Available online at www.sciencedirect.com

SCIENCE DIRECT®

Lingua 115 (2005) 1481–1496

ELSEVIER

Lingua

www.elsevier.com/locate/lingua

Intuitions in linguistic argumentation

Thomas Wasow[a,*], Jennifer Arnold[b,1]

[a]Department of Linguistics, Stanford University, Stanford, CA, USA
[b]Department of Brain and Cognitive Science, University of Rochester, Rochester, NY, USA

Available online 1 October 2004

Abstract

Generative grammarians have relied on introspective intuitions of well-formedness as their primary source of data. The overreliance on this one type of data and the unsystematic manner in which they are collected cast doubt on the empirical basis of a great deal of syntactic theorizing. These concerns are illustrated with examples and one more detailed case study, concerning the English verb-particle construction.

Keywords: Intuitions; Introspection; Methodology

> [L]anguage should be analysed by the methodology of the natural sciences, and there is no room for constraints on linguistic inquiry beyond those typical of all scientific work.
>
> Neil Smith, ‘‘Foreword’’ to Chomsky (2000: vii).

1. Introduction

The first conjunct of the quote above expresses a sentiment few linguists would disagree with. The second conjunct hints that some critics seek to saddle linguistics with arbitrary discipline-specific methodological restrictions. Quite the contrary, we argue, standards of data collection and analysis that are taken for granted in neighboring fields are widely

* Corresponding author. Tel.: +1 650 723 2472; fax: +1 650 723 5666.
E-mail address: wasow@csli.stanford.edu (T. Wasow).

[1] New address: Department of Psychology, University of North Carolina, Chapel Hill, NC, USA.

0024-3841/$ – see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.lingua.2004.07.001

ignored by many linguists. In particular, intuitions have been tacitly granted a privileged position in generative grammar. The result has been the construction of elaborate theoretical edifices supported by disturbingly shaky empirical evidence.

Two types of intuitions have played a central role in linguistic research over the past half century. The first, which we will call 'primary intuitions', are simply introspective judgements of a given linguistic expression's well-formedness or of its meaning. The second, which we dub 'secondary intuitions', are intuitions about why a given expression is (or is not) well-formed or has the meaning it has.

We have no quarrel in principle with the use of primary intuitions as evidence for theoretical claims. The way they are used in practice, however, is another matter. In Section 2, we discuss how primary intuitions are used in linguistic argumentation and identify two major problems: the way they are collected, and the overreliance on this one type of evidence.

Secondary intuitions are obviously important in helping investigators formulate interesting hypotheses, but we argue in Section 3 that they do not themselves constitute evidence for or against theoretical claims.

Section 4 presents a case study to illustrate the methodological points made in the previous two sections. In particular, we describe a series of investigations we carried out to try to test systematically an intuitive claim made by Chomsky (1955/1975). We conclude that the issue is too complex to be determined simply through introspection.

2. Primary intuitions

A central goal of linguistics is to characterize explicitly the knowledge of language represented in what Chomsky calls the ''mind/brain'' of a speaker. This knowledge manifests itself in the many ways we can and do use language. The most common use of language is conversation; another (at least in literate societies) is writing. Yet another way we can use language is by making introspective judgements about the well-formedness or meanings of expressions. Although most non-linguists rarely make such judgements consciously, it is not difficult to explain the task and to elicit such judgements, even from speakers with little formal education.

2.1. Variation across speakers

The robustness of many judgements of well-formedness is striking—so much so that it initially may seem sufficient to obtain a single native speaker's intuitions. Asking 20 English speakers to judge the well-formedness of *The cat is on the mat* or **Mat the on is cat the* seems pointless: we can be confident that they will all respond alike. Unfortunately, a great many of the crucial examples cited in the syntactic literature are not nearly so clearly good or bad. This is often acknowledged by authors who prefix their examples with some number of question marks.

Moreover, what one speaker finds unequivocally well-formed another speaker may find unequivocally ill-formed. This is clear with known dialect differences, such as those illustrated in (1).

(1). a. %Chris might can go.
b. %Pat's a Red Sox fan, and so aren't we.
c. %He don't like that.

And generations of introductory syntax teachers have discovered that there are critical example types from the literature about which students' judgements are divided. For instance, many speakers find nothing wrong with examples like (2), whose ungrammaticality has frequently been cited as evidence for traces.

(2). %Who did you wanna meet your parents?

Examples like (1) and (2) show that even seemingly robust primary intuitions may not be shared by everyone.

The existence of individual and dialect variation is not incompatible with the use of primary intuitions as evidence for grammatical hypotheses. But it raises questions about the generality of some of those hypotheses. For example, the existence of speakers who accept (2) poses a challenge to anyone who maintains that such sentences are ruled out by principles of Universal Grammar.

In other disciplines in which the evidence exhibits analogous variability, investigators strive to test many individuals and build hypotheses based on what is common across individuals. This is certainly the practice in cognitive psychology and most biological research. Generative grammarians, on the other hand, typically provide judgements as if they held for all speakers, without checking to see how variable they are.

