Learning Functional Prepositions

John Stewart

Graduate Center, City University of New York

Recommended Citation

This Dissertation is brought to you by CUNY Academic Works. It has been accepted for inclusion in All Dissertations, Theses, and Capstone Projects by an authorized administrator of CUNY Academic Works. For more information, please contact deposit@gc.cuny.edu.
LEARNING FUNCTIONAL PREPOSITIONS

by

JOHN STEWART

A dissertation submitted to the Graduate Faculty in Linguistics in partial fulfillment of the requirements for the degree of Doctor of Philosophy,

The City University of New York

2015
This manuscript has been read and accepted for the Graduate Faculty in Linguistics to satisfy the dissertation requirement for the degree of Doctor of Philosophy.

Virginia Valian

___________________________________________

Date Chair of Examining Committee

Gita Martohardjono

___________________________________________

Date Executive Officer

Marcel Den Dikken

___________________________________________

Martin Chodorow

___________________________________________

Charles Yang

___________________________________________

Supervisory Committee

THE CITY UNIVERSITY OF NEW YORK
Abstract

LEARNING FUNCTIONAL PREPOSITIONS

by

John Stewart

Adviser: Distinguished Professor Virginia Valian

In first language acquisition, what does it mean for a grammatical category to have been acquired, and what are the mechanisms by which children learn functional categories in general? In the context of prepositions (Ps), if the lexical/functional divide cuts through the P category, as has been suggested in the theoretical literature, then constructivist accounts of language acquisition would predict that children develop adult-like competence with the more abstract units, functional Ps, at a slower rate compared to their acquisition of lexical Ps. Nativists instead assume that the features of functional P are made available by Universal Grammar (UG), and are mapped as quickly, if not faster, than the semantic features of their lexical counterparts. Conversely, if Ps are either all lexical or all functional, on both accounts of acquisition we should observe few differences in learning.

Three empirical studies of the development of P were conducted via computer analysis of the English and Spanish sub-corpora of the CHILDES database. Study 1 analyzed errors in child usage of Ps, finding almost no errors in commission in either language, but that the English learners lag in their production of functional Ps relative to lexical Ps. That no such delay was found in the Spanish data suggests that the English pattern is not universal. Studies 2 and 3 applied novel measures of
phrasal (P head + nominal complement) productivity to the data. Study 2 examined prepositional phrases (PPs) whose head-complement pairs appeared in both child and adult speech, while Study 3 considered PPs produced by children that never occurred in adult speech. In both studies the productivity of Ps for English children developed faster than that of lexical Ps. In Spanish there were few differences, suggesting that children had already mastered both orders of Ps early in acquisition. These empirical results suggest that at least in English P is indeed a split category, and that children acquire the syntax of the functional subset very quickly, committing almost no errors. The UG position is thus supported.

Next, the dissertation explores a “soft nativist” acquisition strategy that composes the distributional analysis of input, minimal a priori knowledge of the possible co-occurrence of morphosyntactic features associated with functional elements, and linguistic knowledge that is presumably acquired via the experience of pragmatic, communicative situations. The output of the analysis consists in a mapping of morphemes to the feature bundles of nominative pronouns for English and Spanish, plus specific claims about the sort of knowledge required from experience.

The acquisition model is then extended to adpositions, to examine what, if anything, distributional analysis can tell us about the functional sequences of PPs. The results confirm the theoretical position according to which spatiotemporal Ps are lexical in character, rooting their own extended projections, and that functional Ps express an aspeuctual sequence in the functional superstructure of the PP.
To the memory of my mothers, Julia and Joan, who valued education above all else.
L'ho creata dal fondo di tutte le cose
che mi sono più care, e non riesco a comprenderla.
– Cesare Pavese, *Incontro*

For if I know that ten is more than three, and then someone were to say: “No, on the contrary, three is more than ten, as is proved by my turning this stick into a snake” – and if he were to do just that and I were to see him do it, I would not doubt my knowledge because of his feat. I would merely wonder how he could do such a thing.

– Al-Ghazali, *Deliverance from Error*
Acknowledgments

I first met Virginia Valian when she gave a guest lecture on language acquisition during the introductory course of the Linguistics program. I was struck by the clarity of her exposition and the iron logic of her central argument – that a resolutely anti-nativist stance entails a universal claim, while soft nativism makes merely existential claims – and felt that, somehow, this really matters. She once told me that language acquisition, as an academic field, has everything in it: philosophy, psychology, language, statistics, sociology. Perhaps that is why I soon resolved to study acquisition, specifically, through large-scale corpus analysis, under Virginia's direction.

This dissertation would never have come into being if it had not been for Virginia's intellectual guidance, her encouragement, her unvarnished and quite direct critique, and I dare say the stoic patience she exhibited faced with the vagaries of a part-time student. It is with the deepest gratitude that I thank her.

Marcel Den Dikken co-taught that first course, and I soon came to appreciate his humor and the thoroughness of his instruction. Those years of Syntax I and II, morphology, and the advanced syntax seminars, have in retrospect an almost orpheic quality for which I feel pointed nostalgia. Marcel took us on a voyage into the bowels of the language system itself, and it was wonderful. Throughout, his positive, funny, supportive attitude helped us when faced with the inevitable moments of self-doubt and exhaustion. Marcel is a terrific teacher and I will always feel very grateful for what I learned from him. His thorough and helpful comments to dissertation drafts, as well as his positive feedback, were invaluable in propelling the work forward.

I am grateful to Martin Chodorow for his course in statistical natural language processing, through
which I acquired, and gained confidence in using, the skills required for large-scale corpus analysis; 
for heading up the committee for my second qualifier; and of course for his contributions to this 
dissertation, from the earliest ideational stages, which imprinted the final form of the statistical 
analyses and the computational model.

Special thanks must go also to Charles Yang for agreeing to act as an external reader and for 
participating in the doctoral defense. I have long admired his work from afar and feel honored by 
his attention.

I would like to thank Bill McClure for pinch-hitting for my second qualifier and for the terrific 
semantics course sequence. My gratitude extends to the other Linguistics faculty members whose 
courses I took, Ricardo Otheguy and Diane Bradley, and also downstairs to the Education 
Psychology program, where I expended nine credits on crucial courses in statistics taught by Jay 
Verkuilen and David Rindskopf.

I owe unpayable debts to Gita Martohardjono, for enabling me to switch from the MA to the 
doctoral program, and for smoothing out the many collisions between life's knots and those of the 
institution. Also of course to Nishi Bissoondial for her guidance through all key institutional 
transitions.

Above all, it is to my wife, Laura Kraber, that I owe everything – not just this work, but my life itself. 
By shouldering more than a single human should, Laura gave the gift of time, without which none of 
this could have unfolded. I pray I can prove my love in return.
## Table of Contents

Abstract................................................................................................................................................................. iii

Acknowledgments...................................................................................................................................................... vii

Introduction............................................................................................................................................................ 1

Leibniz, Empiricism, and Prepositions..................................................................................................................... 1

Summary of the Argument........................................................................................................................................ 7

Structure of the Thesis............................................................................................................................................. 11

Part I. Theoretical Notions...................................................................................................................................... 14

Chapter 1. Adpositions and Features.................................................................................................................... 14

Introduction............................................................................................................................................................ 14

General Notions.................................................................................................................................................... 15

Adpositions............................................................................................................................................................ 15

Features................................................................................................................................................................. 20

Feature Formalisms............................................................................................................................................... 21

Types of Features.................................................................................................................................................. 26

Montague................................................................................................................................................................. 30

Abney (1987)......................................................................................................................................................... 34

Van Riemsdijk (1990).......................................................................................................................................... 36

Grimshaw (1991, 2000)....................................................................................................................................... 38
<table>
<thead>
<tr>
<th>Chapter 6. Learning Functional Prepositions</th>
<th>241</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td>241</td>
</tr>
<tr>
<td>Methods</td>
<td>242</td>
</tr>
<tr>
<td>Analysis</td>
<td>243</td>
</tr>
<tr>
<td>English Prepositions</td>
<td>243</td>
</tr>
<tr>
<td>Spanish Prepositions</td>
<td>252</td>
</tr>
<tr>
<td>The Acquisition of Functional Ps</td>
<td>259</td>
</tr>
<tr>
<td>Conclusion</td>
<td>264</td>
</tr>
</tbody>
</table>

| Chapter 7. Conclusion                      | 269 |
| The functional sequence of P               | 269 |
The acquisition of functional categories...........................................................................................274

Speculative Postscript.......................................................................................................................279

Appendix.................................................................................................................................................281

Source Code...........................................................................................................................................281

Prolog program for learning English and Spanish pronouns............................................................281

Prolog program for learning English and Spanish functional Ps......................................................286

Bibliography...........................................................................................................................................290
List of Tables

Table 1: Properties of functional and lexical categories ................................................................. 40
Table 2: Types of categories in (Zwarts, 1997) .............................................................................. 42
Table 3: The restrictions on syncretism in Classical Armenian, from (Caha, 2012) ....................... 65
Table 4: Description of English CHILDES Corpora .................................................................... 127
Table 5: Description of Spanish CHILDES Corpora ............................................................... 127
Table 6: PP Counts by language ..................................................................................................... 131
Table 7: P counts by MLU group (English) ................................................................................... 131
Table 8: P counts by MLU group (Spanish) ................................................................................ 132
Table 9: Classification of English and Spanish Ps ......................................................................... 139
Table 10: Transcriber-Marked P Substitutions ............................................................................. 150
Table 11: Errors in Sample of PPs ................................................................................................. 151
Table 12: Sequenced P errors ......................................................................................................... 152
Table 13: Sample Novel PPs ......................................................................................................... 178
Table 14: Top 10 novel child PPs .................................................................................................. 179
Table 15: Aggregate counts of novel PP tokens ........................................................................... 183
Table 16: Novel P+head tokens by MLU band (English) ................................................................. 184
Table 17: Novel P+head tokens by MLU band (Spanish) ............................................................... 184
Table 18: Proportions of Novel PPs (English) ............................................................................... 186
Table 19: Proportions of Novel PPs (English; Revised) ................................................................. 189

Table 20: Proportions of Novel PPs (Spanish) .................................................................................. 191

Table 21: Proportions of Novel PPs (English, Manchester corpus) .................................................... 192

Table 22: Classification of candidate feature matrices ........................................................................ 207

Table 23: Counts of English subject pronouns .................................................................................. 214

Table 24: Features of English subject pronouns .................................................................................. 217

Table 25: Table of Spanish subject pronouns ..................................................................................... 218

Table 26: English experientially-derived feature assignments ............................................................. 228

Table 27: Spanish experientially-derived feature assignments ............................................................. 228

Table 28: Surface feature assignments to English pronouns ................................................................. 235

Table 29: P Counts in CHILDES child-directed speech (English) ......................................................... 247

Table 30: Differences in ranks of vector dimensions for "over" and "on" .............................................. 253

Table 31: LEX-RATE scores for pairs of P vectors .............................................................................. 255

Table 32: Frequency of Ps in Spanish Google Books corpus ................................................................ 256

Table 33: Features of Spanish spatial P (compiled from Romeu, 2014) .............................................. 257
List of Figures

Figure 1: Feature geometry of Russian cases........................................................................................................24
Figure 2: Feature geometry of phi features........................................................................................................24
Figure 3: Example of polytree....................................................................................................................................26
Figure 4: Structure of locative and directional PP in (Den Dikken, 2010)..........................................................52
Figure 5: Relative frequencies of prepositions selected by "research" over time.................................................68
Figure 6: Bayesian network for Bannard et al's UB-PCFG (Bannard et al., 2009, p.17285).........................116
Figure 7: Distribution of sessions over child MLU (English)..............................................................................127
Figure 8: Distribution of sessions over child MLU (Spanish)............................................................................128
Figure 9: Child emissions of selected Ps relative to adult (English).................................................................136
Figure 10: Child emissions of selected Ps relative to adult (Spanish)...............................................................137
Figure 11: Relative frequencies of lexical and functional Ps by MLU band (English children).............139
Figure 12: Relative frequencies of lexical and functional Ps by MLU band (Spanish children)..............140
Figure 13: Relative frequencies of lexical and functional Ps by MLU band (Spanish adults)................141
Figure 14: Relative frequencies of lexical and functional Ps by MLU band (English adults)..................142
Figure 15: Adjusted KL distance by MLU group (English)............................................................................169
Figure 16: Adjusted KL distance by MLU group (Spanish).........................................................................171
Figure 17: Path diagram of mixed-effects logistic regression.........................................................................181
Figure 18: Proportion of novel PPs for varying sample sizes.............................................................................187
Figure 19: Dendogram for theory-based assignments ................................................................. 204

Figure 20: Dendograms for distribution vectors ........................................................................ 205

Figure 21: Dendogram of English subject pronoun distributions ............................................ 213

Figure 22: Dendogram of Spanish subject pronoun distributions ........................................... 215

Figure 23: Dendogram of English pronoun features (H&R) .................................................. 218

Figure 24: Dendogram of Spanish pronoun features (H&R) .................................................. 219

Figure 25: Assignment of abstract features to morphs ........................................................... 224

Figure 26: Feature bundles after assignment of abstract features ......................................... 226

Figure 27: Dendogram for surface features of English pronouns ......................................... 239

Figure 28: Dendogram of English P distributions ................................................................. 247

Figure 29: Dendogram of Spanish P distributions ................................................................. 257

Figure 30: Dendogram of Spanish spatial P theoretical (Romeu) feature vectors .................. 260

Figure 31: Dendogram of English P distributions ................................................................. 275
Introduction

Leibniz, Empiricism, and Prepositions

The centuries-old philosophical debate between rationalists and empiricists on the nature of mental representations finds early expression in the contrast Leibniz marked between his theory of language and that of Locke (*Nouveaux essays sur l'entendement humain*, “New Essays on Human Understanding”, 1704). As we shall see, the subject of prepositions played a central role in that contrast.

Leibniz, the rationalist, concurs with his empiricist counterpart that the final cause of language lies in the expression of human sociality, of communication. But this final cause is not sufficient for language: though God endowed humans with the physiological tools for articulated sound, that ability alone does not fully explain the phenomenon. After all, other animals also vocalize, yet they do not thereby speak: “they lack something invisible” (Leibniz, 1921, p. 222; the oddly literal translations from French are mine). That invisible thing is of course reason.

Once formed, language also enables man to reason for himself, as much because of the means words give man to recall abstract thoughts for himself, as because of the utility of [written] characters ( *caractères*) and deaf thoughts ( *pensees sourdes*) for reason; for it would take too long if one had to explain everything and always substitute definitions for terms (p. 223).

Thought depends on words and symbols for the basic mechanisms of reason: abstraction and recollection. Language makes reason manifest and enables its functioning. Since reason is also the basis of language (that invisible something), language and reason are mutually enabling faculties.
The phrase *pensees sourdes* “deaf thoughts” is somewhat puzzling. Remnant and Bennett translate it as “blind thoughts”, as Leibniz himself states he has so translated the Latin phrase *cognitiones caecae*. Yet in repairing Leibniz’s apparent error the translators might mask an intent. It is not that the thoughts are silent or invisible; the issue is not whether they can be perceived. Rather, it is whether the thoughts themselves can perceive. Deaf thoughts cannot hear. In the context of a dialogue with empiricist notions of language centered on vocal communication, a “deaf thought” – a thought that cannot perceive vocal messages – is an autarkical thought that carries on, uncaring of communication, for other ends.

A *pensee sourde* must also not hear itself. According to Micheletti (2010) the phrase was something of a cliché among the Cartesians – another reason for Leibniz to employ it rather than the direct translation from Latin – for whom it bore the negative connotations of “clandestine” thoughts that are detached from ideational content and not transparently available to consciousness. To the Cartesians the possibility of such thinking represented a significant philosophical problem: are we always aware of all our thoughts? Leibniz here breaks with the Cartesians by identifying deaf thoughts with computational and symbolic processing in general. In a letter dated 1697 Leibniz wrote:

> In fact I agree [with the Cartesians] that there is thinking to which neither the spirit, nor images, nor figures correspond and some of this thinking is distinct […] they are] deaf thought[s], like in algebra, where one thinks in symbols in place of objects. So, often, to abbreviate, words are used thinking of them without analyzing them, because that is not necessary in such circumstances (Micheletti, 2010, p. 2).

