# The Forests behind the Trees

John Nerbonne
Center for Language and Cognition, Groningen
P.O. Box 716
University of Groningen
NL 9700 AS Groningen
Netherlands

Tel. +31 50 363 58 15
FAX +31 50 363 68 55
http://www.let.rug.nl/nerbonne/
j.nerbonne@rug.nl

**Abstract:** Syntactic databases are increasingly available and are put to a variety of uses, including serving as organized reference material for descriptive and theoretical syntacticians. Dense databases recording fine variation within a single language area, so-called "microvariation", play a prominent role with respect to this use. In addition the large collections allow syntactic variation to be studied quantitatively in dialectology and in the analysis of second-language, pidgin and creole varieties. The large collections enable exploratory, "data-mining" approaches, and are well positioned to detect statistical tendencies that may be imperfect, and therefore not universal. Finally, some researchers have hypothesized that syntactic features may be more stable over long periods of time than lexical or phonetic features and are investigating whether syntactic structure bears a signal of historical relatedness. This work too requires quantitative analysis that is only possible with large, systematic collections. This article introduces a special issue of *Lingua* devoted to presenting and exploring research using large syntactic databases.

Keywords: Syntax, Diachronic Syntax, Syntactic Variation, Syntactic Database, Aggregation, Datamining,

## 1. Introduction

Syntactic databases are increasingly available and are put to a variety of uses. They serve as the organized reference material for descriptive and theoretical syntacticians, where both coarse-grained (typological) as well as fine-grained (microvariational) collections have their purposes. They provide the syntactic counterparts to traditional dialect atlases, which have concentrated on lexical and phonetic variation, thereby enabling studies in quantitative dialectology. Databases of syntactic properties in second-language, pidgin and creole varieties provide material for further variationist research, enable exploratory, "data-mining" approaches, as well as the analysis of statistical tendencies that may be imperfect, and therefore not universal. Finally, some researchers have hypothesized that syntactic features may be more stable over long periods of time than lexical or phonetic features and are investigating whether syntactic structure bears a signal of historical relatedness. This work too requires quantitative analysis that is only possible with large, systematic collections. This article introduces a special issue of *Lingua* devoted to presenting and exploring research using large syntactic databases.

The papers report on research enabled by the use of substantial collections of syntactic material, with various sorts of research questions being asked from theoretical syntax, typology, historical linguistics, dialectology and variationist linguistics, and using various sorts of collections, some further instantiations of the typological collections, others inspired by dialectological atlases and yet others—the majority—designed for particular studies. We report on a range of collections and uses here, emphasizing the novel sorts of research that has been done, and in particular those with a perspective that requires the large data collections.

A great deal of syntactic theory has been developed from rather little data, intuitively judged with respect to its well-formedness, and involving almost no formal organization. This has been a regular cause for comment and concern, most of which revolves around discussions about the reliability of intuitive judgements (Schütze, 1996), and not on the narrowness of the empirical base. Students of syntax are taught to "think in trees" for the field emphasizes the methods of analyzing individual sentences, phrases and constructions. But evidence is mounting that intuitions may fail to consider the range of uses of a construction, thereby rushing to negative judgments (Bresnan, 2007).

The narrow empirical base is also at odds with the one long-standing explicit goal of syntactic theory, the characterization of what is syntactically possible in human language. Accordingly, two papers in the current special issue focus on the opportunities for theoretical syntax which exist once one begins to exploit large collections, Barbiers's paper on theoretical benefits accruing to the use of a Dutch database of microvariation, and Zwart's on "data-mining" a broader collection of syntaxes. Zwart's paper uses a "typological" perspective which has always been broader that that of syntactic theory (Comrie and Smith, 1977; Comrie, 1989), and has long advocated the use of larger collections of data. In both cases the database of syntactic information allows the research to easier access to data which is then analysed in a standard fashion. In some sense these papers go against the grain

suggested by the saying "you can't see the forest for the trees", but trees are interesting in their own right, and what could be better place than forests to find them? Further, in both cases the authors are concerned with the *range* of possible syntactic constructions, and look therefore beyond the analysis of single constructions.

