overt marking highlights the dependency and may act as a processing 'crutch' which helps to develop a generalized representation of the construction. Furthermore, LDD questions are often discussed in the context of other constructions with long distance dependencies, and readers are encouraged to note the syntactic parallels. It is conceivable that drawing attention to relationships between different constructions contributes to the development of more abstract representations, just as in second language acquisition explicit instruction about a particular aspect of the grammar can sometimes 'jump start' implicit learning (Ellis 2005).

The purpose of this study is to determine whether these differences in experience and/or linguists' beliefs about language affect their acceptability judgments. This will be done by comparing judgments given by professional linguists with those obtained from linguistically naive informants in an earlier study by Dąbrowska (2008). In addition, in order to determine whether there are any systematic differences between linguists of different theoretical orientations, I will compare judgments given by generative and cognitive-functional linguists.

2. Method

Most acceptability judgment experiments conducted by theoretical linguists, particularly generativists (e.g., Featherston 2005; Bard et al. 1996; Sorace and Keller 2005) use a method known as magnitude estimation. In such experiments, participants are presented with a standard stimulus, or modulus, which is assigned a particular value (e.g., 40) and asked to judge new stimuli relative to the standard. If a new stimulus is twice as good as the modulus, it should be given a rating of 80, if it is only a quarter as good, 10, and so on. The method is widely used in psychophysics, and was popularized in linguistics by Bard et al. (1996) and Cowart (1997).

An alternative is to ask participants to judge well-formedness on a Likert-type scale by assigning a numerical value to each item, say, any number from 1 to 5 or -3 to $+3$. In such experiments, the researcher typically provides examples of items at the top and bottom ends of the scale, so the task essentially involves deciding whether the test sentence is more like sentence A or like sentence B in acceptability. This method is preferred in most other social sciences, and has been used in linguistics by Bybee and Eddington (2006), Tremblay (2005), Cowart (1990), and Fanselow and Frisch (2006), among others. The main advantage of this method is that the task is much more natural: it is easier for participants to decide if a particular stimulus sentence is closer to the 'good' or 'bad' end of the scale than to decide whether it is three times as good or only half as good as the modulus. However, the method also has its disadvantages. Since it uses a fixed number of values, it may not be sensitive enough to pick up some fine contrasts. Secondly, it is not clear whether a Likert scale is an interval or an ordinal scale, i.e., whether the distances between various points on the scale are of equal magnitude – in other words, whether the distance between 1 and 2 is the same as that between 3 and 4. For this reason, some researchers (see, e.g., Jamieson 2004) object to the use of parametric tests in such cases, since parametric tests assume that the measurements are interval. However, Jaccard and Wan (1996), Labovitz (1967), Kim (1975), and others have argued that parametric tests are quite robust, so that violations of the intervalness assumption have relatively little impact on the results of the test, and the use of parametric tests with data obtained using Likert scales has now become standard (Blaikie 2003; Pell 2005).[4]

This study uses a Likert-type scale because of its greater naturalness, and the results are analyzed using ANOVA and t-tests. Since this is somewhat controversial, a second analysis which compared the median ratings using non-parametric tests (Mann-Whitney U and Wilcoxon Signed Ranks for between- and within-participants comparisons respectively) was also carried out to validate the conclusions. The results of these tests are reported only when they are at variance with those obtained using parametric tests.

[4] It should also be pointed out that data obtained from magnitude estimation experiments may also be ordinal (see Sprouse 2007).

2.1. Stimuli

The present study is a partial replication and extension of Dąbrowska (2008), which investigated prototypicality effects in questions with long-distance dependencies in naive informants. In the original study, linguistically untrained participants were asked to rate the acceptability of prototypical, less prototypical, and unprototypical LDD questions, grammatical controls (the corresponding declaratives) and ungrammatical controls (*that*-trace violations, sentences involving dependency between a WH word and a gap in a complex NP, negatives lacking an auxiliary, and sentences in which tense/agreement was marked on the auxiliary as well as the main verb). Prototypical LDD questions had the form *WH do you think* + complement clause or *WH did you say* + complement clause. Less prototypical questions departed from the prototype in one respect: that is to say, the main clause contained a lexical subject instead of *you* (WH-Subject), *will* or *would* instead of *do* (WH-Auxliary), *believe, suspect, claim,* or *swear* instead of *think* or *say* (WH-Verb)*, that* instead of a null complementizer (WH-Complementizer), or an extra complement clause (WH-Long). Unprototypical sentences departed from the prototype in all these respects. Examples of sentences in each condition are given in Table 2.

All sentences were 12 words long (13 if they contained an overt complementizer), with seven words intervening between the WH word and the gap.[5] All contained 2 subordinate clauses (either two complement clauses or a complement clause and an adverbial clause). There were four items in each condition, for a total of 72 sentences.