So language gave humans the capacity for a kind of symbolic reasoning that labors on independently of signification and reference; and this capacity is the basis for efficient higher level thought. Later
Leibniz will suggest that if this computational capacity could be associated with an ambiguity-free lexicon, it would be possible to mechanically compute the correctness of propositions – a logic language.

If deaf thoughts are the products of the human capacity for symbolic processing – necessary for reasoning – then study of their surface forms will reveal the mechanisms of reason. Leibniz is particularly interested in the elements that connect and relate symbols in discourse, which he calls “particles”.

The Lockean speaker explains that particles connect ideas and propositions, and that “good speech” depends on their proper usage. Leibniz counters that given complete propositions and ideas appropriately ordered, the listener can always fill in the missing particles (linking the ideas), so at a discursive level particles are less essential to “the art of speaking well”. In other words, the interesting particles are not those that bind the discourse; rather, says Leibniz, they are those that bind elements within ideas.

Leibniz divides particles into three groups:

- **Prepositions**: particles that connect the “parts of ideas”, themselves ideas.
- **Adverbs**: particles that connect ideas within propositions.
- **Conjunctions**: particles that connect propositions.

Conjunctions and adverbs were the units identified by Locke; prepositions are Leibniz’s essential contribution. The implication is that because they are required for the construction of molecular ideas, prepositions are more essential to the “art of speaking well” than other particles. Indeed
Leibniz seems to hold prepositions in high regard. He dismisses grammatical gender as irrelevant to meaning, excluding gender from his “philosophical grammar”, and analyzes case inflection as incorporation of prepositions into nouns (note the identification of preposition with case affixes). The synthetic activity of prepositions, building ideas out of ideational components, is the central activity of the computational function in thought. Prepositions are therefore essential to the creative work of pensees sourdes and to reason itself.

If the function of prepositions is to connect the elements of ideas, what is their meaning? Leibniz asserts that the meaning of particles in general (i.e. functional categories) cannot be stated by means of definitional language because they lack referential content; nor does their full meaning carry over in translation when rendered by a single corresponding term in the target language (e.g. mais for but); nor can a single general definition capture all their senses. “One finds that it is rather by examples and synonyms that one explains [particles] than by distinct notions” (p. 280). Or by paraphrases. Nevertheless, Leibniz insists, the number of distinct meanings for a given particle are finite.

Perhaps one reason why the meanings of particles are best expressed by examples and paraphrases lies in their origin. The Lockean voice states the (today familiar) hypothesis on the genealogy of abstract substantives, identifying their origin in metaphoric transpositions of experience of nature – e.g. esprit originally meant “breath” and so on. In a rather striking gesture, Leibniz extends this idea to functional elements:

It will be well however to consider the analogy of things sensible and insensible that has served as the foundation of tropes: one will understand better by considering a well-understood example as that provided by the usage of prepositions, like to, with, of, before, in, out, by, for, on, toward, which are all taken from place, from distance, and movement, and transferred afterwards to all sorts of changes, orders, sequences, differences,
congruences […] as such analogies are extremely variable and not dependent on any determined notions, languages vary greatly in the usage of these particles and cases (p. 225).

Paraphrasing: the basis of figures of speech is the analogy between sensible and insensible things. We can best understand this analogy, Leibniz teaches, by reflecting on words with which we language users are extremely familiar because they’re everywhere, namely prepositions. The abstract meanings of prepositions that denote general relations, such as ordering, similarity, difference, and change, are based on the spatiotemporal meanings of their homophones and/or etymological predecessors. The name of the relation between, e.g. a Figure and a Ground in space is established as the name of a conceptual relation. And because the analogies are arbitrary – there is no necessary mapping from an item denoting some specific spatiotemporal relation to the abstract notion of ordering, for example – languages vary in how they materialize them. Hence prepositions in a new language are hard to learn.

Leibniz’s dialogue reads as remarkable prefiguration of modern debates about prepositions and, more generally, about the relationship of language to mind. Reason and language are complicit, stereotypically human faculties; there is a combinatorial faculty that produces syntheses that become meaningful only when attended to (interpreted); such syntheses are built from simpler ideas; prepositions – not interesting to empiricists – are the connectors in synthetic thoughts (the markers of relations, we would say today); prepositions and case are different forms of the same underlying function; and there are two sets of meanings of prepositions, linked by analogy: the literal, referring to space and time, and the figural, representing relations in other domains. In some broad sense, the subject of this dissertation is Leibniz’s doctrine of prepositions.
For languages that seemingly have functional adpositions (Ps\(^1\)), how and when might children learn to use them? The present work aims at shedding light on this question while also contributing to the ongoing debate on the nature of the category P. Two problem areas are delineated by this motivating question. For first language acquisition, there are questions around what it means for a category to have been acquired and the mechanisms by which functional categories in general are acquired. For linguistic theory the question is what, if anything, does evidence from acquisition tell us about the status of P as a category? Do functional adpositions actually exist? What if all Ps are functional, as Grimshaw (2000) and others hold? How would we even know – that is, what are the criteria for classifying linguistic units into the mutually-exclusive orders of the functional and the lexical?

The thesis has three broad objectives:

1. To determine from records of child speech whether there are differences in production between functional and lexical Ps;

2. To investigate a “soft nativist” acquisition strategy that by coordinating (i) distributional analysis of input, (ii) pragmatic data, and (iii) minimal a priori linguistic knowledge, yields a mapping of morphemes to the feature bundles of nominative pronouns for English and Spanish;

3. By extending the analysis in (2) to adpositions, to examine what, if anything, distributional analysis can tell us about the functional sequences (f-seqs) of PPs and how adult-like

\(^1\) Following the example of Den Dikken (2010), throughout I will abbreviate “adposition” to P. Den Dikken points out that all adpositions may in fact be base-generated as prepositions.
competence with functional Ps may be learned.

Rather than follow the typical format of introducing theoretical and background notions, this introduction limits itself to sketching out the argument and the structure of the dissertation. The relevant background information is distributed to the individual chapters where it matters.

Summary of the Argument

1. Empiricist theories of language acquisition predict that because the rules for functional elements are more abstract than those of lexical elements, children require more data, more contexts, and thus more time in order to acquire them. The acquisition of functional elements is more gradual, proceeding by ongoing generalization from proto-syntactic templates. UG-based theories of acquisition instead claim that because a priori aspects of syntax are available from birth, children are able to learn the functional vocabulary of their language relatively quickly.

2. Adpositions present a particularly rich topic for investigating these claims, as the mixed-class nature of the P category triggers specific and testable predictions about its acquisition. In contemporary theoretical work on adpositions we see a disjunction between those who view the full PP as the extended projection of the noun, implying that all adpositions are functional, versus others who identify the PP as a projection of one or more lexical adpositions. If it is true that P is a split category, then this distinction entails that, on empiricist assumptions, the latter are learned more slowly than the former, even though the two sets of Ps are otherwise phonologically and semantically comparable (as Leibniz suggests). And if the empiricist prediction is proven empirically false, then either the
linguistic theory is wrong (P is not mixed) or the learning model is wrong — with all that is implied philosophically.

3. One way of determining whether a child “has” a functional category is to analyze grammatical errors in the child's speech. These can be broadly classified as either errors of omission – a required function word was not produced – or of co-mission, such as when a word from a different category is substituted for the required form. It is far from clear what omissions tell us about children's grammar. If a child drops a function word some of the time, and appears to do so in some systematic fashion because of e.g. production constraints, then such errors may in fact constitute evidence that the child does understand the syntactic function of the deleted material. Errors of co-mission stand a better chance of indicating incomplete grammatical knowledge, especially if they occur systematically in young children. When it comes to prepositions in at least English, Spanish and French, the evidence from recored discourse suggests that children produce vanishingly few usage errors in their native tongue. Errors of commission are extremely rare.

4. The correct production of function words in appropriate contexts does not immediately imply that the syntactic understanding enabling it matches adult language. Perhaps children produce multi-word expressions as, in effect, unanalyzed idioms or as mostly fixed formulae containing limited variable material, rather than fully analyzed syntactic constructs. A better measure of syntactic understanding consists in determining the productivity of children's utterances. For functional elements we must determine whether the variety of contexts in which children use such elements matches that of adult usage.
Two CHILDES corpus analyses were carried out to investigate changes in the productivity of children's use of lexical and functional Ps over the course of acquisition. One analysis measured the degree of similarity in the distributions over P complements between children and adults. The other attempted to measure the rate of child-produced PPs that were novel with respect to the adult corpus. On both measures, and in both English and Spanish, functional P developed as quickly or faster than lexical P. These studies also suggest that productivity with lexical Ps is tied to vocabulary growth in general.

5. The thesis then turns to modeling the process of acquisition of Ps. In the evaluation of models of lexical acquisition it is common to measure success by means of a criterion of grouping by natural classes. Does the method assign words to groups in a manner consistent with our theoretical understanding of word classes? This is essentially the unsupervised part of speech (POS) tagging problem in computational linguistics: the induction of categories such as verb, noun, preposition, article etc. When applied to functional categories, the tagging problem is fundamentally miscast when understood as the grouping of phonological words into the traditional categories of grammar (pronoun, article, preposition, etc.). Evaluation by this criterion is arbitrary given the variable and descriptive character of the categories — a finer-grained classification scheme might improve some scores, depress others. From the perspective of the learner, category information is fundamentally inadequate for the acquisition of functional morphemes. What must be learned are the specific syntactic behaviors of individual items. Indeed, the very status of category in theories of language remains unresolved: category is then at most one among many features on lexical items; the challenge is to learn them all. For functional morphemes the acquisition task must in some (theory-dependent) manner reflect the reality that given a
fully-specified syntactic context, incorporating all relevant features, a language offers exactly one exponent.

Therefore, the learning problem must instead be formulated in terms of the linguistic atoms whose theoretical status is better grounded: features syntactic, morphological, and semantic. Rather than model the distribution of functional morphemes across categories, better to first work out a model for learning the feature assignments within individual categories.

6. The thesis explores a model of acquisition in the domains of subject pronouns and adpositions. The former are few in number, realize clear semantics, and the analysis of their features is mature; whereas the nature of adpositions remains today an open area of active research. A review of recent work on Ps suggests that, at least in generative linguistics, there is some convergence on the features of functional P. Congruent proposals for the functional sequence of spatial P posit a rich functional structure above the DP. Directional (spatial) Ps, non-structural case affixes and functional Ps appear to constitute varying morphosemantic strategies for expressing a single functional sequence representing aspectual relations. Significant disagreement remains, however, about whether all Ps are functional.

7. A computational model of the acquisition of English and Spanish nominative pronouns is developed. The model orchestrates 1) induction from distributional data; 2) deduction from a lightweight UG endowment — knowledge about feature geometries or functional sequences; 3) semantic clues from pragmatic contexts. The algorithm makes specific predictions about the types of clues required to map the phonological representations of subject pronouns to their featural content. Because of the current insecurity about the
nature of the P category, the algorithm is developed and evaluated on English and Spanish nominative personal pronouns.

8. Finally, the methods developed in (8) are applied to functional P data in English and Spanish. The relationships between Ps induced by the distributional analysis confirms the functional sequences developed in the literature, and also supports the hypothesis that P is a hybrid category.

The empiricist argument with respect to the acquisition of P can be summarized as:

Major premise: children acquire adult-competence with lexical words before functional morphemes.

Minor premise: there are functional P$s and lexical P$s.

Conclusion: children learn lexical P$s before functional P$s.

Should the conclusion be demonstrated false, then one or both of the premises must be false. If children do not differentially learn adpositions, yet the category is indeed split, then there is reason to question the assumptions of the empiricist project. The model of acquisition of functional categories examined in this thesis represents a useful counter-proposal, in that it formulates specific and mild assumptions about our native linguistic endowment, assumptions that nevertheless enable a learner to accurately map exponents to their morpho-syntactic-semantic features.

Structure of the Thesis

Part I deals with theoretical notions. Chapter 1 reviews contemporary theories of P along with a selection of historical predecessors. It begins with a general discussion on the properties of P, concretized as a summary of Hagège (2010), followed by an equally general discussion about
features in grammar. A selection of works on functional P are then reviewed, starting with Montague (1988) and Abney (1987), through van Riemsdijk (1990), Zwarts (1997) and Grimshaw (2000). The theory of Distributed Morphology is then introduced. Next, the PP analyses in Svenonius (2010) and Den Dikken (2010) are juxtaposed. The key contrast is that Den Dikken assigns path and directional Ps to the order of lexical items, while Svenonius posits for P positions in a single extended projection of the nominal Ground. The chapter ends on a summary of nanosyntactic work on prepositions and case, including Pantcheva’s analysis of directional P (Pantcheva, 2010), Caha’s functional sequence of case (Caha, 2011, 2012), and a critical summary of Romeu’s dissertation on spatial P in Spanish (Romeu, 2014). A guiding theme throughout the chapter is the question of the feature constitution of prepositions. If acquisition of functional elements is understood as learning the specific feature constitution of individual items, then in order to properly characterize the task of learning functional Ps we must fully specify their features.

In Chapter 2 the thesis turns to language acquisition. This chapter examines the theoretical underpinnings of constructivist approaches to acquisition through a consideration of Piaget’s *Genetic Epistemology* (Piaget, 1970) and Tomasello’s work, focusing mostly on (Tomasello, 2003). Goldberg’s case for Construction Grammar is also reviewed. The chapter ends on a critical reflection of constructivism in acquisition.

Part II is concerned with the empirical study of children’s patterns of P usage. Chapters 3 and 4 report on three analyses carried out on child and adult corpus data in CHILDES. In each case the objective was to determine whether “developmental” differences in usage over time of functional and lexical prepositions are evident in child production data. The studies looked at errors --precious few of which were found --, the relative frequencies of productions of Ps, and the frequency at
which children produced P+predicate pairs that were not attested in adult utterances. In these studies functional Ps showed no developmental lags compared to their lexical counterparts.

In Part III the focus turns to modeling the process of learning functional P. Chapter 5 presents a computational model of the acquisition of nominative pronouns in English and Spanish. Since their features are well-established, we have in pronouns a useful benchmark for evaluating the success of a learning procedure. The model brings together distributional analysis of adult language and the geometry of morpho-syntactic features pioneered by Harley and Ritter (H&R) (2002). It is shown that while the learner does require data from experience to correctly map the phonological expression of pronouns to their features, the amount of extra information required is small and computable predictable from the model. The model was simulated as a Prolog program whose source appears in the Appendix.

Finally, Chapter 6 applies the distributional analysis of Chapter 5 to PP data in English and Spanish. The resulting patterns of clusters are interpreted as support for Caha’s f-seq of case and Den Dikken’s hypothesis about the lexical essence of spatial Ps. Spanish spatial P clusters show local support for the structures hypothesized by Romeu, but not for his overall architecture. The Prolog learner of Chapter 5 is adapted to the simpler problem of learning a linear, cumulative f-seq of the Caha variety. Applied to English data, the simulation is able to learn the functional Ps of English with minimal external information.

The conclusion pulls together the threads laid out in the previous chapters, essentially restating the logic presented in the Summary of the Argument above.
Part I. Theoretical Notions

Chapter 1. Adpositions and Features

Introduction

This first chapter provides an overview of theoretical treatments of adpositions and introduces the formal notion of features. The guiding question is this: if the acquisition of Ps entails learning the features of individual Ps, what can linguistic theory tell us about such features? *What are the features of P?*

We begin with general remarks about adpositions, by way of a summary of Hagège (2010), and equally broad remarks about the role of features in linguistic theory. Montague’s treatment of prepositions is then briefly touched to draw out the semantic criterion by which he distinguishes functional from lexical Ps. To wit, the meaning of functional Ps like *about* is intensional, while lexical Ps such as *in* are extensional.