2.2. Marginal intuitions

A more serious problem than the existence of variation in judgements across speakers is that individual speakers are often unsure of their judgements. Contextual factors of various kinds can influence listener reactions to sentences, and this might make a critical difference in judging borderline examples. If linguists consult their own intuitions to test their hypotheses, they have a stake in the outcome of their introspection, which could easily sway their judgements regarding marginal examples. Even when linguists elicit intuitions from others, the predictions of their hypotheses might influence responses, via the 'clever Hans' phenomenon.

These are not novel observations (see Schütze, 1996, for a detailed literature review), nor is it hard to see how to collect intuitions in a way that addresses these problems. In a nutshell, consulting the primary intuitions of native speakers is a type of psychological experiment. Hence, such data collection should be subject to the usual methodological expectations of experimental psychology. In particular:

- The number of subjects should be large enough to allow testing the results for statistical significance.
- The order of presentation of stimuli (that is, linguistic examples) should be randomized.

- The subjects should be ignorant of the hypotheses being tested, preferably with double-blind presentation of stimuli.
- The data collected should be subjected to appropriate statistical analysis.

Unfortunately, such basic precautions are almost unheard of in generative linguistics. Consequently, the journals are full of papers containing highly questionable data, as readers can verify simply by perusing the examples in nearly any syntax article about a familiar language.

We recognize that there are situations in which it is difficult or impossible to follow the methodological suggestions we have laid out here, for example, when there are few available speakers of a language. In these cases it is necessary to weigh the feasibility of alternative techniques against the validity of the data that would result from such techniques. In any case, these situations should be the exception, not the norm.

Even if intuitions were always collected carefully and systematically, they would be only one source of evidence. As noted above, there are many types of linguistic behavior that can – and should – be used as evidence of our knowledge of language. It is surely ‘‘typical of all scientific work’’ (as Smith put it) to test hypotheses against multiple types of evidence using multiple methods.

Psycholinguists have of course employed a variety of paradigms, including reaction-time experiments of various sorts, sentence completion tasks, and eye-tracking, among others. But while debates in psycholinguistics are heavily influenced by developments in linguistics proper, the influence in the reverse direction is minimal. Theoretical linguists formulate hypotheses that they test using only intuitions. Occasionally, they may cite another type of experiment in support of their positions, as when Chomsky described (1968: 65) psycholinguistic ‘‘results [that] showed a remarkable correlation of amount of memory and number of transformations in certain simple cases.’’ But evidence other than intuitions is brought in only as supporting evidence. When the ‘‘derivational theory of complexity’’ that Chomsky cited approvingly turned out not to hold more generally (see Fodor et al., 1974, for a summary), very few syntacticians saw this as a reason to modify their grammatical theories. Rather, it was taken to show that various performance factors obscured the effects of competence grammar in the laboratory experiments.[1]

For reasons that have never been made explicit, many generative grammarians appear to regard primary intuitions as more direct evidence of linguistic competence than other types of data. But there is no basis for this belief. Since knowledge of language is not directly observable, linguists should use every type of evidence available to help us infer what is in speakers’ minds.

Some argue that primary intuitions are cleaner than other forms of data because they somehow escape the semantic and pragmatic dimensions of language use. But making judgements of well-formedness is a type of language use, albeit a somewhat unusual one. Consulting primary intuitions unavoidably involves attempting to assign a meaning and to

[1] The notable exception to this was Bresnan, who suggested (1978, 1982) that all types of evidence about language use are relevant to syntactic theory. In particular, she argued eloquently that the failure of the derivational theory of complexity was a problem for syntacticians, not just for psycholinguists.

imagine a context in which the expression under consideration might be used. By leaving all contextual factors up to the imagination, the use of primary intuitions regarding sentences in isolation is arguably more subject to irrelevant interference than an experimental method that explicitly controls context.

We hasten to add that we are not arguing against the use of primary intuitions in linguistic argumentation. But when they are used, they should be treated as a form of experimental data and evaluated as such. That requires collecting intuitions from multiple speakers (when feasible), with careful attention to the manner of presentation of the stimuli. Moreover, other types of data should play a role in theoretical discussions. Primary intuitions are a legitimate form of evidence for linguistic hypotheses, but they should have no privileged status relative to other forms of evidence.

An old but instructive example of the dominance of informal intuitions in the methodology of generative grammar is provided by the phenomenon that Langendoen et al. (1973) ‘‘dative questions’’. Fillmore asserted (1965: 29–30) that sentences like (3), in which the first object of a double object construction is questioned, are ungrammatical (and, accordingly, he prefixed them with asterisks).

(3). a. Who did I buy a hat?
b. Who did you give this book?

Langendoen and his collaborators (henceforth LKD) probed the primary intuitions of 160 native English speakers on examples like (3b) by asking them to insert *to* into sentences like (4) without changing the meaning.

(4). a. Who(m) did you offer the man?
b. Who(m) did you show the woman?

According to Fillmore’s claim, participants should have uniformly placed *to* between the verb and the following NP. Instead, however, many of the responses placed *to* at the end of the sentence.

