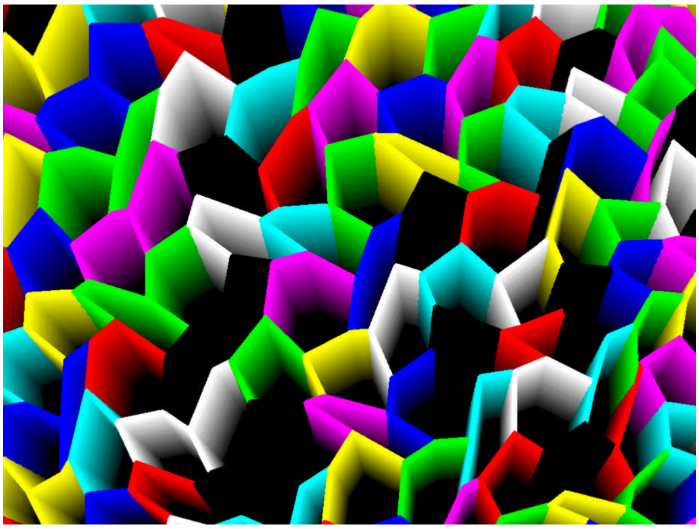
Constraints on language variation: In search of the Language Blueprint

(short name: The Language Blueprint)



Research Priority Area

Amsterdam Center for Language and Communication
Institute for Logic, Language and Computation

University of Amsterdam

Contents

Cover design: Labyrinth, Wolfgang W. Daeumler, Artmetic mono-algorithmic digital art (http://www.artmetic.de)

1. Introduction

Natural languages exhibit a tremendous amount of variation. This variation manifests itself in all aspects of the structure of languages, in the ways languages convey meaning, and in the ways they are used. Any adult confronted with an unfamiliar language will have great difficulty in acquiring that language, let alone understand its structure. Yet any infant anywhere in the world, irrespective of its genetic descent, will learn the language it is exposed to without even being aware of its structure. The human language faculty is tremendously flexible, and accepts a whole array of systems.

Notwithstanding this enormous variety, languages show a remarkable degree of similarity, which takes the form of a set of constraints on linguistic variation. Together this set of constraints defines the Language Blueprint: the basic layout of any system of human communication. The search for this blueprint is the major task of linguistics, which thus provides a window on the human mind. Finding the blueprint is also a major prerequisite for practical applications such as improving language teaching, knowledge base construction, language therapy, and speech recognition. These applications crucially hinge on knowledge of language systems.

The Amsterdam Center for Language and Communication (ACLC) and the Institute for Logic, Language, and Computation (ILLC) at the University of Amsterdam (UvA) apply a novel and integrated strategy in order to significantly increase our understanding of the nature of this blueprint. A key feature of the UvA approach is that constraints on linguistic variation are studied from the widest possible range of perspectives, both descriptive and theoretical, in order to ensure that the findings are not accidental, but are truly representative of the basic parameters that govern the organization of natural languages. The next sections elaborate on this approach. First the concept of constraints on linguistic variation will be illustrated, followed by an introduction on the types of variation. After discussing the ways in which variation can be modeled and explained, a number of examples are given.

2. Constraints

Constraints on linguistic variation can take different shapes. One of these is the absolute universal, which states, for instance, a property that all languages or language phases share. An example of such an absolute universal would be:

> No language in the world has a single word that means *not everything.*

This statement correctly predicts that languages do not have a word such as *nevrything.*

Constraints may also take the form of implicational hierarchies, which systematically define the range of variation allowed within a certain domain. A simple example of an implicational hierarchy is the following:

m ⊂ n ⊂ ñ

In general terms, a hierarchy predicts that if a feature at a certain position within the hierarchy is present, the features that are located to the left of it in that hierarchy will also be. Thus, the hierarchy above says, for instance, that if a child has acquired the *n* as a meaningful sound, it will also have acquired the *m* as a meaningful sound; and if it has acquired the *ñ* as a meaningful sound, it will also have acquired the *n* and the *m* as meaningful sounds. The other way round, if the child hasn't acquired the *m* as a meaningful sound, it will not have acquired the *n* and the *ñ* either; etcetera.

Yet another form constraints may take is that of a statistical generalization, which states a property that languages overwhelmingly, or in most of their phases, demonstrate. An example would be:

> If a language changes from having the verb in the middle position in declarative sentences to having it in final position, it generally changes from having prepositions to having postpositions as well.

Facts such as these are important as they raise the question how the few counterexamples to this general tendency may be explained.