The second portion of the chapter summarizes a few significant moments in the evolution of the theoretical treatment of adpositions. The narrative arc is likely artificial but I hope useful nonetheless. By the time we reach the last work considered, Juan Romeu’s (2014) dissertation on the nanosyntax of spatial P in Spanish, specific and testable hypotheses about the features of functional P will have emerged. These are tested in the empirical work reported in Part III of the thesis.
General Notions

Adpositions

As a descriptive work, Hagège's text on adpositions (2010), the product of an essentially typological and functionalist perspective, matches only imperfectly the generative work I will review further on. The monograph nevertheless constitutes an invaluable resource, thanks to the breadth, completeness and clarity of its descriptions. For this reason I have elected to introduce the topic of adpositions by summarizing this particular text.

Language, writes Hagège, maps concepts, i.e. structures of semantic classes (the categories of traditional philosophy), to grammatical and morphological structures. Language then communicates these structures by physical signs. The mapping of semantic classes to grammatical categories is traditional: events to predicates (nominal and verbal), entities to arguments, properties to modifiers, and so on. To Hagège these are the “full” words of traditional Chinese grammars. Importantly, he does not commit to an abstract dictionary-like lexicon that specifies grammatical categories for each atomic entry. The mappings between concept and grammatical category instead process (more or less complex) words already embedded in sentences; category is therefore a reflex of syntactic configuration.

Opposed to “full” words are the “empty” words that connect them. The function of P is to mark the relationships between a term and the phrase that depends on that term. The first is a head, the second is the complement of the governing P. The head is typically either a predicate or a non-predicative noun, while the complement is usually nominal; the P itself heads the PP.

Language has two other methods for marking relations between heads and their dependents: word
order and case affixes. Speaking functionally — and therefore, according to Hagège, syntactically — Ps and case inflection are “near-identical” and therefore in complementary distribution (See also Fillmore, 1968). As “competing systems”, Ps and case affixes exhibit significant contrasts in meaning and morphology. The linguistic analysis of these systems must therefore proceed through considerations of their morphosemantics.

The case system of a given language typically consists of a small number of obligatory, monosyllabic morphemes arranged in well-defined paradigms that are more resistant to historical change than is the case for Ps. Compared to Ps, case affixes evince greater degrees of grammaticalization. We therefore observe a cline of grammaticality from adverbs (“full” words) to cases (“empty” morphemes); Ps lie somewhere between these two poles, and indeed are themselves distributed widely across this cline. Diachronic change may occur in either direction. In some languages case prefixes and suffixes evolved from prepositions and (especially) postpositions respectively; in others, Ps emerge with the breakdown of the case system, e.g. in Romance languages.

Case affixes typically render more abstract grammatical functions, Ps the more concrete. To abstract is to identify an entity by a subset of its properties; an abstract universal is then a property shared by a class of entities. Since all propositions relate arguments to predicates, the linguistic function of such relating is then an abstract universal of propositions. Languages mark “core” arguments, subjects (S) and direct (DO) and indirect objects (IO), via case, word order, or Ps, though the first two cases (“Structural Case” in generative diction) are often realized via word order

---

2 Hagège reverses Aristotle’s understanding of the relative semantic “fullness” of adpositions and case affixes. See *Poetics* XX.

3 I use “abstract” here in Hegel’s sense when he remarks that at an execution for murder, the “common populace” will understand the condemned only as a murderer, whereas “one who knows men” will take into account the biographical, social and economic contexts of the crime — abstraction is easy, it is the concrete that is hard.
or affixes only. Ps also mark “peripheral complements” such as time and space adverbial phrases, comitative, comparative and manner adjuncts, and so on. Hagège invokes Keenan and Comrie’s (1977) accessibility hierarchy to indicate the tension between adpositional versus affixal/word-order function-marking strategies. By rewriting the accessibility hierarchy ranking as

\[ S > DO > IO > \text{Peripheral complements}, \]

Hagège asserts that if a language marks a given function by Ps then it so marks all functions lower in the hierarchy (See Lestrade, 2010 for the complementary generalization for case affixes). This universal division of labor between Ps and case implies that Ps tend towards the more concrete aspect of function marking. In a language that has both case and adpositions, therefore, case tends to express functions “above” the meanings denoted by Ps. Tendentially concrete, Ps correlate better than case affixes with *contextual meanings*, aspects of the specific discursive situation. Time and its dual space are such concrete abstractions, as they involve specific determinations: if the event occurred here, then it did not occur there. Languages often employ Ps to link spatiotemporal descriptions to the sentence core. In Hagège’s words:

An essential characteristic of all linguistic units, whether “content” or function words, which is not unrelated to the nature of meaning as a fuzzy, prototypical rather than discrete, contextual rather than absolute, phenomenon, namely polysemy, is especially developed in Ps, where local, temporal, and other semantic categories apparently melt together more easily than in case affixes (p. 277)

The contextual aspects of the meaning of Ps and their associated lower degree of grammaticality/abstraction, implies a broad loosening of their syntactic constraints compared to case affixes. As Hagège notes, there is sometimes a choice in what P to use; Ps may separate from their complements; are sometimes optional; second language learners often produce usage errors
with Ps; unlike case affixes, P choice may become “target of official support on the part of political authorities” (p.35), and function as stylistic markers. These and other behaviors of Ps point to their greater semantic specificity, at once “empty” and “full” words. To Hagège this fact indicates an essential aspect of language:

[T]he human mind, in the language construction process, does not only meet the need for words with a cognitive content (“full words”) and tools to link them together and indicate the functions they serve in relationship to one another (“empty” words). The human mind is also able to imagine and make words that are both “empty” tools and “full” lexical units (p. 150).

In Western philosophy, an interpretation of Kant’s doctrine of the Schematism suggests that words that are both “empty” and “full” are those that denote relations within space and time (Kant, 1998). As the a priori (universal) forms of intuition distinct from abstract logical concepts, space and time are for Kant the domains targeted by (the seemingly contradictory) closed-class lexical items. It is interesting to note that in GB theory the category P has traditionally been described by two negative features, [-N, -V], a “neither-nor” specification that says little about what Ps are – and perhaps falls out of GB’s 2×2 categorial paradigm rather than from first principles. Hagège instead describes the category as having two seemingly opposed values, [empty, full], which in some sense amounts to a conjunction “both-and”.

In terms of its distribution in syntax, the PP is strikingly promiscuous, perhaps more so than any other phrasal type. Hagège analyzes its distribution into three broad configurations: P phrases appear as complements and modifiers of verbal predicates; as complements and modifiers of nouns; and as predicates. (He also discusses cases where Ps head phrases that serve information structure.)

(1) The spy saw the cop with the telescope.
(2) Bill kisses Mary on the nose in the park.

(3) Richard depends on Bill for excitement.

Here (1) is clearly ambiguous about who had the telescope; the PP either marks a possessive relation between the cop and the telescope (NP attachment) or an instrumental relation between the verb and the telescope (VP attachment). (2), a straightforward sentence of English in which it is understood that it is Mary's nose Bill kisses, could however be construed as the event of kissing occurring on the famously giant concrete nose in City Park. In (3) the PP “on Bill” is understood as an obligatory complement of the verb depend.

Again, the point is the very wide distribution of PPs. In addition to Hagège’s five configurations, we note that PPs also surface in construction with adjectives (4) and adverbs, as modifiers or modifiee (5).

(4) John talked until he was blue in the face.

(5) Mary was very on point.

Summarizing, the points from Hagège’s exposition that are relevant to this dissertation are:

- An adposition marks the relation between a head and a phrase (grammatical function).
- Ps and cases are different morphological realizations of the same field of semantic meanings and grammatical functions.

---

4 The first two configurations, reanalyzed in terms of 2×2 structural characteristics, i.e. argument vs modifier and VP vs NP attachment, have been the bane of computational linguists working on syntactic parsers (C. T. Schütze, 1995). The so-called “PP attachment ambiguity” problem is a significant source of errors for broad-coverage parsers (Villavicencio, 2002). Sentence (5) proves particularly lethal for the software parsers — the state of the art Stanford invents a predicative AP headed by adverb “very” while the HPSG-based ERG parse crashes — yet is perfectly interpretable for English speakers.
Languages manifest a wide variety of case- and/or P-based strategies for expressing relational meanings, e.g. part/whole, figure/ground in space, open/closed events in time.

Ps are often more contentful than case morphemes; they are typically able to make finer distinctions within the discursive context, distinctions that might lean on world knowledge.

PPs distribute very widely, i.e. may be taken as complements by many different kinds of heads.

Features

As a very general working definition, features are atoms that mark contrasts between representations of linguistic objects; they are the abstract universals of language. Features are essential to linguists because they enable the formulation of generalizations across heterogeneous elements (Corbett, 2012). As such, features have appeared in virtually every aspect of linguistic research – though not always in a uniform or formally transparent manner. In David Adger's words:

Features […] are the basic building blocks of syntax, and the ways that they may be combined, and the ways in which they may relate to each other, are what gives rise to the observed phenomena (Adger, 2003, p. 22)

Features abstract away differences, determining sets of linguistic objects that can then be cross-classified and targeted by succinct rules of the grammar. The universal claims typically made for features imply they are made available by UG. In Minimalism, for example, features are taken as primary objects of the theory, beyond mere descriptive devices. Yet it is precisely their abstract nature that may limit the scientific usefulness of features. Hegel held that the understanding that formulates its rules in terms of abstractions, separately from the concrete complexes in which they
are realized, will inevitably miss the true phenomenon. Boeckx, Tiago and Leivada (2013) indeed argue that “features are at the moment substance-less and as such they cannot inform a theory of language”. The principal target of Boeckx et al’s critique is the “featuritis” they see as having taken hold of the cartographic project, effect of what they view as merely empirical generalizations over cross-linguistic data, yet the broader point is implied: the psychological and indeed ontological status of features — whether features are necessary for an explanation of language — remains as hazy as ever.

That said, one cannot fail to appreciate the tremendous usefulness of features as descriptive devices. In what follows the focus is on the features that are relevant to syntax and meaning, i.e not to the phonological system.

*Feature Formalisms*

If the mere presence of a feature, e.g. AGENT⁵, triggers the execution of some rule, then that feature can be represented as a privative attribute, either present or absent. Privative features are contrastive only on the basis of their presence or absence, and signal instructions to the other modules of language. In models such as the morphosyntactic geometry of Harley & Ritter (2002) the instruction (to the relevant system) when a certain privative feature is lacking, the elsewhere condition, is specified as a default; the feature’s presence is thus associated with the marked condition.

If instead one or more rules must explicitly refer to the absence of a feature, and if we want to utilize absence only as an underspecification mechanism, then binary-valued features are required, e.g. Van Riemsdijk’s [+/- DIRECTIONAL]. In effect this amounts to a ternary logic, since a feature can

---

⁵ Following Corbett’s convention, I will write the names of features in capital letters.
be positive, negative, or unspecified/indeterminate. Binary features then introduce the distinction between a feature as such and its possible values, e.g. the values “+” and “−”. It is not clear, however, that syntax, unlike phonology, makes much use of the negative values of features (setting aside category features, e.g. -V and -N); we may want different sorts of values. As Corbett (2012a) and Corbett (2012b) indicate, binary features are just special cases of multivalued features; and if features can take other features (and even sets of other features) as their values, perhaps even recursively, then we have quite complex feature structures. The feature system can therefore grow into a powerful representational mechanism. How capable a system do we need to capture the facts of natural language?

Most syntactic frameworks seek to subsume agreement phenomena into the feature system. The intuition is that if in some language the subject of a sentence agrees in number with the verb, then somewhere we need a rule that says, roughly, “set the number of the subject to that of the verb”. If number is modeled as a feature, then that one number agreement rule already suggests the distinction between the feature itself and its possible values. Privative systems that merely operate on sets of atoms cannot make that distinction. If the language represents three distinct NUMBER values, a separate copying rule would be required for each in a private system; it is therefore more efficient to group the relevant features and target the group (or “class”) once in a rule. For this reason purely set-based privative formalisms are sometimes judged insufficient (Adger, 2010).

Given at least two at least partly independent sets of (“orthogonal”) features, two feature classes, their Cartesian product defines a space of possible realizations, a paradigm. When the same form appears in more than one cell we have the phenomenon of syncretism. Syncretism appears to

---

Various terms appear in the literature for the feature/value distinction. Where Corbett simply has “feature” and “value”, Adger and Svenonius define a predicate over features, a set, as “class” (but also “attribute”), reserving “value” for qualities of feature tokens, so there are three levels, class:feature:value.
follow certain patterns cross-linguistically. Caha (2008) cites typological literature showing that
syncretism between case features frequently appears constrained by a certain ordering of the cases.
This suggests that the simple model of feature classes as sets of independent features might allow
too much freedom to the theory. Within a class it seems there are limits to the co-occurrence of
features.

A paradigm space that is mechanistically constructed by Cartesian productions the risk of generating
combinations that are not realized in practice or indeed are nonsensical, e.g. when logical
dependencies between feature values reduce the paradigm space. That is, if for feature F1 a value -X
implies that F2 cannot take -Y, then there is no cell \{F1= -X, F2 = -Y\}. The theorist might simply
exclude cells on conceptual grounds, but the elegance of the theory then suffers.

Implicational hierarchies such as feature geometries provide a more principled approach to
modeling the observed dependencies within paradigms, since they encode these dependencies
directly. A feature geometry can be thought of as a polytree, a type of directed acyclic graph
(DAG), in which nodes represent features and directed edges encode dependency: an edge A -> B
means that if A is present, B must also. The graph is a polytree if it contains no trails from some
node back to itself. Feature geometries can be constructed over privative or valued features; in the
latter case the parents of a node will represent all possible values of the dependent node (see Figure
1 for an example from a paper on Russian case (Wiese, 2004)). A well-known example of the
former is Harley and Ritter’s (2002) feature geometry of phi-features (Figure 2); see also (Cowper,
2005).
Such graph-theoretic devices allow us to exclude combinations like \{-X, -Y\} above by ensuring that no path lead from -X to -Y. Now, it is interesting to compare the affordances of feature geometries to the previously discussed feature:value (or class:feature:value) structures. Geometries of privative features can be understood as structuring the value spaces of non-terminal features of the polytrees.
Taking Harley and Ritter’s model, sets of the dependents under the Participant, Individuation and Class (which Harley and Ritter non-theoretically term “group” nodes,) can, in some sense, represent the range of particularizations of the more abstract entities as marked entities. For example, Feminine and Masculine features all share a Class feature and an Individuation feature; Class and Individuation therefore could also be interpreted as predicates of the concrete (marked) GENDER features, i.e. as “second order features” in the sense of Adger (2010). Compared to the simple two-level, cross-cutting class property of features, geometries allow for arbitrary levels of nesting of increasingly abstract parent classes, yet also provide a graphical syntax for declaring linguistically attested constraints on feature classification. Adger and Svenonius, though aware of this possibility, argue that their feature class approach is descriptively more powerful than the DAG because any such graph can be represented as a set of sets of features, where each node in the graph corresponds to the set of all nodes that it (reflexively) dominates; the reverse, however, is not true, e.g. a feature system which consists of the classes \{A, B\} and \{B, C\}. To take an example from the literature, in Chomsky (2000), C and v are phase heads, while v and V are \(\theta\)-assigners. (p.14)

The argument is that while in a DAG a node may not have multiple parents, languages do show evidence of such configurations – DAGs are too constrained. This is true of trees, a specific sort of DAG, but not of polytrees. In principle there is no reason a node in a feature geometry might not imply more than one higher-level feature. The polytree in Figure 3 below, from Koller and Friedman (2009), a seminal text in graph theory, clearly shows that multiple parents are possible.
Caha has argued instead that feature geometries have too few constraints (2012). In his investigation of case syncretism he shows that a specific geometry of case features attributed to Williams predicts patterns of syncretism that are not attested, whereas a linear functional sequence covers the data correctly. Whether this means all feature geometries over-generate is an empirical question, in part dependent on whether the syncretism criterion applies in other domains.

