Available online at www.sciencedirect.com
ScienceDirect
Lingua 119 (2009) 1679–1706
Lingua
www.elsevier.com/locate/lingua

Evidence for syntax as a signal of historical relatedness

Giuseppe Longobardi [a,*], Cristina Guardiano [b]

[a] *Laboratorio di Linguistica e antropologia cognitiva, DSA, Università di Trieste, Italy*
[b] *Dipartimento di Scienze del Linguaggio e della Cultura, Università di Modena e Reggio Emilia, Italy*

Received 15 January 2007; received in revised form 9 September 2008; accepted 9 September 2008
Available online 7 January 2009

Abstract

In addition to its theoretical impact, the development of molecular biology has brought about the possibility of extraordinary historical progress in the study of phylogenetic classification of different species and human populations (especially cf. Cavalli Sforza et al., 1994, among others). We argue that parametric analyses of grammatical diversity in theoretical linguistics, stemming from Chomsky (1981), can prompt analogous progress in the historical classification of language families, by showing that **abstract syntactic properties are reliable indicators of phylogenetic relations**. The pursuit of this approach radically questions the traditional belief in the orthogonality of grammatical typology and language genealogy, broadly supporting Nichols' (1992) program, and ultimately contributes to establishing **formal grammar as a population science** and historical linguistics as an important part of cognitive inquiry.

Keywords: Comparative methods; Parametric syntax; Determiner phrases; Language taxonomy; Phylogenetic algorithms; Cognitive anthropology

1. Introduction

The contribution of formal syntactic theories to modern linguistics is still widely regarded as focused on synchronic generalizations rather than on classical evolutionary problems. This article sums up the results of an ongoing research project, which suggest that theoretical syntax may provide unexpected evidence for phylogenetic issues typical of the historical-comparative paradigm. The level of analysis tentatively used in the research is not that of surface patterns, but that of the more abstract grammatical parameters investigated since Chomsky (1981). On these grounds, we will contend that formal grammar, along the model of molecular genetics, can be construed as a potential contribution to the study of the human past; and that, in turn, the reality of parameter systems, as cognitive theories of grammatical variation and its implicational structure, receives radically new support precisely from their success with historical issues.

* Corresponding author.
E-mail addresses: longbard@units.it (G. Longobardi), cristina.guardiano@unimore.it (C. Guardiano).

0024-3841/$ – see front matter
doi:10.1016/j.lingua.2008.09.012

2. Classification and historical explanation in biology and linguistics

2.1. Biology and linguistics

Since the 19th century, evolutionary biology and historical linguistics have followed parallel paths, attempting to classify human populations and languages, respectively, into genealogically significant families, so explaining the distribution of their resemblances and reconstructing the origins of their diversity.

Such shared interest in historically connected lineages has long suggested the opportunity of also sharing some procedures of comparison and reconstruction.[1] In particular, both disciplines have been confronted with the problem of identifying and adequately evaluating relevant patterns of similarities/differences.

2.2. Biological classifications

The most traditional classifications of species, or populations within a species, are based on externally accessible evidence, the so-called morphological characters (e.g. anthropometric traits in the case of human populations, such as shape and size of body and skull, color of skin, hair, eyes, etc.). Such features are not completely adequate taxonomic characters, because they are often highly unstable through time, as subject to strong evolutionary selection on the part of the environment.

Phylogenetic, hence typically *historical*, investigation underwent a revolution, over the past few decades, on the grounds of purely *theoretical* progress in biology, namely the rise of molecular genetics. The newly available molecular evidence has one great advantage: it is less subject to change driven by natural selection and, therefore, is more likely to retain genealogical information.[2]

Furthermore, genetic polymorphisms, i.e. the *comparanda* of molecular classifications, exhibit a very useful formal property: they are drawn from a *finite* and *universal* list of *discrete* biological options. The practical benefit of this property for taxonomic purposes will become apparent when we discuss analogous aspects of linguistic evidence.

2.3. Linguistic classifications

As in biological classifications, phylogenetic relatedness among languages has also been traditionally investigated on the most externally accessible elements, which are, in this case, sets of words and morphemes (whether roots, affixes or inflections); we will term such entities *lexical* in a broad sense, as they are saliently characterized by Saussurean arbitrariness (nearly infinite possibilities of pairing sound and meaning for each language). Precisely for this reason, lexical items, when resembling each other in form and meaning, seem able to provide the best probative evidence for relatedness. Linguistic classification was only rarely supported through the comparison of entities less accessible to superficial observation and apparently less arbitrarily variable across languages, such as *grammatical* principles, in particular *syntactic* rules.[3]

Basically, two methods of identifying genealogical relatedness have been proposed in linguistics, both based on lexical comparison in the sense defined above: the classical comparative method and Greenberg's mass or multilateral comparison. Their respective advantages and drawbacks are discussed directly.

2.3.1. The classical comparative method

Phylogenetic linguistics has two basic goals: establishing a *relative* taxonomy among three or more languages and establishing *absolute* historical relatedness between two (or more) languages. Therefore, the main problems for comparative methods in linguistics are that of measuring language distance/similarity and that of identifying a sufficient number of similarities so improbable as to go beyond chance and call for historical explanation (prove some form of relatedness).

[1] Even abstracting away from the more complex question of possibly matching results: cf. Cavalli Sforza et al. (1988), Barbujani and Sokal (1990).

[2] "...the major breakthrough in the study of human variation has been the introduction of genetic markers, which are strictly inherited and basically immune to the problem of rapid changes induced by the environment." (Cavalli Sforza et al., 1994:18).

[3] Syntax will be understood here, again in a broad sense, as a set of generalizations combining words and their meanings into well-formed and interpretable sentences.

The classical comparative method can yield neat conclusions on language relatedness, which are immune from the need of serious mathematical evaluation, as they are based on few highly improbable phenomena, like agreements in irregular morphology and, especially, recurrent (optimally 'regular') sound correspondences. Such phenomena patently provide what Nichols (1996:48) terms *individual-identifying* evidence, i.e. that whose ''probability of multiple independent occurrence among the world's languages is so low that for practical purposes it can be regarded as unique and individual'', essentially the equivalent of a haplotype in evolutionary genetics.[4] In principle, even a single well-attested regular sound correspondence could provide such evidence, defining a language family (a haplogroup) with certainty.

This way, the classical comparative method has largely solved the problem of identifying *comparanda* safe from chance similarity (i.e. has provided reliable cognacy judgments), without having to resort to especially sophisticated measurements. Therefore, the method, undoubtedly one of the greatest achievements of the human sciences, has the major epistemological advantage of providing a sharp and much-needed *demarcation* criterion between science and pseudo-science in etymology and historical linguistics.

The other side of the coin is that it is limited by the very conditions warranting its success as a demarcation criterion: it necessarily narrows the scope of inquiry to sets of languages and chronological spans in which such improbable (hence, necessarily rare) phenomena as recurrent correspondences ('sound laws') are recognizable. It has offered spectacular *a posteriori* proofs of the relatedness of language families whose state of cognation was relatively easy to suspect already before the systematic application of the method itself. On the contrary, it has not been equally useful as a heuristic, nor as a proof, for long-distance grouping of such families into deeper stocks, nor (perhaps precisely because it does not need to care for sophisticated measurements to prove relatedness) has always been effective in identifying lower *taxa*, that is, family-internal articulation.

2.3.2. *Mass comparison*

The most notable attempt to overcome this practical limit, aiming at more far-reaching, long-range taxonomic conclusions, is Joseph Greenberg's (1987, 2000, among other works) highly controversial *multilateral* or *mass* comparison.

The method is still based on lexical data, but does not rely on the criterion of exact sound correspondences to identify cognate sets: Greenberg notices that exceptionless sound laws and rigorous reconstruction procedures were discovered only *after* the best-assessed language families, such as Uralic or Indo-European, had been identified. Greenberg's proposal is that the lack of exact sound correspondences, i.e. the reliance on mere phonetic and semantic resemblance, can be compensated for by comparing lists of words not just in pairs of languages, but across larger sets at the same time. This should reduce the risk of mistaking accidental similarity for etymological cognacy: for, if the probability of chance agreement between two languages on a certain item is $1/n$, the probability of the same agreement in three languages is $(1/n)^2$, in four it is $(1/n)^3$, etc.

Similarly, Greenberg has claimed that reconstruction of protolanguages and of precise diachronic steps is not a necessary prerequisite to hypothesize phylogenetic taxonomies, and that consideration of synchronic similarities/differences may be sufficient.

The method has the advantage that, given a sufficiently universal list of meanings, it can in principle be applied to any set of languages, no matter how remote, and not just to those which exhibit recognizable sound correspondences.

The critical disadvantage of Greenberg's method is that it fails to provide, let alone justify, any precise measure of similarity in sound and meaning. This has two serious consequences. First, it is hard to establish mathematically accurate relative taxonomies, no less than with the classical comparative method. Second, although Greenberg's method, unlike the classical one, should crucially be warranted by explicit non-trivial probabilistic calculations, in fact

[4] Such evidence should then characterize a unique individual protolanguage, rather than a set of languages or a language type. The statistical threshold for individual-identifying was obtained by Nichols (1996) multiplying the probability of occurrence of individual languages, calculated in the order of 0.001, with a conventional level of statistical significance, considered to be 0.05 or 0.01, reaching a probability between one in twenty thousand and one in a hundred thousand. She assumes then that ''a probability of occurrence of one in a hundred thousand or less is individual-identifying at a statistically significant level, and a probability of one in ten thousand is at least interesting and borderline useful'' (Nichols, 1996:49).

it is unable to specify the amount and degree of similarities beyond which resemblance becomes non-accidental (individual-identifying in Nichols' terms) and proves absolute relatedness. The relevant probabilistic questions, often of hardly manageable complexity, have been mostly raised by other scholars and in general have received answers which do not support Greenberg's position.[5]

Thus, epistemologically, mass comparison does not yield so sharp a demarcation criterion as one founded on recurrent sound laws: it remains unclear how it may be safe from chance similarity.

To conclude:

(1) a. the classical comparative method has severe restrictions of applicability
b. mass comparison has been unable to yield satisfactory proof of absolute historical relatedness.
c. neither method has proved particularly apt to provide exact measuring of taxonomic distances (also cf. sect. 4.3.2 below)

2.4. Phylogenetic issues and theoretical linguistics

In view of the skeptical and sometimes harshly critical replies to Greenberg's proposals, it is not hazardous to conclude that the 20th century has hardly seen a major widely accepted progress in comparative methods based on the lexicon, in particular as heuristics for novel genealogical conclusions.[6] It is natural, then, to begin to look at linguistic domains beyond the lexicon, as especially suggested by Nichols (1992), Heggarty (2000), and McMahon and McMahon (2005).

As noted, in biology, the impasse resulting from the limits of morphological traits as taxonomic characters was overcome by accessing more sophisticated evidence, provided by independent theoretical developments of the discipline. *Mutatis mutandis*, analogous theoretical progress has been made in linguistics since the 1950s with the rise of typological and formal approaches to syntax (Chomsky, 1955, 1957; Greenberg, 1963). In particular, the theory of Principles&Parameters, developed since Chomsky (1981) within the general framework of cognitive science, has tried to tie together insights from both approaches about grammatical universals and variation. As a result, theoretical syntax, studying the mind as a system of computation of abstract symbolic entities, has not only made available a new level of evidence, but also one including culturally variable data, most suitable for comparison and classification.[7]

On the analogy of the theoretically induced progress in biological classifications, we think it natural in linguistics to ask if syntax can now serve genealogical purposes better than lexical methods.

3. Lexical comparison and grammatical typology

3.1. Humboldt's problem

Asking this question means challenging a tacit assumption in much linguistic practice, namely that the classification of languages based on lexical arbitrariness (which is assumed to be genealogically relevant) and that based on syntactic properties (alleged to be only typologically significant[8]) are essentially orthogonal (cf. section 4.1 below).[9] Thus, we will state our guiding issue as follows:

[5] Cf. Ringe (1992, 1996); also cf. discussions in Joseph and Salmons (1998), among others.

[6] Many linguists concerned with phylogenetic issues would subscribe to Nichols' (1996:65) claim: ''What linguistics needs now are heuristic measures that will be valid in situations where comparativists cannot expect to have reliable intuitions, measures that will detect relatedness at time depths at which face-value individual-identifying evidence has disappeared and the standard comparative method cannot apply''.

[7] See the remarks on this point in Roberts (2007).

[8] In fact, with the inspiring exceptions of the mentioned Nichols (1992) and Dunn et al. (2005). Cf. the remarks in sections 3.2 and 4.4 for the difference in level of abstraction between the evidence used in these works and in the present research.

[9] For a recent statement cf. Newmeyer (2005:102): ''. . .parameter settings do not 'correspond' in the way that cognates do. We can reconstruct *pater* (*patrem*? L&G) as the ancestor of *père* and *padre* because the similarities of form and meaning that have been passed down for 2000 years allow us to conclude that in some relevant sense they are all the 'same' form. Nothing like that can be concluded from the fact that two languages, related or not, share the same value for, say the Ergativity Parameter''.

(2) Are syntactic and lexical classifications of languages significantly isomorphic?

Since the very inspiration for raising this question is rooted in Wilhelm von Humboldt's original distinction between several possible levels of language classification, we will conventionally, and rather anachronistically, refer to (2) as *Humboldt's problem*.[10] Humboldt's problem admits in principle of two basic answers:

(3) a. Syntax and lexicon provide analogous taxonomic results
b. Syntax provides taxonomic results radically at odds with those of the lexicon

For concreteness, we take (3)a to mean that, given two genealogical trees for the same set of languages, built on lexical and syntactic evidence respectively, at least a clear majority of subportions of the two trees will overlap. Now, if this is the case, three logical possibilities arise:

(4) a. Syntax provides weaker insights, i.e. the same taxonomic results, but only up to a shallower level of chronological depth (climbing back the past, the threshold of uncertainty is reached 'more quickly')[11]
b. Syntax and lexicon provide the same tree
c. Syntax provides stronger insights, i.e. more ancient *taxa* can be safely identified (the threshold of uncertainty is reached 'later', i.e. further back in the past)

We will address Humboldt's problem by devising and empirically evaluating a comparison method based on the renovated insights and data provided by syntactic theory over the past 20 years.

3.2. On the notion of grammar

A clarification is in order, at this point, with respect to the notions of *grammar* and *syntax* adopted here and the various concepts of *grammatical* evidence used in other phylogenetic studies. In Greenberg (1987) grammatical elements taken into consideration for taxonomic purposes are essentially inflectional and derivational morphemes; Greenberg states that 'the separation of lexical evidence and grammatical evidence is of course to some extent arbitrary' (Greenberg, 1987:271): this is obvious since his so-called grammatical elements are, in fact, just closed-class lexical items, whose probative value in comparisons relies on the usual idiosyncratic sound-meaning relationships. Similarly, in Nichols (1996) 'grammatical evidence' is, in fact, morphological material, again characterized by the arbitrary pairing of sound and meaning.

Nichols' (1992) pioneering work and, more recently, Dunn et al. (2005), instead, remarkably apply phylogenetic concerns and methods to 'language structure': for instance, in the latter study, what is considered is 'sound system and grammar' (Dunn et al., 2005:2072), encoded in a data matrix of 125 binary features identified as 'features that would typically be described in a published sketch grammar' (Dunn et al., 2005:2073). The description of linguistic characters in the supporting online material shows that several features are actually generalizations spanning over different lexical items (thus *grammatical* in our sense) and that a subset deals with syntactic properties. Therefore, such results, along with Nichols' (1992) original insights, are an encouraging basis for the enterprise presented in this article.[12] However, our inquiry still differs, in several cases, from theirs as for the exact nature of the evidence involved: we will try to explore the historical significance not of surface generalizations, but of syntactic parameters, which should encode the rich implicational relations supposedly connecting distinct observable phenomena at the level of abstract cognitive structures.[13]

[10] Cf. Humboldt (1827, 1836 *passim*). Also cf. Morpurgo Davies (1996:163 ff.) for discussion.