2.2. Participants

206 linguists working in linguistics or English language departments at various UK universities were contacted by e-mail and invited to complete the questionnaire. 29 of those contacted responded, 27 of whom were native speakers of English. The questionnaire was also distributed to all delegates at the 2006 meeting of the Linguistic Association of Great Britain. 13 of the delegates (including 11 native speakers) responded. In this article, I report on the responses given by the 38 native speakers.

The nonlinguists were 38 second- and third-year undergraduate students studying English literature at the University of Newcastle, who were asked to complete the questionnaire after a lecture. All were native speakers of English.

[5] To ensure that sentences in the declarative and interrogative condition contained the same number of words and the same number of content and function words, the declaratives corresponding to questions with *do* began with a conjunction (*and, so*, or *but*).

Table 2: Examples of sentences used in the experiment

Condition	Example
Experimental sentences	
WH-Prototypical	What do you think the witness will say if they don't intervene?
WH-Subject	What does Claire think the witness will say if they don't intervene?
WH-Auxiliary	What would you think the witness will say if they don't intervene?
WH-Verb	What do you believe the witness will say if they don't intervene?
WH-Complementizer	What do you think that the witness will say if they don't intervene?
WH-Long	What do you think Jo believes he said at the court hearing?
WH-Unprototypical	What would Claire believe that Jo thinks he said at the court hearing?
Grammatical controls	
DE-Prototypical	But you think the witness will say something if they don't intervene.
DE-Subject	And Claire thinks the witness will say something if they don't intervene.
DE-Auxiliary	You would think the witness will say something if they don't intervene.
DE-Verb	So you believe the witness will say something if they don't intervene.
DE-Complementizer	So you think that the witness will say something if they don't intervene.
DE-Long	So you think Jo believes he said something at the court hearing.
DE-Unprototypical	Claire would believe that Jo thinks he said something at the court hearing.
Ungrammatical Controls	
**That*	*What did you say that works even better?
*Complex NP	*What did Claire make the claim that she read in a book?
**Not*	*Her husband not claimed they asked where we were going.
*DoubleTn	*His cousin doesn't thinks we lied because we were afraid.

Note that the groups differed not just in level of explicit linguistic knowledge, but also age (the naive participants were aged from 20 to 24, while the linguists' ages ranged from 25 to 65) and the amount of education (most of the linguists had PhDs; all had the equivalent of at least a master's degree). One could also argue that the linguistically 'naive' participants were considerably less 'naive' than the average native speaker. However, the purpose of the study was to determine whether linguists' judgments can be regarded as representative of the population as a whole; from this perspective, it does not really matter who the control group is – although one would expect to find larger differences between groups if the 'naive' participants were, say, coal miners or taxi drivers.

2.3. Procedure

The naive participants in Dąbrowska's (2008) study were told that the questionnaire was a study of native speakers' intuitions about English sentences and that their task was to rate the acceptability of each sentence on a scale from 1 to 5, where 1 is 'very bad' and 5 is 'fine'. (Only the endpoints of the scale were labelled.) It was explained to them that the researcher was interested in their initial reaction, and that they should read each sentence carefully, but not spend too much time thinking about it. These instructions were followed by an example of a 'very bad' sentence (*Did the man who arrive by train is my cousin?*) and a sentence with a '5' rating (*Will the girl who won the prize come to the party?*).

The linguists tested in the present study were given the same sentences to judge, but slightly different instructions. They were told that the researcher had already obtained judgments from naive informants, but would also like to collect analogous data from a control group of linguists, and were asked to base their decisions as far as possible on their intuitions rather than the explicit knowledge they had acquired about English by virtue of being linguists. These instructions were followed by the same anchoring examples as in the original study. At the end of the questionnaire, participants were asked about their theoretical orientation (generative, cognitive/functional, or other) and native speaker status (native, non-native).

The rationale for this change in instructions was to allow a more meaningful comparison of the two groups' responses, since the same instructions (e.g., "rate the acceptability of the following sentences") would probably have been interpreted differently by linguists and non-linguists. Of course there is no guarantee that the change in wording resulted in both groups interpreting the task in the same way; however, the explicit injunction *not* to rely on what they had learned in the course of their linguistic training (and the fact that they were asked for acceptability, as opposed to grammaticality, judgments) should, if anything, result in linguists performing more like the naive informants than they might otherwise have done, thus making it more difficult to detect differences between groups.

3. Results and discussion

3.1. Linguists v. nonlinguists

The mean ratings for each sentence type given by two groups are presented in Table 3 and also graphically in Figure 1. To facilitate comparison, the ratings in the figure have been arranged from the highest (the WH-Prototypical condition) to the lowest (*not), using the nonlinguists' judgments as the baseline. As can be seen from the figure, there are some broad similarities between the two groups: linguists and nonlinguists alike gave higher ratings to sentences that linguists would describe as 'grammatical', and both groups judged prototypical questions as more acceptable than less prototypical and unprototypical questions. Declarative counterparts of prototypical LDD questions were also judged to be better than declarative counterparts of non-prototypical questions, although in this case the difference in acceptability was appreciably smaller.