Types of Features

Svenonius (2006a) suggests a classification of features according to which a feature is either active only within a single module of language, or it is active at an interface between modules. An internal feature gives rise to representational contrasts in its domain but not in any other. For syntax this means that the presence of a feature forces a determinate word order, or requires particular morphological realizations via agreement or government (Kibort, 2010). When instead a feature is relevant to two or more modules, it is an interface feature. Svenonius argues that certain values of interface features (as opposed to the features themselves) might not have
effects in some domain; such features are **uninterpretable** within those domains. For example, NUMBER on nouns is an interface feature because it is relevant to meaning. But in a language that has, say, subject-verb agreement, forcing the expression of NUMBER on the verb, such instances are uninterpretable since the semantics ignores them. Similarly for GENDER: while there are semantic effects associated with gender inflection, GENDER markers on e.g. adjectives under agreement are semantically transparent.

Setting aside the question of interpretability, we might ask whether there are any internal features at all. Morphological features such as declension classes in Latin, which determine the forms of the output but have no effect in syntax, are candidates. But since declension classes interact with interface features such as CASE they are not immune from the operations of syntax. Moreover if morphology in general is subsumed to syntax, phonology and the lexicon (as in Distributed Morphology), then the very possibility of internal morphological features evaporates.

It remains for theory to firmly establish whether there are syntax-specific features that play no role at the syntax-semantics interface. Svenonius (2006), for one, concludes that all syntactic features are shared with the semantics module. In (2012) he simply takes “semantic feature”, “syntactic feature” and “syntactico-semantic feature” to constitute synonymous terms. Let us consider the sorts of features that are often taken to be at least syntactic.

- Generally accepted as morphosyntactic features are the phi-features, PERSON, NUMBER, GENDER. These clearly can have semantic effects.

- CASE, the feature expressed by functional adpositions, also touches all three (or two) modules. CASE is active in agreement; and spatiotemporal values of inherent CASE.
features, along with their metaphoric transpositions, e.g. ablative and dative, are implicated in meaning, as are the semantics of participants, the domain of structural Case (Grimm, 2011).

- Corbett views PART OF SPEECH, i.e. category, as uncontrovertially internal to syntax. Adger also lists CATEGORY as a feature, taking values from \{N, V, A, P\} or [+/- N] and [+/- V]. Others, e.g. Embick (2012) understand categories to constitute mere metalinguistic labels for groupings of primitive (non-categorial) features. A different perspective posits a conceptual distinction between category and feature. Adger and Svenonius point out that in GB and Minimalism categories are involved in processes of projection and selection, whereas features are targeted by agreement mechanisms (Adger & Svenonius, 2011, p. 6). On this view categories are labels for the nodes in the functional sequences of UG, while features constitute the payload of functional heads. In contrast, the distinction between category and feature is collapsed in cartographic and nanosyntactic approaches (Starke, 2009), according to which each node in a functional sequence corresponds to a single feature – features project. Since on this view features are interpreted by the semantic module, CATEGORY disappears as an internal feature of syntax.

- From outside GB there is the SLASH feature to account for movement and long-distance dependencies. Minimalism would reject this feature as a violation of Chomsky’s Inclusiveness principle, whereby the derivation is barred from adding features to those present in lexical items feeding it.

- The selectional feature SELECT specifies sets of features with which the bearer must
unify as a requirement for grammaticality. Though this feature is central to Categorial Grammar, it is operationally different from more atomic features like GENDER since it drives the derivation itself. As with CATEGORY, in Minimalism syntactic selection falls out of universal functional sequences (and lexical semantics) rather than from an explicit feature.

- Van Riemsdijk has [+/-PROJ] and [+/- MAX] to encode the X-bar status of nodes in the tree (Corbett mentions BAR also). These would appear to act only in the syntax. These violations of Inclusiveness are, however, only possible within non-Minimalist frameworks, since X-bar theory itself is incompatible with Inclusiveness.

- Basic conceptual notions derived from experience are not syntactic but we exclude them from the discussion because of their a posteriori character. Researchers working in cognitive linguistics might characterize the dimensions of conceptual structures as complex features, such as \{RED, GREEN, BLUE\} for color, or take species-wide properties of our embodied nature as the basis for semantic features, e.g. ALIVE, WOMAN/MAN. Nevertheless these belong to the conceptual system, not the linguistic.

- From Kant, relations in space and time are known by means of the universal and a priori forms of spatiotemporal intuition. TENSE and ASPECT are therefore semantic. According to Kibort (2010) TENSE and ASPECT are only semantic, while in Minimalism, because they represent important nodes in functional sequence of the clause, they class with the interface features. Adger (2003), on the other hand, groups TENSE with the morphosyntactic features (contra Corbett), and Adger and Svenonius (2010) add ASPECT and MODALITY.
Corbett also lists VOICE, MOOD, POLARITY, EVIDENTIALITY and others as semantic-only. Again, in Minimalism some of these name functional projections in the clause, though Cinque states that NEG (i.e. POLARITY) cannot be associated with a single projection.

We see then that if Svenonius is right in reasoning that all syntactic features are also semantic (the reverse is not always true), and if we accept, say, the cartographic model of the clause, then all features (other than phonological) lie on the syntax-semantic interface. If in addition we dispense with a separate module responsible for morphology – declension classes would then move to the phonology – we can then collapse these distinctions into a unified set of morpho-syntactic-semantic features.

Based on the understanding of synsem features as at once morphosyntactic and semantic I offer the following definition: a feature is an a priori atom of language that (i) is relevant to syntax and (ii) instructs the semantic module to apply a specific function to the meaning of the syntactic complement of the node bearing the feature. The problem now is to determine the features present in items belonging to the category P.

Theories of P

---

7 Caha (2012) similarly writes “I will be assuming a view on feature meaning such that each feature operates on the semantics of its complement” (p. 17). It seems more modular to distinguish between the feature as that which operates, as Caha states, versus the feature as the trigger of a semantic function.
Montague

In *The Proper Treatment of Quantification in Ordinary English* (PTQ) Montague assigns prepositions to the syntactic type that takes term phrases (noun phrases and proper names) to yield modifiers of intransitive verbs. The corresponding semantic type in the intensional logic is rather complex in its expression: the function from properties of properties of individuals to functions of properties of individuals to sets of individuals. The semantic type can be shown to map mechanically from the syntactic definition: properties of individuals are (intensional) noun phrases, while sets of individuals are predicates, verb phrases and intransitive verbs in PTQ. A function from properties of individuals, which are intensions of sets of individuals, to sets of individuals, is in effect a predicate modifier. Thus a preposition takes the intension of the properties of individuals – which is another way of characterizing an individual, i.e. a nominal referent – to return a predicate modifier, mirroring the syntactic type.

In PTQ Montague only lists two prepositions in his fragment of English: *in* and *about*. The selection appears intentional, as if intending to contrast a lexical preposition with a functional one. Indeed, the contrast emerges in the specification of the condition that constrains the semantic interpretation of prepositions in PTQ. This is Condition (8), one of Montague’s more fiendishly opaque meaning postulates (Dowty, Wall, & Peters, 1981, pp. 224 – 225). In the symbolic formulation that follows, $P$ ranges over “sublimation concepts” (Dowty et al., 1981, pp. 220 – 221), intensions of all the properties of individuals; $Q$ ranges over properties of individuals (intensions of predicates); and $x$ and $y$ over individuals. $G$ represents a relation-in-intension between individuals and predicate modifiers. Montague asserts:

$$
\exists G \forall P \forall Q \forall x \sqcap [\delta(P)(Q)(x) \leftrightarrow P{^\lambda y [[^\ast G](\ast y)(Q)(x)]}]$$

where $\delta$ translates in'
This is the “condition of extensionality (or extensional first-order reducibility) for prepositions”.

Without entering into full detail (Dowty et al (1981) provide a detailed explanation of Montague’s condition on transitive verbs, which is similar to (8)), we first note that the condition applies only to *in*, not *about*. That is, it is a constraint on lexical prepositions. The modal operator $\Box$ states that it is necessarily the case that its argument is true. The argument itself consists of a mutual entailment, the left of which straightforwardly describes a sequence of function applications. So for *John walks in a park*, its translation into IL would roughly read:

$$\text{(6)} \quad \text{in'}(\text{[a park]'})(\text{walk}')(j)$$

which is the compositional semantic translation corresponding to the syntactic analysis

$$\text{(7)} \quad \text{[John [walks in [a [park]]]]}.$$ 

The right hand side of the formula in Condition (8) falls out of the fact that Montague’s grammar allows the noun phrase *a park* to be generated by a second route, by “quantifying in”. (The same is true of the subject *John*, hence Montague specifies a separate condition for extensional transitive verbs.) The NP *a park* can be introduced into the derivation as an operator binding a variable that Montague’s symbolizes with an indexed third-person pronoun he. The NP takes the PP as an argument, thus valuing the variable as the NP. The derivation might be glossed as:

$$\text{(8)} \quad \text{[[a park, [John [walks in [it]]]]]}$$

The term $\mathcal{D}\{\lambda y[[\mathcal{G}](y)(Q)(x)]\}$ captures this alternative derivation, where $\mathcal{D}$ represents the term phrase *a park* and $y$ the variable. The crucial difference between this term and the simpler left hand
formula constructed via function application, $\delta(\mathcal{P})(Q)(x)$, is that in the term on the right $\mathcal{P}$ is in an extensional positions. In Montague's system the arguments of functions are always intensions, so when in' takes $\mathcal{P}$ as its argument, the latter must be an intension (which it is by definition). But the brace notation is a specific type of function application where the functor is first cast to its extension (i.e. committed to some index):

$$(9) \quad P\{x\} = [^\mathcal{P}]\langle x\rangle$$

On this reading the existence of the complement of the preposition is therefore presupposed. Since the condition states that the intensional and extensional meanings entail each other, Montague is asserting that in in English always has extensional meaning, denoting the situation at the current index. By contrast, about does not fall under Condition (8); a book about unicorns does not imply that unicorns exist. More generally, Montague seems to suggest that spatiotemporal adpositions are necessarily extensional while functional prepositions can go either way, depending on the derivation 8.

On Montague's account about behaves like an intensional relator because that is the default for functions – the extensional character if in is the marked case.

Now, as Den Dikken has pointed out (p.c., April 9, 2015), the semantic type of the preposition is not determined by the unit itself, but by its use in context. On their spatiotemporal readings, in and about both have extensional interpretations, while on their metaphoric, functional readings they are intensional, as in (10) and (11):

(10) Mary is interested in unicorns.

(11) Mary runs about the unicorn farm.

---

8 Since in of course has many uses beyond literal spatiotemporal senses, Montague would have to assume there is another sense of in not constrained by Condition (8).
Since the predicate as a whole informs the intended sense of the preposition, Montague's analysis suggests a semantic criterion for the functional/lexical contrast for Ps: in spatiotemporal contexts, the P is extensional, while it is intensional in non-spatiotemporal (i.e. metaphoric) contexts. The latter are also the stereotypically functional (abstractly relational) uses of Ps. Thus, Montague's analysis suggests the semantic ratios: lexical:functional :: extensional:intensional :: spatiotemporal : relational.

_Abney (1987)_

Abney contrasts functional elements with “thematic” elements (preferring “thematic” to “lexical” since he has a unified lexicon that also stores function words). Stereotypically, complementizers and inflections are functional while verbs and nouns (and adjectives, adverbs, possibly Ps) are thematic. Thematic elements are the linguistic signs that “survive when language is reduced to bare-bones” (p.44), as in early child speech; syntactically they are distinguished by a [+- N] feature (he does not adopt the [+- V] feature typical of GB). Functional elements are distinguished from thematic elements by a [+- F] feature.

Abney sketches out general characteristics for the natural class of functional elements, emphasizing that the characteristics are descriptive but not definitional. Functional elements are:

- Members of closed classes;
- morphologically and phonologically dependent; unstressed; sometimes null;
- take a single, non-argumentative, obligatory complement from which they cannot separate; and
• mark grammatical or relational features, rather than act as bearers of descriptive content.

The characteristics of [-F] elements are complementary to those of [+F] elements. Thematic items form open classes; bear stress; may take multiple optional complements that may undergo movement; and are vectors of referential meaning. The essential distinction between these natural classes lies in the nature of their complements. Abney writes:

The primary property of functional elements is this: they select a unique complement, which is not plausibly either an argument or an adjunct of the functional element [...] they do not describe a distinct object from that described by their complement (p. 38).

Unlike other properties that distinguish the two classes, this “primary property” is definitional for functional elements. It ultimately rests on a semantic distinction, though he argues there is a syntactic reflex: C selects IP, I selects VP; they each take exactly one required complement whose category label is not typically argumentative (i.e. not DP, PP, CP). The key, according to Abney, is that the functional element does not add referential content to the expression. Abney reminds us he is operating squarely within the Aristotelian tradition: functional elements are the meaningless sundesma and arthra of Aristotle's Poetics, distinct from the meaningful thematic elements, nouns, verbs and adjectives. “Descriptive content — what functional elements lack — is a phrase’s link to the world” (ibid). In model-theoretic terms, Abney represents the distinction via semantic types: thematic elements are of type \(<e, t>\) first-order predicates from entities to truth values, while functional elements are “functors”, higher-order predicates taking predicates as arguments\(^9\), such as generalized quantifiers (e.g. determiners) whose type is \(<<e, t>, <e, t>, t>>\) (Dowty, 1979, p. 108).

\(^9\) It is not clear that the asemos/semantikos “meaningless/meaningful” distinction amounts simply to one of order in the predicate calculus. If modification is taken to operate by functional application, then adjectives would also group with functional elements.
The semantic distinction between functional and lexical elements is relevant to the question of the categorial status of Ps, as Abney’s diagnostics give no clear reading for the class of Ps as a whole. Ps constitute a more-or-less closed class and are generally phonologically dependent; but they can strand and usually take two arguments. Unfortunately the semantic criterion as Abney formulates it fails to clarify, since the meaning of lexical Ps, at least according to the model-theoretic interpretation of locative Ps, necessarily involves descriptive content. Compare the denotations of \textit{gray} and \textit{in}, from Heim and Kratzer (1998, p. 66):

\begin{align}
(12) \quad [\text{gray}] & = \lambda P \lambda x[P(x) = 1 \text{ and } x \text{ is gray}] \\
(13) \quad [\text{in}] & = \lambda y \lambda P \lambda x[P(x) = 1 \text{ and } x \text{ is in } y]
\end{align}

The verification formulae “x is gray” and “x is in y” capture the descriptive content of the terms. In both cases the truth conditions are world-specific, suggesting that the adjective and the preposition are thematic in character. So in the end Abney groups Ps with verbs as [-F], [-N], non-functional and non-nominal, but with a question mark. “P seems to straddle the line between functional and thematic elements; one might wish to treat it an unspecified for [+/ - F]” (p.43).

\textit{Van Riemsdijk (1990)}

Van Riemsdijk begins by laying out his Categorial Identity Thesis (CIT). Like Abney, Van Riemsdijk maintains that projecting functional elements usually only take a single obligatory argument. But while Abney argues that functional projections are transparent to the descriptive content of their complements, Van Riemsdijk makes the stronger claim that, per the CIT, functional elements bear specific categorial features that must match those of their complements. Each function type selects a single lexical category. Nouns and verbs, conversely, are quite free in their selection; noun phrases,
verb phrases, clauses, PPs are in principle all candidates as complements to thematic heads, suggesting that constraints on their selection originates elsewhere — in the semantics perhaps, or other functional material.