[11] Even in this case the pursuit of syntactic comparison might be useful for deciding in controversial cases of dialectal classification.

[12] Nichols (1992) is a precursor of this work also for her explicit proposal of regarding structural comparison as a population science. For a first mathematical attempt at founding historical syntax in this sense cf. Niyogi and Berwick (1996).

[13] For a first attempt of comparison between the results provided by lexical, phonetic and syntactic data on the same geolinguistic domain, now also cf. Spruit (2008).

4. Syntax as historical evidence?

4.1. Syntactic comparison

Syntax has indeed never been central to the discovery of genealogical relatedness among languages. For instance, relationships among Indo-European varieties have hardly ever been supported exclusively or prevailingly by syntactic evidence. Apparently, there are two main reasons for this:

(5) a. It is difficult to identify precise syntactic *comparanda*
b. Syntax is not as variable as the lexicon, hence similarities are less probative

Objection (5)b has to do with probabilistic considerations addressed in section 6.2 below. Here we will be concerned with (5)a.

As observed by Watkins (1976), in order to establish syntactic correspondences among languages, one must adopt the same strict requirements which characterize the comparative method in other modules, i.e. one must compare items clearly falling into equivalence classes. Some surface syntactic patterns occasionally provide sufficient correspondences of form and meaning, as suggested by Watkins himself; however, these phenomena are often not so systematic as to allow for measuring distances and for supporting relatedness hypotheses.

In agreement with Roberts (1998), we suggest that Principles&Parameters theory does in principle provide the required systematic *comparanda*, namely parameter values.

The possibility of an efficient lexically blind system of (morpho-)syntactic comparison, precisely the parametric comparison method (henceforth, PCM), was first suggested in Longobardi (2003), Guardiano and Longobardi (2005).

4.2. Some basic properties of parametric data

In Principles&Parameters theory, parameters are conceived as a set of open choices between presumably binary values, predefined by our invariant language faculty, Universal Grammar (UG), and closed by each language learner on the basis of his/her environmental linguistic evidence (*triggers*, in Clark and Roberts' 1993 terms, or *cues* in Lighfoot's 1991 sense). Therefore, setting the value of a parameter is an operation of *selection* rather than *instruction*, in the perspicuous terminology adopted by Piattelli Palmarini (1989) pursuing the biological analogy. Open parameters would define the space of variation of biologically acquirable human grammars, closed parameters specify each of these grammars. Thus, grammar acquisition should reduce, for a substantial part, to parameter setting, and the core grammar of every natural language can in principle be represented by a string of binary symbols (e.g. a succession of 0,1 or +,−; cf. Clark and Roberts, 1993), each coding the value of a parameter of UG.[14] Such strings can easily be collated and used to define exact correspondence sets.

4.3. Potential advantages of PCM

4.3.1. Formal properties

Thus, parameters, like genetic markers of molecular biology, form a *universal* list of *discrete* options. Because of these properties, PCM may share one of the useful formal features of molecular genetic tools and is likely to enjoy some advantages over other linguistic taxonomic methods. In particular, it combines the two strengths of the classical comparative method and of multilateral comparison.

Like the latter and unlike the former, it is in principle applicable to any set of languages, no matter how different: since parameters are drawn from a universal list, PCM only needs to collate their values in the languages under comparison. It does not need to search for highly improbable phenomena (individual-identifying agreements between such languages, like e.g. sound correspondences) and rely on their existence, as a prerequisite to be applied.

[14] In the current minimalist view of language variation, parameters are ideally regarded as properties of functional heads, essentially following Borer (1984). Cf. Newmeyer (2005:53–69), for a survey of the history of the parametric approach since its appearance in the 1980s, and Boeckx and Piattelli Palmarini (2005) for the best epistemological presentation of the roots of the minimalist program.

Like the classical method, though by completely different means, PCM overcomes the intrinsic uncertainty about the appropriate identification of *comparanda* which undermines mass comparison. For, owing again to the universal character of parameters, in PCM there cannot be any doubt about what is to be compared with what. As, in genetics, a blood group, for instance, must be compared to the same blood group in another population, and not to other genetic polymorphisms, so the value of a parameter in a language must and can be collated with the value of exactly the same parameter in other languages. Of course, agreement in one binary parameter proves nothing by itself (as opposed to the probative value of even a single regular sound law of the classical method); but, unlike Greenberg's resemblances, parameters, owing to their discrete nature, lend themselves to precise calculations: the probability of agreements in large numbers of parameters chosen from exactly predefined sets can, in principle, be objectively computed.

4.3.2. Measuring syntax and lexicon

The other key factor of progress in evolutionary biology, beyond a new domain of taxonomic characters, was precisely the introduction of objective mathematical procedures. Their application guarantees the replicability of experiments under controlled variable conditions, so allowing one to test the impact of every minimal variation and progressive enrichment of the input data.[15]

In linguistics, the statistical treatment of data and the use of quantitative methods have successfully been proposed in such fields as sociolinguistics, dialectology, and corpus linguistics. These methods and, more specifically, the computational tools devised in biology to automatically generate genealogical trees begin now to be exploited in historical linguistics.[16]

Most quantitative experiments in phylogenetic linguistics have exclusively or prevailingly focused on lexical databases.[17] Given the poor role of syntax in traditional comparative linguistics, this is understandable. Nonetheless, parametric comparison lends itself to such procedures much better than lexical comparison. For, the fact that the values of a parameter are discrete settings (in principle binary ones, hence equidistant) enables PCM to also overcome the common failure of the two lexical methods: it can provide exact measuring of how close or distant two languages are, allowing for mathematically grounded taxonomies. Parametric comparisons, indeed, yield clear-cut answers: two languages may have either identical or opposite values for a given parameter.[18] The lexicon is different: in a given language any meaning may be arbitrarily coded by a huge and hardly definable set of minimally different phonological expressions. In such a virtual continuum of possibilities it is hard to specify how similar a word must be to the word of another language to begin to count as *relevantly* similar, not to speak of making probabilistic evaluations about the individual-identifying value of whole sets of resemblances.

Further problems with the use of lexical data as a reliable input for quantitative phylogenetic treatments are related to the frequent vagueness of cognacy judgments and to the unstable relation between form and meaning, and are especially acute when the comparanda are not drawn from limited closed-class paradigms (e.g. inflectional categories, cf. Nichols, 1996 for some discussion).[19] Briefly, there are at least four common sources of vagueness (mathematical uncertainty):

(6)
- a. partial correspondence of form (e.g. Italian *prendo*, English *get*)
- b. partial correspondence and non-discreteness in meaning comparisons (e.g. the classical case of German *Hund* vs. English *dog/hound*)
- c. correspondence of form without correspondence of meaning (English *clean*, German *klein*)
- d. similarity of meaning shifts with no correspondence of form (e.g. Italian *fegato*, Greek συκώτι)

Parameters neutralize even such difficulties virtually by definition, since they do not encode a form-meaning relation and, for the reasons discussed above, formal correspondences are exact, in principle, within any language set

[15] Furthermore, as stressed by Koyré (1961), in the history of science, the search for accuracy of measurement and a precise mathematical structure of the data has often been an *a priori* methodological decision, eventually rewarded by its empirical success.

[16] See for instance Lohr (1998), Ringe et al. (2002), McMahon and McMahon (2003, 2005), Nerbonne and Kretzschmar (2003), the papers collected in Clackson et al. (2004) and those in *Transactions of the Philological Society*, 103:2 (2005).

[17] See for instance Embleton (1986), Boyd et al. (1997), Gray and Atkinson (2003), Warnow et al. (2004), McMahon (2005). As remarked, Dunn et al. (2005) and Spruit (2008) use some syntactic data within their structural evidence.

[18] Unless that parameter is made irrelevant for one of the languages by the interaction with independent properties; cf. section 4.4 below.

[19] For a summary of the debate see McMahon and McMahon (2005).

whatsoever, even between pairs of languages so dissimilar (e.g. Wolof and Norwegian) that no serious cognacy judgment can be imagined on their core vocabularies.

4.3.3. Substantive properties

Parameters are promising characters for phylogenetic linguistics in at least two other respects.

Like many genetic polymorphisms, they are virtually immune from natural selection and, in general, from environmental factors: e.g. lexical items can be borrowed along with the borrowing of the object they designate, or adapted in meaning to new material and social environments; nothing of the sort seems to happen with abstract syntactic properties.[20]

Furthermore, parameter values appear to be *unconsciously* and rather *uniformly* set by all speakers of the same community in the course of acquisition, therefore they are largely unaffected by deliberate individual change, which, according to Cavalli Sforza (2000), may influence the history of other culturally transmitted properties[21]; therefore, parameters promise to be better indicators of general historical trends than many cultural features.

4.4. The implicational structure of linguistic diversity

A problem for accurate quantitative treatments might be, however, that observable syntactic properties are often not independent of each other. The difficulty can be controlled for. Two levels of considerations are in order, one related to the classical concept of parameter, since Chomsky's first proposals at the end of the 1970s, the other to more recent empirical and theoretical work.

As for the first point, parametric hypotheses already intrinsically encode a good deal of the implicational structure of language diversity, in the very formulation of many parameters: as a matter of fact, Principles&Parameters theory, inspired by Greenberg's discovery of implicational universals, regards parameters as abstract differences frequently responsible for wider typological clusters of surface co-variation, often through an intricate deductive structure. In this sense, the concept of parametric data is not to be simplistically identified with that of syntactic pattern.

A parameter will be such only if all the grammatical properties supposed to follow from it typologically co-vary; conversely, it will be satisfactorily defined only if no other property significantly co-varies with them. This is a necessary, though not sufficient, condition to ensure that we focus on cognitive structures (i.e. components of I-language, in Chomsky's 1986 terms), not just on generalizations over surface extensions of such structures (parts of E-language). In fact, patterns such as e.g. the traditional N-Gen/Gen-N have already proved at best epiphenomenal at the parametric level: there exist several unrelated types of both constructions and, most importantly, they follow from the combinations of more abstract and independent parameters.[22] Thus, they might even turn out to be misleading, if used to arithmetically assess areal or genetic relatedness, or just typological similarity, although we will not address this issue in the present experimentation.[23]

The second pervasive implicational aspect of parametric systems, potentially challenging independence of characters, has been highlighted in Baker (2001) and by the present research: one particular value of a certain parameter, but not the other, often entails the irrelevance of another parameter. Therefore, the latter parameter will not be set at all and will represent completely implied information in the language, deducible from the setting of the former. These entailments affect both single parameters and entire formal classes of parameters, termed *schemata* in Longobardi (2005). Unsettable parameters in this sense will have to be appropriately disregarded for assessing degrees of relatedness. The symbolism we adopt to represent this aspect of such databases will be presented in section 5.3. In section 6.1, we will discuss how it can be encoded in the measuring of distances in a way to tentatively neutralize its negative effects on successive computations.

---

[20] On syntactic borrowing in general, cf. Thomason and Kaufman (1988).

[21] Cavalli Sforza (2000:176): "There is a fundamental difference between biological and cultural mutation. Cultural 'mutations' may result from random events, and thus be very similar to genetic mutations, but cultural changes are more often intentional or directed toward a very specific goal, while biological mutations are blind to their potential benefit. At the level of mutation, cultural evolution can be directed while genetic change cannot."

[22] Cf. at least Longobardi (2001b), Crisma (in press), on this topic.

[23] A good parametric theory could actually contribute to solving the problem noted in Nerbonne (2007), namely that sometimes it is hard to decide *how many* surface properties typologically cluster together.

5. The empirical domain

5.1. The choice of the parameters

A crucial step in this research line is the choice of the parameters to be compared: a completely random choice or one based on external criteria (for instance, choosing those which have been studied in some recent literature) runs the risk of accidentally producing unbalanced results; two languages might look alike or different precisely on that subset by pure chance, rather like Spanish *mucho*, *día*, *haber* and English *much*, *day*, *have*. In principle, the only way to avoid introducing spurious variables into the experiment would be pursuing exhaustiveness, i.e. considering all parameters. On the other side, at this stage, a practical choice must be made: UG parameters number at least in the hundreds, although we are too far from being able to make precise estimates.

A viable approach, in our view, is trying to be exhaustive relatively to a limited subdomain, possibly intrinsically well defined within syntactic theory itself (and sufficiently vast to hopefully be representative). Here, the only fully arbitrary choice is selecting the module. This should also help us to better avoid a risk pointed out by Nichols for lexical mass comparison: randomly choosing *comparanda* from a larger sample poses serious probabilistic problems, as ''a set of elements has much greater individual-identifying value when taken as a closed set rather than when taken as a group of trials in a larger sample'' (Nichols, 1996:62).

The suggested approach is actually the historical application of the general strategy proposed in Longobardi (2003) under the label 'Modularized global parameterization' (MGP).

5.2. Modularized global parameterization

In order to realistically investigate the properties of parameters, either theoretically or historically, it is important to study a sufficiently broad set of them in a congruous number of languages. Most crosslinguistic analyses within the generative framework have instead been devoted to single or few parameters investigated within a narrow number of languages at a time.[24] MGP has been proposed precisely with the goal of overcoming the drawbacks of this situation. Such a strategy aims to attain at the same time the depth of analysis required by parametric hypotheses and sufficient crosslinguistic coverage. In particular, a sound parametric testing ground should involve:

(7) a. a sufficient number of parameters, possibly subject to reciprocal interactions, but relatively isolated from interaction with parameters external to the set
b. a sufficient number of languages
c. a sufficiently fine-grained analysis of the data[25]

In MGP these goals can be neared at the acceptable cost of narrowing the study to a single syntactic module and trying to be as exhaustive in that module as allowed by our best current linguistic understanding.

The MGP strategy in principle requires the elaboration of a complex tool consisting of:

(8) a. a set of parameters, as exhaustive as possible, for the module chosen
b. a set of UG principles defining the scope and interactions of such parameters
c. a set of triggers for parameter values
d. an algorithm for parameter setting

In the execution of MGP used for the present genealogical experiment, the module chosen is the nominal domain, more technically the internal syntax of the Determiner Phrase (DP). The DP, besides presumably meeting condition (7)a, has a further advantage for historical purposes: it is less rich than the clausal domain in informational structure, an area often regarded as a typical source of diachronic reanalysis. Let us consider to what extent the database collected conforms to MGP requirements.

[24] Cf. Newmeyer (2005:50 ff.): ''researchers have not attempted a comprehensive treatment of parameters and their settings''. Also cf. Chomsky (1995:7): ''the P&P model is in part a bold speculation rather than a specific hypothesis''. Notice, however, the exception of Baker (2001), who discusses a system of hierarchical relations among parameters connected to polysynthesis.

[25] Descriptions in terms of, say, Dixon's (1998) Basic Linguistic Theory are normally insufficient to determine the setting of parametric values.

5.3. Database

5.3.1. Parameters

As for (8)a, 63 binary parameters have been identified within the DP domain, listed in the first column of Table A (Fig. 1 in the Appendix). Parameters have been tentatively selected on the basis both of existing proposals and of novel empirical investigation over the collected database. As a matter of fact, many parameters of Table A represent current assumptions within generative or typological literature, sometimes with slight – and mostly irrelevant for our purposes – variants in their formulation.[26]

Values for the 63 parameters have been hypothesized in 23 contemporary and 5 ancient languages, each represented as a vertical string of values in a column of Table A. The 28 languages were chosen from the Indo-European ones with six exceptions. They are the following: Italian (It), Salentino[27] (Sal), Spanish (Sp), French (Fr), Portuguese (Ptg), Rumanian (Rum), Latin (Lat), Classical Greek (ClG), New Testament Greek (NTG), Grico[28] (Gri), Modern Greek (Grk), Gothic (Got), Old English (OE), Modern English (E), German (D), Norwegian (Nor), Bulgarian (Blg), Serbo-Croatian (SC), Russian (Rus), Irish (Ir), Welsh (Wel), Hebrew (Heb), Arabic (Ar), Wolof (Wo), Hungarian (Hu), Finnish (Fin), Hindi (Hi), and Basque (Bas).