LKD also conducted a second experiment, with 109 participants. In this study, participants again read sentences like (4), but this time were instructed to write an answer, using a full sentence with the same verb as occurred in the question. If examples like (3b) were truly ungrammatical, the responses should consistently have treated the postverbal NP in the question as the goal, e.g., *I showed the woman my daughter.* Again, many of the responses indicated that the participants had interpreted the stimuli in the supposedly impossible way, e.g., *I showed my daughter the woman*. Based on these two studies, LKD concluded (p. 469) that ‘‘at least one-fifth’’ of their participants found examples like (3b) acceptable.

How did LKD’s results influence subsequent syntactic literature on dative questions? Not at all. A cursory examination of introductory syntax texts and theoretically-oriented surveys of English grammar published after LKD’s paper revealed that most of them simply did not discuss dative questions. We found three works that did discuss them, namely Culicover (1976: 300),[2] Wexler and Culicover (1980: 275), and Jacobson (1982: 194).[3] All of them repeat the claim that sentences like (3) are ill-formed, giving similar examples marked with asterisks.[4]

In short, the standard unsystematic use of primary intuitions is so entrenched among generative grammarians that contradictory evidence from other sources – including more rigorously collected intuition data – is simply ignored.[5]

Another type of evidence that has been largely ignored in the theoretical syntax literature is usage. With the increasing availability of large on-line corpora of both written and spoken text, it is possible for linguists working on certain languages to check whether their primary intuitions are in accord with what people actually say and write. As in the case of psycholinguistic data, however, usage data gets almost no attention from generativists.

The literature on idioms provides a good illustration of how ignoring usage and instead relying exclusively on intuitions can lead to dubious results. Nunberg et al. (1994) cited a number of claims in the idioms literature that proved inaccurate when tested against corpus data. Riehemann (2001) provides more.

Specifically, Riehemann cites Jackendoff (1997: 170) as claiming that the idiom *raise hell* is syntactically inflexible, specifically unpassivizable. But her search of a large *New York Times* corpus turned up numerous examples of this idiom in non-canonical forms, including the following:

(5). a. So much hell was raised that the biologists threw up their hands in surrender.
b. . . . the internal investigation was reopened “in part because of the hell that Plitman raised about Newcomb’s role in Leatherneck.”
c. Few folks in the Apple speculated on the hell that would have been raised by George Steinbrenner if the Yankees had been similarly robbed at Camden Yards.
d. . . . but how much hell can you raise at Jack-in-the-box?
e. All that was really raised was a little hell and a lot of dust.

[2] The second edition of this book, published in 1982, contains the same discussion on p. 337.

[3] They are also mentioned in Greenbaum (1996), a descriptive grammar that boasts on its cover, “Based on the evidence of real English”. In contrast to the more theoretical works, Greenbaum states (p. 65) that “the indirect object can be questioned by *who(m)* or *what*” and gives *Who do easterly winds bring this extreme cold?* as an example.

[4] Jacobson marks her example (*Who did John give the book?*) with a question mark followed by an asterisk.

[5] A referee questions the relevance of the LKD experiments, on three grounds: (i) that sentences like (4) are “pragmatically bizarre”; (ii) that only about 20% of the participants gave responses that contradict the earlier claims in the literature; and (iii) that there are British/American dialect differences on this point. None of these objections would justify the dismissal of their data. There is nothing unusual, much less bizarre, about showing one person to another. Moreover, the experiments were prompted by LKD’s observation that many of their subjects in a pilot study “had no objections at all to D[ative]Q[uestion]s of any sort.” (LKD: 462). With regard to (ii), if the hypothesis under consideration were a probabilistic one, perhaps an 80% success rate would be acceptable, but this is a very low standard for categorical hypotheses. Finally, we have already addressed the issue of dialect differences; and we doubt that many of the CUNY and Rutgers undergraduates in the LKD experiments were British.

One might object that these examples merely show that Jackendoff's idiolect differs from those of the writers and editors at the *New York Times*. Perhaps. But then the idiosyncracy of Jackendoff's idiolect should render questionable any theoretical claims based on it. Moreover, if it is in principle impossible to question someone else's primary intuitions (as suggested by this objection), then they differ from every other sort of evidence used in science, where replicability of experimental results is normally taken as an essential standard of evaluation.

Riehemann also provides counterexamples to the following claim, first put forward by Koopman and Sportiche (1991) and repeated by Richards (2001):

(6). If X is the minimal constituent containing all the idiomatic material, the head of X is part of the idiom.

This formulation of (6) is somewhat vague about whether it is meant to apply to tokens or types. That is, does every occurrence of every idiom have to conform to (6), or is it only a claim about the canonical forms of idioms? The former (token) interpretation is easily falsified by examples of raised idiom chunks (e.g., *the tables appear to have turned*), so the authors probably intended the latter (type) interpretation.

Even on this interpretation, however, there are counterexamples to (6). Consider, for example, negative polarity idioms like *born yesterday, have a leg to stand on, know . . .. from Adam*, etc. The canonical forms of these idioms presumably have the word *not* in them. Riehemann found that 19 out of 28 occurrences of *born yesterday* contained (contracted or uncontracted) *not*. But the minimal phrase containing both *not* and the rest of such idioms is usually headed by an auxiliary verb, e.g., *was not born yesterday*. And the auxiliary verb is clearly not "idiomatic material".