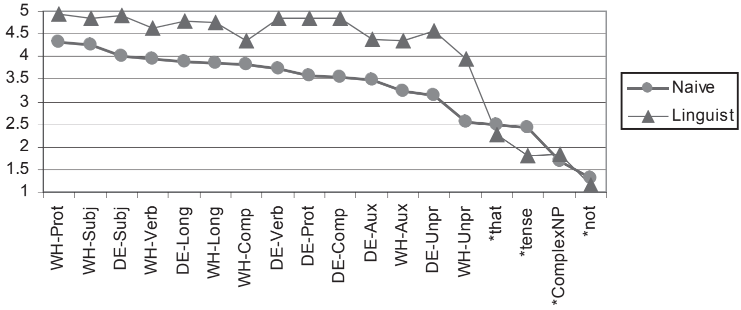

Figure 1: Comparison of linguists' and naive informants' judgments

However, there are also some clear differences. First, while both groups judged the WH and DE sentences as better than the ungrammatical controls, for linguists, there is a very sharp drop between the least acceptable grammatical sentence (WH-Unprototypical), which was rated 3.95, and the most acceptable ungrammatical sentence (*that), rated 2.28, while the nonlinguists' ratings show a continuum of acceptability, with virtually identical ratings for WH-Unprototypical (2.54), *that (2.50), and *Tense sentences (2.41).[6] Secondly, while both groups show some evidence of prototypicality effects, they are much more pronounced for the nonlinguists. Thirdly, there is a great deal more

[6] The reasons for the differences in ratings of the different types of grammatical sentences are discussed at length in Dąbrowska (2008).

Table 3: Mean, standard deviation and range for each group and condition

Condition	Naive speakers			Linguists		
	Mean	SD	Range	Mean	SD	Range
WH-Prototypical	4.31	0.63	2.75	4.93	0.19	0.75
WH-Subject	4.25	0.59	2.50	4.86	0.27	1.00
WH-Verb	3.93	0.71	2.75	4.64	0.52	2.25
WH-Auxiliary	3.23	0.83	3.25	4.34	0.74	2.50
WH-Complementizer	3.84	0.84	3.25	4.36	0.69	2.50
WH-Long	3.85	0.76	3.25	4.75	0.40	1.50
WH-Unprototypical	2.54	0.75	3.25	3.95	0.97	3.25
DE-Prototypical	3.57	0.85	3.75	4.86	0.26	0.75
DE-Subject	4.00	0.63	2.75	4.90	0.23	1.00
DE-Verb	3.74	0.78	3.25	4.85	0.26	1.00
DE-Auxiliary	3.49	0.66	3.25	4.39	0.54	1.75
DE-Complementizer	3.53	0.79	3.75	4.85	0.29	1.00
DE-Long	3.89	0.75	3.25	4.78	0.42	1.75
DE-Unprototypical	3.14	0.90	3.50	4.57	0.67	2.50
**that*	2.50	0.75	3.00	2.28	0.92	3.25
*DoubleTn	2.41	0.95	3.25	1.82	0.98	3.75
*Complex NP	1.69	0.56	2.00	1.83	0.90	3.50
**not*	1.31	0.49	1.75	1.15	0.34	1.00

individual variation in the non-linguists' group, as evidenced by the higher standard deviations and greater ranges. (Note, however, that the relatively small amount of variation in the linguists' group is partly due to ceiling and floor effects.)

The difference in the linguists' and nonlinguists' sensitivity to grammaticality was further explored by means of a 2 × 2 split plot ANOVA. The analysis revealed a main effect of Grammaticality (participants gave higher ratings to grammatical than to ungrammatical sentences), $F(1.74) = 1151.03$, $p < 0.001$, η_p^2 (partial eta squared) $= 0.94$, and Group (overall, linguists tended to give higher ratings), $F(1.74) = 18.70$, $p < 0.001$, $\eta_p^2 = 0.20$. The latter effect was qualified by a Group × Grammaticality interaction (see Figure 2): $F(1.74) = 78.01$, $p < 0.001$, $\eta_p^2 = 0.51$. Linguists' mean rating for grammatical sentences (4.64) was significantly higher than the nonlinguists' (3.67), $t(74) = 9.32$, $p < 0.001$. For ungrammatical sentences, linguists' mean rating (1.77) was slightly lower than the nonlinguists' (1.98); the difference approaches significance: $t(74) = 1.77$, $p = 0.080$. (When non-parametric tests are used, the last difference is also significant: Mann Whitney $U = 544.5$, $p = 0.038$.)