3. Crosslinguistic variation

The general limitations on crosslinguistic variation are studied within the field of linguistic typology. While at a superficial, e.g. lexical, level the differences between languages are beyond comparison, the language system that is behind it allows for important broad generalizations. Constraints on crosslinguistic variation are uncovered by comparing languages in samples representative of the languages of the world. The bulk of research into constraints on linguistic variation has concentrated on phonology, morphology, and syntax, but more recently the areas of semantics, pragmatics, and the lexicon have also come to the fore. It is especially in the light of this development that collaboration between ACLC en ILLC adds significantly to the general research enterprise.

Normally typological samples are restricted to languages of the oral modality and exclude sign languages. This is a rather unfortunate situation, since, in order to find the

Language Blueprint, one wants to generalize across modalities and define constraints on linguistic variation at the highest level of abstraction. The exclusive focus on a particular modality may blur our understanding of the truly universal properties of language. There is no doubt that it is only at the highest level of generalization that we can find the principles that are responsible for the tremendous facility with which human beings communicate by means of language.

Just as there are large differences between the oral languages of the world, there are also enormous differences between the sign languages of the world. It is to be expected that the modality used affects the variation in oral languages on the one hand and in sign languages on the other. In our search for the Language Blueprint we have to exclude the modality-specific restrictions in order to arrive at the truly universal features of language.

4. Variation in time

Language systems change over time, either as the result of internal pressure within the language system, or due to external pressure, i.e. contact with other languages. The former situation occurs when, for instance, the erosion of a grammatical element in a language triggers the introduction of a new element. The latter situation occurs when e.g. a language adapts its word order pattern to that of a dominant co-extensive or neighbouring language. In both cases any change within an existing language system is expected to lead to a new system that is compatible with the Language Blueprint. Diachrony should therefore mirror typology, in the sense that language change occurs along the lines of the same constraints as those describing crosslinguistic variation.

Language change may also be more abrupt. This happens when groups of speakers with different linguistic backgrounds unfamiliar with the languages of the other groups are forced to communicate among each other. In these circumstances pidgins develop, which in the course of time may grow into full-fledged creole languages. Again, the development from pidgin to creole should be consistent with the Language Blueprint and the constraints on which it is based.

5. Variation in acquisition

When a young child acquires a language, the learning process passes through numerous intermediate stages. No matter the degree of complexity of that stage, the language system of the child at that particular moment in time should be an instance of a possible language. Each of these intermediate stages should therefore represent a system that is consistent with the

Language Blueprint, and every change in the system as a result of further acquisition should lead to a system that is again compatible with that blueprint.

The acquisition of a second language, whether from birth or at a later stage, poses further challenges to research in linguistic variation. Here again the different intermediate phases that the second language learner passes through should be compatible with the overall constraints on linguistic variation. Interference from the first language adds a level of complexity here that is interesting to study from a crosslinguistic acquisitional perspective.

6. Modeling

Once constraints on linguistic variation have been uncovered, they can be integrated into a model of language. By systematically integrating constraints on linguistic variation that have been detected in previous linguistic research into consistent and rigorously formalized models of language, powerful tools are developed that show the interaction between these constraints, while at the same time generating significant new hypotheses regarding further interactions.

Existing formal models differ among themselves depending on the kind of constraint that they assume to be the basic determinant of linguistic structure. Formal models of language start from the assumption that language is primarily an autonomous system. Functional models of language take the communicative instrumentality of language as their point of departure. In the latter case the interaction between a linguistic expression and its context of interpretation becomes especially relevant. Examples of models employed at the UvA are Generative Grammar, Functional Discourse Grammar, Dynamic Semantics, and Game Theory. Several of these models employed actually had the University of Amsterdam as their birthplace.

Given the differences in their basic assumptions, formal and functional approaches to linguistic modeling provide complementary contributions to the search for the Language Blueprint. The integration of these contributions is facilitated by the fact that over the last decade there has been a convergence in research methods across frameworks, in that the study of linguistic variation, the starting point of this research programme, has become common to all. Joining forces has synergetic effects that, as we know from other domains of research, will strongly enhance the results of the research.