The basic alternative states of each parameter are encoded as '+' and '−' in Table A. It is important to bear in mind that such symbols have no ontological, but just oppositional value. All parameters in Table A exhibit at least one contrast in value over our language sample (see the rows corresponding to each parameter), with the exception of parameters 1, 2, and 24, which however define characteristics known to clearly distinguish other languages presently under investigation, but not yet comprised in the sample.

As a general guiding criterion, we decided to build a crosslinguistic morpho-syntactic difference into Table A as a parameter if and only if it appeared to entail any of three types of surface phenomena: the position of a category, the variable form of a category depending on the syntactic context, or the presence of obligatory formal expression for a semantic distinction (i.e. the obligatory valuing of an interpretable feature). Thus, we did not encode as a parameter differences in pure morpho-phonological representation which, as far as we know, do not produce, even indirectly, any of the three manifestations above (e.g. the presence/absence of gender marking on adjectives).

Within the chosen DP module, further subdomains can be distinguished: the status of various features, such as Person, Number, Gender (param. 1–6), Definiteness (roughly 7–16), Countability and related concepts (17–24), and their impact on the syntax/semantic mapping; the grammar of genitive Case (25–31); the properties of adjectival and relative modification (32–41); the position of the head noun with respect to various elements of the DP and the different kinds of movements it undergoes (42–50); the behavior of demonstratives and other determiners, and its consequences (51–55 and, in a sense, 60–63); the syntax of possessive pronouns (56–59).

5.3.2. Principles and implications

With respect to (8)b, what we are able to explicitly provide, within the limits of this work, is a set of theorems, which derive from the relevant UG principles and express *partial implications* among parameter values. It is often the case that setting a parameter A on one value leaves the choice for a parameter B open, but setting A on the other value necessarily determines the value of B as well.

Now, similarities or differences among languages must not be overstated by considering completely redundant information. In order to encode the irrelevance of such settings, i.e. their dependence on the value of another parameter, a '0' has been used in these cases in Table A; the implication giving rise to 0 has been indicated next to the name of the implied parameter in the first column, in the form of (conjunctions or disjunctions of) valued implying parameters, identified by their progressive number. E.g. '+5, −17 or +18' in the label of parameter 19 means that 19 can only be valued when 5 is set on + and either 17 is set on − or 18 is on +; otherwise 19 will receive a 0.

Sometimes, implications are virtually analytical in the formulation of a parameter: for instance, every time a given parameter refers to the behavior of, say, the feature 'definiteness' (e.g. parameters 12, 14, 15...), its setting becomes irrelevant in languages where such a feature is not grammaticalized, that is, if the language does not display a + value for parameter 7.

[26] Cf. for example Bernstein (2001) and Longobardi (2001b), and Plank (2003), as basic overviews of the literature in the two approaches.
[27] The Italo-Romance variety spoken in the provinces of Brindisi and Lecce (represented here by the dialect of Cellino San Marco).
[28] A Greek variety spoken South of Lecce (represented here by the dialect of Calimera).

Some other times, instead, implications are formulated on the basis of empirically observed correlations: for instance, following Greenberg's Universal 36, it is assumed that a language will grammaticalize gender (value + for parameter 3) only if it grammaticalizes number (value + for parameter 2): as a consequence, languages not displaying positive evidence for the feature 'number' will receive a 0 for (the absence of) grammaticalization of gender; the implication will be signalled in the label of parameter 3 as '+2'.

Such implications as carefully made explicit in Table A are, for practical purposes, the most important consequences of UG principles on the variation domain under study. A fuller treatment of the theory and observations presupposed by our parameters could only be contained in a dedicated monograph.

5.3.3. Triggers

As for (8)c, for each parameter a set of potential triggers has been identified and used to build up a questionnaire for data collection, using English as a metalanguage. In defining the notion of trigger, we follow Clark and Roberts (1993:317): "A sentence σ expresses a parameter p_i just in case a grammar must have p_i set to some definite value in order to assign a well-formed representation to σ". Such a sentence (or phrase) σ is thus a *trigger* for parameter p_i. The structural representations of the literal translation(s) of the utterances contained in our questionnaire should be able to set the relevant parameter to a specific value in a given language. Two languages have been set to opposite values of a parameter only if their triggers for that parameter in the questionnaire differ in structural representation in at least one case. The contents of the questionnaire, called Trigger List, will be made available on the project's website, in construction at www.units.it/linglab, along with the answers provided by our informants: for modern languages, each value assigned in Table A and not warranted by reliable literature has been checked with at least one linguistically trained native speaker.[29]

5.3.4. Parameter setting

With respect to (8)d, again, a realistic approach would require the explicit statement of the conditions under which a parameter is set to + or to −, together with an acquisitionally plausible order of setting. Again, it is impossible to pursue such a task in the limits of the present research. The current list of parameters in Table A simply reflects the practical ordering condition that a parameter always follows other parameters on which its setting depends.

6. Elaboration of data

6.1. Coefficients and distances

The information collected in Table A can now be elaborated in numerical terms.

The first operation is to compute the number of identities and differences in the parameter settings of each pair of languages. Such computations are represented in the form of ordered pairs of positive integers (or zero) $<i;d>$, called *coefficients*. Table A contains a robust number of 0s. As noted, 0s cannot be taken into account for measuring relatedness; therefore, even if only one of the languages of a pair has a 0 for a certain parameter, that correspondence is not counted at all for that pair. Also a few empirically uncertain states, indicated in Table A by a '?', are counted like 0s for the purposes of these computations. For these reasons, the sum of *i* and *d* in a coefficient does not necessarily equal 63, and rarely is it the same for different language pairs.

As such, coefficients are not a practical measure to rank the distances instantiated by different pairs of languages and answer questions like: 'Are Italian and French closer than English and German?'. Therefore, coefficients must be reduced to a monadic figure, suitable for a uniform ranking of distances.

The simplest distance between any two strings of binary characters is the *Hamming distance* (Hamming, 1950), which amounts to the number of differences between the two strings. Yet, it is conceptually (and even empirically) wrong to reduce our coefficients to the number of differences (or alternatively to that of identities), because, as they are computed over non-uniform totals, both figures contain relevant information. The computation of *i-d* (call it *algebraic coefficient*) is misleading too, as the result obscures the respective weight of identities and differences.

[29] The parameter values assigned to the ancient languages in our sample rely on specific research done on written *corpora*, such as Crisma (1997, in press) for Old English (XI century prose), Guardiano (2003) for Classical (IV century BC) and New Testament Greek, Gianollo (2005) for Latin (I century BC-I century Christian era), and on further personal work by the authors.

Thus, we adopted another form of reduction of coefficients into a single figure, which will be called a *normalized Hamming distance*. It results from dividing the number of differences by the sum of identities and differences of each pair, thus making the differences proportional to the actually compared parameters (i.e. $d/(i+d)$).[30] Two languages with identical settings for all valued parameters will then have distance 0, two with opposite settings for all valued parameters will have distance 1, all other cases falling in between. Such a distance turns out to be conceptually and empirically more satisfactory, even if not yet completely adequate in at least the theoretical case of pairs of languages with no differences at all, for which the distances so computed uniformly equal 0, irrespectively of the number of identities: the distance between two languages exhibiting, for example, 45 identities and 0 differences and that between two other languages exhibiting 15 identities and 0 differences would be identical.[31] Nonetheless, in practice, already very satisfactory reconstructions are attainable by means of the normalized Hamming distance (cf. section 7).[32] Table B (Fig. 2 in the Appendix) contains both the coefficients and the normalized Hamming distances for the 378 pairs generated by the 28 languages of Table A.

6.2. Probability of chance agreement

The probability of two languages coinciding by chance in the value of a specific binary parameter (ideally assuming both values to be unmarked, i.e. equiprobable[33]) is 1/2, which, as such, has no probative value at all. The probative value becomes more significant as the number of *comparanda* increases: indeed, 63 binary independent parameters generate 2^{63} languages. The probability for two languages to coincide in *all* the values of 63 chosen parameters is $1/2^{63}$, which is highly significant. However, in the real world, we expect most language pairs to agree in but a subset of the parameters compared. Therefore, one must calculate the probabilistic significance of such *partial* agreements. Sticking to the usual assumption that parameters are binary, a rough formula to begin computing the probability of partial agreement is the following: suppose n is the total number of relevant parameters in the sample and h the number of disagreeing (or agreeing, for that matter) parameters; the probability of such an event will be

(9) $$P = \frac{\binom{n}{h}}{2^n} = \frac{\frac{n!}{h!(n-h)!}}{2^n},$$

where $n!$ is the product of the first n integer numbers, and similarly for $h!$ and $(n-h)!$. So, for example, in the case of 44 agreements (i.e. identities), 6 divergences (i.e. differences) and 13 irrelevant comparisons out of 63 independent parameters (or of 6 agreements and 44 divergences, which has the same probabilistic value, but is unattested in our domain for principled reasons, cf. 7.1), the probability of this event will be computed in the following two steps:

(10) a. $$\binom{50}{6} = \frac{50 \times 49 \times 48 \times 47 \times 46 \times 45}{6 \times 5 \times 4 \times 3 \times 2} = \frac{11\ 441\ 304\ 000}{720} = 15\,890\,700$$

b. $$\frac{15\ 890\ 700}{2^{50}} = \frac{15\ 890\ 700}{1\ 125\ 899\ 906\ 842\ 620} = \frac{1}{70\ 852\ 757} = 0.000000014113777$$

7. Evaluation of results

Table D (Fig. 3 in the Appendix) lists the 378 pairs in increasing order of distance; for each pair, the table also reports the probability of chance agreement and indeed the normalized Hamming distance, in parallel columns.

[30] This distance is essentially equivalent to a so-called Jaccard distance (Jaccard, 1901).

[31] Dubious is also the more common case of pairs with a high number of identities and a very limited (though non-null) number of differences: for example, the distance between two languages with 44 identities and 2 differences is twice the distance between two languages with 45 identities and 1 difference, probably running counter the historical linguist's intuition.

[32] Hopefully, the exploration of other distance metrics, such as those elaborated by dialectometry for lexical data since Séguy (1971) and Goebl (1982), might even improve on the present results (cf. also Spruit, 2008). On this topic, also cf. Cavalli Sforza and Wang (1986).

[33] Assuming equiprobability, i.e. neglecting markedness, is an obviously false, but very useful, idealization: since no solid theory of markedness exists to date, it would be risky to assign impressionistic weights to the various parameter values. On the contrary, if parametric comparison turns out to be empirically successful already under the idealization of equiprobability, it will be surprising for an eventually more realistic approach, assigning correct weights to different parameter values, to fail to match, or improve on, this result.

Conceptual and empirical criteria of adequacy have been applied to Table D: the empirical criteria evaluate the correlation of our results with independently known historical variables, such as lexically based taxonomies; the conceptual ones are independent of any specific historical evidence, thus they allow one, in principle, to ascertain the degree of potential failure of the system irrespectively of contextual information.

7.1. Conceptual tests

In order for our results to be probative, they must present both a significant and plausible distribution of differences and similarities between languages.

Hence, a first obvious conceptual test is the following:

(11) The pairs in Table D should be scattered across different degrees of similarity.

For, if all or most pairs were concentrated around the same distance, say 0.5, no significant taxonomic conclusion would be available in syntax.

Since the coefficients range from <40;1> (distance 0.024) to <13;18> (distance 0.64), assuming many intermediate values, we believe that such a criterion is satisfied.

The second conceptual test is related to Nichols' (1996) notion of *individual-identifying* evidence:

(12) The probability of chance resemblance for the most similar languages must attain individual-identifying levels.

The probabilistic threshold in order for a value to begin to be individual-identifying, defined by Nichols (see fn. 4), lies around $1/(2 \times 10^4)$. Among the 378 pairs of Table D, 116 show probabilities ranging from less than $1/(3 \times 10^{12})$ to $1/(2 \times 10^4)$, that is low enough to satisfy Nichols' requirement.

The third conceptual test is based on an assumption formally stated for the first time in Guardiano and Longobardi (2005) as the *Anti-Babelic Principle*:

(13) Anti-Babelic Principle: similarities among languages can be due either to historical causes (common origin or, at least, secondary convergence) or to chance; differences can only be due to chance[34] (no one ever made languages diverge on purpose).

The Anti-Babelic Principle should have long been an obvious founding postulate of historical linguistics. However, it has been formulated just so recently because it may only have measurable effects in domains of discrete variation drawn from a universal list, such as parametric syntax, hardly in an infinitely variable field like the lexicon.

Its main consequence for our model is that negative algebraic coefficients must essentially be due to chance. In a system of binary equiprobable differences, the Anti-Babelic Principle predicts that two completely unrelated languages should exhibit a distance closer to 0.5 than to 1.

For example, such coefficients as <46;5> and <5;46> would have the same (negligible) probability of being due to chance (1/958,596,125), therefore both would call for an explanation. However, historical explanations can be advocated only for the former, not for the latter. Thus, in a realistic system of parametric comparison, negative coefficients like <5;46> or, more generally, pairs where the differences largely exceed the identities should not exist.

According to the Anti-Babelic Principle, we expect pairs of languages known to be related to exhibit coefficients of a clear form $i > d$, while other pairs should tend toward $i = d$, i.e. a distance of 0.5. An extremely encouraging result of our data is that amongst the 378 pairs of languages in Table D only 7 (i.e. 1.85%) display a negative coefficient, with distances between 0.51 and 0.64, and chance probability between 1/7 and 1/16.

[34] This statement needs obvious qualification for the case where a language moved away from the bulk of properties of its family as a consequence of contact with external varieties: see fn. 37 for an example. However, this simplistic formulation is sufficient to make the crucial point of the present argument.

7.2. Empirical tests

7.2.1. Distribution of distances

An obvious procedure to evaluate the reliability of the results produced by new taxonomic methods is comparing them with independently established phylogenetic relations.[35]

Thus, to test the validity of PCM, the relations between pairs in Table D have been compared with those arrived at by traditional taxonomies, usually based on quantitatively unanalyzed, but abundant and sometimes precise, lexical and phonological data. The knowledge of the different degrees of historical relatedness among the 28 languages of Table A gives us the possibility to define three main types of relations:

(14) a. Strong relations: a relation between a pair of languages is *strong* iff one of the languages derives from the other or both derive from a common ancestor within a time span of (presumably) at most ±4000 years.[36]
b. Looser relations: a relation between a pair of languages is *looser* iff it is not strong but both languages derive from a safely established common ancestor (e.g. Proto Indo-European).
c. Weak relations: a relation between a pair of languages is *weak* iff the pair does not instantiate an independently safely assessed genealogical relation.

The pairs falling into each type are signalled by progressive order of dark shading in Table D: those belonging to the first type are not shaded; those belonging to the second type are shaded in pale grey; those of the third type are variously shaded (relations involving Semitic, Uralic, and IE among each other in darker grey, relations involving Wolof and Basque in black). The distribution of shades in Table D is immediately suggestive, as most strong pairs cluster in the topmost part of the table, while weak relations tend to occur from the middle to the bottom. In more detail:

(15) a. Out of the 378 pairs in Table D, 48 instantiate *strong* relations. 39 out of 48 occur among the top 50 of the table (which also include 2 pairs which are independently known to have undergone lexically conspicuous interference: It-Gri, Sal-Gri), and 46 within the first 93 pairs. Furthermore, such 46 strong relations show values of chance probability lower than $1/(2 \times 10^4)$, that is they all satisfy the requirement proposed by Nichols for a value to be considered individual-identifying (see fn. 4).[37]
b. Of the 145 *weak* relations, none occurs in the topmost 112 positions and 90 occur among the last 100 pairs. Only one reaches a (low) individual-identifying level of similarity.