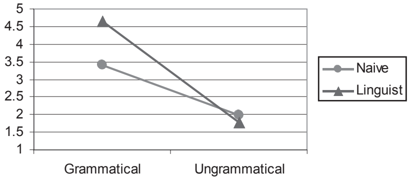

Figure 2: Grammaticality effects

The effects of prototypicality were further analyzed using a 2 × 2 × 2 split-plot ANOVA, with the within subject factors of Prototypicality (Prototypical, Unprototypical)[7] and Construction (Declarative, Interrogative) and the between-subjects factor of Group (Linguists, Nonlinguists). The analysis revealed a main effect of Prototypicality (both groups judged the prototypical variants of both constructions to be more acceptable than the unprototypical variants), $F(1,74) = 150.11$, $p < 0.001$, $\eta_p^2 = 0.670$, qualified by three interactions: Prototypicality × Construction, $F(1,74) = 84.48$, $p < 0.001$, $\eta_p^2 = 0.533$ (in both groups, prototypicality effects were greater for questions than for declaratives); Prototypicality × Group, $F(1,74) = 10.47$, $p = 0.002$, $\eta_p^2 = 0.124$ (the Prototypicality effect was larger in nonlinguists than in linguists); and Prototypicality × Construction × Group: $F(1,74) = 8.55$, $p = 0.005$, $\eta_p^2 = 0.104$ (the difference between linguists and nonlinguists was due primarily to the latter group giving particularly low ratings to unprototypical questions: see Figure 3). There was also a main effect of Group (as discussed earlier, the linguists' ratings for all grammatical sentences were higher than the nonlinguists'), $F(1,37) = 91.45$, $p < 0.001$, $\eta_p^2 = 0.553$, and Construction, (overall, questions were given slightly lower ratings than declaratives), $F(1,74) = 3.66$, $p = 0.060$, $\eta_p^2 = 0.047$. (However, the Wilcoxon Signed Ranks test suggests that the difference is not significant.) This was qualified by a Construction × Group interaction, $F(1,74) = 10.29$, $p = 0.002$, $\eta_p^2 = 0.122$. The linguists gave slightly higher ratings to declaratives than to interrogatives ($t(37) = 4.77$, $p < 0.001$) while the naive informants had a slight preference for interrogatives, although in this case the difference was not statistically significant. (The results of the Wilcoxon Signed Ranks test also suggest an interaction between construction type and group, although in this case the results were significant for naive informants: $Z = -2.222$, $p = 0.023$ but not for linguists: $Z = -1.633$, $p = 0.102$.) There are two possible (mutually non-exclusive) reasons for this

[7] As explained in the Method section, the stimuli set also contained five types of "less prototypical" sentences, i.e., those which involve only one departure from the prototypical variant. For the sake of clarity, these will be analyzed separately below.

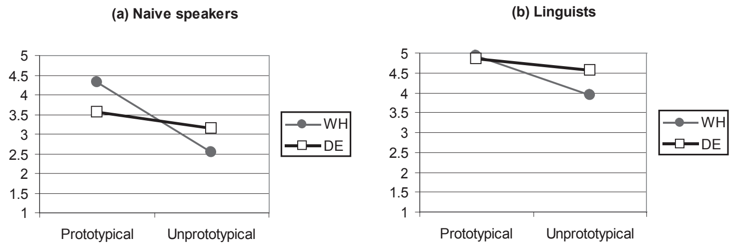

Figure 3: Prototypicality effects

interaction. The linguists' preference for declaratives over interrogatives could be due to a tendency to regard the latter as less canonical than the former, since in many syntactic theories, interrogatives are derived from underlying structures with a declarative word order. Another possibility is that the naive speakers' relatively low ratings for declaratives were attributable to the presence of the initial conjunction, which they might have objected to either because of prescriptivist notions ("You should not begin a sentence with *and* or *but*") or simply because they found it difficult to imagine a context in which it would be appropriate to produce a declarative beginning with *and, but,* or *so*.

Summarizing the results so far, we may say that linguists' judgments are sensitive to grammatical structure and relatively insensitive to lexical content, while the opposite is the case for the nonlinguists. Linguists' greater sensitivity to purely structural properties could be regarded as a virtue (they attend to the interesting, or theoretically relevant aspects of the sentences) or a vice (their judgments are distorted by theoretical commitments). Many linguists regard lexical effects as syntactically irrelevant and uninteresting – just an additional complication which should be "controlled for where possible, discounted when encountered" (Featherston 2005: 702). This is a reasonable position when we are dealing with general phenomena which do not apply to a small number of quirky lexical items. But in this case, as Dąbrowska (2008) argues, we are dealing with a very different case: long-distance WH questions are fully acceptable only with very specific lexical content (main verb = *think* or *say,* main auxiliary = *do,* and so on): thus, they may be more like idiom chunks than productive syntactic patterns.