Spruit, Heeringa and Nerbonne use the same database as Barbiers, but approach it from a dialectological perspective, noting that geographic coherence is every bit as influential in syntax as in phonology or the lexicon. They go on to ask whether the geographic influence follows the same paths in syntax as it does in pronunciation and vocabulary. Szmrecsanyi and Kortmann also pose variationist questions in examining a collection of English varieties from around the world, asking about their natural division into classes of varieties, and also which syntactic variables characterize the natural classes they find. These two papers both ask questions about the properties of the aggregate collection, essentially the range of variation in it, whether there are relative clusters of similar varieties, and what defines these linguistically. So it is fair to say that both focus on the properties of forests rather than those of trees. Both of these papers, like the historical ones (below), would be inconceivable without extensive computational support. Not only is the data digitized, but the analytical procedures, the statistical analysis of the results and their visualization all require significant processing.

Dunn turns to a syntactic database in his work in historical linguistics on Island Melanesian because phonological and lexical evidence of shared Papuan history is scarce, and Longobardi and Guardiano postulate that syntactic features are more resistant to change than the other linguistic features more commonly used in historical analysis. Their program is to add to the techniques of historical reconstruction in linguistics, which rely primarily on phonetic and morphological evidence. The fact that they both deal with large, digitized sets allows them to exploit novel computational tools for inferring historical developments. Like the variationist investigations, the historical ones examine an aggregate of languages in an attempt to infer a shared history. They make sense only at a high level of aggregation.

The present collection is not representative of all ongoing work on large syntactic collections, in particular that on corpora, both annotated and unannotated. We admire this work but find that there are enough fora. Appeals to the web as a source of data are becoming commonplace, and indeed result in striking insights (Bresnan, 2007), while very exciting syntactic work is being carried out using automatically parsed corpora (Bouma, 2008) and tools are being developed to support linguistically informed search in large corpora (Bouma and Kloosterman, 2007). This line of work already has its own series of *International Workshop on Treebanks and Linguistic Theories*. The sixth was held December 7-8, 2007 in Bergen, Norway (http://tlt07.uib.no/), and the eighth has been announced for Jan. 23-24, 2009 in Groningen (http://www.let.rug.nl/tlt/). The new journal *Corpus Linguistics and Linguistic Theory* is also focused on just this juncture.

We provide a brief preview of the papers in the rest of this introduction.

## 2. The Papers

### 2.1 Typology and Syntactic Theory

"Relevance of typology to minimalist inquiry" by Jan-Wouter Zwart proceeds from the minimalist program (MP, Chomsky 1995), and finds that his research benefits from the organized access to a large collection of data. Zwart sees an opportunity to extend the usual narrow empirical base of theoretical grammar, which normally consists of few intuitive judgments of grammaticality, while he makes use of information on 214 languages, many of which are genealogically unrelated, in the tradition of typology (Comrie 1989, Dryer et al., 2003). As we shall note below, Barbiers uses a dense set of varieties, involving many very similar dialects, while Zwart's collection is more comprehensive and involves dozens of unrelated language families.

Syntax clearly deals with ordered sequences of words at some level, but MP explains some aspects of order on the basis of the interface between the structure-building module and a phonological module. Given this background, Zwart asks whether the order sensitivity is also present in the structure-building operation Merge. The issue is difficult to address because 'order' is understood abstractly, *not* as concerning the succession of words as pronounced or written (which words come earlier and which later), but rather as concerning the status of the input elements once Merge has applied. Normally the input elements remain recognizable even after the merger has been effected, so Zwart then asks whether the two elements have the same status with respect to subsequent syntactic operations. If they do, then Merge may be said to produce unordered output. If one input has a distinguished status, this is equivalent to construing the output as ordered.

Given how abstract the research question is, any answer would seem to require extensive preliminary analysis and consideration of a large range of alternative analyses, some potentially quite complex, and these are not the sorts of activities which benefit immediately from access to a large data collection.