The results obtained through PCM largely overlap with those suggested by traditional lexical comparative practice: this leads us to the preliminary conclusion that a positive answer to Humboldt's problem is likely to be possible.

7.2.2. Phylogenetic algorithms

A second empirical test for PCM consists in the elaboration of phylogenetic hypotheses through computational algorithms. As remarked above, such methods have been devised by computational evolutionary biology and have been increasingly applied to linguistic data over the last 15 years, all relying on lexical and/or phonological evidence, for example the list of words in Dyen et al.'s (1992) Indo-European database.[38] Fig. 4 in the Appendix represents a first attempt to generate a computational phylogeny using purely syntactic data, i.e. our parametric distances of Table D, as an input: to generate the tree we relied on a distance-based program, *Kitsch*, contained in

[35] As suggested, e.g., by McMahon and McMahon (2005).

[36] Among the strong relations, we have included those between Finnish and Hungarian and between Arabic and Hebrew. The latter decision is based on the fact that the variety of Arabic analyzed is the Standard one, still somewhat based on Classical Arabic from the first millennium of the Christian era, while that representing Modern Hebrew draws from the Biblical language (12th–6th cent. BC). The relationship of all the ancient varieties of the sample with one another were considered strong, with the exception of Old English, thus tentatively locating Proto Indo-European within the 4th millennium BC.

[37] The only two strong relations occurring relatively low in Table D (actually below the 145th position) both involve the oldest language of a subfamily and the modern variety of that subfamily known to have undergone the sharpest effects of interference (ClGr-Gri, Got-E).

[38] For an overview of some applications of computational cladistics to linguistic phylogenies cf. Wang (1994) and McMahon and McMahon (2005).

Felsenstein's PHYLIP package (Felsenstein, 2004a, 2004b). The database was subjected to bootstrapping through 1000 re-samplings.[39]

Kitsch – like all similar phylogenetic algorithms – imposes some conditions on the output not necessarily appropriate to our linguistic data: it only produces binary-branching trees and treats all taxonomic units as leaves (i.e. with no acknowledgment of possible mother–daughter relationships among the languages of the sample). Such properties may in principle represent sources of possible grouping problems; yet, the tree generated from our syntactic distances meets most of our expectations, again based on the genealogical hypotheses traditionally established.

Basque, one of the most commonly proposed isolates, is the first outlier; second comes Wolof (indeed, no long-distance hypothesis has ever attempted to closely connect the West Atlantic family to any of the European or Mediterranean languages); both are solidly recognized as external to a node coinciding with a *lato sensu* Nostratic grouping, a split suggested, though hardly proven, by long-distance comparativists.

Then the next outmost bifurcation singles out the (West) Semitic subgroup.[40] The Uralic (Finno-Ugric) family is correctly singled out as well, as opposed to the Indo-European unity. Within the latter cluster, the branching is overwhelmingly the expected one, at least at the taxonomic levels on which there is independent general agreement: the Celtic, Slavic, Germanic, (homeland) Greek families are correctly detected, as well as some relative similarity of Slavic and Indic (the ''*satəm* varieties''). Within the Romance group, a Western Romance unity as opposed to Rumanian is captured, with Latin occurring intermediately, as well as the plausible unity of the two Iberian varieties.

The internal articulation of the Germanic group resulting from the computational elaboration is apparently questionable if compared to the traditionally accepted reconstruction, according to which a West Germanic unity – Old English, English and German – would be expected as opposed to Gothic (East Germanic) and Norwegian (North Germanic). However, two plausible factors may affect the output: the position of English, paired with Norwegian, might correctly reveal actual historical events, like the Scandinavian influence on English and the Norman conquest, whose traces are very obviously manifested also in the vocabulary of modern English, keeping it more removed from Old English and German; then, the two ancient varieties, chronologically closer to the common source, will naturally attract each other, and indirectly affect the position of German.

Such results show anyway that parametric comparison brings to light various types of definitely historical information, though it is hard, at this stage, to single out genetic from areal sources of similarities. Future empirical research in PCM might identify parameters whose values are easy to borrow from genetically stable ones. Some hints are already available: for example, Bulgarian and Rumanian are likely to have come to share as an areal feature their + for parameter 12, which produces the peculiar noun–article constructions; but they continue to be well-behaved Slavic and Romance languages, respectively, with opposite values for parameter 45. It is also possible to argue that this persistence in 45 makes the two languages very different in other subtler surface properties, which go beyond the simplest noun–article phrases; incidentally, such differences would normally escape non-parametric, less formal analyses, somewhat overstating the impression of interference. Other diachronically stable parameters might be similarly identified.

Still, two clear errors are visible in the topology of the tree. The first affects the node merging Russian and Serbo-Croatian together and excluding Bulgarian, thus failing to recognize the plausible South-Slavic unity; it is possible that our parameter sample accidentally underrepresents the difference between Serbo-Croatian and Russian. It remains to be seen if the result would still hold under an extension of parametric comparison to non-nominal domains and/or if this discrepancy can be partly imputed to areal influences, again, affecting Bulgarian as an effect of substrate or membership in the Balkan *Sprachbund*.

Factors of areal influence obviously explain the other mistake, i.e. the appearance of Grico within the Romance family (actually clustering with the Balkan Romance language), rather than in a node with homeland Greek. Needless to say, any pure tree-like representation is insufficient to formalize admixture of this sort.[41]

As noted in fn. 37, at least two of the controversial points (involving the relations between Classical Greek and Grico, and Gothic and English) might be due to the combined effect of time span and admixture with external varieties.

[39] Cf. Rigon (forthcoming) for further experiments.

[40] The remnant cluster happens to coincide with Greenberg's (2000) proposed Eurasiatic family.

[41] This representational problem may perhaps be avoided through the use of network algorithms (as suggested among the others by McMahon and McMahon, 2005).

Furthermore, it must be recalled that a program like Kitsch is devised to best classify coeval entities. Therefore, a further experiment was attempted: another tree was drawn by means of Kitsch, under the same conditions, though leaving out the non-contemporary languages of the sample (see Fig. 5 in the Appendix). Here the results are sharply improved, with Grico recognized as clustering with Greek, and English with German.

Given PCM's sensitivity to some salient contact effects, it is significant that, in spite of the strong similarities in DP syntax pointed out in the literature,[42] no particular support arises for the hypothesis of prehistoric Semitic substrate influence on Insular Celtic, explicitly raised by Vennemann (2002).[43] Even singling out from Table D data on absolute relatedness, the relations between the two Celtic and the two Semitic languages fall far short of the threshold of individual-identifying probability (ranging from Irish-Hebrew 1/865, dist. 0.256, to Welsh-Arabic and Irish-Arabic 1/55, dist. 0.33).

Apart from few problems, the computational elaboration of parametric data with *Kitsch* yields a phylogeny of the whole sample (out of the 53!! possible rooted bifurcating ones[44]) which is largely in agreement with the commonly accepted reconstructions of traditional historical linguistics. This result, notwithstanding the relatively small number of traits compared, supports the effectiveness of PCM and its potential for quantitative historical linguistics. Notice that, perhaps surprisingly, the fact that it has been achieved simply on the basis of distances among attested languages without hypotheses about ancestral states, turns out to support, in a completely different context, one of Greenberg's methodological claims, namely that reasonably successful phylogeny can apparently dispense with preliminary reconstruction of protolanguages and intermediate stages. After all, as pointed out by E. Stabler (personal communication), the situation is not different from what is often found in evolutionary biology.

8. Lexical and syntactic phylogenies

8.1. Distances and trees

In order to compare parametric and lexical phylogenetic results more precisely (of course with the provisos of section 4.3.2 on the accuracy limits of lexical figures), we performed a further experiment. In the lexical database used by McMahon and McMahon (2005), there are 15 languages also represented in our syntactic database (i.e. the modern standard Indo-European varieties); thus, it is possible – using the same taxonomic units (that is, completely overlapping samples of languages) – to minimally compare genealogical trees produced through the same algorithm from two different inputs, the lexical distances and the parametric ones.

In order to bridge the quantitative mismatch between the number of lexical characters (Swadesh's lists with 200 words with cognacy judgments ultimately derived from Dyen et al., 1992) used by McMahon and McMahon and that of our syntactic characters (the 63 parameters of Table A), distances have all been calculated using the formula of the normalized Hamming distance ($d/(i + d)$); the matrices represented in Fig. 6 in the Appendix reveal that, even under such normalization, syntactic distances are considerably smaller than the lexical ones, most of which show clear 'Babelic' values (that is higher than 0.5): this suggests that syntactic differentiation proceeds more slowly than lexical differentiation from the same common source.[45]

Analogous results are produced by Kitsch and shown in Fig. 7 (tree from lexical distances) and Fig. 8 (tree from parametric distances). The topologies of the two trees largely overlap, with just minimal rearrangements basically revolving around the position and structure of the Slavic group. However, it is relevant to notice that in the outputs of Kitsch – as in those of most phylogenetic algorithms – branch length signals the *amount of evolution* occurred between two nodes (cf. Felsenstein, 2004a); thus, in the light of the information provided by the distances, we expect the branches of the lexical tree to be longer than those of the syntactic one. Although the editing of the two tree images has reduced them to the same size for reasons of space, the ruler placed below the two trees signals the actual original proportion of branch length from the root to the leaves. Figs. 7 and 8, in which trees are not bootstrapped for better comparability, confirm our expectation. The distance from the root to the leaves suggested by Kitsch is almost four times longer for the lexical tree.

[42] E.g. cf. Duffield (1996), Rouveret (1994), Roberts (2005:94).
[43] Cf. Roberts (2004) for the suggestive proposal of using PCM to address this question.
[44] Cf. Felsenstein (2004a:19–36), Cavalli Sforza et al. (1994:32). n!! is the semifactorial of n, in this case the product of all odd integers up to 53.
[45] Anyway, no pair of languages shows a syntactic distance bigger than their lexical distance.

Thus, this further quantitative experiment substantiates the hypothesis that syntax is more conservative than the lexicon.[46]

8.2. Syntax and distant relationships

This conclusion is also supported by Fig. 9 in the Appendix, displaying a scatter plot of lexical distances versus syntactic distances for all possible pairs of languages in Fig. 6: each pair is represented by a circle whose co-ordinates are given by the two distances.

The vertical (horizontal) dashed line denotes the mean of lexical (syntactic) distances. Two clouds of points are clearly discernible in the scatter plot: one in the bottom-left part of the graphic, containing pairs with both distances smaller than the respective mean, the other in the right part of the graphic, containing pairs with lexical distance bigger than the mean.

The first group or cloud contains all and only the pairs where both languages belong to the same Indo-European subfamily, i.e. strong relations, non-shaded in Table D, while in the second group only pairs where the two languages do not belong to the same subgroup are found. In the first cloud, the increase of the two distances is essentially proportional. Only one clear outlier is visible, on the rightmost part of the bottom-left area, namely the pair formed by Irish and Welsh, which have a relatively small syntactic distance and a relatively big lexical one.[47]

The second cloud is more compact, that is circles are closer to each other. In particular, lexical distances show little variability, whereas crucially a greater degree of variability continues to be exhibited by syntactic distances.

The visual impression can be confirmed by measuring variability with the aid of standard deviation: 0.035 for lexical distances versus 0.051 for syntactic distances. Alternatively, variability can be measured with respect to the mean distance by means of the coefficient of variation (ratio of the standard deviation to the mean). This gives 0.044 for lexical distances, and 0.22 for syntactic distances, clearly emphasizing that syntactic distances are much more capable of discriminating among languages which are far from each other. On the other hand, the first cloud suggests an inverted configuration; indeed, the standard deviations are 0.12 for lexical distances and 0.041 for syntactic distances. However, the coefficients of variation are here 0.36 for the lexical distances and 0.47 for the syntactic distances, showing that relative variability is still greater with syntactic distances, also for languages which are close to one another.

The moral seems to be, then, again that syntactic divergence from a common ancestor is slower; but also and most noticeably, that syntax continues to remain a potential good indicator of relative taxonomy among sets of distant languages whose vocabularies display too few cognates to make solid clusters identifiable. In principle, long-distance relationships could thus be hinted at, even if not proved by individual-identifying evidence, by syntactic properties better than by lexical ones.

9. Humboldt's problem: answers

The evidence provided by PCM shows that the taxonomies obtained through syntax and vocabulary closely resemble each other: thus, the diachronic persistence of syntactic properties is sufficiently robust to allow for plausible genealogical reconstructions. Then, positive answers to Humboldt's problem appear at hand: testing the phylogenetic hypotheses based on syntactic characters against those known from traditional comparison warrants answer (4)b. Furthermore, in light of the preliminary results described in section 8 above, there seems to be no reason to exclude the possibility that the eventually correct answer to Humboldt's problem will be (4)c.

The impression of historical irrelevance of syntax is likely to be an artifact of the poor number and quality of grammatical variables traditionally considered, and fades away when a sufficiently large set of formally analyzed syntactic polymorphisms is taken into account.

Thus, the belief in the orthogonality of grammatical typology and historical taxonomy is recalled into question by a new level of evidence – parametric syntax – and the search for a systematic and mathematically accurate description of the facts.

The same two factors, as pointed out, have caused a major breakthrough in contemporary natural sciences: the choice of a more sophisticated domain of entities, though more remote from common sense and observation, such as genetic markers, has indeed provided biologists with a better object of evolutionary investigation, as well as with the

[46] This is precisely the expectation of diachronic theories in the spirit of Keenan's notion of Inertia: cf. Keenan (1994, 2000, 2002), Longobardi (2001a).

[47] I. Roberts (personal communication) suggests this may reflect the wealth of Latin loanwords in Welsh.

opportunity of a more articulated mathematical representation of the data. It is suggestive to think of formal syntactic data as potentially playing the same role in historical linguistics.

10. Conclusions: parametric syntax as cognitive anthropology

PCM is a new method to classify languages on the basis of syntactic characters, whose results lend themselves particularly well to mathematical and computational evaluation. It exploits parametric theories to formally measure grammatical diversity and suggests that the taxonomies so derived are likely to have not only typological value, but also some historical (genealogical) significance (also cf. Nerbonne, 2007:4). Such first results of PCM may point to three sorts of consequences: for theoretical linguistics, historical linguistics, and neighboring disciplines, respectively.

First, the historical success of PCM indirectly provides evidence of an unprecedented type for Principles&Parameters models of grammatical variation, on which the method is based.

Second, PCM suggests the possibility of a full historical paradigm in formal syntax, beyond the simple description of scattered cases of diachronic syntactic change. Through parametric linguistics, successfully applied to relatedness issues, historical concerns may be reestablished as central in the study of language and in the wider paradigm of modern cognitive science. A crucial breakthrough in 19th century science was the development of a whole scientific paradigm instantiated by the classical comparative method and by historical linguistics in general. This success was warranted precisely by phylogenetic discoveries: as a matter of fact, hardly any serious theory of etymology and of phonological change would have been conceivable, if the sound shape of the lexicon evolved in so chaotic a way as to display no salient cues on the genealogical relations of languages. Similarly, we think that no serious historical paradigm in syntax could be warranted, if grammatical characters evolved in so chaotic a way as to provide no hints on language relatedness.