Moreover, the effects are not purely lexical. The naive informants' judgments show clear interactions between lexical content and grammatical structure: that is to say, the same lexical substitution has a different effect on acceptability judgments for questions and for declaratives. This is shown graphically in Figures 4a, 5a, 6a, 7a, and 8a; a full statistical analysis of these results can be found in Dąbrowska 2008. As shown in the figures, in naive informants,

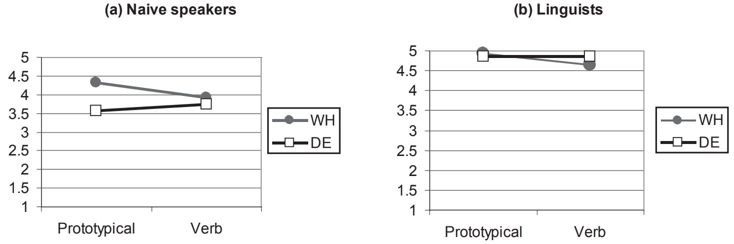

Figure 4: The verb manipulation

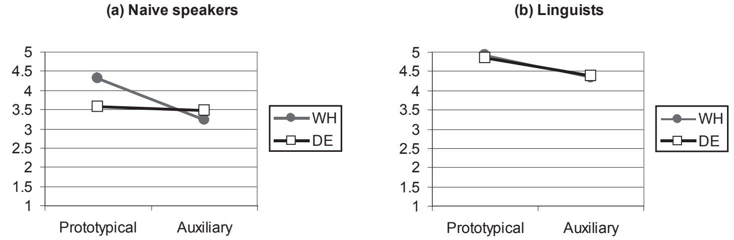

Figure 5: The auxiliary manipulation

the use of a different main clause verb, a modal auxiliary or the addition of an overt complementizer results in considerably lower acceptability ratings for questions but has no effect on declaratives (in fact, declaratives with *believe, suspect, claim,* and *swear* were judged slightly *better* than declaratives with *think* and *say*, although the difference is not statistically significant). Adding an additional complement clause also makes questions with LDDs less acceptable while having no effect on declaratives (Figure 7a). This effect may be partially attributable to processing difficulty (the WH word must be held in working memory while the rest of the sentence is being processed; the more clause boundaries intervening between the filler and the gap, the greater the processing load), although prototypicality probably also plays a role (see Dąbrowska 2008 for discussion). Finally, we also have an interaction between construction type and the lexical status of the subject (Figure 8a) – but in this case, the manipulation makes a difference only for declaratives: that is to say, changing *you* to a lexical subject makes the declarative sentence *more* acceptable, which is almost certainly a pragmatic effect (it is odd to assert what the addressee thinks or says).

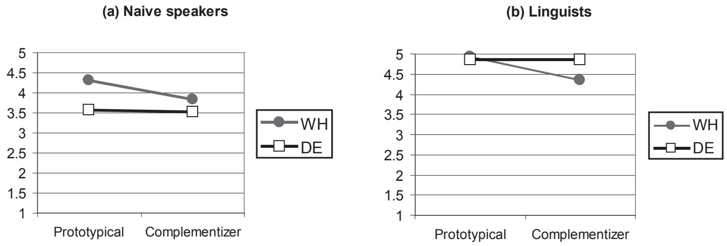

Figure 6: The complementizer manipulation

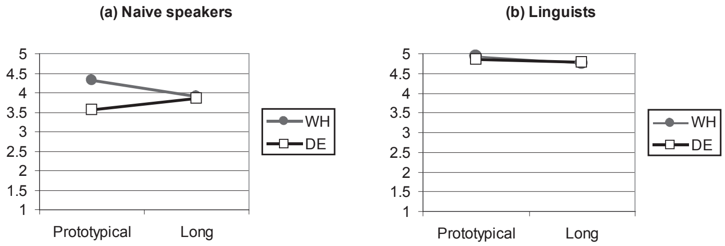

Figure 7: Additional complement clause

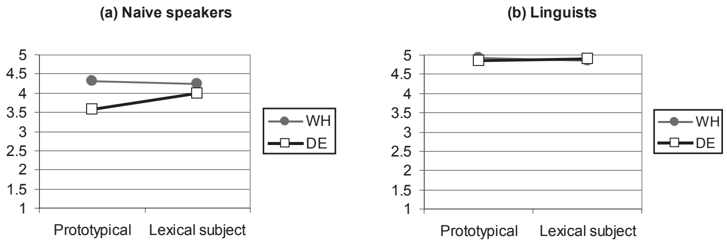

Figure 8: The subject manipulation

In linguists, we see a similar pattern for the verb manipulation (Figure 4b) and for sentences containing complementizers (Figure 6b).[8] In the remaining sentence types, however, the interactions are absent: linguists judge declara-

[8] This is somewhat surprising, since LDD questions in the syntactic literature usually do contain complementizers (cf. Table 1), so this is the place where we should be least likely to find any effects.

tives and interrogatives with lexical subjects to be as good as their 'prototypical' counterparts (Figure 8b), are equally likely to accept prototypical questions, questions with 'very long' dependencies, and their declarative counterparts (Figure 7b), and show the same decrease in acceptability for both constructions when a modal auxiliary is added (Figure 5b).