But Zwart sets his sights on coordination, focusing in this paper on symmetric structures, and avoiding structures where order follows from temporal succession (*He came in and he sat down*) or a causal relation (*He won the game and the match*). Coordination is an interesting choice as one would imagine the subparts of coordinate structures to *more* like each other in grammatical status than the subparts of other grammatical structures, and not the easiest choice if one wishes to demonstrate that grammatical structures are ordered in the relevant sense. But Zwart examines the 214 languages in his sample, and shows that while one finds both medially conjoined [α & β] (where the ampersand represents the coordinating conjunction), as well as the finally marked [α β &], nonetheless, all the apparent examples of medial conjunction are in fact best analysed as finally marked, i.e. [α (& β)] rather then [(α &) β]. He provides evidence for this in the form of alternative expressions. For example, English is superficially medially conjoined, but if conjuncts are separated, then the conjunction is invariably expressed with the second element, and never with the first, thus *I saw Bill yesterday[… and Dave]* but never **I saw [Bill and …] yesterday Dave.*

The paper nicely illustrates how a sharp theoretical question can lead to empirical questions and novel empirical results. Because the empirical question fortunately did not require a great deal of preliminary analysis or attention to a large range of analytical options, it could be answered on the basis of a substantial reservoir of syntactic information.

**2.2 Microvariation**

Sjef Barbiers shares Zwart's theoretical perspective, the Minimalist Program (MP, Chomsky 1995), from which he examines the hypothesis that minor syntactic differences can cause substantial syntactic variation in his contribution, "Locus and limits of syntactic microvariation". He evaluates this hypothesis based on four case studies, involving complementizer drop, ONE-insertion, strong reflexives and doubling in Wh-chains, using material from the *Syntactic Atlas of the Netherlandic Dialects* (SAND), a database comprised of information on 100 syntactic variables at 267 sites, also used by Spruit et al. (this volume). As a theoretical syntactician, Barbiers is intrigued by dialect variation for two reasons: first, because it is controlled variation, where many syntactic variables remain constant, and second, because it exposes syntactic theory to challenges it does not face when the primary data is the intuitive judgments of the investigator. With respect to the latter, Barbiers notes that a number of the criticisms that have been levelled at methodology in generative syntax are successfully met using dialectological methodology.

The sorts of analyses the author develops are like those developed by other MP syntacticians, and in that respect the paper does not present a radical break with earlier work. But the paper underscores the value of the large and dense data collection with hundreds of rather comparable varieties of Dutch, even for very abstract theoretical work.

Barbiers works within the MP postulate that grammars (rule systems) are (largely) universal so that differences in syntax distributions reflect underlying lexical differences and superficial phonological differences. There are varieties of Dutch in which three elements can appear at the beginnings of embedded questions, an interrogative pronoun, followed by an interrogative complementizer, and finally followed by a general complementizer:

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Vertel | niet | **wie** | **of** | **dat** | ze | geroepen | hebben. |
| tell | not | who | if | that | they | called | have |

'Don't tell me who they have called.'

Barbiers sees this as reason to postulate that the general grammar underlying all the SAND varieties allows three positions at the beginning of the subordinate clause, and then asks: what must the lexical properties of these words be in order to explain the range of data in the atlas. Focusing on the question of why one or both complementizers may be left unpronounced, Barbier appeals to recoverability: a lexical element can be silent if its morphosyntactic features are a subset of the features of a locally available lexical element. As Barbiers notes, without the large syntactic database, the empirical basis of this work would be thin, and the conclusions tenuous.

The examination of ONE-insertion, or common noun phrase anaphora (the analogous English construction is seen in *a blue one*), likewise leads to an analysis where crucial lexical items must vary in order to explain the syntactic variation, as does an analysis of questions in which elements of embedded clauses are questioned, sometimes leading to multiple interrogative pronouns in single sentences. The case of the reflexives is likewise amenable to a lexical analysis, but requires that crucial lexical properties themselves conform to syntactic principles. In this case as well, Barbiers illustrates how insightful it is to compare minimally different ranges of syntactic behaviour, i.e. the dense concentration of variability present in the dialect syntax atlas.

Barbiers closes with a speculation that the detailed work with the varieties is likely to turn up cases in which some of the possibilities licensed even by the best analyses are not instantiated, and indicates that he would be willing in such a case to draw a distinction between ungrammatical and unrealized options. He suggests that this is analogous to noting that some phonotactically well-formed sequences may be uninstantiated.