If categorial features are shared between lexical and thematic types, a separate feature is needed to distinguish the two. Van Riemsdijk therefore also adopts the feature [+/- F]10. Since he retains both N and V features, the resulting 2×2×2 table creates eight possibilities: the usual GB lexical categories {V, N, A, P}, all branded [-F], and their [+F] functional correlates, which he writes as the lower case letter corresponding to the lexical category (n, v, p, and presumably a, though this is not taken up). The functional elements are then:

\[
\begin{align*}
\{ & +N, -V, +F \} \\
\{ & -N, +V, +F \} \\
\{ & -N, -V, +F \}
\end{align*}
\]

Note that while the functional elements associated with nouns and verbs map to previously established categories, D and I respectively, functional P maps only to itself, meaning some subset of P. Van Riemsdijk identifies postpositions and circumpositions in German (and Dutch) as examples of functional P. The suggestion is that postpositional and circumpositional PPs consist each of a projection of a “little” p taking a PP as its complement, where the p head-moves to the right of the PP complement when it is a postposition, while for circumpositions, p is base-generated in situ. For German and Dutch, at least, the general form of the PP therefore has the following structure:

(14)  [pP [PP P NP] p].

10 Van Riemsdijk also proposes features [+/- PROJECTION] and [+/- MAXIMAL] to characterize X-bar structural positions. These violate the Inclusiveness Condition, as they presumably emerge from the derivation.
This result makes sense, argues Van Riemsdijk, because it fits with the general purpose of functional elements, which is to decorate the lexical head with additional morphosyntactic features, such as definiteness or gender for nouns. The job of the functional prepositions he investigates — all spatial — is to “express certain locational dimensions where the lexical prepositional head does not do so itself.” (Henk van Riemsdijk, 1990, p. 239) In a move that prefigures the work in vector-based spatial semantics later in the 90s (Zwarts & Winter, 2000), Van Riemsdijk identifies (but only “partially”) the feature contribution of German ‘little p’ as [+/- DIRECTIONAL], [+/- ORIENTATION] with subfeatures [+/- UP] and [+/- IN], and [+/- PROXIMAL].

Van Riemsdijk’s gloss of functional elements as carriers of features develops the link between syntactic features and function words. Abney also related the two notions, e.g. by arguing that D is the site of φ-features, but as we saw he also underlined the relational responsibility of functional items. A question thus emerges: what connection is there, if any, between the features borne by functional elements and their function as relators?


Jane Grimshaw’s crucial notion of _extended projection_ (2000) clarifies the structural relationships and dependencies between thematic and functional units. Like Van Riemsdijk, Grimshaw notes that projections of lexical terminals (l-heads) are stereotypically wrapped in projections of functional elements (f-heads). It is a theory of percolation:

The defining property of an X-bar projection is that it is the domain through which information flows. The defining property of a head is that it determines properties of the phrase that it is the head of. This follows from the theory of projection, in which properties of heads project, or percolate, up the tree to the entire phrase (p. 123).
The patterns of distribution of f-heads with respect to l-heads are controlled by the hypothesis that certain f-heads share categorial features with certain l-heads, independently of their belonging to distinct categorial classes. Thus \{N, D, P\} constitute a natural class, sharing nominal categorial features, and similarly for \{V, I, C\} in the verbal domain. Because categorial features percolate upwards from the l-head to the functional projections above, all categorial features must match or the sentence crashes.

Functional heads are distinguished from their lexical category peers by a non-categorial F feature indexed by an integer set at zero for l-heads, 1 for the lowest f-heads (sisters to the lexical phrases), and up for the higher f-heads. Thus V and N each bear a \{F0\} feature, I and D \{F1\}, C and P \{F2\}. Grimshaw then defines the concept of projection as a hierarchy in which the highest node shares categorial features with a lower terminal and no intervening node is lexical (i.e. F0). An extended projection is one in which the F value of the top node is greater than that of the lexical head, i.e. where the projection of the l-head is the complement of an f-head. The F index increases monotonically through the extended projection.

L-heads theta-mark their complements, while f-heads can only take complements of their category, to form extended projections:

If information projects from all of the heads of an extended projection we expect consistency within a projection for all projected features. Projected features must agree throughout the extended projection (NP-DP-PP, and VP-IP-CP), and wherever morphology records the value of these features they will be visible (p. 125).

On Grimshaw’s assumption the PP node constitutes the apex of the nominal extended projection, parallel to CP in the verbal domain. Prepositions are therefore functional items – a radical reinvention that
breaks with GB theory, in that P no longer belongs to the pantheon of lexical categories and no longer roots its own projection. But we must then ask, why can P be stranded (at least in English)? Grimshaw simply notes that “perhaps this should be related to the proposal[that] P can be either functional or lexical” (p. 128). So the problem of the categorial status of Ps is not quite resolved.

Zwarts (1997)

Zwarts’ essay contributes two innovations to the theoretical treatment of Ps, vectored through a reanalysis of the distinction between lexical and functional categories. First, rather than viewing the distinction as exclusive, he follows Grimshaw in disentangling the thematic order from the functional into orthogonal, binary-valued dimensions – though unlike Abney and Van Riemsdijk, he does not mark an item’s being functional via an explicit F feature. Rather, and this is the second innovation, an item is functional if its syntactic type marks it as a functor over phrases of lexical categories. This decomposition of the thematic/functional classification sets up four types of categories: lexical and functional, lexical and nonfunctional, nonlexical and functional, nonlexical and nonfunctional. Zwarts initially places P in the last group, nonfunctional and nonlexical.

Zwarts’ assignment criteria for the separate lexical and functional dimensions are built on standard notions. Lexical categories are simply the open classes that admit new members via productive morphological processes or borrowings (see also Talmy (2000) for whom the distinction amounts to that between open and closed classes).

---

11 Den Dikken (p.c., April 9, 2015) has pointed to a problem with the stranding criterion: in English the infinitival marker to, standardly analyzed as a realization of T(ense), can be stranded, as in go, Mary wanted to.
Table 1: Properties of functional and lexical categories

<table>
<thead>
<tr>
<th>Functional</th>
<th>Nonfunctional</th>
</tr>
</thead>
<tbody>
<tr>
<td>Take a single, obligatory argument</td>
<td>Varying argument structure</td>
</tr>
<tr>
<td>Map phrases to phrases</td>
<td>Constituents of phrases</td>
</tr>
<tr>
<td>Small morphemes</td>
<td>Varying sizes</td>
</tr>
<tr>
<td>No stranding</td>
<td>Can be stranded</td>
</tr>
<tr>
<td>Do not participate in morphology</td>
<td>Can participate in morphology</td>
</tr>
</tbody>
</table>

As units belonging to a closed class, Ps are nonlexical – necessarily so since the inherently limited range of pure spatiotemporal meanings constrains the number of possible signs. Yet Ps generally also fit the criteria for nonfunctional elements: some Ps can be used intransitively, as particles or in predicate position, as in “the doctor is in” and so on. We therefore have a “neither-nor” description of Ps, perhaps not accidentally parallel to the [-V, -N] feature matrix of GB.

Zwarts retains GB’s 2×2 table of categories over features [+/- V] and [+/- N], but now defines as lexical those categories that have at least one positive feature value, thus N, V, and A, leaving P as nonlexical due to its [-V, -N] features. Elements that have category features, whether positive or negative, Zwarts terms (somewhat confusingly) categorial, so P, along with the three lexical classes, is categorial. Function words are instead characterized as operators over categories, F(X) where X is a feature bundle [+/-N, +/-V]. Thus D, a function having nouns as its domain, is of type F([+N, -V]). Since f-morphemes map phrases to phrases they have no role in morphological processes.

We end up with a 2×2×2 classification scheme: lexical/nonlexical X functional/nonfunctional X categorial/noncategorial. Not all combinations obtain since an item’s being lexical (positive categorial feature) entails its being categorial, but the obverse is not true. The only nonlexical,

---

12 I refer to functional elements variously as f-morphemes (Heidi Harley & Noyer, 2000), f-heads, and f-morphs.
nonfunctional category is \( P \), which is categorial, so the combination \{nonlexical, nonfunctional, noncategorial\} is also excluded. The resulting five combinations allowed are:

**Table 2: Types of categories in (Zwarts, 1997)**

<table>
<thead>
<tr>
<th>Traditional Names</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lexical Nonfunctional Categorial</td>
<td>Verbs, Nouns, Adjectives ([+V, -N])</td>
</tr>
<tr>
<td>Nonlexical Nonfunctional Categorial</td>
<td>Prepositions ([-V, -N])</td>
</tr>
<tr>
<td>Lexical Functional Categorial</td>
<td>Modals ( V(+V, -N) )</td>
</tr>
<tr>
<td>Nonlexical Functional Categorial</td>
<td>Functional Prepositions ( P(N) = [-V, -N] ([-V, +N]) )</td>
</tr>
<tr>
<td>Nonlexical Functional Noncategorial</td>
<td>f-morphemes ( D = F(N) = F([-V, +N]) )</td>
</tr>
</tbody>
</table>

Zwarts takes seriously the stranding criterion: functional categories cannot be stranded, because they belong to the extended projections of lexical categories whose lines must not be broken. Since, in Dutch, locative Ps can be stranded, they must not be functional – which does not make them lexical either, on Zwarts’ hypothesis. The explanation is that as \( P \) is categorial it projects its categorial features upwards, and is not tied to the categorial features of its complement. In other words, \( P \) is the base of its own extended projection. Zwarts’ analysis of prepositions thus breaks with Grimshaw, at precisely the point where she stated her reservations – that the facts of stranding are problematic for her hypothesis concerning the functional status of \( P \).

Facts from Dutch point to a class of Ps that strand but do not participate in productive morphology and cannot be used intransitively. Zwarts suggests these are *functional prepositions*. They retain the categorial feature matrix of standard Ps, but behave as operators, taking obligatory nominal complements like all other f-morphemes; their category is \( P(N) \). But because they are categorial, functional \( P \) establish their own extended projections and therefore can strand \(^{13}\).

---

\(^{13}\) A question here is whether these prepositions theta mark their complements, as Grimshaw would require if they are not embedded in the extended projections of lexical categories. It would be somewhat odd to view theta-marking
Finally, there are unambiguously functional Ps that cannot strand at all, take an obligatory complement, are not active in morphology and so on. The category corresponding to these Ps is F(N), where F is the generic, noncategorial operator. Truly functional Ps therefore belong to the extended projection of N.

We see then how Zwarts’ system is able to differentiate between three different types of Ps, avoiding the dichotomist constraints of prior representations, which required an all-or-nothing decision about whether Ps are lexical or functional – or perhaps a “mixed category”, whatever that might entail. To get there Zwarts needs the syntax of his category metalanguage to allow category feature matrices to become operators over feature matrices. P(N) is shorthand for [-V, -N]( [-V, +N] ), a categorial operator P taking a nominal complement. Similarly, Zwarts speculates that the (allowed) class of lexical and functional categories has as its members the modal verbs, whose category V(V) = [+V, -N]( [+V, -N] ) entails they strand (true in English) because categorial. This proposal, however, mistakenly predicts that modals are an open class; it also suggests that their semantics are encyclopedic rather than logical, which is debatable.

**Distributed Morphology**

Recall that Abney rejected the [+/- V] categorial feature of GB theory. When applied to Zwarts’ model, in which F is an operator, Abney’s exclusion of [+/- V] leaves us with a single categorial feature [+/- N]. Nouns and adjectives are reduced to [+N]; verbs and Ps to [-N] and are now both nonlexical, in Zwarts’ sense, as they no longer contain a positive categorial feature. If we also switch the feature model from binary to multivalued (Corbett, 2012), and restrict the lexical category

---

14 Zwarts’ analysis of modals might not rescue the stranding criterion from the counter-example of infinitival to (see p. 40 footnote 11), depending on whether or not one regards modals and infinitival to as competing for the same position. English modal predicates ought to X and had to X suggest they do not so compete.
feature value to just privative N, then the lexical/nonlexical distinction disappears. Nouns, adjectives, lexical prepositions and verbs have their category feature valued as N.

Nothing in the syntax of Zwarts’ category labels prevents an operator from taking an arbitrary category as its domain, so in principle we could have e.g. \( F(F(N)) \), functions over determiners to a phrase – a Kase phrase (KP) perhaps (Bittner & Hale, 1996). In fact the whole system of syntactic labels could be recast using a type-logical formalism as in Categorial Grammar, consisting only of the elements \{N, NP\} and /, in increasingly complex composite types formed by recursion.

Zwarts and Van Riemsdijk explicitly state that functional categories add feature specifications to the feature bundle percolating up the extended projection. When fully specified, we take the feature bundle of all functional morphemes to be unique (Marantz, 1995). So for the Spanish plural, feminine, indefinite article *unas* “some+FEM” we might have a specification \([\text{num}=\text{pl}, \text{gender}=\text{fem}, \text{def}=\text{undef}]\)\([\text{cat}] = N\), where the argument restricts the selection.

Having reduced the syntactic mark of open-class words to a single label, there is little reason to posit a single lexicon incorporating both open- and closed-class words. Distributed Morphology (DM) (Halle & Marantz, 1993; Heidi Harley & Noyer, 1999; Marantz, 1995) re-architects the language faculty by splitting the traditional lexicon into three components: the Lexicon proper, listing the language-specific sets of universal features that the syntax manipulates as units; the Vocabulary, which maps feature bundles in syntactic contexts to their phonological expressions; and the Encyclopedia, the repository of all experiential knowledge and the forms of its expression – open-class idioms (everything in the Encyclopedia is an idiom).

Since all the syntax sees when it manipulates items belonging to the traditional lexical categories is
the single element N, or Root in DM parlance, there is no need for the full set of non-syntactic meanings associated with a given root to enter the derivation. Equally for phonological properties, as these too play no role in the syntax. Accordingly, the doctrine of Late Insertion declares that encyclopedic and phonological material is added to the structure after the syntax has completed its work. Morphological processes are converted either to syntactic manipulation or to post-syntactic operations at Spell-Out.

DM defines functional morphemes, f-morphemes, as the bundles of universal (syntactic-semantic) bundles whose insertion into the tree at Spell-Out is forced. The choice of morpheme is fully determined by the Vocabulary. Since Vocabulary items may overlap in their feature bundles, it is possible for several f-morphemes to match features in the target bundle. The resulting competition between Vocabulary entries is regulated by the Subset Principle, which establishes that a candidate item cannot be specified for features not found in the target, and the candidate with the largest number of matching features wins.

The notion that for f-morphemes Spell-Out is deterministic (not subject to choice) I take as a fundamental insight into the nature of the lexical/functional divide. In much that follows I will employ the determinism of insertion as the basic criterion for identifying the order of the vocabulary items in question. Whenever there is unforced choice – whenever the structuring of syntactically-relevant features underspecifies their Spell-Out – by hypothesis a Root is present.

L-morphemes (the “l” may stand for “licensed”, see Harley and Noyer (2000)) instead are encyclopedic items whose insertion is not dictated by the syntax – the grammaticality of a sentence is not affected by choice of l-morpheme. DM re-interprets the traditional lexical categories in terms of Roots and their governing f-morphemes. Namely, a noun is a Root whose nearest c-
commanding f-morpheme (its “licenser”) is a D element; a verb is a “little v” light verb taking a Root complement that is itself dominated by Tense and Aspect nodes, and so on. Verbs amount to functional elements that introduce events in the semantics, plus encyclopedic content derived from the incorporation of nominal elements with attendant restrictions on event predicates.