The differences between linguists and nonlinguists were further explored with five additional $2 \times 2 \times 2$ ANOVAs examining the effects of group, construction and the various prototypicality manipulations (subject, auxiliary, verb, complementizer, and additional complement clause). The analysis confirmed the observations made above. There were three significant three-way interactions: group × construction × subject: $F(1,74) = 7.15$, $p = 0.009$, $\eta_p^2 = 0.09$; group × construction × auxiliary $F(1,74) = 13.19$, $p < 0.001$, $\eta_p^2 = 0.15$; and group × construction × extra complement clause $F(1,74) = 16.46$, $p < 0.001$, $\eta_p^2 = 0.18$. The other three-way interactions (group × construction × verb, group × construction × complementizer) were not significant. The non-parametric analysis suggests a similar pattern of interactions, with one exception: the effect of adding an additional complement clause resulted in a decrease of the acceptability of questions in both groups (for linguists, $Z = -2.460$, $p = 0.014$; for non-linguists, $Z = -3.558$, $p < 0.001$) but to lead to no significant differences for declaratives (for linguists, $Z = -0.905$, p = 0.366; for non-linguists, $Z = -1.868$, $p = 0.062$).

3.2. Functional v. generative linguists

As explained in the Method section, the linguists who participated in the study were also asked about their theoretical orientation. 17 participants identified themselves as cognitive/functional, 16 as generative, and 5 as other. This section compares the responses given by the first two groups.

As we can see from Figure 9, the judgments given by generative and cognitive/functional linguists are fairly similar, although the latter tend to be a little closer to nonlinguists than the generativists. The only sentence type in which there was a significant difference between the two groups of linguists was Complex NP violations, which were given considerably higher ratings by the generativists than by the functionalists, $t(31) = 3.37$, $p = 0.002$. (Note that this difference remains significant even after the Bonferroni adjustment for multiple comparisons: $0.002 \times 18 = 0.036$.)

Why should we find such group differences in the acceptability sentences with Complex NP violations? One possibility is that this is due to the amount of exposure to such sentences: since Complex NPs and other island phenomena are widely discussed in the generative literature, linguists exposed to such ungrammatical example sentences eventually begin to accept them. Some sup-

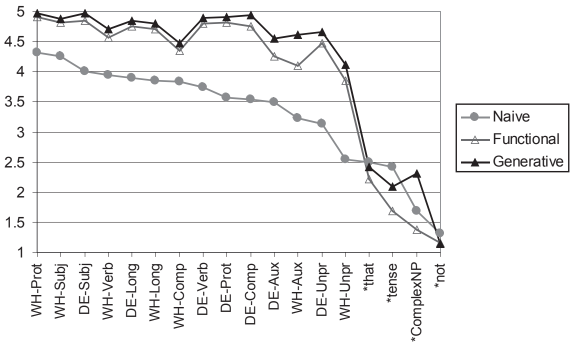

Figure 9: Cognitive/functional and generative linguists' responses (Naive informants' judgments have been added for comparison)

porting evidence for such an explanation can be gleaned from a study by Snyder (2000), who found that naive participants were more likely to accept complex NP and other island violations at the end of the experiment (after they had read a number of sentences instantiating the ungrammatical pattern) than at the beginning.[9]

Such an explanation, however, raises another question, namely, why we don't get analogous differences for sentences with *that*-trace violations, which are also quite frequent in the generative literature. Earlier studies by Snyder (2000) and Hiramatsu (1999) both found *that*-trace sentences to be impervious to satiation, so there appears to be a genuine difference in how repeated exposure affects judgments about these two sentence types. Snyder proposes two possible explanations for the finding that some structures satiate while others do not: satiable sentences such as Complex NP violations may reflect processing limitations rather than linguistic constraints; or the differences in satiability may reflect the fact that the two sentence types involve violations of a different type of constraint. A possible functional explanation would be that complex NP and other island violations are normally rejected because it is difficult to 'see' the intended meaning; however, once one is aware of it, acceptability ratings improve with exposure. *That*-trace violations, on the other hand, are rejected because they are garden-path sentences: the complementizer *that* is preferentially interpreted as subject of the complement clause, and it is very difficult to overcome this bias. Some support for this suggestion can be gleaned from Kandybowicz's (2006) observation that *that*-trace violations are

[9] Note, however, that Hiramatsu (1999) found no satiation effect for complex NP sentences.

acceptable when the complementizer and the gap occur in different prosodic units, since this reduces the likelihood of misparsing.

4. Conclusion

Usage-based models emphasize the effect of linguistic experience on speakers' mental grammars. In the course of language acquisition, learners initially extract lexically specific schemas; more general patterns develop later in acquisition as a result of exposure to a diverse set of exemplars and are gradually entrenched through repeated use (Bybee 2006; Tomasello 2003). One implication of this claim is that general schemas may not emerge when the learner is not exposed to diverse exemplars. Since 'real life' questions with long-distance dependencies are quite stereotypical, ordinary language users could get by with only the lexically specific patterns, or a relatively weak general schema (Dąbrowska 2004, 2008; Verhagen 2005). Linguists, on the other hand, are exposed to a much wider range of LDD questions, and hence have more opportunities to develop general schemas. These differences in experience should result in slightly different grammars and corresponding differences in grammaticality/acceptability judgments. Thus, usage-based theories predict that prototypicality effects for LDD questions will be absent or attenuated in linguists, particularly in generative syntacticians, who encounter non-prototypical LDD questions relatively frequently.