Finally, PCM promises to make a new tool for the investigation of our linguistic past, hopefully able to overcome the limits of the classical comparative method and the issues raised by Greenberg's more questionable mass comparison: in this sense, it might eventually join traditional comparative linguistics, archeology, and genetics in the 'New Synthesis' approach to the study of human history and prehistory.[48]

Acknowledgements

Some preliminary results of this research have been presented to audiences at the XIV Colloquium in Generative Grammar, 2004, Porto; TiLT, 2004, Budapest; DiGS VIII, 2004, Yale; Workshop on The Structure of Parametric Variation, 2005, Newcastle; The Fifth Asian GLOW, 2005, Delhi; Digital Humanities Conference, 2006, Paris; Conference on Biolinguistics: Language Evolution and Variation, 2007, Venice; BCI Summer School, 2007, Trieste; Workshop on Formal models of linguistic diversity, 2007, Trieste; XXXII SIG Conference, 2007, Verona; Workshop on Bantu, Chinese and Romance nouns and noun phrases, 2007, London; XVIII International Congress of Linguists, 2008, Seoul. They also made the object of individual presentations in Paris, Potsdam, Rome, Los Angeles, Cambridge (Mass.). We are grateful to all such audiences.

We want to thank, for discussion and comments at various stages of this project, Luca Bortolussi, Luca Cavalli Sforza, Lucio Crisma, Francesco Guardiano, John Nerbonne, Andrea Novelletto, Andrea Sgarro, Marco René Spruit, Ed Stabler, and two anonymous referees. Special gratitude for invaluable help and advice is due to Paola Crisma, Gabriele Rigon, and Serena Danesi. Paola Crisma has also designed the basic structure of the website allowing a revealing and efficient encoding of the parametric data. Gabriele Rigon assisted us with most of the computational experiments.

Part of this research was accomplished while either author was visiting the Linguistics Department at UCLA, to which we are very grateful.

We are extremely indebted to our informants and consultants, who agreed to complete a Trigger List for us and/or to discuss the most intricate issues concerning the languages already included or others in the process of analysis: Birgit Alber, Manuela Ambar, Marcello Aprile, Rocco Aprile, Zlatko Anguelov, Judy Bernstein, Željko Bošković, Ana Castro, Ariel Cohen, Ricardo Etxepare, Franco Fanciullo, Abdelkader Fassi Fehri, Judit Gervain, Manuel Leonetti, Io Manolessou, Lanko Marušic, Matti Miestamo, Maria Polinsky, Sanja Roić, Alain Rouveret, Paweł Rutkowski, Roumyana Slabakova, Donca Steriade, Maryam Sy, Øystein Vangsnes, Akira Watanabe, Marit Westergaard, David Willis. Of course, errors are all ours and we take all the responsibility for the parametric analysis of data.

[48] Cf. Cavalli Sforza (2000) and Renfrew (1987).

Appendix

TABLE A	It	Sal	Sp	Fr	Ptg	Ru	Lat	CIG	NTG	Gri	Grk	Got	OE	E	D	Nor	Blg	SC	Rus	Ir	Wel	Heb	Ar	Wo	Hu	Fin	Hi	Ba
(1) ± gramm. person	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
(2) ± gramm. number +1	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
(3) ± gramm. gender +2	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	+	-
(4) ± variable person on D +1	-	-	+	-	-	-	?	+	?	-	+	?	?	-	-	-	-	-	-	?	?	-	+	-	-	-	-	+
(5) ± feature spread to N +2	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	-
(6) ± numb. on N (BNs) +5	+	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	0	+	+	+	0
(7) ± gramm. partial def	+	+	+	+	+	+	-	+	+	+	+	+	+	+	+	+	+	-	-	+	+	+	+	+	+	-	-	-
(8) ± gramm. def +7	+	+	+	+	+	+	0	+	+	+	+	-	+	+	+	+	+	0	0	+	+	+	+	+	+	0	0	0
(9) ± free null partitive Q +6	-	-	-	0	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	-	+	-	0
(10) ± gramm. dist. art. -5 or -6 or +7	-	-	-	-	-	-	0	-	-	-	-	-	-	-	-	-	-	0	0	-	-	-	-	+	-	0	0	+
(11) ± gramm. top. art. -10	-	+	-	-	-	-	0	-	-	-	-	-	-	-	-	-	-	0	0	-	-	-	-	0	-	0	0	0
(12) ± def checking N +7	-	-	-	-	-	+	0	-	-	-	-	-	-	-	-	+	+	0	0	-	-	-	-	-	-	0	0	0
(13) ± def spread to N +12	0	0	0	0	0	-	0	0	0	0	0	0	0	0	0	+	-	0	0	0	0	0	0	0	0	0	0	0
(14) ± def on attrib. +7, -12	-	-	-	-	-	0	0	+	+	-	+	+	-	-	-	0	0	0	0	-	-	+	+	-	-	0	0	0
(15) ± def on relatives +7	-	-	-	-	-	-	0	-	-	-	-	-	-	-	-	-	-	0	0	-	-	-	+	+	-	0	0	0
(16) ± D-controlled infl. on N +5	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	0	-	-	-	0
(17) ± gramm. cardinal nouns	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	?	+	+	?	+	+	+	?	-	+	-	-
(18) ± gramm. cardinal adjectives +17	0	0	0	0	0	0	0	0	0	0	0	0	+	0	0	0	?	-	+	?	-	-	+	?	0	+	0	0
(19) ± plural spread from cardinals +5, -17 or +18	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	?	0	+	?	0	0	+	0	-	-	+	0
(20) ± gramm. mass-to-count	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	?	+	-	+	-	-	-	-	-
(21) ± N-to-predicate incorporation	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-	+	-
(22) ± gramm. partial count -5 or -6 or +7, -21	+	+	+	+	+	+	0	-	-	+	+	0	+	+	+	+	+	0	0	-	-	-	-	0	0	0	0	+
(23) ± gramm. count +22	+	+	+	+	+	+	0	0	0	+	+	0	-	+	+	+	-	0	0	0	0	0	0	0	0	0	0	-
(24) ± count checking N +21 or +22	-	-	-	-	-	-	0	0	0	-	-	0	-	-	-	-	-	0	0	0	0	0	0	-	-	0	-	-
(25) ± prepositional Gen	+	+	+	+	+	-	-	-	-	-	-	-	-	+	+	+	+	-	-	+	+	+	+	-	-	-	-	-
(26) ± free inflected Gen -25	0	0	0	0	0	-	+	+	-	-	-	-	-	0	0	0	0	-	-	0	0	0	0	-	-	-	-	-
(27) ± GenO +25 or -26	-	-	-	-	-	-	0	0	+	+	+	+	+	-	+	?	-	+	+	+	+	-	-	+	-	+	+	+
(28) ± GenS +25 or -26	-	-	-	-	-	+	0	0	-	-	-	+	+	+	+	+	-	-	-	-	-	+	+	-	+	+	+	+
(29) ± postpositional Gen +27 or +28	0	0	0	0	0	-	0	0	-	-	-	-	-	+	-	+	0	-	-	-	-	-	-	-	-	-	+	?
(30) ± Gen over DemP	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	?
(31) ± poss checking N	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	+	+	-	-
(32) ± structured APs	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+
(33) ± feature spread to structured APs +32	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+	+	+	?	+	+	0	-	+	+	-
(34) ± feature spread to pred. APs	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	+	+	+	+	+	?	+	+	+	+	+	+	+
(35) ± numb. on A +6, +33 or +34	+	+	+	0	+	+	+	+	+	+	+	+	+	0	+	+	+	+	+	+	?	+	+	0	+	+	+	0
(36) ± D-controlled infl. on A +33	-	-	-	-	-	-	-	-	-	-	-	-	+	+	0	+	+	-	-	-	?	-	-	0	0	-	-	0
(37) ± DemP over relatives	+	+	+	+	+	+	?	?	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	?	-
(38) ± free APs in Mod +32	+	+	+	+	+	-	+	+	+	-	+	+	-	-	-	-	-	-	-	-	-	-	+	0	-	-	-	-
(39) ± APs in Mod -38	0	0	0	0	0	-	0	0	0	-	0	0	+	+	-	-	+	+	+	-	-	+	0	0	-	-	?	?
(40) ± overt Mod° -32 or +38 or +39	-	+	-	-	-	0	-	-	-	0	-	-	-	-	0	0	-	-	-	0	0	-	-	+	0	0	?	?
(41) ± adjectival Gen	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	-	-	-	-	-	-	-	?	-
(42) ± N-raising with pied-piping	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-	-	-	-	+
(43) ± N over ext.arg. -42	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	-	+	+	+	+	+	0	0	+	+	+	+	0
(44) ± N over GenO +26 or +27, -30 ,+43	0	0	0	0	0	0	-	-	+	+	+	+	+	0	+	0	0	+	+	+	+	0	0	+	0	-	0	0
(45) ± N over As +32, (-26 or -27, +43) or +44	+	+	+	+	+	+	0	0	-	+	-	-	-	0	-	0	-	-	-	+	+	0	0	0	-	0	-	0
(46) ± N over M2 As +45	+	+	+	+	+	+	0	0	0	+	0	0	0	0	0	0	0	0	0	+	+	0	0	0	0	0	0	0
(47) ± N over M1 As +46	-	+	-	-	-	-	0	0	0	?	0	0	0	0	0	0	0	0	0	+	+	0	0	0	0	0	0	0
(48) ± N over high As +47	0	+	0	0	0	0	0	0	0	?	0	0	0	0	0	0	0	0	0	-	-	0	0	0	0	0	0	0
(49) ± N over cardinals +42 or +48	0	-	0	0	0	0	0	0	0	?	0	0	0	0	0	0	0	0	0	0	0	+	+	?	0	0	0	-
(50) ± strong D (person) +1, +8 or +28	+	+	+	+	+	+	0	+	+	+	+	0	-	-	-	-	+	0	0	-	-	+	+	-	?	0	0	+
(51) ± NP over D	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	+
(52) ± strong deixis	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	+	+	+	+	+	+	+
(53) ± strong anaphoricity +52	+	+	-	+	+	-	-	-	-	+	-	-	+	+	+	+	-	+	+	0	0	?	-	+	+	+	+	+
(54) ± DP over Dem -51, +52	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	0	+	?	0	-	-	-	0
(55) ± D checking Dem -5 or -6 or +7, +52	+	+	+	+	+	+	0	-	-	+	-	?	+	+	+	-	+	0	0	0	0	+	-	+	-	0	0	+
(56) ± D-checking poss -5 or -6 or +8, +50 or -28	-	-	+	+	?	-	0	-	-	-	-	0	0	0	0	0	-	0	0	+	+	-	-	+	?	0	0	-
(57) ± feature spread on poss +35 or +34 or +33	+	+	+	+	+	+	+	+	+	-	-	+	+	0	+	+	+	+	+	-	?	-	-	+	-	-	+	-
(58) ± feature spread on postp. Gen +29, +57	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-	0	0	0	0	0	0	0	0	0	0	+	0
(59) ± enclitic possessives	-	-	-	-	-	-	-	+	+	+	+	-	-	-	-	-	?	-	-	-	-	-	-	?	-	-	-	-
(60) ± Consisten. Princ. +51 or (-44 or ... or -47, +A-Compl)	+	0	+	+	+	+	?	?	?	?	-	?	+	+	+	+	-	-	-	0	0	0	0	?	+	+	0	+
(61) ± null-N-licensing art -5 or-6 or -12, +50 or +51	-	-	+	-	+	0	0	+	+	-	-	0	0	0	0	0	0	0	0	0	0	-	-	?	?	0	0	+
(62) ± obl. def. inh. +7,-22, (-25, +26) or +27,+42 or +45 or -50	0	0	0	0	0	0	0	0	0	0	0	0	-	0	0	0	0	0	0	-	+	+	+	-	0	0	0	0
(63) ± gramm. geogr. art. -5 or-6 or +7, -22 or -23 or +45	+	+	-	+	-	+	0	+	+	+	0	0	0	0	0	0	-	0	0	?	?	-	-	0	-	0	-	-
	It	Sal	Sp	Fr	Ptg	Ru	Lat	CIG	NTG	Gri	Grk	Got	OE	E	D	Nor	Blg	SC	Rus	Ir	Wel	Heb	Ar	Wo	Hu	Fin	Hi	Ba

Fig. 1: Table A.