**2.3 Dialectology**

"Associations among linguistic levels" by Marco Spruit, Wilbert Heeringa and John Nerbonne examines the geographic structure in the Syntactic Atlas of the Netherlandic Dialects (SAND), a database comprised of information on 100 syntactic variables at 267 sites. Spruit (2005) showed that syntactic variation shows roughly the same degree of geographical cohesion that lexical and phonological variation shows, rather to the surprise of some syntactic theorists, who expected to see the structural constraints in syntax to be so dominant as to mask whatever geographic coherence might be present. Heeringa (2004), working with a phonological atlas, measured the degree to which pronunciation differences are associated with geographic distances. Since the same collection included lexical variation, that was estimated as well for the current paper.

But differences in syntax, pronunciation and lexis can each correlate with geography without it being the case that they genuinely correlate with each other. Naturally one expects to see a certain degree of superficial correlation, but once one corrects the superficial correlations for the intervening variable, geography, one sees whether what remains is of interest. It could be the case for example, that lexical variation is dependent on livelihood, agricultural vs. manufacturing. If there are then geographic tendencies with respect to these livelihoods, then lexical traces are likely to be found geographically as well. If pronunciation or syntax follows other lines, perhaps confessional or educational, then we might find that correlations among the linguistic levels largely vanish once one controls for geography.

Spruit et al. note that most linguists would agree that lexical variation is most volatile, since, after all, new words are added to languages daily. They would likewise agree that the lexical level shows the least structural cohesion, while there have been conjectures that syntax and phonology may be structurally interdependent. That conjecture predicts that the phonology-syntax correlation should be higher than either of the correlations involving the lexicon. The authors therefore ask whether syntax and pronunciation more strongly influence one another than either—taken separately—influences or is influenced by lexical distance.

From a second, social point of view we are reminded that when we "explain" linguistic variation on the basis of geography, we are in fact operationalizing a variable which we might better call 'social distance' (or 'inverse social closeness') in a convenient, measurable way. If we add to this the plausible assumption that there are *no* strong structural ties between any two of the linguistic levels, something the authors regard as a reasonable null hypothesis, then we are still interested in the residual correlations between the linguistic levels as indicators of the degree to which simple geographic distance is not doing its job of operationalizing (the inverse) of social contact. Then, if there are significant correlations among different levels, beyond those geography can explain, and especially if these are at roughly compatible levels, we should conclude that extralinguistic, non-geographical influences were at work.

These questions clearly require large collections of syntactic (and other) material. Given the well-known noise in dialectal data, it simply makes little sense to attempt answers on the basis of the behaviour of a small number of variables. Indeed, before the development of large data reserves and good quantitative techniques, we submit that the Spruit et al. paper would have been impossible.

### 2.4 World Englishes

Benedikt Szmrecsanyi and Bernd Kortmann's paper "The morphosyntax of varieties of English worldwide: A quantitative perspective" analyses the spoken English varieties described in the recent *Handbook of Varieties of English* (hence HVE, Kortmann et al. 2004), the first large-scale comparison of the varieties of English morphosyntax. It compares 46 varieties on the basis of 76 non-standard morphosyntactic features. Unlike the SAND database analyzed by Spruit et al. (this volume) the HVE contains information on non-native and pidgin and creole varieties, and each feature was recorded in each variety as frequent, infrequent, or undocumented.

Szmrecsanyi and Kortmann are particularly interested in whether features may be classified as "vernacular universals", which Chambers (2003) hypothesized to be common to native and non-native varieties, and in particular to the English of second-language learners, and to pidgins and creoles. They contrast Chambers's categorization with Mair's (2003) notion of "universals of new Englishes", features common to post-colonial varieties. As a first result, they note that while there are *no* absolutely universal features (underscoring the need for the broad empirical base they proceed from), several features nonetheless occur in over 80% of the database's varieties.

Particularly innovative in Szmrecsanyi and Kortmann's piece is the use of exploratory statistics in investigating relations among features. So they examine all of the nearly 3,000 potential equivalences among features, i.e. pairs of features that are always either both present or both absent in a given variety. For example, *ain't* may be both a negated form in the paradigm of *be* (*he ain't heavy*), but also in the paradigm of *have* (*he ain't got a dime to his name*). They likewise examine all of the nearly 6,000 implicational relations, showing for example that *ain't* as a general negative auxiliary

(*he ain't know why*) only occurs where *ain't* is established as a main verb (in the *be* and *have* paradigms, just adduced).