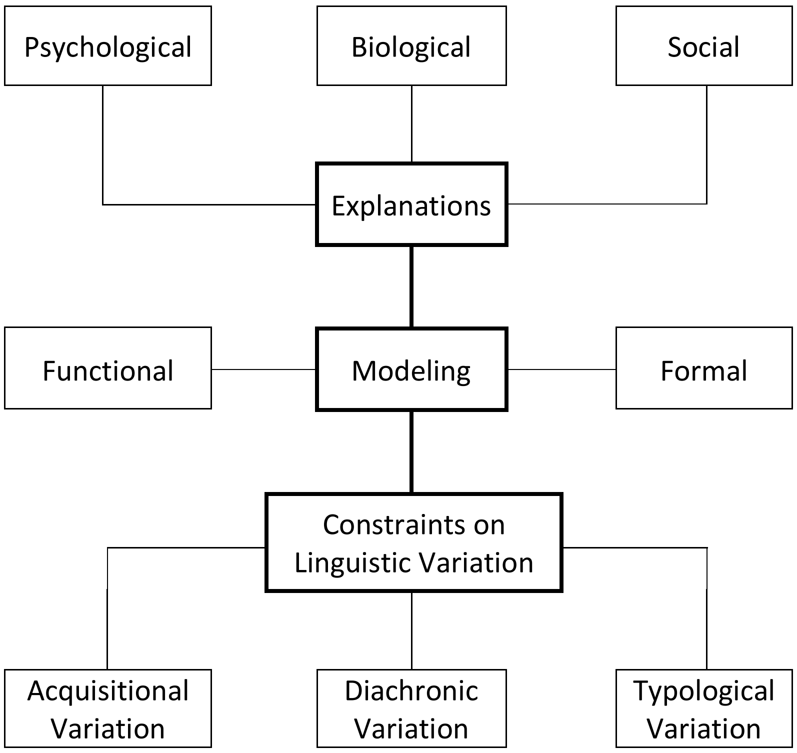
7. Explanations

An important general property of constraints on linguistic variation is that, all other things being equal, they predict which features are less likely to occur in language systems. This increasing markedness of features needs an explanation. Explanations given to constraints on linguistic variation may be (i) psychological (e.g. cognitive development, iconicity), (ii) biological (e.g. spoken versus signed modality, processing speed), or (iii) social (community size, multilingualism). The first type of explanation is the reason that linguistics is solidly grounded in the cognitive sciences. The second type of explanation links linguistic research to the neurosciences and anatomy, and the third to sociology.

8. In search of the Language Blueprint

Taking the notion of constraint as the unifying concept and as the basic building block of the search for the Language Blueprint, UvA researchers are in a position to engage in collaborative research enterprises, in which individual (sets of) constraints are studied from the various perspectives described in the previous sections and summarized in Figure 1.

Figure 1. *The Language Blueprint programme*



By applying this research strategy, assumptions that a hypothesized constraint is indeed part of the Language Blueprint can be tested much more thoroughly than in any other strategy. The strategy furthermore ensures cooperation between research groups that normally operate within their own research tradition, and thus stimulates the exchange of scientific results, so as to give them a broader significance. The UvA is one of the few places in the world in which such a strategy can be applied successfully, given the presence of established scholars and promising young researchers in all of the fields mentioned above, both formal and functional, and concerning both the language system and language use.

9. Some specific examples

In this section we illustrate the program by means of four examples. The first two are supposed to be accessible to non-specialists; the latter two are unavoidably of a more specialist nature.

9.1. Lexical categories

It has long been assumed that all languages have at least verbs and nouns. New language data have cast doubt on this assumed universal. These data suggest the following implicational hierarchy:

$$V \subset N \subset Adj \subset Adv$$

This hierarchy should be read as follows: If a language has an opposition between adverbs and adjectives, it also has an opposition between adjectives and nouns (N) and between nouns and verbs (V). If a language has an opposition between adjectives and nouns, it also has an opposition between nouns and verbs, etc. The hierarchy shows that nouns and verbs are the most central lexical categories, but that languages may also survive with verbs only.

Explanations for this hierarchy may for instance refer to the acts of reference and predication (a communicative explanation), or to the embedded (syntactic) positions of adjectives and adverbs (a cognitive explanation). These different explanations may lead to different theories.

Whatever the explanations and the theory, the cross-linguistically based implicational hierarchy, if it is of sufficient generality to qualify as part of the Language Blueprint, should be equally capable of handling other types of variation. Given that the more to the right a category/opposition is positioned in the hierarchy, the less likely it is to occur, and vice versa, we might, formulate the following hypotheses: (i) variation in time: parts-of-speech oppositions (if not available within a language at a certain stage of its development) are introduced into a language from left to right in the course of time. This has in fact been hypothesized for the history of the Indo-European languages; (ii) variation in time: if a new language emerges, for instance a creole language, parts-of-speech oppositions will be introduced from left to right; (iii) acquisitional variation: parts-of-speech oppositions more to the left on the hierarchy are acquired earlier than those more to the right. In the earliest stages infants do not seem to distinguish parts of speech at all; (iv) attritional variation: in language attrition parts-of-speech oppositions more to the right disappear earlier from the language system than those more to the left.