Another clear example discussed by Riehemann is *from/out of the frying pan into the fire*. Riehemann found these two prepositional phrases occurring idiomatically as complements to a variety of different verbs (including *be, go, leap, move, step, get, throw*, and *take*). It is clear that the verbs are not "idiomatic material", but they are the heads of the VPs that are the minimal phrases containing both PPs.

A similar case is *the cat . . . out of the bag*. The two parts of this idiom must cooccur for the idiomatic interpretation to be possible, but they occur as parts of another phrase headed by something else, usually a verb.[6] The idiom is often listed as including the verb *let* (see, e.g., Spears, 1992), but Riehemann found that 23 out of 48 occurrences of the idiom in the *New York Times* corpus did not contain *let*.[7]

Of course, constituent structure and headedness are theoretical notions, so it is possible that (6) could be maintained in the face of such data by defining these notions in ways that made (6) tenable. Whether a plausible version of (6) is salvageable is beyond the scope of

[6] We say "usually" because it seems to us that sentences like *With the cat out of the bag, the plan had to be abandoned*are well-formed, though we do not have attested examples of such usages. Our point still goes through if the minimal phrase containing both parts of this idiom is always headed by a verb.

[7] Most (but not all) of the others had some form of *be*, so (6) might be maintained by claiming that there are two distinct idioms, *let the cat out of the bag* and *the cat be out of the bag*. But this misses an obvious generalization and still doesn't cover all the data. Further, such a move would render (6) essentially unfalsifiable, since any counterexamples could be handled simply by multiplying idioms.

this paper. Our point is simply that Koopman and Sportiche's analysis could have been more convincing if they had checked – and tried to accommodate – usage data.

Examples of dubious factual claims based on primary intuitions could easily be multiplied. Resolving theoretical controversies often involves examining complex sentence structures, and primary intuitions about these are influenced by many extragrammatical factors. In the absence of independent support from carefully collected and analyzed data, conclusions based on the usual sort of informal primary intuitions should be taken with a substantial grain of salt.

3. Secondary intuitions

Investigators in every discipline have intuitions about what constitutes a plausible explanation. General considerations of parsimony and elegance (often only vaguely articulated) are ‘‘typical of all scientific work’’ and play an important role in the process of discovery. They do not, however, constitute empirical evidence, and their role should be subordinate to primary data. Unfortunately, this is not always the case in linguistics.

For example, the notion that distinct principles should have no overlap in coverage has been repeatedly invoked in arguments for particular formulations of grammatical hypotheses. The first instance of this we are aware of was Chomsky's modification of his Subject Condition (1973: 250) to apply only in subjacent domains, so as to prevent examples like the following from being ruled out by both the Subject Condition and the Subjacency Condition:

(7). *What did that John saw surprise Mary?

The idea that conditions ‘‘with overlapping empirical coverage ... are wrongly formulated’’ was later made explicit as ‘‘a working principle’’ (Chomsky, 1995: 5). Note that this working principle is not about cases in which one principle is dispensable because it is subsumed under another. So it is not merely an instance of Ockham's razor. Rather, it concerns cases in which both principles are independently motivated, but apply to some cases in common.

On one possible interpretation this working principle would be very odd, for it would rule out overdetermination of ungrammaticality. Many non-linguistic facts are overdetermined. For example, a person submerged in freezing water for more than a few minutes will die, from some combination of lack of oxygen and hypothermia. And it is not hard to invent linguistic analogues that nobody would argue were the kind of overlapping empirical coverage Chomsky proposed to disallow. A non-sentence like **Who did you think Pat and was talking* is impossible because: (i) it involves a filler-gap dependency in which the gap is a coordinate conjunct (cf. **Who did you think Pat and were talking*); and (ii) the subject and verb of the subordinate clause do not agree (cf. **Did you think Pat and someone was talking*). It seems unlikely that Chomsky was proposing to exclude this sort of overlapping coverage.

A more plausible interpretation of Chomsky's working principle is that it rules out cases in which a single phenomenon is covered by two distinct principles. However, the judgement that the existence of such overlap is undesirable is essentially an aesthetic one. Investigators in all disciplines are guided in part by such non-empirical considerations, but they do not constitute evidence for or against an analysis, and they should carry little, if any, weight as arguments for one analysis over another. Moreover, the determination that a given case of overlapping coverage involves one phenomenon rather than two is highly subjective. It is, in fact, a case of what we are calling secondary intuitions: it is an intuition about how to analyze the unacceptability of a given sentence, not about the acceptability itself. In the absence of a well defined and well justified method of individuating linguistic phenomena, claims that principles have overlapping coverage should carry little, if any, weight.

Another instance of using a secondary intuition as an argument is the claim that one can tell by the degree of ill-formedness of an example what constraint it violates. Chomsky employed this form of argumentation, suggesting (1986: 80) that **from which city did you meet the man* is not bad enough to be a violation of the Empty Category Principle, and seems more like a Subjacency violation. Similar claims appear elsewhere in the literature.

At first glance, this sort of argumentation seems quite reasonable. After all, it is common for scientists to use variations in the strength of an experimental effect in inferring an explanation. To be persuasive, however, such arguments have two prerequisites: (i) differences in the effect strength need to be carefully documented; and (ii) a theory is needed of how the purported difference in causes would yield the observed differences in effects. Arguably, the theory of barriers satisfies (ii), though the relative weakness of Subjacency violations remains essentially a stipulation. In any event, (i) is not satisfied: claims about degrees of unacceptability are consistently simply assertions, based on casual introspection.