The results reported here are broadly consistent with this prediction, and thus provide some indirect support for two claims made by proponents of usage-based approaches: that mental grammars are shaped by linguistic experience, and that readjustments to the grammatical system may occur in adulthood. This interpretation of the results, is, of course, controversial, and some findings do not fit very easily with usage-based predictions: specifically, it is unclear why linguists' judgments should show prototypicality effects for LDD questions with complementizers, since the relevant example sentences found in the literature often do contain complementizers (cf. Table 1).

It is also possible that the differences between the two groups of informants are due to linguists' beliefs about language – specifically, the conviction that there are two classes of sentences, gramatical and ungrammatical, and that grammatical rules are fully general and apply 'across the board'. It follows that if prototypical instances of a construction are grammatical, so, too, should less prototypical exemplars (cf. Featherston 2005: 702; Schütze 1996: 47, 121). This belief could lead linguists to impose the grammatical/ungrammatical dichotomy on the data. As pointed out in the introduction, expectations have been shown to influence observations in a variety of research contexts; there is no reason to expect linguists to be immune to such biases.

Clearly, further research will be needed to establish whether the observed differences between linguists and nonlinguists are indeed due to differences in experience. Ideally, these should target unusual structures which occur relatively frequently as examples in the syntactic literature and use online measures such as sentence matching (Freedman and Forster 1985; Gass 2001). This will make it possible to determine whether any observed differences are attributable to linguistic or metalinguistic knowledge, since the latter presumably does not affect online performance.

Whatever the final verdict will turn out to be, the research reported in this article indicates that linguists' judgments of the same sentences differ in systematic ways from those of naive informants, even when they are asked to behave like ordinary language users.[10] There may also be differences between linguists of different theoretical orientations, although further research will be necessary to confirm these findings. What is clear is that syntacticians cannot simply rely on their own intuitions and assume that they are representative of the community at large.

Northumbria University
ewa.dabrowska@northumbria.ac.uk

References

Bard, Ellen, Dan Robertson & Antonella Sorace. 1996. Magnitude estimation of linguistic acceptability. *Language* 72. 32–68.

Barlow, Michael & Suzanne Kemmer. 2000. *Usage-based models of language*. Cambridge: Cambridge University Press.

Blaikie, Norman. 2003. *Analyzing qualitative data*. London: Sage Publications.

Borsley, Robert D. 1999. *Syntactic theory. A unified approach*. London: Arnold.

Bradac, James J., Larry W. Martin, Norman D. Elliott, and Charles H. Tardy. 1980. On the neglected side of linguistic science: multivariate studies of sentence judgment. *Linguistics* 18. 967–995.

British National Corpus, *The,* version 2 (BNC World). 2001. Distributed by Oxford University Computing Services on behalf of the BNC Consortium. URL: http://www.natcorp.ox.ac.uk/.

Bybee, Joan. 2006. From usage to grammar: The mind's response to repetition. *Language* 82. 529–551.

Bybee, Joan & David Eddington. 2006. A usage-based approach to Spanish verbs of becoming. *Language* 82. 323–354.

Carnie, Andrew. 2002. *Syntax: A generative introduction*. Oxford: Blackwell.

Chomsky, Noam. 1965. *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Chomsky, Noam. 1977. On *wh*-movement. In Peter W. Culicover, Thomas Wasow and Adrian Akmajian (eds.) *Formal syntax*, 71–132. New York: Academic Press.

Cordaro, Lucian & James R. Ison. 1963. Psychology of the scientist: X. Observer bias in classical conditioning of the planarian. *Psychological Reports* 13. 787–789.

[10] A number of linguists who participated in the study commented that they found it extremely difficult to ignore their metalinguistic knowledge and try to behave like naive informants.

Cowart, Wayne. 1990. Interpreting reflexives in coordinate NPs: English for a non-syntactic analysis of NP coordination. *Eastern States Conference on Linguistics (ESCOL)* 7. 55–66.

Cowart, Wayne. 1997. *Experimental syntax: Applying direct object methods to sentence judgments*. Thousand Oaks, CA: Sage Publications.

Dąbrowska, Ewa. 2004. *Language, mind and brain. Some psychological and neurological constraints on theories of grammar*. Edinburgh: Edinburgh University Press.

Dąbrowska, Ewa. 2008. Questions with 'unbounded' dependencies: A usage-based perspective. *Cognitive Linguistics* 19. 391–425.

Dąbrowska, Ewa. In preparation. Prototype effects in questions with unbounded dependencies.