	It	Sal	Sp	Fr	Ptg	Rum	Lat	ClG	NTG	Gri	Grk	Got	OE	E	D	Nor	Blg	SC	Rus	Ir	Wel	Heb	Ar	Wo	Hu	Fin	Hi	Bas	
It		49; 3	48; 5	49; 2	50; 2	45; 5	28; 3	36; 8	38; 9	45; 5	40; 10	33; 8	38; 9	38; 6	40; 7	37; 7	38; 8	28; 8	29; 8	32; 8	28; 9	34; 10	31; 14	21; 12	32; 12	24; 11	26; 8	18; 17	**It**
Sal	0,0577		44; 8	45; 5	46; 5	42; 7	27; 4	34; 10	36; 11	44; 6	38; 11	31; 10	35; 11	35; 8	38; 8	35; 8	36; 9	27; 8	28; 8	32; 9	28; 10	32; 13	29; 17	22; 11	30; 13	23; 11	26; 8	18; 17	**Sal**
Sp	0,0943	0,1538		46; 5	50; 2	43; 7	29; 2	37; 7	38; 9	40; 10	40; 10	34; 7	37; 10	36; 8	38; 9	35; 9	38; 8	26; 10	27; 10	33; 7	29; 8	32; 12	32; 13	20; 13	31; 13	22; 13	24; 10	19; 16	**Sp**
Fr	0,0392	0,1000	0,0980		47; 3	41; 7	25; 4	32; 10	34; 11	41; 7	36; 12	30; 9	35; 10	36; 7	37; 8	34; 8	34; 10	25; 9	26; 9	30; 8	27; 9	30; 12	27; 16	22; 11	29; 13	22; 11	23; 9	17; 18	**Fr**
Ptg	0,0385	0,0980	0,0385	0,0600		43; 6	28; 3	35; 8	37; 9	42; 7	38; 11	33; 8	38; 9	38; 6	40; 7	37; 7	38; 7	28; 8	29; 8	32; 7	28; 8	33; 10	30; 14	21; 11	33; 11	24; 11	26; 8	19; 15	**Ptg**
Rum	0,1000	0,1429	0,1400	0,1458	0,1224		28; 3	35; 7	38; 8	45; 6	39; 10	35; 6	39; 9	35; 9	40; 8	39; 8	39; 8	29; 9	30; 9	31; 10	27; 11	33; 10	29; 14	20; 13	35; 11	29; 9	29; 7	21; 15	**Rum**
Lat	0,0968	0,1290	0,0645	0,1379	0,0968	0,0968		31; 2	29; 4	25; 7	28; 5	29; 4	26; 7	21; 7	24; 7	24; 6	24; 4	26; 6	27; 6	20; 6	16; 7	20; 8	21; 8	12; 10	20; 10	23; 9	25; 5	13; 9	**Lat**
ClG	0,1818	0,2273	0,1591	0,2381	0,1860	0,1667	0,0606		43; 2	35; 10	40; 5	34; 5	31; 11	26; 12	29; 12	29; 10	31; 8	24; 9	25; 9	26; 9	22; 10	29; 12	31; 11	15; 16	26; 13	21; 12	23; 8	15; 15	**ClG**
NTG	0,1915	0,2340	0,1915	0,2444	0,1957	0,1739	0,1212	0,0444		41; 8	46; 3	40; 4	37; 10	27; 14	34; 11	30; 11	34; 8	31; 6	32; 6	31; 9	27; 10	31; 13	32; 13	21; 13	30; 13	24; 12	26; 8	16; 17	**NTG**
Gri	0,1000	0,1200	0,2000	0,1458	0,1429	0,1176	0,2188	0,2222	0,1633		45; 6	34; 9	40; 9	34; 10	41; 8	35; 10	35; 10	31; 7	32; 7	36; 6	31; 8	33; 12	28; 17	25; 10	35; 11	29; 9	28; 8	21; 15	**Gri**
Grk	0,2000	0,2245	0,2000	0,2500	0,2245	0,2041	0,1515	0,1111	0,0612	0,1176		39; 5	38; 12	30; 15	36; 13	32; 13	36; 9	31; 8	32; 8	31; 9	26; 11	32; 12	34; 11	21; 15	32; 13	25; 13	26; 10	19; 17	**Grk**
Got	0,1951	0,2439	0,1707	0,2308	0,1951	0,1463	0,1212	0,1282	0,0909	0,2093	0,1136		39; 5	28; 10	36; 6	31; 7	29; 8	30; 7	31; 7	28; 9	25; 9	28; 10	28; 11	19; 12	30; 11	25; 11	27; 7	15; 12	**Got**
OE	0,1915	0,2391	0,2128	0,2222	0,1915	0,1875	0,2121	0,2619	0,2128	0,1837	0,2400	0,1136		37; 8	44; 5	37; 8	35; 9	34; 6	36; 5	33; 8	30; 9	32; 12	28; 16	26; 9	35; 10	30; 9	29; 6	21; 12	**OE**
E	0,1364	0,1860	0,1818	0,1628	0,1364	0,2045	0,2500	0,3158	0,3415	0,2273	0,3333	0,2632	0,1778		40; 5	38; 5	31; 10	23; 11	24; 11	26; 9	26; 9	30; 10	24; 16	18; 14	32; 10	21; 13	23; 8	19; 13	**E**
D	0,1489	0,1739	0,1915	0,1778	0,1489	0,1667	0,2258	0,2927	0,2444	0,1633	0,2653	0,1429	0,1020	0,1111		41; 5	33; 11	29; 9	30; 9	33; 7	31; 6	30; 12	25; 17	23; 10	35; 10	27; 11	28; 7	18; 15	**D**
Nor	0,1591	0,1860	0,2045	0,1905	0,1591	0,1702	0,2000	0,2564	0,2683	0,2222	0,2889	0,1842	0,1778	0,1163	0,1087		32; 11	25; 10	26; 10	28; 8	25; 8	28; 12	25; 15	17; 13	32; 10	25; 11	27; 7	18; 14	**Nor**
Blg	0,1739	0,2000	0,1739	0,2273	0,1556	0,1702	0,1429	0,2051	0,1905	0,2222	0,2000	0,2162	0,2045	0,2439	0,2500	0,2558		31; 4	31; 4	27; 10	23; 10	33; 8	27; 13	17; 15	28; 13	21; 12	24; 7	16; 16	**Blg**
SC	0,2222	0,2286	0,2778	0,2647	0,2222	0,2368	0,1875	0,2727	0,1622	0,1842	0,2051	0,1892	0,1500	0,3235	0,2368	0,2857	0,1143		40; 1	25; 7	23; 7	26; 9	20; 14	19; 8	25; 12	29; 10	29; 5	15; 13	**SC**
Rus	0,2162	0,2222	0,2703	0,2571	0,2162	0,2308	0,1818	0,2647	0,1579	0,1795	0,2000	0,1842	0,1220	0,3143	0,2308	0,2778	0,1143	0,0244		25; 7	22; 8	25; 10	22; 13	19; 8	25; 13	30; 10	30; 5	15; 13	**Rus**
Ir	0,2000	0,2195	0,1750	0,2105	0,1795	0,2439	0,2308	0,2571	0,2250	0,1429	0,2250	0,2432	0,1951	0,2571	0,1750	0,2222	0,2703	0,2188	0,2188		39; 1	29; 10	26; 12	20; 10	27; 9	23; 8	19; 9	11; 15	**Ir**
Wel	0,2432	0,2632	0,2162	0,2500	0,2222	0,2895	0,3043	0,3125	0,2703	0,2051	0,2973	0,2647	0,2308	0,2571	0,1622	0,2424	0,3030	0,2333	0,2667	0,0250		27; 10	24; 12	18; 11	24; 10	19; 10	15; 10	9; 16	**Wel**
Heb	0,2273	0,2889	0,2727	0,2857	0,2326	0,2326	0,2857	0,2927	0,2955	0,2667	0,2727	0,2632	0,2727	0,2500	0,2857	0,3000	0,1951	0,2571	0,2857	0,2564	0,2703		40; 7	16; 16	30; 10	25; 9	19; 11	17; 16	**Heb**
Ar	0,3111	0,3696	0,2889	0,3721	0,3182	0,3256	0,2759	0,2619	0,2889	0,3778	0,2444	0,2821	0,3636	0,4000	0,4048	0,3750	0,3250	0,4118	0,3714	0,3158	0,3333	0,1489		14;19	26;14	22;12	16;15	15;19	**Ar**
Wo	0,3636	0,3333	0,3939	0,3333	0,3438	0,3939	0,4545	0,5161	0,3824	0,2857	0,4167	0,3871	0,2571	0,4375	0,3030	0,4333	0,4688	0,2963	0,2963	0,3333	0,3793	0,5000	0,5758		22;10	17;9	16;8	15;13	**Wo**
Hu	0,2727	0,3023	0,2955	0,3095	0,2500	0,2391	0,3333	0,3333	0,3023	0,2391	0,2889	0,2683	0,2222	0,2381	0,2222	0,2381	0,3171	0,3243	0,3421	0,2500	0,2941	0,2500	0,3500	0,3125		31;6	26;9	20;12	**Hu**
Fin	0,3143	0,3235	0,3714	0,3333	0,3143	0,2368	0,2813	0,3636	0,3333	0,2368	0,3421	0,3056	0,2308	0,3824	0,2895	0,3056	0,3636	0,2564	0,2500	0,2581	0,3448	0,2647	0,3529	0,3462	0,1622		25;9	19;9	**Fin**
Hi	0,2353	0,2353	0,2941	0,2813	0,2353	0,1944	0,1667	0,2581	0,2353	0,2222	0,2778	0,2059	0,1714	0,2581	0,2000	0,2059	0,2258	0,1471	0,1429	0,3214	0,4000	0,3667	0,4839	0,3333	0,2571	0,2647		18;8	**Hi**
Bas	0,4857	0,4857	0,4571	0,5143	0,4412	0,4167	0,4091	0,5000	0,5152	0,4167	0,4722	0,4444	0,3636	0,4063	0,4545	0,4375	0,5000	0,4643	0,4643	0,5769	0,6400	0,4848	0,5588	0,4643	0,3750	0,3214	0,3077		**Bas**
	It	**Sal**	**Sp**	**Fr**	**Ptg**	**Rum**	**Lat**	**ClG**	**NTG**	**Gri**	**Grk**	**Got**	**OE**	**E**	**D**	**Nor**	**Blg**	**SC**	**Rus**	**Ir**	**Wel**	**Heb**	**Ar**	**Wo**	**Hu**	**Fin**	**Hi**	**Bas**	

Fig. 2: Table B.

Pairs	Ch. Prob.	NHD
SC, Rus	1/53634713550	0,0244
Ir, Wel	1/27487790694	0,0250
It, Ptg	1/3396379809480	0,0385
Sp, Ptg	1/3396379809480	0,0385
It, Fr	1/1766117500930	0,0392
ClG, NTG	1/35539769787	0,0444
It, Sal	1/203782788569	0,0577
Fr, Ptg	1/57443872798	0,0600
Lat, ClG	1/16268816	0,0606
NTG, Grk	1/30555251488	0,0612
Sp, Lat	1/4618244	0,0645
NTG, Got	1/129591576	0,0909
It, Sp	1/3138741449	0,0943
It, Lat	1/477749	0,0968
Ptg, Lat	1/477749	0,0968
Rum, Lat	1/477749	0,0968
Sal, Ptg	1/958596125	0,0980
Sp, Fr	1/958596125	0,0980
It, Rum	1/531395678	0,1000
Sal, Fr	1/531395678	0,1000
It, Gri	1/531395678	0,1000
OE, D	1/295219821	0,1020
D, Nor	1/51335793	0,1087
ClG, Grk	1/28798128	0,1111
E, D	1/28798128	0,1111
Got, OE	1/16198947	0,1136
Grk, Got	1/16198947	0,1136
Blg, SC	1/656221	0,1143
Blg, Rus	1/656221	0,1143
E, Nor	1/9137868	0,1163
Gri, Grk	1/125034277	0,1176
Rum, Gri	1/125034277	0,1176
Sal, Gri	1/70852757	0,1200
Lat, NTG	1/209920	0,1212
Lat, Got	1/209920	0,1212
OE, Rus	1/2934386	0,1220
Ptg, Rum	1/40257248	0,1224
ClG, Got	1/954840	0,1282
Sal, Lat	1/68250	0,1290
It, E	1/2492146	0,1364
Ptg, E	1/2492146	0,1364
Fr, Lat	1/22604	0,1379
Sp, Rum	1/11272030	0,1400
Sal, Rum	1/6553506	0,1429
Ptg, Gri	1/6553506	0,1429
Got, D	1/838396	0,1429
Gri, Ir	1/838396	0,1429
Rus, Hi	1/105842	0,1429
Lat, Blg	1/13110	0,1429
Fr, Rum	1/3822878	0,1458
Fr, Gri	1/3822878	0,1458
Rum, Got	1/489064	0,1463
SC, Hi	1/61741	0,1471
It, D	1/2237782	0,1489
Ptg, D	1/2237782	0,1489
Heb, Ar	1/2237782	0,1489
OE, SC	1/286452	0,1500
Lat, Grk	1/36193	0,1515
Sal, Sp	1/5984547	0,1538
Ptg, Blg	1/775334	0,1556
NTG, Rus	1/99569	0,1579
It, Nor	1/459079	0,1591
Sp, ClG	1/459079	0,1591
Ptg, Nor	1/459079	0,1591
NTG, SC	1/59119	0,1622
D, Wel	1/59119	0,1622
Hu, Fin	1/59119	0,1622
Fr, E	1/272966	0,1628
NTG, Gri	1/1248287	0,1633
Gri, D	1/1248287	0,1633
Rum, D	1/745927	0,1667
Rum, ClG	1/163021	0,1667
Lat, Hi	1/7535	0,1667
Rum, Nor	1/447556	0,1702
Rum, Blg	1/447556	0,1702
Sp, Got	1/97813	0,1707
OE, Hi	1/21168	0,1714
It, Blg	1/269681	0,1739
Sal, D	1/269681	0,1739
Sp, Blg	1/269681	0,1739
Rum, NTG	1/269681	0,1739
Sp, Ir	1/58975	0,1750
D, Ir	1/58975	0,1750
OE, E	1/163228	0,1778
OE, Nor	1/163228	0,1778
Fr, D	1/163228	0,1778
Ptg, Ir	1/35743	0,1795
Gri, Rus	1/35743	0,1795
It, ClG	1/99260	0,1818
Sp, E	1/99260	0,1818
Lat, Rus	1/7756	0,1818
Gri, OE	1/274014	0,1837
Got, Nor	1/21781	0,1842
Gri, SC	1/21781	0,1842
Got, Rus	1/21781	0,1842
Sal, E	1/60659	0,1860
Sal, Nor	1/60659	0,1860
Ptg, ClG	1/60659	0,1860
Rum, OE	1/167834	0,1875
Lat, SC	1/4740	0,1875
Got, SC	1/13349	0,1892
Fr, Nor	1/37262	0,1905
NTG, Blg	1/37262	0,1905
It, NTG	1/103282	0,1915
It, OE	1/103282	0,1915
Sp, NTG	1/103282	0,1915
Sp, D	1/103282	0,1915
Ptg, OE	1/103282	0,1915
Rum, Hi	1/8232	0,1944
It, Got	1/23015	0,1951
Ptg, Got	1/23015	0,1951
OE, Ir	1/23015	0,1951
Blg, Heb	1/23015	0,1951
Ptg, NTG	1/63872	0,1957
It, Grk	1/109606	0,2000
Sp, Gri	1/109606	0,2000
Sp, Grk	1/109606	0,2000
Sal, Blg	1/39704	0,2000
Grk, Blg	1/39704	0,2000
It, Ir	1/14297	0,2000
Grk, Rus	1/14297	0,2000
D, Hi	1/5110	0,2000
Lat, Nor	1/1808	0,2000
Rum, Grk	1/68504	0,2041
Sp, Nor	1/24815	0,2045
Rum, E	1/24815	0,2045
OE, Blg	1/24815	0,2045
ClG, Blg	1/8936	0,2051
Gri, Wel	1/8936	0,2051
Grk, SC	1/8936	0,2051
Got, Hi	1/3194	0,2059
Nor, Hi	1/3194	0,2059
Gri, Got	1/15598	0,2093
Fr, Ir	1/5621	0,2105
Lat, OE	1/2011	0,2121
Sp, OE	1/27180	0,2128
NTG, OE	1/27180	0,2128
It, Rus	1/3560	0,2162
Sp, Wel	1/3560	0,2162
Ptg, Rus	1/3560	0,2162
Got, Blg	1/3560	0,2162
Lat, Gri	1/1276	0,2188
SC, Ir	1/1276	0,2188
Rus, Ir	1/1276	0,2188
Sal, Ir	1/6277	0,2195
ClG, Gri	1/11029	0,2222
Fr, OE	1/11029	0,2222
Gri, Nor	1/11029	0,2222
Gri, Blg	1/11029	0,2222
OE, Hu	1/11029	0,2222
D, Hu	1/11029	0,2222
It, SC	1/2271	0,2222
Sal, Rus	1/2271	0,2222
Ptg, SC	1/2271	0,2222
Ptg, Wel	1/2271	0,2222
Gri, Hi	1/2271	0,2222
Nor, Ir	1/2271	0,2222
Sal, Grk	1/19322	0,2245
Ptg, Grk	1/19322	0,2245
NTG, Ir	1/4021	0,2250
Grk, Ir	1/4021	0,2250
Lat, D	1/817	0,2258
Blg, Hi	1/817	0,2258
Sal, ClG	1/7090	0,2273
Fr, Blg	1/7090	0,2273
Gri, E	1/7090	0,2273
It, Heb	1/7090	0,2273
Sal, SC	1/1460	0,2286
Fr, Got	1/2594	0,2308
Rum, Rus	1/2594	0,2308
OE, Wel	1/2594	0,2308
D, Rus	1/2594	0,2308
OE, Fin	1/2594	0,2308
Lat, Ir	1/291	0,2308
Ptg, Heb	1/4588	0,2326
Rum, Heb	1/4588	0,2326
SC, Wel	1/527	0,2333
Sal, NTG	1/8080	0,2340
It, Hi	1/946	0,2353
Sal, Hi	1/946	0,2353
Ptg, Hi	1/946	0,2353
NTG, Hi	1/946	0,2353
Rum, SC	1/1686	0,2368
D, SC	1/1686	0,2368
Rum, Fin	1/1686	0,2368
Gri, Fin	1/1686	0,2368
Fr, ClG	1/2989	0,2381
E, Hu	1/2989	0,2381
Nor, Hu	1/2989	0,2381
Sal, OE	1/5275	0,2391
Rum, Hu	1/5275	0,2391

Fig. 3: Table D.