9.2. Pronouns

All known languages employ a device for pronominal reference, pronouns that refer to things introduced earlier in a sentence, text, or earlier in a discourse. In English we find the well-known pronouns *he* and *she*, but we can also detect pronominal elements like *then*, *now*, and *current* and *local*. These pronominal elements are studied from many different perspectives. However, their syntactic, semantic and pragmatic properties remain controversial and deserve further research. Pronouns can illustrate very well how all the parts of Figure 1 are interconnected.

There is a large amount of variation on all levels in the field of pronominal reference. On the typological level, different languages have different kinds of pronouns: for instance, some languages have possessive reflexives (*self's book*, Russian) and some do not (*own book*, English), and some languages do not have reflexive pronouns at all (*He saw him* meaning 'He saw himself', Frisian). Different rules connected to pronouns are acquired at different stages by both L1 and L2 learners. Pronouns develop in time and can change functions, for instance, a reflexive pronoun can develop from an emphatic one. It appears, however, that this type of variation is severely constrained. One obvious generalization is that reflexive and non-reflexive pronouns are usually in complementary distribution (*He saw himself* vs. *He saw him*). Therefore different models have been proposed to formalize these constraints. Constraints on these variations have also been given very different explanations: either as innate principles, or by means of psychological and functional factors such as economy and unambiguous communication. In the former approach, for instance, one principle says that a reflexive must be bound, while a pronominal must be free in some domain, thus accounting for the complementary distribution. An approach along the latter lines claims that reflexives and non-reflexives are not interchangeable because it would create unnecessary ambiguity and ineffective communication. These approaches may overlap.

When theories are well-formulated, they can be tested on the constantly growing pool of acquisitonal, typological and diachronic data. For instance, the two main theories of discourse anaphoric reference can be compared and give us more insight in general linguistic properties once they are applied to sign languages, which provide a clear perspective on properties of pronominal reference because referential indices are overtly expressed in these types of languages. A typological, cross-linguistic, and multi-modal study allows us to uncover possibly universal patterns of the fundamentally linguistic phenomenon of pronominal anaphora.

9.3. Epistemic indefinites

Typology Epistemic indefinites are indefinites that signal that the speaker is unable to identify the individual that is referred to: in (1) there is a student who called, the speaker just does not know who. In some languages, such epistemic indefinites can also be used in negative sentences to express that there is no referent in the first place: in (2) no question has been asked at all. Finally, sometimes epistemic indefinites can also be used to convey an emphatic, so-called *free choice*, meaning: in (3) just any doctor would be a suitable husband for Mary.

(1) Irgendein Student hat angerufen. (#Rat mal wer?)
Some student has called (guess prt who)
`Some student called, I don't know who.'

(2) Niemand hat irgendeine Frage beantwortet.
Nobody has some question answered
`Nobody answered any question (at all).'

(3) Maria muss irgendeinen Arzt heiraten.
Mary must some doctor marry
`Mary must marry a doctor, any doctor is a permissible option.'

The following table illustrates the variety of epistemic indefinites cross-linguistically:

(4)

	Ignorance	**Negation**	**Free choice**
irgendein (German)	yes	yes	yes
algún (Spanish)	yes	yes	no
-*si* (Czech)	yes	no	no

It is tempting to read (4) as an implicational map and formulate a hypothesis of function contiguity (Haspelmath 1997): the possible usages of an indefinite in any language will always express a contiguous area of the map. The following would thus be an example of an impossible distribution:

(5)

	Ignorance	**Negation**	**Free choice**
#	yes	no	yes

The validity of this hypothesis is still a matter of empirical investigation. In particular, it would be interesting to see if it extends to non-Indo-European languages.

Modeling Aloni and Port (2011) and Aloni (2012) proposed to model epistemic indefinites as existential quantifiers triggering an obligatory shift in the current domain of quantification. In this modeling, differences between different indefinite forms are captured in terms of different domain shifts they can induce. One kind of domain shift (conceptual cover shift) will produce ignorance uses and will be available for all epistemic indefinites. Another kind of domain shift (domain widening), yielding negative uses, will be an option only for a subset of the epistemic indefinites. Emphatic free choice uses are then explained as obligatory pragmatic enrichments triggered by domain widening under certain circumstances. This modeling predicts the impossibility of (5): emphatic free choice uses presuppose the same mechanism which generates negative uses, namely domain widening, so whenever an emphatic free choice use is possible for an epistemic indefinite, a negative use is also allowed.