Dąbrowska, Ewa. 2010. The mean lean grammar machine meets the human mind: Empirical investigations of the mental status of rules. In Hans-Joerg Schmid, Sandra Handl and Susanne Handl (eds.) *Empirical approaches to cognitive linguistics*, 151–170. Berlin: Mouton de Gruyter.

Dąbrowska, Ewa, Caroline Rowland & Anna Theakston. 2009. The acquisition of questions with long-distance dependencies. *Cognitive Linguistics* 20. 571–597

Ellis, Nick C. 2005. At the interface: Dynamic interactions of explicit and implicit language knowledge. *Studies in Second Language Acquisition* 27. 305–352.

Fanselow, Gisbert & Stefan Frisch. 2006. Effects of processing difficulty on judgments of acceptability. In Gisbert Fanselow, Caroline Fery, Matthias Schlesewsky & Ralf Vogel (eds.), *Gradience in grammar,* 291–316. Oxford: Oxford University Press.

Featherston, Sam. 2005. Universals and grammaticality: Wh-constraints in German and English. *Linguistics* 43. 667–711.

Freedman, Sandra E. & Kenneth I. Forster. 1985. The psychological status of overgenerated sentences. *Cognition* 19. 101–131.

Gass, Susan. 2001. Sentence matching: A re-examination. *Second Language Research* 17. 421–441.

Hiramatsu, Kazuko. 1999. What syntactic satiation can tell us about islands. *Papers from the Regional Meetings, Chicago Linguistic Society (CLS)* 35. 141–151.

Jaccard, James & Choi K. Wan. 1996. *LISREL approaches to interaction effects in multiple regression*. Thousand Oaks, CA: Sage Publications.

Jamieson, Susan. 2004. Likert scales: How to (ab)use them. *Medical Education* 38. 1212–1218.

Kandybowicz, Jason. 2006. Comp-trace effects explained away. In Donald Baumer, David Montero and Michael Scanlon (eds.) *Proceedings of the 25th West Coast Conference on Formal Linguistics*, 220–228. Somerville, MA: Cascadilla Proceedings Project.

Kim, Jae-On. 1975. Multivariate analysis of ordinal variables. *American Journal of Sociology* 81. 261–298.

Labovitz, Sanford. 1967. Some observations on measurement and statistics. *Social Forces* 46. 151–160.

Langacker, Ronald W. 1988. A usage-based model. In Brygida Rudzka-Ostyn (ed.), *Topics in cognitive linguistics*, 127–161. Amsterdam: John Benjamins.

Langacker, Ronald W. 2000. A dynamic usage-based model. In Michael Barlow & Suzanne Kemmer (eds.) *Usage-based models of language*, 1–63. Stanford, CA: CSLI Publications.

Levine, Robert D. & Thomas E. Hukari. 2006. *The unity of unbounded dependency constructions*. Stanford, CA: CSLI Publications.

Newmeyer, Frederick J. 1983. *Grammatical theory, its limits and its possibilities.* Chicago: University of Chicago Press.

Noseworthy, John H., George C. Ebers, Margaret K. Vandervoort, R. E. Farquhar, Elizabeth Yetisir & R. Roberts. 1994. The impact of blinding on the results of a randomized, placebo-controlled multiple sclerosis clinical trial. *Neurology* 44. 16–20.

Pell, Godfrey. 2005. Use and misuse of Likert scales. *Medical Education* 39 (9). 970.

Radford, Andrew. 2004. *Minimalist syntax: Exploring the structure of English*. Cambridge: Cambridge University Press.

Riemer, N. (2009). Grammaticality as evidence and as prediction in a Galilean linguistics. *Language Sciences* 31. 612–633.

Roberts, Ian. 1997. *Comparative syntax*. London: Edward Arnold.

Schütze, C. T. 1996. *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*. Chicago: University of Chicago Press.

Snow, Catherine, and Meijer, Guus. 1977. On the secondary nature of syntactic intuitions. In Sideny Greenbaum (ed.), *Acceptability in Language*, 163–177. The Hague: Mouton.

Snyder, William. 2000. An experimental investigation of syntactic satiation effects. *Linguistic Inquiry* 31. 575–582.

Sorace, Antonella & Frank Keller. 2005. Gradience in linguistic data. *Lingua* 115. 1497–1524.

Spencer, N. J. 1973. Differences between linguists and nonlinguists in intuitions of grammaticality-acceptability. *Journal of Psycholinguistic Research* 2. 83–98.

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

Tomasello, Michael. 2003. *Constructing a language: A usage-based theory of child language acquisition*. Cambridge, MA: Harvard University Press.

Tremblay, Annie. 2005. Theoretical and methodological perspectives on the use of grammaticality judgment tasks in linguistic theory. *Second Language Studies* 24. 129–167.

Verhagen, Arie. 2005. *Constructions of intersubjectivity: Discourse, syntax and cognition*. Oxford: Oxford University Press.

Wekker, Herman & Liliane Haegeman. 1985. *A modern course in english syntax*. London: Routledge.