The authors also examine the database at a high level of aggregation, using multi-dimensional scaling (MDS), cluster analysis, and principal component analysis (PCA) on all the features and all the varieties. This step in their paper is one of the clearest points in the collection where authors move from examining individual features to large sets of them – taking the step back from the trees to make the forest visible. The MDS result is striking in exposing three sorts of varieties, native English (L1) varieties, English spoken as a second language (L2), and pidgins and creoles (P/C). The L1 and P/C varieties are clearly distinct, and the L2 are intermediate, with overlap both with L1 varieties and with P/C's. The suggestion is that distinctions among L1, L2 and P/C varieties cannot be reduced to a simple set of diagnostic features, but rather rely on statistical tendencies among a large set of features. The cluster analysis also isolates a very clear P/C group, but fails to confirm the MDS finding of a split between L1 and L2.

Szmrecsanyi and Kortmann's application of PCA treats morphosyntactic features as instances and varieties as variables in an effort to see what might be common linguistically to the different sorts of varieties. In this case the L1 varieties were clearly distinct, and the authors make a good case that the first principal component distinguishes levels of morphosyntactic complexity and the second the degree to which varieties prefer analytic means of expression. They indicate how the different sorts of varieties are classified with respect to these two dimensions, and they close with remarks on the wisdom of combining the perspective of the high level of aggregation with that of detailed structural analysis and on the potential interest in exploring databases such as Kortmann et al.'s (2004) from an historical perspective.

**2.5 Phylogenetics**

Michael Dunn's paper "Contact and phylogeny in Island Melanesia" begins with an historical puzzle. The languages of Island Melanesia (see the article for a map) fall clearly into two groups, with 90% of the languages belonging to the Oceanic subgroup of the Austronesian languages, which spread to that area about 3,000 years ago, while the remaining 10% are "Papuan" languages. All the analyses of the languages of the area indicate these two groups, but the indications are of two sorts. There is good linguistic evidence that the Austronesian languages share common ancestry, but the Papuan languages are more or less "the rest", with an uncertain genealogy. The standard methodology in historical linguistics is to look for shared cognates to prove relatedness, but there are few, and all are disputed. This of course makes the search for shared phonological innovations impossible. Are the Papuan perhaps distantly related to the Austronesian languages, or are they remnants of pre-Austronesian inhabitants?

Dunn et al. (2005) began therefore to investigate the structural—syntactic and morphological—similarities among the Papuan languages, adducing statistical evidence that the structural correspondences point to an ancient common ancestor. So in this case the syntactic database is used to research an historical question. The 2005 paper applied a "maximum parsimony" technique from biological phylogeny to infer a tree of relatedness with the fewest assumptions of structural innovation, and the

paper provided evidence of distant relatedness for the Papuan languages—something that would not be possible without the syntactic database. But the maximum parsimony technique has been criticized for not proceeding from a model of evolution, and for not being statistically based, and therefore not able to assign confidence scores to hypotheses.

The present paper applies Bayesian phylogenetic inference, an alternative technical approach that has a rudimentary evolutionary model and also assigns confidence values to hypotheses. It is "Bayesian" in incorporating so-called "prior probabilities" to its estimates in this case some of the experimenter's hypotheses about the correct phylogeny. Bayesian techniques are very popular in many areas of quantitative linguistics (Nerbonne, 2007), and one service of Dunn's contribution is to lay out the fairly complicated Bayesian phylogenetic inference in understandable fashion. Given the rising popularity of phylogenetic inference in historical linguistics, many practitioners will wish understand the techniques in more detail, and this paper provides an excellent high-level sketch.