Gri, Hu	1/5275	0,2391
Grk, OE	1/9274	0,2400
Nor, Wel	1/619	0,2424
It, Wel	1/1105	0,2432
Got, Ir	1/1105	0,2432
Sal, Got	1/1961	0,2439
Rum, Ir	1/1961	0,2439
E, Blg	1/1961	0,2439
Fr, NTG	1/3466	0,2444
NTG, D	1/3466	0,2444
Grk, Ar	1/3466	0,2444
Fr, Grk	1/4040	0,2500
D, Blg	1/2294	0,2500
Ptg, Hu	1/2294	0,2500
E, Heb	1/1297	0,2500
Rus, Fin	1/1297	0,2500
Heb, Hu	1/1297	0,2500
Fr, Wel	1/730	0,2500
Ir, Hu	1/730	0,2500
Lat, E	1/227	0,2500
Nor, Blg	1/1529	0,2558
ClG, Nor	1/865	0,2564
SC, Fin	1/865	0,2564
Ir, Heb	1/865	0,2564
Fr, Rus	1/487	0,2571
ClG, Ir	1/487	0,2571
E, Ir	1/487	0,2571
E, Wel	1/487	0,2571
Hu, Hi	1/487	0,2571
SC, Heb	1/487	0,2571
OE, Wo	1/487	0,2571
ClG, Hi	1/272	0,2581
E, Hi	1/272	0,2581
Ir, Fin	1/272	0,2581
ClG, OE	1/1027	0,2619
ClG, Ar	1/1027	0,2619
Got, E	1/581	0,2632
Sal, Wel	1/581	0,2632
Got, Heb	1/581	0,2632
Fr, SC	1/328	0,2647
ClG, Rus	1/328	0,2647
Got, Wel	1/328	0,2647
Fin, Hi	1/328	0,2647
Heb, Fin	1/328	0,2647
Grk, D	1/2144	0,2653
Rus, Wel	1/183	0,2667
Gri, Heb	1/1223	0,2667
NTG, Nor	1/696	0,2683

Got, Hu	1/696	0,2683
Sp, Rus	1/395	0,2703
NTG, Wel	1/395	0,2703
Blg, Ir	1/395	0,2703
Wel, Heb	1/395	0,2703
It, Hu	1/834	0,2727
Sp, Heb	1/834	0,2727
Grk, Heb	1/834	0,2727
OE, Heb	1/834	0,2727
ClG, SC	1/223	0,2727
Lat, Ar	1/125	0,2759
Sp, SC	1/270	0,2778
Grk, Hi	1/270	0,2778
Nor, Rus	1/270	0,2778
Fr, Hi	1/153	0,2813
Lat, Fin	1/153	0,2813
Got, Ar	1/328	0,2821
Fr, Heb	1/398	0,2857
D, Heb	1/398	0,2857
Nor, SC	1/187	0,2857
Rus, Heb	1/187	0,2857
Gri, Wo	1/187	0,2857
Lat, Heb	1/86	0,2857
Grk, Nor	1/482	0,2889
Sal, Heb	1/482	0,2889
Sp, Ar	1/482	0,2889
NTG, Ar	1/482	0,2889
Grk, Hu	1/482	0,2889
Rum, Wel	1/228	0,2895
D, Fin	1/228	0,2895
ClG, D	1/278	0,2927
ClG, Heb	1/278	0,2927
Sp, Hi	1/131	0,2941
Wel, Hu	1/131	0,2941
Sp, Hu	1/339	0,2955
NTG, Heb	1/339	0,2955
SC, Wo	1/60	0,2963
Rus, Wo	1/60	0,2963
Grk, Wel	1/161	0,2973
Nor, Heb	1/197	0,3000
Sal, Hu	1/240	0,3023
NTG, Hu	1/240	0,3023
Blg, Wel	1/93	0,3030
D, Wo	1/93	0,3030
Lat, Wel	1/34	0,3043
Got, Fin	1/114	0,3056
Nor, Fin	1/114	0,3056
Hi, Bas	1/43	0,3077

Fr, Hu	1/172	0,3095
It, Ar	1/211	0,3111
ClG, Wel	1/67	0,3125
Wo, Hu	1/67	0,3125
E, Rus	1/82	0,3143
It, Fin	1/82	0,3143
Ptg, Fin	1/82	0,3143
ClG, E	1/102	0,3158
Ir, Ar	1/102	0,3158
Blg, Hu	1/125	0,3171
Ptg, Ar	1/153	0,3182
Ir, Hi	1/39	0,3214
Fin, Bas	1/39	0,3214
E, SC	1/60	0,3235
Sal, Fin	1/60	0,3235
SC, Hu	1/74	0,3243
Blg, Ar	1/91	0,3250
Rum, Ar	1/112	0,3256
Grk, E	1/102	0,3333
ClG, Hu	1/68	0,3333
NTG, Fin	1/55	0,3333
Wel, Ar	1/55	0,3333
Sal, Wo	1/44	0,3333
Fr, Wo	1/44	0,3333
Fr, Fin	1/44	0,3333
Lat, Hu	1/36	0,3333
Ir, Wo	1/36	0,3333
Wo, Hi	1/23	0,3333
NTG, E	1/62	0,3415
Grk, Fin	1/51	0,3421
Rus, Hu	1/51	0,3421
Ptg, Wo	1/33	0,3438
Wel, Fin	1/27	0,3448
Wo, Fin	1/21	0,3462
Ar, Hu	1/47	0,3500
Ar, Fin	1/31	0,3529
OE, Ar	1/42	0,3636
ClG, Fin	1/24	0,3636
Blg, Fin	1/24	0,3636
It, Wo	1/24	0,3636
OE, Bas	1/24	0,3636
Heb, Hi	1/20	0,3667
Sal, Ar	1/40	0,3696
Sp, Fin	1/23	0,3714
Rus, Ar	1/23	0,3714
Fr, Ar	1/33	0,3721
Nor, Ar	1/27	0,3750
Hu, Bas	1/19	0,3750

Gri, Ar	1/32	0,3778
Wel, Wo	1/16	0,3793
E, Fin	1/19	0,3824
NTG, Wo	1/19	0,3824
Got, Wo	1/15	0,3871
Sp, Wo	1/15	0,3939
Rum, Wo	1/15	0,3939
E, Ar	1/17	0,4000
Wel, Hi	1/10	0,4000
D, Ar	1/17	0,4048
E, Bas	1/12	0,4063
Lat, Bas	1/8	0,4091
SC, Ar	1/12	0,4118
Rum, Bas	1/12	0,4167
Gri, Bas	1/12	0,4167
Grk, Wo	1/12	0,4167
Nor, Wo	1/9	0,4333
E, Wo	1/9	0,4375
Nor, Bas	1/9	0,4375
Ptg, Bas	1/9	0,4412
Got, Bas	1/8	0,4444
D, Bas	1/8	0,4545
Lat, Wo	1/6	0,4545
Sp, Bas	1/8	0,4571
SC, Bas	1/7	0,4643
Rus, Bas	1/7	0,4643
Wo, Bas	1/7	0,4643
Blg, Wo	1/8	0,4688
Grk, Bas	1/8	0,4722
Ar, Hi	1/7	0,4839
Heb, Bas	1/7	0,4848
It, Bas	1/8	0,4857
Sal, Bas	1/8	0,4857
Blg, Bas	1/7	0,5000
Heb, Wo	1/7	0,5000
ClG, Bas	1/7	0,5000
Fr, Bas	1/8	0,5143
NTG, Bas	1/7	0,5152
ClG, Wo	1/7	0,5161
Ar, Bas	1/9	0,5588
Ar, Wo	1/10	0,5758
Ir, Bas	1/9	0,5769
Wel, Bas	1/16	0,6400

Fig. 3: Table D (*Continued*).

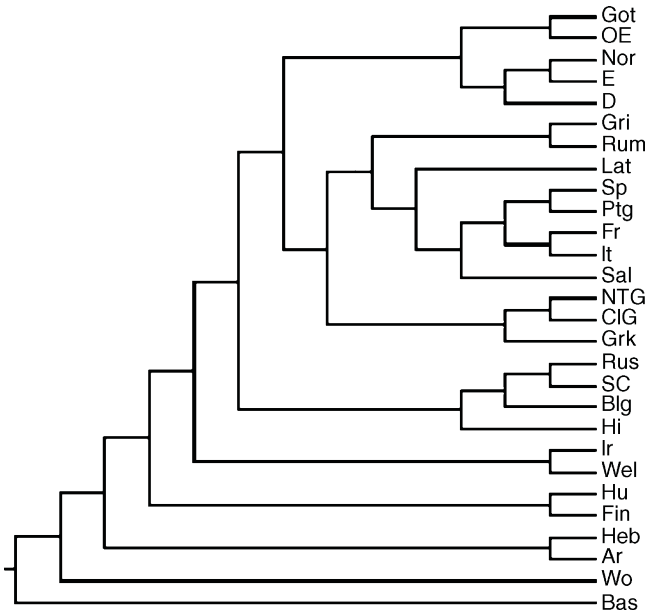

Fig. 4. Bootstrapped tree from Table D (Kitsch, 1000 re-samplings, all languages).

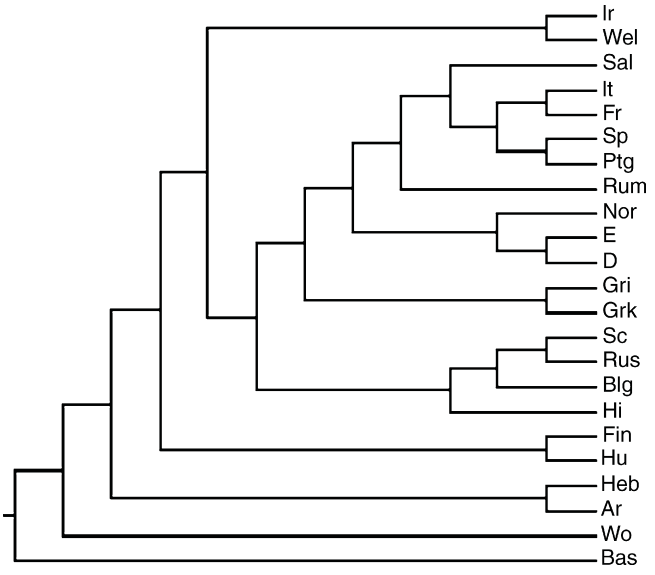

Fig. 5. Bootstrapped tree from Table D (Kitsch, 1000 re-samplings, modern languages).

Lexical Distances

	It	Sp	Fr	Ptg	Rum	Grk	E	D	Nor	Blg	SC	Rus	Ir	Wel	Hin	
It	0,0000	0,2120	0,1970	0,2270	0,3400	0,8220	0,7530	0,7350	0,7540	0,7690	0,7550	0,7610	0,8000	0,7930	0,8180	**It**
Sp	0,0943	0,0000	0,2660	0,1260	0,4060	0,8330	0,7600	0,7470	0,7610	0,7820	0,7680	0,7690	0,8050	0,8130	0,8190	**Sp**
Fr	0,0392	0,0980	0,0000	0,2910	0,4210	0,8430	0,7640	0,7560	0,7700	0,7910	0,7720	0,7780	0,8120	0,8100	0,8240	**Fr**
Ptg	0,0385	0,0385	0,0600	0,0000	0,3710	0,8330	0,7600	0,7530	0,7610	0,7810	0,7660	0,7730	0,8170	0,8040	0,8130	**Ptg**
Rum	0,1000	0,1400	0,1458	0,1224	0,0000	0,8430	0,7730	0,7510	0,7860	0,7980	0,7780	0,7810	0,8370	0,8120	0,8270	**Rum**
Grk	0,2000	0,2000	0,2500	0,2245	0,2041	0,0000	0,8380	0,8120	0,8210	0,8110	0,8210	0,8320	0,8590	0,8670	0,8740	**Grk**
E	0,1364	0,1818	0,1628	0,1364	0,2045	0,3333	0,0000	0,4220	0,4520	0,7720	0,7660	0,7580	0,8170	0,8410	0,8540	**E**
D	0,1489	0,1915	0,1778	0,1489	0,1667	0,2653	0,1111	0,0000	0,3670	0,7690	0,7640	0,7550	0,8060	0,8200	0,8530	**D**
Nor	0,1591	0,2045	0,1905	0,1591	0,1702	0,2889	0,1163	0,1087	0,0000	0,7730	0,7720	0,7580	0,8360	0,8490	0,8520	**Nor**
Blg	0,1739	0,1739	0,2273	0,1556	0,1702	0,2000	0,2439	0,2500	0,2558	0,0000	0,2910	0,3650	0,8180	0,8380	0,8010	**Blg**
SC	0,2222	0,2778	0,2647	0,2222	0,2368	0,2051	0,3235	0,2368	0,2857	0,1143	0,0000	0,3250	0,7960	0,8210	0,8050	**SC**
Rus	0,2162	0,2703	0,2571	0,2162	0,2308	0,2000	0,3143	0,2308	0,2778	0,1143	0,0244	0,0000	0,7820	0,8180	0,8000	**Rus**
Ir	0,2000	0,1750	0,2105	0,1795	0,2439	0,2250	0,2571	0,1750	0,2222	0,2703	0,2188	0,2188	0,0000	0,6450	0,8780	**Ir**
Wel	0,2432	0,2162	0,2500	0,2222	0,2895	0,2973	0,2571	0,1622	0,2424	0,3030	0,2333	0,2667	0,0250	0,0000	0,8760	**Wel**
Hin	0,2353	0,2941	0,2813	0,2353	0,1944	0,2778	0,2581	0,2000	0,2059	0,2258	0,1471	0,1429	0,3214	0,4000	0,0000	**Hin**
	It	**Sp**	**Fr**	**Ptg**	**Rum**	**Grk**	**E**	**D**	**Nor**	**Blg**	**SC**	**Rus**	**Ir**	**Wel**	**Hin**	

Syntactic Distances

Fig. 6. Lexical and syntactic distances among 15 IE languages.

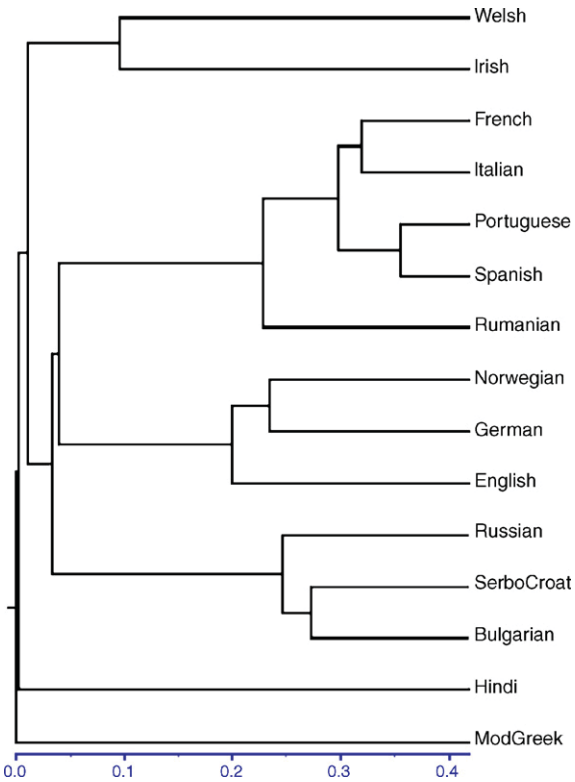

Fig. 7. Tree produced by Kitsch using lexical distances.

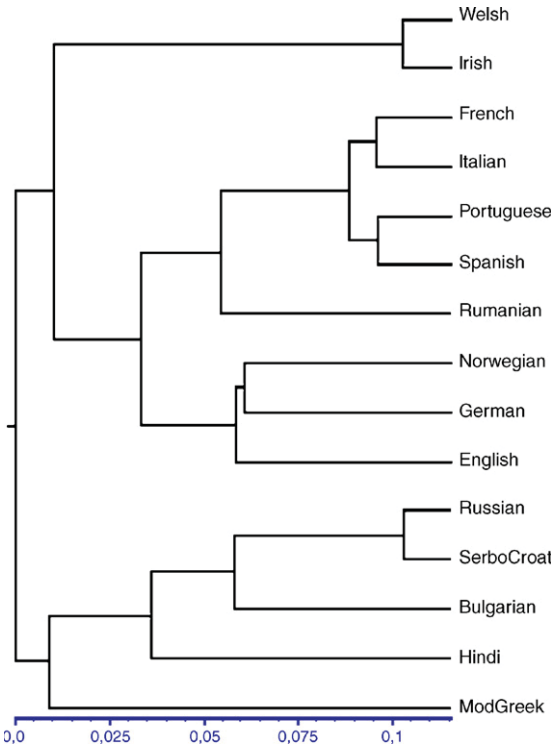

Fig. 8. Tree produced by Kitsch using syntactic distances.