The DM proposal represents a radical reformulation of the lexical/functional divide and of the notion of category. All lexical categories are reduced to the single category Root. Everything else — all other units represented in syntax — are functional in the sense that they map to sets of UG features. Category itself is at best one among other features, though possibly merely a label for feature bundles. This is Embick:

[C]ategories are nothing more than bundles of features, an idea that is assumed (but not always fully appreciated) in many current theories of grammar. Put simply, traditional questions about lexical categories and their universality must be recast, because current theories do not make use of lexical categories in the traditional sense […] the distinction between Roots and functional morphemes replaces the distinction between lexical and functional categories. The Roots are by definition acategorial, and thus bear little resemblance to the members of the traditional lexical categories (Embick, 2012, pp. 73–74)

That said, while the absorption of morphology into syntax and post-syntactic phonological processes has important ramifications for morphologists and phonologists, in practice it is not always clear that DM’s re-architecture visibly alters the day-to-day work of the syntactician. For example, the impact of taking a verb as having “category” [v + Root] rather than V may be mostly theory-internal. When it comes to adpositions, DM practitioners have unfortunately said little.

Harley (2006) has noted that because Bare Phrase Structure does away with bar levels, it cannot reproduce the elegant explanations of one-replacement and unaccusativity afforded by X-bar theory. She goes on to invoke the DM notions of argument-taking acategorial roots and Vocabulary Insertion to update the analysis of these phenomena on Minimalist assumptions.
Harley and Noyer remark that

[T]he ramifications of the L-morpheme Hypothesis (according to which open-class Vocabulary Items always instantiate the same syntactic category) point to the need for continued study of so-called ‘mixed’ categories and the cross-linguistic validity, if any, of traditional part-of-speech labels in universal syntax (Heidi Harley & Noyer, 1999, p. 25)

Supposing that the grammar makes available a category-assigning, light, functional element \( p \) (“little” \( p \)) that applied to a Root returns a preposition, then functional and lexical prepositions might be structurally distinguished by the absence or presence, respectively, of a Root. That is, a functional preposition such as \( \text{of} \) could be inserted for a bare \( p^{16} \), while \([p \text{RootP} [\text{Root [complement]]}]\) would be realized by a lexical preposition. Little \( p \) would then represent an abstract relational unit, perhaps a RELATOR in the sense of Den Dikken (2006), discharging the basic function of Ps that Hagège highlights: to relate a phrase to a head. The descriptive content (typically spatiotemporal) of the relation would instead derive, for a lexical P, from the Root complement of \( p \). Thus Ursini and Akagi (2013) develop a type-logical extension of DM in which overt functional Ps and silent \( p \) heads have the type of relators, while lexical Ps are essentially argument-taking Root morphemes that require (overt or covert) abstract \( p \)-like objects in order to relate to their governors.

The trouble with this analysis, as far as determining the feature constitution of prepositions is concerned, is that little is added to a pre-DM representation based on features [+/- F] and [+/- N]. A finer-grained analysis is required to shed light on the precise behavior of individual prepositions, in order to establish the features required to mark the appropriate structural and semantic contrasts. One such effort has been the cartographic analysis of spatial Positions, to which we now turn.

---

16 In fact \( \text{of} \) may be so light that it evaporates from the derivation altogether. Harley in (2006) treats \( \text{of} \) as a “non-syntactic” reflex of inherent case having no node of its own.
The Cartography of Spatial $P$

By “cartography” I mean the many efforts since at least Pollock’s “explosion” of INFL (1989), that aim at discovering the fine syntactic structure of the extended projections of lexical categories. Work on $P$ by Koopman (2010), Svenonius (2003), (2006b), (2006c), (2010), Botwinik-Rotem (2004), Den Dikken (2010), Asbury (2008), Pantcheva (2010) has brought empirical precision and some theoretical convergence to the study of adpositions. In what follows I am only able scratch the surface of this rich body of work. The approach is to contrast the essays by Svenonius and Den Dikken in *Mapping Spatial PPs* (Cinque & Rizzi, 2010), since these take opposite positions on the question of whether or not all $P$s are functional, leading to sharply distinct views on the feature content of spatiotemporal $P$.

Svenonius (2010)

The model of spatial PP developed by Svenonius is based on Grimshaw’s analysis of extended projection of nouns. Spatial PP forms part of the extended projection of $N$, the latter being the nominal complement of $P$, which is the schematic Ground in Talmy’s sense (2000). A feature of Svenonius’ study of spatial PP is that the syntactic analysis works hand-in-hand with that of its compositional semantics, for which Svenonius draws upon Zwarts and Winter (2000). Very briefly, according to Zwarts and Winter:

1. A function $\text{loc}$ maps objects to regions (sets of points in 3-space) (type $<p, t>$, where $p$ is “point”)

2. $P$ denotes a function from regions to sets of vectors, i.e. vector spaces (type $< <p, t>, <v, t> >$ where $v$ is for “vector”)
3. PP denotes a function \( \text{loc} \)-mapping vector spaces to the objects contained in the region (type \( \langle \mathbf{v}, \mathbf{f} \rangle, \langle \mathbf{v}, \mathbf{f} \rangle \))

The composition of these functions gives us the set of entities (Figures) that are in such-and-such a relation (as described by \( P \)) to the set of entities comprising the ground.

Svenonius develops the following structure for the region of PP he calls Place:

\[ p \text{- Deg - Deix - Loc - AxPart - K - DP } \]

The DP is the Ground of the schema. K stands for “Kase” (case), about which Svenonius says little, other than to observe that there is evidence from several languages for a position above the DP that houses a genitive marker, as in *in front of the barn*. If genitive case can surface there, then it might be true that other oblique cases are inserted in that same position, so that K is associated with “functional prepositions and case markers” (Svenonius, 2010, p. 155) – thus, like Hagège, Svenonius acknowledges the overlapping semantics of oblique cases and functional Ps. It is important to highlight this point: Svenonius asserts that functional prepositions (and their corresponding cases) merge lower than locative \( P \), where, semantically, K denotes a function from Ground DP to regions, corresponding to Zwarts’ function \( \text{loc} \).

AxPart (“axial part”) is a nominal-like functional projection that picks out a subregion of the Ground DP, such as *front in to the front of the car* (Svenonius, 2006b). Loc is a function from regions to vector spaces (corresponding to Zwarts’ \( P \)) that, from what I can tell, for English at least is null. Deix (“deixis”) and Deg (“degree”) pick out the regions pointed to by the vectors, roughly corresponding to \( \text{loc} \).¹⁷

¹⁷ Svenonius glosses \( \text{loc} \) as a function from vector spaces to regions rather than sets of entities.
Finally, a functional projection is required to introduce the Figure, i.e. the “subject” of the PP.

Echoing the VoiceP of Kratzer (1996) but using the “little” notation of DM, Svenonius tops off the locative PP with a $p$ node hosting the adposition proper. “This $p$ is the natural locus of relational notions of containment, attachment, and support, which are commonly expressed by prepositions such as in and on and their counterparts cross-linguistically” (p. 133).

As for directional P, Svenonius theorizes a Path node housing a goal or source (to or from), and a Dir (“directional”) node above Path that in conjunction with Path specifies the set of points traversed. These sit atop the locative structure. So in all we have the f-seq:

(16) $\text{Dir} - \text{Path} - p - \text{Deg} - \text{Deix} - \text{Loc} - \text{AxPart} - K - \text{DP}$

For example, the PP in *the music poured out from deep inside the cave* might receive the following analysis.

(As Svenonius points out, some movement might be required prior to lexical insertion to ensure the right word order, e.g. incorporation of AxPart into $p$.)

(17) $[\text{Dir} \ out \ [\text{Path} \ from \ [p \ in\-] [\text{Deg} \ deep \ [\text{Loc} \ \emptyset \ [\text{AxPart} \ side \ [K \ \emptyset \ [\text{DP} \ the \ cave]]]]]]]]$

*Den Dikken (2010)*

Den Dikken’s analysis of spatial P explicitly extends and solidifies the seminal work on Dutch PPs by Hilda Koopman (2010). Both approach the problem in a directly syntactic manner – there is little discussion of the compositional semantics of PP – by investigating the rather complex system of Dutch adpositions. Koopman takes it as given that P is a lexical category rooting its own extended projection. The criterion appears to be semantic, as she draws a contrast between locative PPs and “empty” (i.e. functional) Ps, which “play a variety of roles as case markers and as Cs” (p. 61). She argues these units belong as much to the category P as locative Ps, and indeed base-generates
functional Ps in the canonical P position.

The contribution by Den Dikken (2010) constitutes, in his words, “an extended plea for the existence of P as a lexical category” (p.117). He lays bare a deep structural parallelism between two extended projections of spatial P and those of V and N. The complete functional sequence Den Dikken assigns to spatial P is:
$P_{dir}$ and $P_{loc}$ are the insertion points for *lexical* directional and locative Ps respectively, the latter taking the Ground DP as its complement. The remaining nodes reproduce Koopman’s labels: the specifiers...
of Place and C(Place), and Path and C(Path), provide landing sites for various merge and move operations exposed by the analysis of Dutch PPs; Deg nodes are for adverbial modifiers. The isomorphism between the extended projections of Place and Path are clear: each packages a lexical node, a functional node that also assigns the type of P involved (Path and Place), and Deg and C nodes parameterized for P type.

Den Dikken then proposes the following re-interpretations of the projections of spatial P. First, Path and Place can be thought of as aspectual nodes, indexed by P type. The applicability of the bounded/unbounded distinction is immediately clear for path-denoting PPs (walk into vs walk around), and, according to Den Dikken, similar arguments can be made for places. Second, rather than specifying separate Deg and Deix (“deixis”) heads as Svenonius does, Den Dikken suggests a single Dx (again, “deixis”) node that also hosts adverbial modifiers via adjunction. The C node topping off the projection remains, as its landing site is required, though its inherent featural content remains “largely obscure”.

The relabeled structure now looks a lot like (abstractions of) the extended projections we find in the nominal and verbal domains. The generic structure for some lexical category X structure is simply:

\[
(18) \quad [\text{CP} \ C \ [DxP \ P \ [AspP \ Asp \ [XP \ X \ \ldots \ ]]]]\]

The heads corresponding to the phrasal projections vary depending on X:

<table>
<thead>
<tr>
<th>CP</th>
<th>DxP</th>
<th>AspP</th>
<th>XP</th>
</tr>
</thead>
<tbody>
<tr>
<td>C_{FORCE}</td>
<td>Dx^{TENSE}</td>
<td>Asp^{EVENT}</td>
<td>[VP V \ \ldots \ ]</td>
</tr>
<tr>
<td>C_{DEF}</td>
<td>Dx^{PERSON}</td>
<td>Asp^{NUM}</td>
<td>[DP N \ \ldots \ ]</td>
</tr>
<tr>
<td>C_{SPACE}</td>
<td>Dx^{SPACE}</td>
<td>Asp^{SPACE}</td>
<td>[PP P \ \ldots \ ]</td>
</tr>
</tbody>
</table>
From here Den Dikken goes on to explore cross-domain similarities for more complex structures, e.g. by extending the analogy of V to P through an analysis of the similarities between nested clauses and the full structure of $P_{\text{dir}}$ over $P_{\text{loc}}$ in Figure 4. Details aside, the broader point of Den Dikken’s speculation is that the homologies across domains strongly support the lexical character of spatial adpositions.

**Discussion**

Since Koopman and Den Dikken deny that locative and directional Ps are functional in the first place, their analyses cannot shed light on the morphosyntactic, as opposed to semantic, feature content of spatiotemporal P. Neither is concerned with the syntax of functional P. Whereas Svenonius, we saw, follows Grimshaw’s lead in situating the entire P category within the extended projection of N. Unfortunately his analysis does not translate to specific information about the inherent features of individual P items. Unless the specific $p$ merged into the structure is sufficiently specified for features that would identify the one Vocabulary item to be spelled out, there is a choice about which P gets output (in or on etc.). But when there’s optionality, there’s a Root. The other option – full specification of the features on the individuals $p$ in the Lexicon (versus the Vocabulary) – certainly remains a possibility, though Svenonius gives us no clues as to what those might be. They cannot be semantic-internal since otherwise we either run into generative semantics, i.e. we would have the syntax dealing in features that do not concern it – a possibility that DM axiomatically excludes\(^{18}\) –, or the Ps would have to be listed in the Encyclopedia, making them Roots. What is

\(^{18}\) Arguably Svenonius is aware of these difficulties. He mentions the “rich descriptive content regarding spatial configuration” proper to spatial P. Then, when remarking on the essential difference between his and Den Dikken’s analysis with respect to the categorial status of P, Svenonius writes: “On the assumption that rich ‘encyclopedic’ or conceptual content can be associated with vocabulary items which are inserted under functional heads, there is no need for a special lexical root at the bottom of a sequence of functional heads” (p. 144). This cannot be right, since idioms from the Encyclopedia are only inserted at Root nodes, i.e. at lexical heads, which is the possibility Svenonius is trying to avoid. He needs the “rich descriptive content” of spatial Ps but his model is predicated on there being no projecting Roots other than the nominal at the bottom. Allowing the system to attach lexical content anywhere
thus required is further unpacking of at least the *p* node (possibly leading to further restructuring of PP) in order to determine whether sufficient structure can be found there that is syntactically justified, not just semantically.

Put differently, assuming with Grimshaw that features percolate up the extended projection, the analyses of PP discussed in this section do offer fairly rich feature content specifications for (spatial) PP as a whole. But note that the features are mostly *categorial*, the labels for the various positions in the structure, not inherent features of the specific lexical items. The fact that a PP has a Deix feature is a consequence of the merger of a head belonging to that category; it has no relation to the inherent features of the *P* head. So again, if the Embick is right in claiming that categories are just shorthand for sets of features, it is necessary to disassemble the categories themselves into fine-grained features.

A way forward suggests itself when we recall Hagège's analysis of the semantics of *P*: *Ps* and case affixes stand in complementary distribution with respect to spatiotemporal relations and their metaphoric displacements. The job of both case and *P* is to mark the semantic roles of arguments with respect to their predicates. We might therefore look to case for the features of *Ps*. The features of spatiotemporal *Ps* might then fall out of an analysis of the spatial cases, and those of functional *Ps* determined as the metaphoric transposition of the spatial elements. This is roughly the strategy of Asbury (2008), among others. Better put, the analysis of the semantics of spatial case yields a paradigmatic space which the Vocabularies of languages then express through some

---

19 Following Lestrade (2010), Grimm (2011) and Hagège (2010) I assume there is no distinction in principle between the marking of core arguments and oblique arguments, i.e. that the structural cases lie on a conceptual continuum with the other cases. In this perspective cases affixes emerge as highly grammaticalized, stereotyped and abstract expressions of argument functions, but adpositions can in principle substitute for all cases. Indeed Hagège presents evidence of languages that use *Ps* for marking core arguments.
mix of spatial affixes and adpositions.

This approach presents two problems for theories of P that view the category as only functional. First, as pointed out by Hagège, adpositional semantics are typically more articulated than is true of case affixes, and indeed a language typically has many more adpositions than cases – languages having an average of 6-7 cases (Lestrade, 2010). (For comparison, the corpus analyses of English and Spanish data described in part II found 86 Ps for English and 59 for Spanish.) Second, the interpretation of spatial cases necessarily presupposes that of the spatiotemporal configurations indicated by stative adpositions, because – this is the crucial point – the paradigmatic space of case incorporates the meanings of static spatiotemporal adpositions.

Consider Kracht's (2002) analysis of spatial language. In Kracht's terminology a locative expression is made up of three components: the Ground DP, a **localizer**, and a **modalizer**. The localizer specifies how several objects are positioned with respect to each other, the **configuration**; the modalizer describes the way an object (the Figure) moves with respect to the configuration. Thus the structure:

\[(19) \quad [\text{ModeP Modalizer} [\text{LocP Localizer DP}]]\]

In Kracht's proposal the number of possible modes is fixed cross-linguistically:

- **static**: the Figure remains in the same configuration over the course of the event;

- **cofinal**: the Figure moves into the configuration by the end of the event;

- **coinitial**: the Figure moves out of the configuration during the event;
• **transitory**: the Figure moves through the configuration during the event;

• **approximative**: the Figure approaches (or moves away from) the configuration during the event.