The competitor to phylogenetic explanation of shared traits is areal explanation, i.e., sharing through language contact. Like Spruit et al. (this volume) Dunn therefore examines his data from an areal perspective as well, showing that there is indeed a correlation between geographic distance and "structural distance", measured in the percentage of shared structural traits. Could it be that the Papuan languages borrowed so extensively from one another and from the Austronesian languages that we are effectively seeing an ancient *Sprachbund*? Of course it is not uncommon that historically related languages are geographically close, as well (Campbell, 1995) points out, so the mere existence of a significant correlation between geographic and structural distance is inconclusive. To choose between the two possible explanations, Dunn compares the structural-geographical correlations among the Papuan languages and among the Austronesian languages on the one hand and those between the "mixed pairs" of Austronesian-Papuan on the other, showing that the correlation among the latter is quite small. This suggests that the structural similarities among the Papuan languages are not the effect of contact in Island Melanesia. Dunn concludes that the structural similarities are most likely the result of common ancestry, while not ruling out that intensive ancient contact may also be the cause.

### 2.6 The Historical Signal

Giuseppe Longobardi and Cristina Guardiano's "Evidence for syntax as a signal of historical relatedness", like Dunn's contribution, inspects evidence from a large syntactic database from the perspective of historical linguistics. As the authors note, scholars have viewed structural coincidence between languages as evidence of typological affinity rather than of historical relatedness. The authors deliberately examine syntactically more abstract levels than Dunn, taking inspiration from the study of evolution in biology. That study, too, began by using superficial traits as indicators, but progressed enormously once DNA was recognized as the basis from which to calculate differences and innovations. Longobardi and Guardiano look to the linguistic characterizations of Chomsky's "Principles and Parameters" program (P&P, Chomsky, 1981) as the most promising source of features with which to contrast languages. They note that the range of parameter values which P&P aims to characterize is intended to cover the entire range of possible human languages. Since

parameters are intended to characterize the range of languages parsimoniously, they must also be independent of one another in principle, making them good candidates as evidence of relatedness (we are avoiding the discussion of statistical independence at this point).

Since the authors assume the correctness of genealogies inferred from lexical data, including pronunciation, they frame their research as asking the degree to which lexical and syntactic properties indicate the same historical developments. Newmeyer (2005) has explicitly denied that the evidence of lexical cognates and the evidence of structural coincidence point in the same direction.

While the P&P program has not developed to the point where the entire range of values is known for all parameters, the authors can still settle on a range intended to characterize the considerable variation in the syntax of determiner phrases (DPs, such as *the five green cars in the street*). They identify 51 binary parameters whose values they determine for a selection of 26 ancient and modern languages. The range of parameters includes agreement features such as number, gender, etc.; the status of demonstratives and determiners; adjectival and relative modification; the grammar of genitives (and of possessives); and the position of the head noun. The language sample chosen is largely known to be related (Indo-European), but it likewise includes languages known to be unrelated, in order to test the hypothesis that syntactic evidence will concur with lexical evidence of relatedness. The degree to which languages differ is estimated using the proportion of instantiated properties for which they differ.

Longobardi and Guardiano test their results by examining the measure of relatedness calculated by their procedure and comparing this to the linguistic relatedness which has been established by scholarship, noting a reasonable fit. As a second empirical test, they submit their data to analysis using a distance-based clustering technique from Felsenstein's PHYLIP package (Felsenstein, 2004). The fit is nearly perfect and the deviations are examined closely. As a final test they compare the syntax-based phylogeny to a lexically based one produced by McMahon and McMahon (2005), deriving a scatterplot showing a linear relationship between the lexical and syntactic measures for related languages (unrelated languages tend to obtain ceiling scores for both measures).

## 3. Looking Forward

We take inspiration from other areas of inquiry in which the analysis of aggregate properties is recognized as contributing scientifically. Macro-economics deals with aggregates such as markets, supplies and prices, and infers from these the properties of individual workers, consumers and goods. Thermodynamics deals with aggregate properties of materials and predicts (mean) properties of molecules based on temperature. Cosmologists study the distribution of lighter isotopes as an indication of the more detailed workings of the big bang, and epidemiologists infer social behaviour in part based on the rate and extent of the spread of disease.