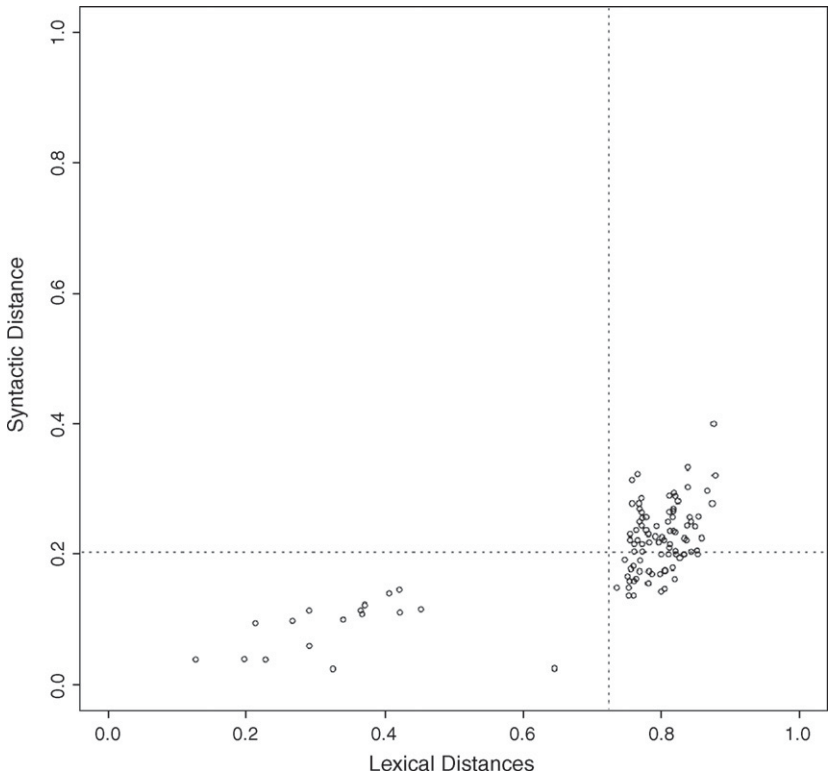

Fig. 9. Scatter plot of lexical and syntactic distances.

References

Baker, M., 2001. The Atoms of Language. Basic Books, New York.

Barbujani, G., Sokal, R.R., 1990. Zones of sharp genetic change in Europe are also linguistic boundaries. Proceedings of the National Academy of Science USA 87 (5), 1816–1819.

Bernstein, J., 2001. The DP hypothesis: identifying clausal properties in the nominal domain. In: Baltin, M., Collins, C. (Eds.), The Handbook of Contemporary Syntactic Theory. Blackwell, Oxford, pp. 536–561.

Boeckx, C., Piattelli Palmarini, M., 2005. Language as a natural object; Linguistics as a natural science. The Linguistic Review 22 (2–3), 447–466.

Borer, H., 1984. Parametric Syntax. Foris, Dordrecht.

Boyd, R., Bogerhoff-Mulder, M., Durham, W.H., Richerson, P.J., 1997. Are cultural phylogenies possible? In: Weingart, P., Mitchell, S.D., Richerson, P.J., Maasen, S. (Eds.), Human by Nature. Between Biology and the Social Sciences. Lawrence Erlbaum Associates, Mahwah, NJ, pp. 355–386.

Cavalli Sforza, L.L., 2000. Genes, Peoples, and Languages. University of California Press, Berkeley.

Cavalli Sforza, L.L., Wang, W.S.Y., 1986. Spatial distance and lexical replacement. Language 62, 38–55.

Cavalli Sforza, L.L., Piazza, A., Menozzi, P., Mountain, J., 1988. Reconstruction of human evolution: bringing together genetic, archeological and linguistic data. Proceedings of the National Academy of Science USA 85, 6002–6006.

Cavalli Sforza, L.L., Menozzi, P., Piazza, A., 1994. The History and Geography of Human Genes. Princeton University Press, Princeton, NJ.

Chomsky, N., 1955. The Logical Structure of Linguistic Theory. Ms. (published in 1975, Plenum, New York).

Chomsky, N., 1957. Syntactic Structures. Mouton De Gruyter, The Hague.

Chomsky, N., 1981. Lectures on Government and Binding. Foris, Dordrecht.

Chomsky, N., 1986. Knowledge of Language. Praeger, New York.

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

Clackson, J., Forster, P., Renfrew, C. (Eds.), 2004. Phylogenetic Methods and the Prehistory of Languages. McDonald Institute for Archaeological Research, Cambridge, UK.

Clark, R., Roberts, I., 1993. A computational model of language learnability and language change. Linguistic Inquiry 24 (2), 299–345.

Crisma, P., 1997. L'articolo nella prosa inglese antica e la teoria degli articoli nulli. Ph.D. dissertation, Università di Padova.

Crisma, P., in press. Triggering syntactic change: the history of English genitives. In: Jonas, D., Whitman, J., Garrett, A. (Eds.), Grammatical Change: Origin, Nature, Outcomes. Oxford University Press, Oxford.

Dixon, R., 1998. The Rise and Fall of Languages. Cambridge University Press, Cambridge, UK.

Duffield, N., 1996. On structural invariance and lexical diversity in VSO languages: arguments from Irish noun phrases. In: Borsley, R., Roberts, I. (Eds.), The Syntax of the Celtic Languages: A Comparative Perspective. Cambridge University Press, Cambridge, UK, pp. 314–340.

Dunn, M., Terril, A., Reesink, G., Foley, R.A., Levinson, S.C., 2005. Structural phylogenetics and the reconstruction of ancient language history. Science 309 (5743), 2072–2075.

Dyen, I., Kruskal, J., Black, P., 1992. An Indoeuropean classification: a lexicostatistical experiment. Transactions of the American Philosophical Society 82 (5).

Embleton, S.M., 1986. Statistics in Historical Linguistics. Brockmeyer, Bochum.

Felsenstein, J., 2004a. Inferring Phylogenies. Sinauer, Sunderland, MA.

Felsenstein, J., 2004b. PHYLIP (Phylogeny Inference Package) version 3.6b. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.

Gianollo, C., 2005. Constituent structure and parametric resetting in the Latin DP: a diachronic study. PhD dissertation, Università di Pisa.

Goebl, H., 1982. Dialektometrie: Prinzipien und Methoden des Einsatzes der Numerischen Taxonomie im Bereich der Dialektgeographie. Österreichische Akademie der Wissenschaften, Wien.

Gray, R., Atkinson, Q., 2003. Language tree divergences support the Anatolian theory of Indo-European origin. Nature 426, 435–439.

Greenberg, J., 1963. Some universals of grammar with particular reference to the order of meaningful elements. In: Greenberg, J. (Ed.), Universals of Language. MIT Press, Cambridge, MA, pp. 73–113.

Greenberg, J., 1987. Language in the Americas. Stanford University Press, Stanford.

Greenberg, J., 2000. Indoeuropean and its Closest Relatives: The Eurasiatic Language Family. Stanford University Press, Stanford.

Guardiano, C., 2003. Struttura e storia del sintagma nominale nel Greco antico: ipotesi parametriche. PhD dissertation, Università di Pisa.

Guardiano, C., Longobardi, G., 2005. Parametric comparison and language taxonomy. In: Batllori, M., Hernanz, M.L., Picallo, C., Roca, F. (Eds.), Grammaticalization and Parametric Variation. Oxford University Press, Oxford, pp. 149–174.

Hamming, R.W., 1950. Error detecting and error correcting codes. Bell System Technical Journal 26 (2), 147–160.

Heggarty, P.A., 2000. Quantifying change over time in phonetics. In: Renfrew, C., McMahon, A., Trask, L. (Eds.), Time-Depth in Historical Linguistics-2. MacDonald Institute for Archaeological Research, Cambridge, pp. 531–562.

Humboldt, W. von, 1827. Über den Dualis. In: Leitzmann, A. (Ed.), Wilhelm von Humboldt, Gesammelte Schriften, im Auftrag der (Königlichen) Preußischen Akademia der Wissenschaften, VI.1, Behr, Berlin, 1907.

Humboldt, W. von, 1836. Über die Verschiedenheit des menschlichen Sprachbaues und ihren Einfluss auf die geistige Entwicklung des Menschengeschlechts. In: Buschmann, E. (Ed.), Gedruckt in der Druckerei der Königlichen Akademie der Wissenschaften, Berlin. [In: Leitzmann, A. (Ed.), Wilhelm von Humboldt, Gesammelte Schriften, im Auftrag der (Königlichen) Preußischen Akademia der Wissenschaften, VII.1, Behr, Berlin, 1907.]

Jaccard, P., 1901. Etude comparative de la distribution florale dans une portion des Alpes et des Jura. Bulletin de la Société Vaudoise des Sciences Naturelles 37, 547–579.

Joseph, B., Salmons, J.C. (Eds.), 1998. Nostratic, Sifting the Evidence. Benjamins, Amsterdam.

Keenan, E., 1994. Creating anaphors. An historical study of the English reflexive pronouns. Ms. UCLA.

Keenan, E., 2000. An historical explanation of some binding theoretic facts in English. Ms. UCLA.

Keenan, E., 2002. Explaining the creation of reflexive pronouns in English. In: Minkova, D., Stockwell, R. (Eds.), Studies in the History of the English Language. Mouton de Gruyter, Berlin, pp. 325–354.

Koyré, A., 1961. Du monde de l''à peu près' à l'univers de la précision. In: Etudes d'histoire de la pensée philosophique, A. Colin, Paris, pp. 311–329.

Lightfoot, D., 1991. How to Set Parameters. MIT Press, Cambridge, MA.

Lohr, M., 1998. Methods for the Genetic Classification of Languages. PhD dissertation, University of Cambridge, UK.

Longobardi, G., 2001a. Formal syntax, diachronic minimalism and etymology: the history of French *chez*. Linguistic Inquiry 32 (2), 275–302.

Longobardi, G., 2001b. The structure of DPs: some principles, parameters and problems. In: Baltin, M., Collins, C. (Eds.), The Handbook of Contemporary Syntactic Theory. Blackwell, Oxford, pp. 562–603.

Longobardi, G., 2003. Methods in parametric linguistics and cognitive history. Linguistic Variation Yearbook 3, 101–138.

Longobardi, G., 2005. A minimalist program for parametric linguistics? In: Broekhuis, H., Corver, N., Huybregts, M., Kleinhenz, U., Koster, J. (Eds.), Organizing Grammar: Linguistic Studies for Henk van Riemsdijk. Mouton de Gruyter, Berlin, pp. 407–414.

McMahon, A., 2005. Introduction. Transactions of the Philological Society 103 (2), 113–119.

McMahon, A., McMahon, R., 2003. Finding families: quantitative methods in language classifying. Transactions of the Philological Society 101 (1), 7–55.

McMahon, A., McMahon, R., 2005. Language Classification by Numbers. Oxford University Press, Oxford.

Morpurgo Davies, A., 1996. La linguistica dell'Ottocento. Il Mulino, Bologna.

Nerbonne, J., 2007. Review of McMahon, A., McMahon, R., Language Classification by Numbers. Oxford University Press, Oxford, 2005. Linguistic Typology 11, 425–436.

Nerbonne, J., Kretzschmar, W., 2003. Introducing computational methods in dialectometry. In: Nerbonne, J., Kretzschmar, W. (Eds.), Computational Methods in Dialectometry; special issue of Computers and the Humanities, vol. 37 (3). pp. 245–255.

Newmeyer, F.J., 2005. Possible and Probable Languages. A Generative Perspective on Linguistic Typology. Oxford University Press, Oxford.

Nichols, J., 1992. Linguistic Diversity in Space and Time. University of Chicago Press, Chicago.

Nichols, J., 1996. The comparative method as heuristic. In: Durie, M., Ross, M. (Eds.), The Comparative Method Reviewed: Regularity and Irregularity in Language Change. Oxford University Press, Oxford, pp. 39–71.

Niyogi, P., Berwick, R., 1996. A language learning model for finite parameter spaces. Cognition 61, 161–193.

Piattelli Palmarini, M., 1989. Evolution, selection and cognition: from 'learning' to parameter setting in biology and in the study of language. Cognition 31, 1–44.

Plank, F. (Ed.), 2003. Noun Phrase Structure in the Languages of Europe (Empirical Approaches to Language Typology, EUROTYP, 20-7). Mouton de Gruyter, Berlin.

Renfrew, C., 1987. Archaeology and Language: The Puzzle of Indo-European Origins. Jonathan Cape, London.

Rigon, G., forthcoming. A quantitative approach to the study of syntactic evolution. PhD dissertation, Università di Pisa.

Ringe, D., 1992. On calculating the factor of chance in language comparison. Transactions of the American Philosophical Society 82 (1), 1–110.

Ringe, D., 1996. The mathematics of Amerind. Diachronica 13, 135–154.

Ringe, D., Warnow, T., Taylor, A., 2002. Indo-European and computational cladistics. Transactions of the Philological Society 100 (1), 59–129.

Roberts, I., 1998. Review of Harris, A., Campbell, L., Historical Syntax in Cross-linguistic Perspective. Romance Philology 51, 363–370.

Roberts, I., 2004. Parametric comparison: Welsh, Semitic and the Anti-Babelic principle. Unpublished hand out, University of Cambridge.

Roberts, I., 2005. Principles and Parameters in a VSO Language. A Case Study in Welsh. Oxford University Press, Oxford.

Roberts, I., 2007. The Mystery of the Overlooked Discipline: Modern Syntactic Theory and Cognitive Science. Ms. University of Cambridge.

Rouveret, A., 1994. Syntaxe du gallois. Principes généraux et typologie. CNRS Editions, Paris.

Séguy, J., 1971. La relation entre la distance spatiale et la distance lexicale. Revue de Linguistique Romane 35, 335–357.

Spruit, M., 2008. Quantitative Perspectives on Syntactic Variation in Dutch dialects. LOT, Utrecht.

Thomason, S., Kaufman, T., 1988. Language Contact, Creolization, and Genetic Linguistics. University of California Press, Berkeley and Los Angeles.

Vennemann, T., 2002. Semitic > Celtic > English: the transitivity of language contact. In: Filppula, M., Klemola, J., Pitkänen, H. (Eds.), The Celtic Roots of English (Studies in Languages 37). Joensuu University Press, Joensuu, pp. 295–330.

Wang, W.S.Y., 1994. Glottochronology, Lexicostatistics, and Other Numerical Methods, Encyclopedia of Language and Linguistics. Pergamon Press, Oxford, pp. 1445–1450.

Warnow, T., Evans, S.N., Ringe, D., Nakhleh, L., 2004. Stochastic models of language evolution and an application to the Indo-European family of languages. In: Clackson, et al. (Eds.), pages not numbered. Download at http://www.stat.berkeley.edu/users/evans/659.pdf.

Watkins, C., 1976. Towards Proto-Indo-European Syntax: problems and pseudo-problems. In: Steever, S., Walker, C., Mufwene, S. (Eds.), Diachronic Syntax. Chicago Linguistic Society, Chicago, pp. 305–326 [Reprinted 1994 in: Lisi Oliver (Ed.), Calvert Watkins. Selected Writings. Innsbrucker Beiträge zur Sprachwissenschaft, Innsbruck, pp. 242–263].