Lestrade (2010) further reduces the list to just the first three **directional** meanings (static, cofinal and coinitial), deriving the other modes from these primitives.

In contrast, there is no limit to the number of possible configurations. Configuration is complex, multidimensional, and culturally-specific, constructed out of a posteriori concepts such as vertical positioning (gravity), contact, containment, and many more. As Lestrade puts it, “there is no uniform set of configurational concepts that is similarly privileged in human cognition and language” (p. 134).

For languages that have them, spatial cases incorporate configuration and modes, as the *names* of many cases show us. Kracht points out these are often built from a preposition plus the suffix *-essive* or *-lative*, the former referring to static mode while the latter is used for the dynamic modes:

<table>
<thead>
<tr>
<th></th>
<th>Static</th>
<th>Coinitial</th>
<th>Cofinal</th>
</tr>
</thead>
<tbody>
<tr>
<td>in</td>
<td>inessive</td>
<td>elative</td>
<td>illative</td>
</tr>
<tr>
<td>at</td>
<td>adessive</td>
<td>ablatice</td>
<td>allative</td>
</tr>
<tr>
<td>on</td>
<td>superessive</td>
<td>delative</td>
<td>sublative</td>
</tr>
<tr>
<td>...</td>
<td>-essive</td>
<td>-lative</td>
<td>-lative</td>
</tr>
</tbody>
</table>

Hagège too exploits this latinate naming procedure to coin novel metalinguistic terminology for presumably unattested cases (which he equates with semantic functions), such as his *secutive* case,
signifying “follow a place or person” (Hagège, 2010, p. 286) – meanings that are expressed via adpositions. This is a paradigm of sorts, where the columns are fixed to the modes made available by UG, whereas the rows are not in principle limited. The point is that the semantic space underlying case/P is finite only in one dimension – Kracht's mode – but not the other, configuration. On these assumptions it not possible to determine a finite set of features underlying case phenomena that could then inform the feature constitution of all Ps. Both the static cases (e.g. inessive) and the dynamic cases (e.g. illative) depend for their interpretation on a configurational component that has all the hallmarks of lexical Roots: complex, multidimensional meaning, cultural specificity, and sensitivity to context.

Kracht's semantic model appears to parallel Den Dikken's syntactic theory of spatial P, in that both locative and directional PPs each embed a lexical P within an extended projection whose functional elements define the aspectual characteristics of the PP as a whole. On this homology the mode function is equated with aspect: static, cofinal, coinitial and the rest. This suggests a working hypothesis: the synsem features that are active in the PP, and that are therefore associated with purely functional Ps, are the features of prepositional aspect. These features have been studied by various researchers operating within the nanosyntactic current of Minimalism, especially Marina Pantcheva, Pavel Caha and Juan Romeu. We now turn to a review of their work.

**Nanosyntax**

Reflecting on the past few decades of syntactic research, Michal Starke writes: “[A]s syntactic structures grew, not only did their terminals become ‘much smaller’, they became submorphemic – smaller than individual morphemes” (Starke, 2009, p. 1). As we shall see below, Pantcheva (2010) analyses the structure of directional expressions as:
A monomorphemic P might spell out the Goal node, or the Source and Goal nodes together, or Route+Source+Goal, but not Route + Goal (skipping Source). Thus the “bundles” corresponding to a lexical item are not constructed as mere sets; rather, lexical items express features arranged in hierarchies as determined by UG. In the following example from the language Lak (Pantcheva, 2010), the prolative suffix, with Route meaning, lexicalizes the whole structure, since the superessive suffix \( j \) is a Place exponent.

(20)  `sa'ra – j – x`^

street – SUP - PROL

‘through the street’

This “surprisingly simple” result, while perfectly plausible to us today, strikes us nevertheless as somewhat scandalous, accustomed as we are to thinking of syntax as the organization of morphemes (if not fully accessorized words) into meaningful phrases. Distributed Morphology, though it defers until after the computation the associations of words/morphemes in flesh to syntactic bones, nevertheless imagines syntactic heads as sets of features – complex entities. But if atomic terminals are smaller than morphemes, then even when abstracted as mere sets of features, morphemes are no longer the atoms of syntactic organization. Structure is built out of more basic elements – call them features. “Syntax projects from single features and nothing else” (ibid, p. 6).
This basic precept of nanosyntax thus amounts to the “explosion” of DM’s syntactic nodes into microstructures in which each feature in the bundle is assigned its own distinct position in the structure. And since nanosyntax thinks of features as morpho-syntactic-semantic, “[t]he important question is what role in syntax each of the features has and how they can be motivated from the point of view of compositional semantics” (Pantcheva, 2010, p. 15).

In reviewing Svenonius (2010) we observed that the structures described by Svenonius are labeled by categories, not synsem-type features. If Embick is right that categories are mere labels for feature sets, then in nanosyntax the notion of category, at least in its traditional sense, may well evaporate as a theoretical primitive. It may persist for purposes of labeling nodes: if e.g. the p node cannot be analyzed into further components, and if it indicates a semantic effect that is not better described by a more specific meaning, then we might as well take it as a value of the traditional category feature. Nevertheless, while the papers by Pantcheva and Caha I have consulted label portions of the tree not under analysis using traditional names (V, DP etc.), none of the nodes of interest are categorial.

The notion of feature bundle has not disappeared from the theory altogether. The Vocabulary of DM – the mapping of sets of features to phonological forms – remains more-or-less unchanged, the main difference being that rather than merely listing individual morphemes, the Vocabulary may contain larger units, even idioms. What is gone is DM’s Lexicon, the sets of features made available by the language as inputs into syntax. Spell-Out still calls for the matching of the feature bundles specified by Vocabulary entries against the features in the syntactic derivation, but instead of lexicalizing the tree one node at a time, portions of the tree, spanning one or more nodes = features, are expressed by the best-fitting exponents. The fit is regulated by three principles:

4. **Superset Principle** (cf the Subset Principle of DM): A lexically stored tree matches a syntactic
node iff the lexically stored tree contains the syntactic node (Starke, 2009, p. 3).

5. **Biggest wins**: Spell-out is cyclic, working its way up the tree, and a later step overrides an earlier one. Therefore the largest possible match wins.

6. **Minimize Junk**: At each cycle, if several lexical items match the root node, the candidate with least unused nodes wins (ibid).

The Superset Principle implies that a morpheme may spell out several syntactic nodes. The second principle ensures that the morpheme matching the most nodes is spelled out at the end of the process, while **Minimize Junk**, which is just a statement of an elsewhere condition, ensures that the exponent that most tightly matches a set of nodes/features (without under-representation) is produced.

The nanosyntactic approach is elegantly demonstrated by Pantcheva’s (2010) analysis of Path expressions across a wide variety of languages. First, based in part on Zwarts’ semantic analysis of paths (Zwarts, 2005), Pantcheva enumerates the three types of expressions denoted by instances of Path: Goal, Source, and Route. Goal expressions are true when the Figure starts at a place away from the Place and ends at (in the relevant sense) the Place. Source expressions are the opposite of Goals, in that they are true when the Figure begins at the Place and then moves away. Routes are true when the Figure begins away from the Place, transitions to being at the Place, then moves away again. Source and Goal paths thus represent one “transition” (not-at -> at), while Routes have two, suggesting that the latter are complex and composed of a Goal path (not-at -> at) and a Source Path (at -> not-at). The central fact of natural language concerning Paths is that there appear to be no spatial markers that chain Source and Goal in that order; no single expression meaning “X started at
P, ended at Q”.

From here Pantcheva marshals a variety of cross-linguistic data to propose the explosion of Path into nodes representing the three path types she analyses. In particular, her evidence suggests a layered structure that orders the nodes by increasing complexity:

7. [Goal [Place … Ground]]

8. [Source [Goal [Place … Ground]]]

9. [Route [Source [Goal [Place … Ground]]]]

There is a monotonic implicational hierarchy here: the presence of a Source node implies that of a Goal, and Path implies both Source and Goal. This is shown nicely by the spatial markers in Quechua, whose ablative suffix (indicating Source) -man-ta contains the allative (Goal) suffix -man (2010, p. 12). Crucially, a strength of this model is that the acquisition task is greatly simplified; we will return to this point in Chapter 6.

The semantic interpretation of the exploded Path follows directly. If Place denotes a region in space (Zwarts & Winter, 2000), then Goal describes the transition to that region. Source simply encodes the reversal of the Goal relation, movement away from the Place; it is therefore an operation of semantic inversion, similar to negation. Route is a second transition, representing movement to the (starting position) of its complement, i.e. the Source, thus capturing the sequence Goal + Source, or not-at -> at -> not-at. Because this sequence is fixed and universal, it impossible to encode (without recursion) the series at -> not-at -> at, as this would require a transition to a Goal, implying that Route take Goal as its complement.
The observation that Source behaves as a negation of sorts and the similarity between Route and Goal (both denote transitions) leads Lestrade to abstract the terms even further: Goal and Route are simply Transition nodes, and Source a Negation, giving the basic structure as:

(21) [Negation [Transition [Place]]]

By this schema a Route would correspond to an additional Transition node merging above the Negation. At this level of abstraction we are transparently dealing with the way in which language describes change – aspect. As it happens, in (2011) Pantcheva tops off the structure with two unambiguously aspectual elements: Scale and Bound. Scale makes a transitional (telic) path non-transitional (imperfective), while Bound indicates whether a path ends (or begins) at the transition point – the aspectual notion of boundedness.

Aside from its value as a cogent demonstration of nanosyntactic thinking, Pantcheva’s deconstruction of Path relates directly to the question concerning the features of P. She has broken out the functional Path head of, say, Svenonius (2010) so that we can now assign distinct positions for English from and to, and furthermore we can assert that the feature set corresponding to from contains at least \{SOURCE, GOAL, PLACE\} \textsuperscript{20} while that of to does not have SOURCE. This constitutes a major step forward, since now to and from need not compete for insertion into a single Path node.

Fundamental issues remain, however. First, in English at least there are many more than three direction-denoting adpositions, e.g. over, by, round, past, through, under\textsuperscript{21}. Observe that Pantcheva mostly

\textsuperscript{20} The features at each node are not the same as the case labels, so this is a simplification. See the discussion of Caha below.

\textsuperscript{21} Some directional Ps in English seem to be derived from primitive forms, e.g. towards, around, along (Den Dikken, p.c., April 9, 2015.). Gravity-dependent Ps like up and down might be analyzed as small-clause resultatives, as in the stroller bounced down the steps.
deals with spatial case markers, which are fewer and more abstract. When applied to adpositional systems the problems of competition and lexical choice persist. Second, the question of the nature of P as a category remains. Consider how Pantcheva’s analysis might be integrated into Den Dikken’s model of the PP, in which Koopman’s Path is re-interpreted as a (spatial) aspectual category. Again, Pantcheva’s Route/Source/Goal distinction is open to an aspectual interpretation, in terms of the ontological characteristics assigned to paths by their descriptions (Zwarts, 2005). So even on her analysis it remains plausible for the richness and complexity of spatial meanings proper to Ps to be encoded in the lexical values at $P_{\text{dir}}$ and $P_{\text{loc}}$, while functional properties such as the aspect of specific uses is determined by the application of aspectual functions to the general lexical meanings of the roots. We can still view to and from as functional, namely denoting aspectual relations, where e.g. the benefactive or goal sense of to results from the type of Figure and Place, and where the lexical P is an empty Root.

Similar issues resurface in Caha’s nanosyntactic investigations of case (Caha, 2009, 2011, 2012). In his analysis of case in Classical Armenian, Caha exploits the patterns of syncretism evident in the Armenian paradigms to derive an f-seq for case in general (not just spatial case). The key observation is that there appears to be an (at least quasi-)universal restriction on how languages exploit the possibility of syncretism. Caha formalizes the observation into a universal principle:

**Universal (Case) Contiguity**

a. Non-accidental case syncretism targets contiguous regions in a sequence invariant across languages.

b. The Case sequence: Comitative (COM) > Instrumental (INS) > Dative (DAT) > Genitive (GEN) > Locative (LOC) > Accusative (ACC) > Nominative (NOM)
This means that syncretism can only hold at a boundary in the f-seq\textsuperscript{22}. If, for example, there is syncretism between the genitive and the nominative, then necessarily there is also syncretism between nominative and accusative. Succinctly: *A-B-A. The case sequence is read as: COM merges above INS, which merges above DAT, and so on.

Synthetically, Caha develops the f-seq by assuming the (a) clause of his principle and by then studying the patterns of syncretism in the target language, shown in Table 3.

\begin{table}[h]
\centering
\caption{The restrictions on syncretism in Classical Armenian, from \cite{Caha, 2012}}
\begin{tabular}{|c|c|c|c|c|}
\hline
 & \textbf{N.A.} & \textbf{spirit (SG.)} & \textbf{N.A.} & \textbf{word (SG.)} & \textbf{N.A.} & \textbf{Tigran (SG.)} \\
\hline
\textbf{NOM} & A & hogi-\textvar{} & ... & bay-\textvar{} & ... & Tigran-\textvar{} \\
\textbf{ACC} & B & hogi-\textvar{} & ... & bay-\textvar{} & ... & Tigran-\textvar{} \\
\textbf{LOC} & A & hogi-\textvar{} & A & bay-i & A & Tigran-ay \\
\textbf{GEN} & A & hogw-oy & B & bay-i & B & Tigran-ay \\
\textbf{DAT} & B & hogw-oy & A & bay-i & B & Tigran-ay \\
\textbf{ABL} & A & hogw-oy & ... & bay-\textvar{\textvar{}} & A & Tigran-ay \\
\textbf{INS} & ... & hogw-oy & ... & bay-iw & ... & Tigran-aw \\
\hline
\end{tabular}
\end{table}

Since NOM is syncretic exclusively with ACC, these two cases share a boundary in the f-seq. The same applies to LOC and ACC (but not LOC and NOM), so they too must be neighbors. And so on. Drawing on evidence from languages from distinct families, Caha makes his case for the f-seq he proposes.

By the nanosyntactic assumption, each position in the f-seq translates to a node in the tree that contributes one feature to the extended projection. And if features percolate upwards as Grimshaw assumed, then we would expect that all the features at node X would also be specified for the next

\textsuperscript{22} Strictly speaking, at this stage of his exposition Caha concludes only that there is a sequence of cases and that their features are cumulative. He later shows that the sequence maps to a functional sequence in the cartographic sense, based on a careful analysis of case attraction in Armenian and Amharic.
node up in the f-seq. That is, if NOM has features \{A\}, then ACC has \{A, B\}, LOC \{A, B, C\} and so on. Feature bundles are cumulative. The feature added at each node is not itself the case represented by that node; rather, the case is the label for the full set of features at each position, the non-terminal projected by the additional feature. Caha sketches out a few ideas about possible synsem contents at each node. First he divides the cases into four zones:

Zone 1: NOM - ACC (structural cases)

Zone 2: LOC - GEN and Partitive (PART) (stative cases)

Zone 3: DAT (goal case)

Zone 4: Ablative (ABL) - INS (source cases) - COM

NOM and ACC are structural, with ACC adding the sense of “dependent”. LOC and GEN denote static location and static ownership. DAT adds a change of state, a transition, such as change of position or of ownership. INS, COM and ABL describe more complex changes in state, along the lines of Pantcheva’s proposal for Source.