If we are to continue in this direction in linguistic research, we will need to acquire the appropriate techniques, especially quantitative analysis, and to adapt them for language analysis; note that the field is moving in this direction for other reasons

anyway. More crucially, though, we will also wish to steward our data more professionally. There is already a shared methodology and an infrastructure for sharing corpus technology (see for example the Linguistic Data Consortium, http://www.ldc.upenn.edu/), but less so for structured linguistic databases. The advantage of relying on a structured database as opposed to drawing on raw corpus material (or, for that matter, on a loose collection of fieldworkers' notebooks) is, of course, that the former essentially organizes preliminarily analyzed linguistic data and thus can – by incorporating the results of time-consuming qualitative data analysis and retrieval – greatly facilitate our formulation and verification of higher-order generalizations. In short we need protocols and platforms for sharing structural databases, and we also need techniques for linking these to corpora.

The present volume shows where scientific advances are being made due to the availability of large syntactic data collections, but they also suggest many more avenues that might be explored. They suggest that we should turn to an aggregate perspective wherever genuine tendencies are not hard and fast rules – in these cases we need the aggregate perspective in order to obtain the statistics we need to confirm the tendency. This likely to be the case in studies of language contact, of language pathology, language acquisition, and perhaps wherever there is substantial variation in the sort of language studied.

**References**


Bouma, Gerlof 2008. Starting a Sentence in Dutch: A corpus study of subject- and object-fronting. PhD thesis, University of Groningen.

Bouma, Gosse, and Kloosterman, G. 2007. Mining Syntactically Annotated Corpora using XQuery. Proceedings of the Linguistic Annotation Workshop (ACL 07), 17-24.

Bresnan, J. 2007 Is syntactic knowledge probabilistic? Experiments with the English dative alternation In: Featherston, S. and Sternefeld, W. (Eds.) Roots. Linguistics in Search of its Evidential Base. Berlin: Mouton De Gruyter, 75-96.

Campbell, L. 1995 The Quechumaran hypothesis and lessons for distant genetic comparison. Diachronica XII(2), 157-200.

Chambers, J. K., 2003. Sociolinguistic theory: Linguistic variation and its social implications. Blackwell, Oxford.

Chomsky, N. 1981. Lectures on government and binding. Foris, Dordrecht.

Chomsky, N. 1995. The minimalist program. MIT Press, Cambridge, MA.

Comrie, B. 1989. Language universals and linguistic typology: Syntax and morphology. University of Chicago Press, Chicago. (1st ed., 1981)

Comrie, B. and Smith, N. 1977. Lingua Descriptive Studies: Questionnaire. Lingua, 42:1-72.

Dryer, M., Haspelmath, M., Gil, D, and Comrie, B. 2003. The world atlas of language structures. Oxford University Press, Oxford.

Dunn, M., Terrill, A., Reesink, G., Foley, R., Levinson, S., 2005. Structural phylogenetics and the reconstruction of ancient language history. Science 309, 2072–2075.

Felsenstein, J. 2004. Inferring phylogenies. Sinauer, Sunderland, MA.

Heeringa, W., 2004. Measuring Dialect Pronunciation Differences using Levenshtein Distance, PhD thesis, University of Groningen.

Kortmann, B., Schneider, E., Burridge, K., Mesthrie, R., Upton, C. (Eds.), 2004. A handbook of varieties of English. *2 vols.* Mouton de Gruyter, Berlin/New York.

Mair, C., 2003. Kreolismen und verbales Identitätsmanagement im geschriebenen jamaikanischen Englisch. In: Vogel, E., Napp, A., Lutterer, W. (Eds.), Zwischen Ausgrenzung und Hybridisierung. Ergon, Würzburg, pp. 79-96.

McMahon, A., McMahon, R. 2005. Language classification by numbers. Oxford University Press, Oxford.

Nerbonne, J. 2007. The exact analysis of text. Foreword in: Mosteller, F., Wallace, D. Inference and disputed authorship: The Federalist. CSLI, Stanford. xi-xx.

Newmeyer, F. J. 2005. Possible and probable languages. A generative perspective on linguistic typology. Oxford University Press, Oxford.

Schütze C. 1996 The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology. University of Chicago Press: Chicago.

Spruit, M., 2006. Measuring syntactic variation in Dutch dialects. In: Nerbonne, J., Kretzschmar, W. Jr. (Eds.), Literary and Linguistic Computing 21(4), special issue, Progress in Dialectometry: Toward Explanation, pp. 493-506.