Acquisitional and diachronic variation This model gives rise to a number of testable predictions with respect to the acquisition and the diachronic development of epistemic indefinites. For example it predicts that with respect to an epistemic indefinite exhibiting all three usages, e.g. German *irgend*-series, its emphatic free choice usage will be acquired or it emerged only after its usage in negative contexts has been acquired/emerged. As for the diachronic perspective, this prediction has been confirmed by the historical corpus study reported in Port (2012). As for acquisition, the prediction will be checked by consulting a German child language corpus.

Explanation According to the described model, emphatic free choice uses of epistemic indefinites clearly come with a very high cost for the interpreter who in order to arrive at the intended interpretation needs to calculate all kinds of pragmatic implicatures and integrate them in the conveyed meaning. Hearer economy then explains why using an epistemic indefinite for free choice is relatively rare for languages and occurs at the right end of the implicational map. It also explains why many languages develop specialized morphology to express free choice. In Spanish, for instance, the availability of a specialised free choice item (*cualquiera*) is used for expressing free choice, which the regular epistemic indefinite *algún* cannot express.

References

Aloni, M. (2012). On epistemic indefinites: a note on emphatic free choice uses. In *Proceedings of Sinn und Bedeutung* 16 .

Aloni, M. and Port, A. (2011). Epistemic indefinites crosslinguistically. In *Proceedings of NELS 41*.

Haspelmath, M. (1997). *Indefinite pronouns*. Oxford University Press, Oxford.

Port, A. (2012). The diachronic development of German *irgend*-indefinites. ms.

9.4. Verbs and agreement

A well-known observation is that in all Indo-European languages (as well as a number of other languages) the finite verb is placed more to the left in a clause if and only if the language has rich agreement morphology. If it does not have such morphology, the verb is place after sentential adverbs. Take the following two examples from English and French:

(1) a. John often talks with Mary
b. Jean parle souvent avec Marie
Jean talks often with Marie
'Jean often talks with Marie.'

English is a poorly inflected language (it only has an overt subject agreement marker in the third person singular: *talk-s*) and it places the verb to the right of the adverb often. French, on the other hand is more richly inflected (think of nous parl-ons ('we speak'), vouz parl-ez ('you speak'), etc.) and places the verb to the left of the adverb *souvent* 'often'. That the position of the finite verb is conditioned by rich subject-verb agreement has long functioned as an indication for a tight connection between syntax and morphology (Kosmeijer 1986, Holmberg & Platzack 1991, 1995, Roberts 1993, Rohrbacher 1994, Vikner 1995, Bobaljik 1995; Bobaljik & Thráinsson 1998 and others), suggesting that morphology drives syntax.

In more recent years, this generalization has been disputed on both empirical and theoretical grounds. Empirically, some data have been put forward that seem to suggest the existence of language varieties that are poorly inflected but still behave as if they were rich with respect to verb placement (e.g. Colloquial French), as well as varieties that show the mirror image (see Bentzen et al. 2007 for some Icelandic dialects). Theoretically, it is nowadays widely assumed that morphology takes place after syntax, i.e. insertion of agreement markers takes place when the syntactic structure, and hence word order, is finished. Consequently, morphology can no longer drive syntactic operations.

Recently, it has been shown that all of the empirical arguments provided against the relation between richness of inflection and position of the finite verb are seriously flawed (cf. Thráinsson 2009, Koeneman & Zeijlstra 2010) and that, at least for all languages that have been studied in this respect so far, the relation between morphological richness and syntactic verb placement is exceptionless.

Moreover, one can observe that the notion of richness that is required to correctly describe the variation in verb placement shows a striking parallel with the poorest pronoun inventories attested in human languages. An example of such a minimal system is shown in (2):

(2) Kuman

<table>
<tr><th></th><th>SG</th><th>PL</th></tr>
<tr><td>1st</td><td><i>na</i></td><td><i>no</i></td></tr>
<tr><td>2nd</td><td colspan="2"><i>ene</i></td></tr>
<tr><td>3rd</td><td colspan="2"><i>ye</i></td></tr>
</table>

Although Kuman only has four pronouns, it minimally distinguishes (i) between 1st and 2nd person (ii) between 2nd and 3rd person and (iii) between singular and plural (albeit only in the 1st person) (cf. Greenberg 1963, universal 42). If a language shows the same minimal distinctions in the subject agreement paradigm, that language will behave like French (i.e. as in 1b). If not, it behaves as English (i.e. as in 1a). This suggests that there is a tight connection between rich agreement and subjecthood.