The failure to satisfy (i) is of course really about the use of a kind of primary intuition, not about the related secondary intuition. But in this case, the intuitions are fine-grained ones of relative acceptability, rather than the simple binary intuitions we discussed in the preceding section. Hence, the failure to observe normal scientific standards of data collection renders the results even more questionable.

4. A case study

This section presents a case study that illustrates some problems with depending on intuitions (both primary and secondary) as the sole source of data. We consider an old claim of Chomsky's concerning the roles of constituent length and complexity in explaining the position of verb particles. Understanding what factors influence constituent ordering in this construction is relevant to theories of both grammatical competence and performance (see Wasow, 2002, for further discussion). Our data reveal a somewhat different pattern of results from the one predicted by Chomsky's original intuitions, demonstrating subtleties not directly accessible to intuitions.

The following passage is from the founding document of generative syntax, *The Logical Structure of Linguistic Theory*:[8]

> While . . .. both [*the detective brought in the suspect*] and [*the detective brought the suspect in*] are grammatical, in general the separability of the preposition is deter-

[8] This monumental work was completed in 1955 but remained unpublished for 20 years. We list it in the references as Chomsky (1955/1975). The passage quoted occurs on p. 477 of the published version.

> mined by the complexity of the NP object. Thus we could scarcely have ... *the detective brought the man who was accused of having stolen the automobile in*
>
> It is interesting to note that it is apparently not the length in words of the object that determines the naturalness of the transformation, but, rather, in some sense, its complexity. Thus 'they brought all the leaders of the riot in' seems more natural than 'they brought the man I saw in.' The latter, though shorter, is more complex ...

This passage is of interest in the present context because it involves appeals to both primary and secondary intuitions. The primary intuitions in the first paragraph concern the well-formedness of the examples, and those in the second paragraph concern the relative naturalness of two sentences. Chomsky simply assumed that readers would share these intuitions. The secondary intuitions are that complexity, not length, is the relevant factor, and that *the man I saw* is more complex than *all the leaders of the riot*. The second of these intuitions is offered as evidence in support of the first (in conjunction with the primary intuition about the relative acceptability of the examples), but, in the absence of any attempt to characterize complexity objectively, the second paragraph is little more than the assertion of a (secondary) intuition.

To the best of our knowledge, these assertions about the verb-particle construction have gone untested in the intervening decades. We attempted to rectify this situation through a combination of a questionnaire study and corpus analysis. For our questionnaire study, we assumed (as Chomsky evidently did in the passage under discussion) that NPs containing subordinate clauses are more complex than NPs that do not contain clauses. We then constructed pairs of NPs that were identical in length and similar in meaning but differed on this criterion of complexity, e.g., *everything we said* versus *all our instructions* and *what follows* versus *the consequences*. We used these NP pairs to construct quadruples of example sentences, crossing NP complexity with constituent order, along the following lines:

(8). a. The children took everything we said in.
b. The children took in everything we said.
c. The children took all our instructions in.
d. The children took in all our instructions.

We also constructed similar quadruples of sentences involving two other alternations in English, the dative alternation and heavy NP shift. This was done to explore whether the role of (this criterion of) complexity in influencing ordering was the same across constructions. (9) and (10) give examples of quadruples used to test these other two alternations.

(9). a. The company sends what Americans don't buy to subsidiaries in other countries.
b. The company sends subsidiaries in other countries what Americans don't buy.
c. The company sends any domestically unpopular products to subsidiaries in other countries.
d. The company sends subsidiaries in other countries any domestically unpopular products.

(10). a. Nobody reported where the accident took place to the police.
b. Nobody reported to the police where the accident took place.
c. Nobody reported the location of the accident to the police.
d. Nobody reported to the police the location of the accident.

Each questionnaire consisted of 12 test sentences, randomly interspersed among 20 filler sentences (constructed so as to be marginal in acceptability). Each questionnaire included exactly one sentence from each of the 12 quadruples, so that no participant would see more than one variant of any sentence. Participants were asked to rate the acceptability of each sentence on a four-point scale: 4 for 'fully acceptable', 3 for 'probably acceptable, but awkward', 2 for 'marginal, at best', and 1 for 'completely unacceptable'. The instructions said, 'Rely on your own intuitions of what sounds good, not on what you think is correct according to the experts.'

Each questionnaire was filled in by 22 participants (almost all of them Stanford undergraduates), making a total of 88 completed questionnaires. Each subject saw only one questionnaire. One participant gave scores of 1 to all but one of the sentences on the questionnaire, so this individual's responses were discarded. To keep the numbers of participants balanced, we randomly selected one participant from each of the other questionnaires and discarded that participant's data. Thus, results were computed based on 21 responses to each questionnaire.

The results for the verb-particle construction provide some support for Chomsky's intuitions. Scores for all examples with the particle adjacent to the verb (what we will call the 'joined' ordering) were very high, averaging 3.3 when the NP was complex and 3.4 when the NP was simple. When the particle followed the object NP (what we will call the 'split' ordering), the mean scores fell off, particularly if the NP was complex: 2.8 for simple NPs and 1.8 for complex NPs. An analysis of variance revealed that the interaction between complexity and ordering was significant ($P < 0.001$) by subjects, but not by items ($P > 0.1$).

Interestingly, there was considerable variation in the responses. In particular, although the mean score for split examples with complex NPs was far lower than any of the others, 17% of the responses to such sentences were scores of 3 or 4. That is, about one-sixth of the time, participants judged such examples to be no worse than awkward. When primary intuitions are so variable, it is necessary to test a sizeable sample in order to get a reliable picture of what is going on. This sort of variability is the rule, not the exception, when it comes to judgements of acceptability.

The questionnaire results for the heavy NP shift examples were cleaner than those for the verb-particle examples. When the NP was simple, participants gave higher scores to the canonical (V-NP-PP) order, and when the NP was complex, they gave higher scores to the shifted (V-PP-NP) order. Mean scores were quite high across the four conditions, ranging from 2.8 to 3.4. And the interaction between complexity and order was significant both by subjects ($P < 0.001$) and by items ($P < 0.05$). These results support the notion (implicit in the label 'heavy NP shift') that the shifted ordering is preferred when the NP contains a subordinate clause. On the other hand, the 3.0 mean score for shifted examples with simple NPs undermines the claim made by Ross (1967, Rule 3.26) that only NPs dominating S can occur in the shifted position. This illustrates nicely the need to test secondary intuitions empirically.

The results for the dative alternation were relatively difficult to interpret, because half of the stimuli manipulated the complexity of the theme NP while the other half manipulated the complexity of the goal NP. Consequently, we got only half as many responses in each of these conditions as for the other two alternations. Since our primary concern in the present context is the verb-particle construction, we will not discuss the dative alternation data in any detail.[9] Suffice it to say that these data also revealed a preference for complex constituents to come late.[10]

This questionnaire study provides some empirical support for Chomsky's suggestion that a particular criterion of complexity influences preferences in word order. However, the judgements collected in the study exhibit interesting patterns that are not accessible to casual introspection. Moreover, Chomsky's claim was not only that complexity was relevant to the ordering preferences, but that length was not. Even if the questionnaire data had shown unequivocally that complexity influenced ordering, it would not follow that length does not. It is entirely conceivable that the position of the verbal particle could be sensitive to both the length and the complexity of the NP object. To investigate this issue, we conducted corpus studies using parsed corpora made available through the Treebank of the Linguistic Data Consortium. In particular, we extracted data from two parsed written corpora, *The Wall Street Journal* and Brown, and one spoken corpus, the Switchboard. In these corpus studies, we examined the verb-particle construction and the dative alternation.

We extracted all VPs immediately dominating a particle (in either the split or joined construction) and all VPs exhibiting either form of the dative alternation. Double object forms were identified simply by selecting VPs directly dominating two NPs, and prepositional forms were identified through the tag PP-DTV, used in recent versions of the Treebank to mark PPs headed by *to* that introduce a goal argument. Examples whose direct objects were personal pronouns were excluded, since they are obligatorily split (in the case of the verb-particle construction) or nearly obligatorily prepositional (in the case of the dative alternation). We inspected all remaining examples to make sure that they were in fact instances of the alternations we were interested in. The resulting datasets included 1393 examples of dative alternation (856 from written sources, 537 from spoken) and 3268 examples of verb-particle constructions (2264 from written sources, 1004 from spoken). We then coded the relevant NPs in all of the examples on a three-level complexity scale: 'complex' for NPs containing verbs: 'prepositional' for noncomplex NPs containing PPs; and 'simple' for all other NPs. We also coded these same NPs for length (in words). Each example VP was coded for its constituent ordering (split versus joined verb-particles, and theme-first versus goal-first, in the dative alternation).[11]

We used a logistic regression to examine the degree to which complexity and length accounted for ordering in each dataset (Verb-Particle and Dative Alternation) A logistic regression can identify whether a given factor (e.g., length) accounts for variance in a dataset, given the presence of other factors (e.g., complexity) in the model. In addition to

[9] See Wasow (2002: Chapter 2) for more discussion.

[10] The interaction of complexity and order was significant by subjects ($P < 0.001$) in both conditions and by items for the goal manipulation ($P < 0.05$) but not the theme manipulation ($P > 0.1$).

[11] All of these codings were done by computer programs written by Adam Yarnold. We checked the outputs for accuracy.

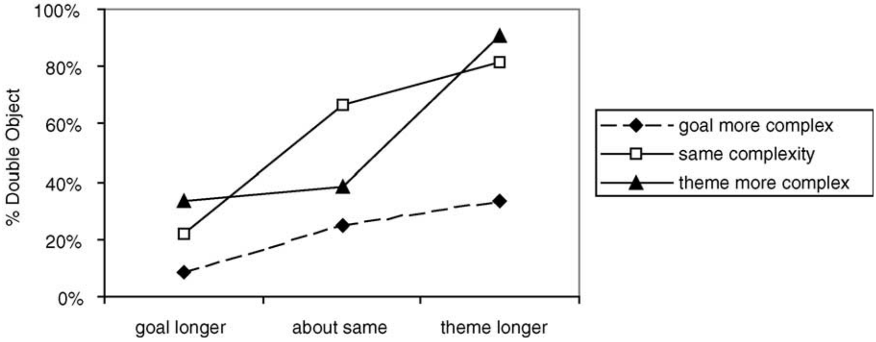

Fig. 1. Dative alternation in switchboard. Brown and *Wall Street Journal* Corpora.

length and complexity, 'corpus' (written versus spoken) was included as a dependent variable.