The generalization holds in a bi-directional way: if and only if a language variety has rich agreement does it place the finite verb between the subject and sentential adverbs. The typological consequence is that out of four possible language types, only two are realized:

(3)

	+SU-V(finite)-ADV	***SU- ADV- V(finite)-***
+Rich	French, Icelandic, Yiddish, Italian	
-Rich		English, Norwegian, Swedish, Faroese

This creates a theoretical paradox: although morphology cannot drive syntax, there is still a 1:1 relation between the richness of morphology and the application of syntactic operation. However, this paradox can be resolved if it is assumed that overt morphological distinctions function as a cue for language learners to determine what formal features their target grammar contains and that these features in turn are the drivers of syntactic operations. Without rich morphology language learners can never acquire the feature that would make the verb appear to the left of adverbs; with rich morphology language learners must acquire such a feature. Hence, morphology *does* drive syntax but only indirectly, via acquisition. Under this account, the absence of languages with rich morphology and without verbal movement, and of languages without rich morphology but with verbal movement, follows from acquisition. Such languages are simply unlearnable.

This account therefore reduces a universal correlation between syntax and morphology to the application of a single language learning procedure. Therefore, it makes strong predictions in exactly the three domains under investigation, not only for language variation (as we have seen) but also for language acquisition and language change.

In the course of time, languages tend to weaken their morphological markers. Consequently, if the loss of verbal morphology renders a language's agreement system from rich to poor, it is expected to place its verb in a lower position. This is indeed the case. Danish, Swedish and Norwegian, for instance, used to be languages with rich verbal agreement and after their verbal morphology became poor, they started to place verbs more to the right, following adverbs.

Finally, this type of research makes strong testable predictions for language acquisition: since determining the exact position of the finite verb is dependent on the richness of verbal morphology, a child will only put the verb in the correct position of it has also acquired the verbal inflectional paradigm of its target language. Research indicates that this is indeed the case (cf. Meisel 1994), but more acquisitional research is needed before substantive claims can be made.

References

Bentzen, Kristine, Gunnar Hrafn Hrafnbjargarson, Þorbjörg Hróarsdóttir & Anna-Lena Wiklund. 2007. Rethinking Scandinivian verb movement *Journal of Comparative Germanic Linguistics*, 10: 203-33.

Bobaljik, Jonathan. 1995. *Morphosyntax: The Syntax of Verbal Inflection*. PhD. dissertation, MIT.

Greenberg, Joseph. 1963. Some universals of grammar with particular reference to the order of meaningful elements, in *Universals of grammar*, ed. by Joseph H. Greenberg, 2nd edition, 73-113. Cambridge, MA: The MIT Press.

Holmberg, Anders & Christer Platzack. 1991. On the role of inflection in Scandinavian syntax, in *Issues in Germanic syntax,* ed. by Werner Abraham, Wim Kosmeijer & Eric Reuland. Berlin/New York: Mouton de Gruyter.

Koeneman, Olaf & Hedde Zeijlstra. 2010. Resurrecting the Rich Agreement Hypothesis: weak isn't strong enough, appear in Movement in minimalism, Proceedings of the 12th Seoul Conference on Generative Grammar, ed. by An, Duk-Ho & Soo-Yeon Kim.

Kosmeijer, Wim. 1986. The status of the finite inflection in Icelandic and Swedish, *Working Papers in Scandinavian Syntax* 26,1-41.

Meisel, Jürgen. 1990. Grammatical development in the simultaneous acquisition of two first languages, in *Two first languages: Early grammatical development in bilingual children*, ed. by Jürgen

Roberts, Ian. 1993. *Verbs and Diachronic Syntax*, Kluwer: Dordrecht.

Rohrbacher, Bernhard. 1994. *The Germanic languages and the full paradigm*. PhD. dissertation, University of Massachusetts.

Thráinsson, Höskuldur. 1996. On the (non)-universality of functional projections, in *Minimal Ideas: Syntactic Studies in the Minimalist Framework*, ed. by Werner Abraham, Samuel David Epstein, Höskuldur Thráinsson & Jan-Wouter Zwart, 253-281. Amsterdam/Philadelphia: John Benjamins.