In the verb-particle construction, all complex direct object NPs[12] occur after the particle – that is, VPs with complex object NPs categorically exhibit the joined ordering. Despite this fact, the logistic regression indicated that length was a significant predictor of ordering ($P < 0.001$) but complexity was not ($P > 4$).[13] An inspection of the data suggests that this may be a ceiling effect: almost all NPs longer than four words occur in the joined ordering, and virtually no NPs shorter than five words are complex. Hence, the behavior of complex NPs in these datasets is predictable on the basis of length alone.

Interestingly, the number of prepositional NPs shorter than five words is not negligible: 86 written and 59 spoken. And these NPs are no more (or less) likely to appear in the split construction than simple NPs of the same length. This suggests either that PPs do not increase the complexity of NPs that contain them or that Chomsky was wrong about particle position being sensitive to complexity rather than length. So, while the questionnaire data might be taken as support for one of Chomsky's secondary intuitions (that complexity influences ordering), the corpus data argue against both that claim and the claim that length does not influence ordering.

The apparent ceiling effect that made the corpus data on verb-particles hard to interpret does not occur with the dative alternation data. That is because there are two postverbal NPs, and what is relevant to their ordering is their relative weight (see Hawkins, 1994, and Wasow, 1997, for arguments to this effect). Fig. 1 summarizes the results for the dative alternation.

The horizontal axis here represents the difference in length between the two NPs, and 'about the same' means that they are either the same length or differ in length by one word. A logistic regression indicates that ordering is significantly influenced by both length ($P < 0.001$) and complexity ($P < 0.05$).[14]

[12] There are 98 complex NPs in the spoken corpus and 300 complex NPs in the written corpus.

[13] There was also a significant effect of corpus: 90% of the utterances were joined in the written items, but only 62% were joined in the spoken items.

[14] The effect of corpus was also a significant predictor of the dative alternation data: 87% of the written items occurred in the double object construction, 67% of the spoken items.

The corpus studies described so far uniformly indicate that length plays a role in constituent ordering, contrary to Chomsky's secondary intuition. Moreover, the verb-particle data fail to support Chomsky's claim that complexity influences ordering.

One potential problem with these analyses, however, is that they are predicated on a particular operationalization of the notion of 'complexity', a notion which Chomsky invoked but did not attempt to characterize precisely. A defender of the claim that complexity, but not length, influences ordering might argue that our three-level complexity scale is simply too crude. Perhaps a finer-grained measure of complexity could predict ordering as well as the combination of this coarse complexity measure and length, thus vindicating Chomsky's secondary intuitions. In order to explore this possibility, it is necessary to select a complexity measure that is at once fine-grained, theoretically motivated, and objectively measurable. The obvious candidate is number of nodes, argued for by Hawkins (1994), and automatically computable from the Treebank data.

We reanalyzed the data discussed above, using node counts instead of the three-way classification of complexity. In counting nodes, we accepted the Treebank parses, but excluded empty nodes from our counts. Thus, for example, in the VP *offering special discount packages to big customers, called Tariff 12*, the Treebank includes an empty VP node in the NP *discount packages*, corresponding to the extraposed phrase *called Tariff 12*; but we coded this NP as simple, with a node count of five[15] (and a word count of three).

Using relative nodes as a measure of complexity still did not vindicate Chomsky—the data patterns were identical to the analyses with complexity.[16] Thus, these data provide no support for Chomsky's claim that length does not matter, and only weak evidence that complexity does.[17]

These explorations into length and complexity are not the final word on what determines constituent ordering. Considered in isolation, length appears to have a very significant influence on ordering (as first noted by Behaghel, 1909/1910). By the same token, any of a number of metrics of complexity, taken by themselves, would be good predictors of

[15] This is actually one more node than would have appeared if there had been no empty node. The Treebank assigned the following structure to the NP in question:

```
(NP (NP (JJ special)
(NN discount)
(NNS packages))
(VP (-NONE- *ICH*-1)))
```

The outermost NP bracket would not have been included if the empty VP at the end were not present. In short, extraposed elements added one to the node count, but were ignored for purposes of word count and complexity coding.

[16] The only exception is that relative length was only a marginally reliable predictor of dative alternation ($P = 0.06$), along with relative number of nodes ($P < 0.05$).

[17] A closer inspection revealed that the correlation between length and number of nodes in these datasets was extremely high (0.99 in both datasets). And given the lack of agreement among syntacticians about the details of tree structures, the node counts we employed in this study must be considered only approximate measures of NP complexity. These considerations weaken the force of the length versus nodes comparison. What remains consistent, however, is that every quantitative study including length as a factor indicates that it influences order, contrary to Chomsky's intuition.

ordering. Our questionnaire data suggest that there is an effect of complexity distinct from length. But they are entirely consistent with the hypothesis that both length and complexity are relevant to ordering, a possibility our corpus studies generally support.

The empirical issue we have been discussing – whether length, complexity, or both influence postverbal constituent ordering in certain constructions – needs to be investigated using a wide range of empirical methods. We have shown how questionnaire data and corpus studies can be brought to bear on the issue. Other types of evidence, such as acquisition data, eye-tracking, reaction times, or brain studies, could be employed. This question is not one that can be resolved simply by introspection. It requires a precise characterization of what is meant by complexity; in the ideal case, such a metric would follow from a well-motivated theory of linguistic structure or processing. In addition, the predictions of the metric must be carefully tested, using well-controlled empirical methods, including, but not limited to, systematically collected intuitions of well-formedness.

5. Conclusion

Disciplines differ considerably in the relative emphasis they place on data collection versus theory construction. In physics, there is a clear division of labor between experimentalists and theorists. Linguistics, too, has subfields (including psycholinguistics and sociolinguistics) in which theories tend to be data-driven and others (notably generative grammar) that focus almost exclusively on the formulation of elegant theories, with little attention devoted to careful data collection. Unfortunately, the findings of the experimentalists in linguistics very rarely play a role in the work of generative grammarians. Rather, theory development tends to follow its own course, tested only by the unreliable and sometimes malleable intuitions of the theorists themselves. The theories are consequently of questionable relevance to the facts of language.

In sum, linguistic inquiry should be subject to the methodological constraints typical of all scientific work.

Acknowledgements

We are grateful to Susanne Riehemann for helpful discussions of this paper. The paper has also benefited from the comments of Bob Borsley and two reviewers. Any remaining shortcomings are, of course, solely our responsibility.

References

Behaghel, O., 1909/1910. Beziehungen zwischen Umfang und Reihenfolge von Satzgliedern. Indogermanische Forschungen 25, 110–142.

Bresnan, J., 1978. A realistic transformational grammar. In: Bresnan, J., Halle, M., Miller, G. (Eds.), Linguistic Theory and Psychological Reality, MIT Press, Cambridge, MA, pp. 1–58.

Bresnan, J., 1982. The Mental Representation of Grammatical Relations, MIT Press, Cambridge, MA.

Chomsky, N., 1955/1975. The Logical Structure of Linguistic Theory, University of Chicago Press, Chicago.

Chomsky, N., 1968. Language and the mind. Psychology Today 48–68. Reprinted in: D.D. Bornstein. Readings in the Theory of Grammar. Winthrop Publishers, Cambridge, MA, pp. 241–251.

Chomsky, N., 1973. Conditions on transformations. In: Anderson, S.R., Kiparsky, P. (Eds.), A Festschrift for Morris Halle, Holt, Rinehart, and Winston, New York, pp. 232–286.

Chomsky, N., 1986. Barriers, MIT Press, Cambridge, MA.

Chomsky, N., 1995. The Minimalist Program, MIT Press, Cambridge, MA.

Chomsky, N., 2000. New Horizons in the Study of Language and Mind, Cambridge University Press, Cambridge.

Culicover, P., 1976. Syntax, Academic Press, New York.

Fillmore, C., 1965. Indirect Object Constructions in English and the Ordering of Transformations, Mouton, The Hague.

Fodor, J.A., Bever, T.G., Garrett, M.F., 1974. The Psychology of Language, McGraw-Hill, New York.

Greenbaum, S., 1996. The Oxford English Grammar, Oxford University Press, Oxford.

Hawkins, J.A., 1994. A Performance Theory of Order and Constituency, Cambridge University Press, Cambridge.

Jackendoff, R.S., 1997. The Architecture of the Language Faculty, MIT Press, Cambridge, MA.

Jacobson, P., 1982. Evidence for gaps. In: Jacobson, P., Pullum, G.K. (Eds.), The Nature of Syntactic Representations, D. Reidel, Dordrecht, pp. 187–228.

Koopman, H., Sportiche, D., 1991. The position of subjects. Lingua 85, 211–258.

Langendoen, D.T., Kalish-Landon, N., Dore, J., 1973. Dative questions: a study in the relation of acceptability to grammaticality of an English sentence type. Cognition 2, 451–477.

Nunberg, G., Sag, I.A., Wasow, T., 1994. Idioms. Language 703, 491–538.

Richards, N., 2001. An idiomatic argument for lexical decomposition. Linguistic Inquiry 32, 183–192.

Riehemann, S.Z., 2001. A Constructional Approach to Idioms and Word Formation. Stanford University Dissertation.

Schütze, C.T., 1996. The Empirical Base of Linguistics, University of Chicago Press, Chicago.

Spears, R.A., 1992. NTC's American Idioms Dictionary, National Textbook Company, Lincoln, IL.

Wasow, T., 1997. Remarks on grammatical weight. Language Variation and Change 9, 81–105.

Wasow, T., 2002. Postverbal Behavior, CSLI Publications, Stanford.

Wexler, K., Culicover, P., 1980. Formal Principles of Language Acquisition, MIT Press, Cambridge, MA.