Caha is suggesting points of contact between his analysis of case and work on P in general (see also Krifka, 1998 for a unified treatment of change across spatial and non-spatial domains). Setting aside the structural cases in the first zone, we see links between the remaining zones and the analyses of PP discussed previously. Zones 3 and 4 map to Path (or Asp\textsuperscript{PATH}), while Zone 2 cases, since they denote states, presumably including part/whole structures, may relate to Place, e.g. Svenonius’ K for GEN and Place for LOC, or Asp\textsuperscript{PLACE} in Den Dikken’s model – see Kiparsky (1998) on the relationship between partitive case and aspect. On the latter analysis PPs like \textit{out of ten students} might involve \textit{out} in Path, possibly in Asp\textsuperscript{PATH} as a Source expression depending on whether or not extra
descriptive content is intended (in which case it would merge as $P_{dir}$), and $of$ in $Asp^{PLACE}$.

These considerations suggest the merger of the functional sequence(s) of $P$ and that of case, as argued for by Asbury in her dissertation (Asbury, 2008). That is, the oblique cases and the functional shells of spatiotemporal Ps (and cases) are in reality expressions of the same underlying universal features, and thus compose into the same positions in the tree. One immediate result is that case heads are merged in higher than is traditionally held and re-asserted by Svenonius, that is, higher than a function applying to DPs (Caha, 2011). And if, with Hagège and others, we take functional Ps as competing with oblique cases as realization strategies for a single set of synsem features, then we can take Caha’s case hierarchy as also a hypothesis about the f-seq of functional prepositions. In fact, it may be possible to incorporate Pantcheva’s analysis of path $P$ into this same sequence: Goal mapping to dative case and Source to instrumental (in Latin the ablative can also express instrumental meanings). Route meanings, such as “by way of”, “through”, “past” are sometimes expressed using prolative (e.g. Finnish) and perlative cases, which also have an instrumental sense. Whether or not Route corresponds to comitative case is not clear, complicated by the fact that comitative and instrumental markers often coincide (e.g. English “with”, Spanish “con”). For now I will simply assume the two are distinct and that the case sequence accommodates Pantcheva’s decomposition without remainder.

So much for functional $P$. The question of the status of basic $P$ exponents like $in$ remains. The standard criteria for identifying functional elements seem to not do their job when applied to adpositions: even the stringent feature-based criteria that have emerged in Distributed Morphology and Nanosyntax – competitive post-syntactic insertion governed respectively by the Subset and

---

23 In a footnote to (2011) Caha states his proposal is “sympathetic” and “compatible” with Pantcheva's work, understood as the decomposition of Mode.
Superset Principles, implying that whenever there is ambiguity is the selection of an exponent, a lexical Root is present – fail to provide us with a clear answer. After all, it is not clear that there is much optionality in the choice of many Spatial Ps, even though these represent the obvious candidates for lexical Ps (and explicitly so for Koopman (2000) and Den Dikken). How many ways are there of saying “X in Y” without using the morpheme in? On the other hand, one senses an insistent halo of optionality around even the best candidates for functional P. The idiosyncratic uses of abstract Ps, in languages that have them, are known to be difficult to master for L2 learners, and even native speakers might disagree on particular uses – is it research on or research in or research about or research into? Doubtless the choice depends in part on the communicative intent, but as Figure 5 (generated using Google's ngram viewer) shows, the relative preferences for this example do change over time, and synchronically there is always choice.

![Figure 5: Relative frequencies of prepositions selected by "research" over time.](image)

Practitioners of nanosyntactic approaches regard Ps as uniformly functional. Pantcheva seems to assume that Place is topped by the \( p \) of Svenonius (2010). Caha produces a detailed argument from the ban on remnant movement to defend his proposal that
sequences never occur other than in recursive P structures. That P is never lexical is a pillar of his argument; if P could move on its own, leaving DP behind, then P-over-K as in out of the pan could be modeled without recursion – and indeed such an analysis would not be dissimilar from traditional ones of into that base-generate to in Path, in in Place, then displace the latter over the former.

The “merged” case + spatial P f-seq that emerges might be sketched as follows, where Degree, Deixis and structural Case are left out for simplicity:

(23) COM/Route - INST/Source - DAT/Goal - LOC – GEN-PART/Stative - p - AxPart - DP

The basic problem with this scheme, again, is that there are too many actual Ps in language for their insertion to be deterministic. In particular, if p is where simple locative Ps like in and on are inserted, then the syntactic system cannot determine which one must be selected. The choice is semantic, contradicting the claim that the whole f-seq is an extended projection of the Ground DP. We can find unique locations for functional Ps: to in DAT, from in INST, of in GEN/PART, with in COM and/or INST. But spatiotemporal terms are underdetermined by the structure. The same is true for AxPart. Svenonius houses a wide variety of terms at that position, but fails to specify how the syntax might select between front, -side, below, under, -bove and so on.

A theory that does not try to decompose spatial Ps into syntactically relevant semantemes would instead introduce one or two roots into the P system. In Den Dikken’s approach, for example, if we decompose his AspSPACE into portions of the f-seq above, we get:

(24) CPSPACE - DxSPACE - AspSPACE = [COM - INST - DAT] - Pdir - DP
Lifting the pressure to find synsem features to discriminate between all possible spatial senses results in simpler structures that, as Den Dikken emphasizes, find reflexes in the verbal and nominal domain. This view, I suggest, is consistent with a long tradition of epistemology going back at least to Kant, according to whom spatial relations, while universal, belong to the faculty of Sensibility and not to the logical categories of the Understanding – they are of a fundamentally different sort.

Recent nanosyntactic work by Romeu (2014) does attempt a decomposition of individual spatial Ps in Spanish and to a degree in English. In Chapter 6, in which I present evidence from a distributional analysis of English and Spanish corpora to help inform the question concerning the status of P, Romeu’s analysis forms the basis of a predicted distributional pattern that is then submitted to a distributional algorithm for verification. A summary of his model concludes the present chapter.

Romeu takes the nanosyntactic framework as previously described as his starting point, but crucially supplements it with a notion of modification that is central to his decomposition of Spanish P. Because he factors several synsem features out of the main f-seq, he is able to reduce the f-seq to just three nodes:

- **Rel(ación)** “Relation”: A totally abstract position that introduces the Figure in its specifier. Rel might seem close to \( p \) in Svenonius (2010), Romeu intends relation in the most general sense, along the lines of Den Dikken’s (2006) RELATOR while for Svenonius \( p \) is still specific to locative P.

- **AxPart**: Here Romeu remains close to Svenonius’ notion. AxParts are functional elements
resembling Ps, Ns and adverbs that specify the relevant subregion of the Ground.

• \textit{Reg(ión)} “Region”: this is Zwarts’ \textit{Loc}, a function from DPs to regions (sets of points). By factoring the notions of Degree and Measure (and Deixsis) into modifiers and determining that projective terms (in Talmy’s sense) are inherently measurable, Romeu no longer requires the transformation of regions to vectors and back that Zwarts and others needed in order for the restrictions denoted by those notions to apply. The universal spine of the PP is thus:

\begin{center}
\includegraphics[width=0.5\textwidth]{diagram.png}
\end{center}

Modifiers are elements that in principle merge as specifiers of \textit{Rel}, \textit{Reg} and \textit{AxPart}. Modifiers may themselves be modified in their own Spec. Romeu is careful to claim that the set of modifiers are universals, thus skirting the heterodoxy of generative semantics. Right from the outset he wants to make it clear that modifiers are not a backdoor mechanism by which lexical content is manipulated by the syntax:

[I]t is important to signal that these modifiers are \textit{not} [my emphasis] where encyclopedic meaning is encoded. In opposition to generative semantics, this thesis contends that encyclopedic meaning is not present in the syntactic-semantic structure. Modifiers do not encode encyclopedic meaning, but rather general semantic notions that restrict the properties of the semantic primitives of heads, and that can condition the grammatical \textit{[c-]selection} properties of the element with which they combine. (Romeu, 2014, p. p.30,
The last point, that modifiers decorate functional heads in the f-seq with features that then affect the selections made by those heads, constitutes one of the strengths Romeu sees in his approach. He argues that when cartographic analysis explodes single nodes into segments of functional heads, the selectional restrictions of the node in question may no longer apply to the right complement because of newly discovered intervening material. The mechanism of modification, paired with the upward percolation of features, is able to shift the selectional properties to the appropriate head without upsetting the overall f-seq.

A second strength of his approach, Romeu claims, is that since modifiers may potentially merge anywhere in the f-seq (subject to semantic restrictions, some of which he discusses), it becomes easier to represent recurring aspects of meaning without forcing complex duplications of structure throughout the f-seq. A further advantage is that modifiers enable us to make a distinction between units that represent type-altering functions and those that do not change the semantic type but merely restrict it. The spine of the f-seq is constituted by functions of the former group; the rest are best modeled as modifiers.

Now, if modifiers can attach anywhere and do not alter the semantic type of the modifiee, one might ask: how do the compositional semantics handle them? Since they do not lie in the f-seq modifiers cannot be functions. There are two possible mechanisms: (i) modifiers merge through intersective modification, or (ii) as arguments of functions. In the first case there would need to be one modifier of each type for each modifiable semantic type, since intersective modification is only possible with items of the same type. But in this case, what other than a label would they have in

---

24 In his text Romeu does not explicitly make the case for partitioning semantic functions this way; thus I may be stretching the point somewhat — hopefully in the spirit of his proposal.
common? That is, Dis-junto applied to Rel may mean something quite removed from the same applied to Reg. In the second case we would have to either hypothesize a separate f-seq head for each modifier, minimally two (a one-place [no modifier] and a two-place functional head) if all modifiers have the same semantic type, more if the modifiers have varying semantic types; and/or hypothesize a null modifier for the base cases. So from a semantic angle, the modification strategy entails a (minor) explosion of functional elements.

The principal modifiers proposed by Romeu for spatial P are:

- **Con-junto** “Con-joint”: the modified element is included in or coincides with another (the Ground).

- **Dis-junto** “Dis-joint”: the modified element is the second in an interval defined by its juxtaposition with another identified element.

- **PuntoEscalar** “ScalarPoint”: the modified element belongs to a scale, and is a point in that scale, not the scale itself. Which point is defined by a modifier of *PuntoEscalar* itself.

- **Dispersion** “Dispersion”: the modified element is divided into multiple points.

- There are also modifiers for Deixis, Measure and Degree but I will not touch on these, as they do not affect the feature makeup of core adpositions.

Let us now view the system in action by examining a few examples of Romeu’s analyses. The simplest case is Spanish *de* “of”, which is the most bleached of the language’s Ps, denoting relation in a very general sense, similar to English *of* but of wider distribution. *De* simply lexicalizes *Rel.*
Spanish *en* “in” expresses coincidence in a highly abstract manner, more broadly than English *in*. For example:

(26) El dibujo está en la pared

  The drawing is in the wall

  “The drawing is on the wall”

In English, with *in*, the sense is the drawing was stuffed into a crevice in the wall, or plastered over into the wall; while in Spanish the interpretation is the standard one. Since *en* adds the general sense of spatial coincidence to that of relation, its structure adds the modifier *Con-junto* in the specifier of *Rel*.

For a more complex structure we turn to the analysis of *para*. This is a highly polysemous item in Spanish, incorporating senses benefactive (*es para ella* “it’s for her”), purposive (*es para escribir* “it’s for writing”), directional (*el tren para Barcelona* “the train to/for Barcelona”; *va para la iglesia*, “goes toward the church”) and many others. While in its directional senses *para* often competes with *hacia*

```
\begin{center}
\begin{dependency}
  \依存头{
    \依存前{RelP}{},
    \依存后{\textit{Con-junto}}{Rel'},
    \依存前{\textit{Rel}}{}
  }
\end{dependency}
\end{center}
```
“toward”, in that they are often interchangeable, Romeu draws out subtle differences tied to their behavior when selected by verbs yielding resultative interpretations, and shows how his framework is able to represent the relevant distinctions. The structure of para composes its etymological roots in por, which contains the modifier Dispersión indicating a set of points, and a, which signals a transition by means of Dis-junto. Together these adornments, both modifications of Rel, support the interpretation that the Figure will establish a relation with the Ground/Goal, which must be represented (by Dis-junto) and that the Figure is on a path facing the Goal (Dispersión).

Hacia too signifies a transition, in that in the prototypical case the Figure starts out separated from the Goal, but here the aspect is unbounded – the Figure never coincides with the Goal, hence it remains facing the Goal. The fact of this orientation is fundamental to its meaning, Romeu suggests, but cannot be modeled as Dispersión since the latter implies a path that does terminate in the Goal. Instead, Romeu modifies Rel with what would appear to be a semantic-internal feature, [faz][face].

It seems to me that the semantic nature of the additional modifier faz is a symptom of the difficulties faced by theories of spatial P that commit to the view that all Ps are functional. In this example we see that the lexicalization of the structure (choosing between para and hacia) is determined by whether the second [Spec, Rel] contains faz or Dispersión – a contrast involving what is

---

25 Romeu references Tortora (2008) in suggesting that there is a meaning of futurity in Spanish a (as in Italian), tied to the feature Dis-junto, while a bare Rel bears a relation to the past, in that Rel specifies the initial point of an interval and Dis-junto enforces the availability of a (future) Goal.
likely a purely semantic feature, contradicting Romeu's insistence that modifiers are not encyclopedic. Another example of such contagion arises from the attempt to extend the analysis of *en* to English *in*. Because the latter requires a notion of “inside-ness” (not mere coincidence), in (2012) Romeu specifies an additional modifier [interiority]. Now, languages and cultures differ in how they interpret interiority (Talmy, 2000), thus any such universal notion must be highly abstract; indistinct, perhaps, from the aspectual notion of boundedness. But then we lose precisely the featural specificity required for deterministic Spell-Out. So despite the initial protestations that this theory is not generative semantics, elements such as *faz* force us to question whether it is truly possible to model all spatial contrasts in language solely by means of synsem features. For one, the suspiciously lexical-seeming AxParts category seems to act as a bin for all less abstract Ps. For another, even when augmented by Romeu's machinery the theory is still unable to identify a structural distinction between pairs of essential Ps (Svenonius, 2010). Romeu assigns exactly the same structure, a bare Rel, to *de* and to *sobre* “on, over, above”, admitting that “[t]he difference between the two is due to *sobre*’s being related to certain conceptual information not relevant to the syntax” (p. 136). Why not hypothesize a modifier such as [superiority] to capture the distinction?

That said, Romeu’s work is invaluable since it constitutes a set of specific and testable Vocabulary entries for spatial Ps. We will return to this in Chapter 6.

**Summary**

In exploring ways in which the field has theorized the features of P, this chapter has described an evolution of sorts, from the basic [-N, -V] feature bundle of GB to the articulated structures we find in recent cartographic and nanosyntactic treatments of adpositional phrases. The latter agree that PPs contain functional positions above (and sometimes below) the position of P, but differ on
whether P roots its own extended projection or occupies a functional projection within the extended projection of N.

Three proposals have emerged:

1. If we take Pantcheva's model of directional P and Caha's functional sequence of case expressions to both constitute the structure of the semantic Mode of Ps (using Kracht's terminology), and then merge these with the structure of Place suggested by Svenonius, we have the following extended projection (23) repeated here:

   (23) \[ \text{COM/Route} - \text{INST/Source} - \text{DAT/Goal} - \text{LOC} - \text{GEN/PART} - p - \text{AxPart} - \text{DP} \]

   In this model the p position hosts spatiotemporal stative Ps, while functional Ps lexicalize the nodes labeled with cases/directionals above it. I have suggested that because the common element between case-type meanings and directional meanings is the aspectual component – how the grammar assigns relevant structures to the situation (events, space, time, objects) – the features expressed by functional Ps be referred to more generally as features of adpositional aspect.

2. Koopman, Den Dikken and, working from somewhat different assumptions, Lestrade (and Kracht) instead hold that because of the complexity of configurational meanings, the spatial P items expressing them are non-functional. Den Dikken nevertheless posits the existence of aspectual nodes in the extended projections of locative and directional P. Expanding those positions with the Pantcheva/Caha sequence yields:

   (24) \[ \text{CP}^{\text{PATH}} - \text{Dx}^{\text{PATH}} - \text{Asp}^{\text{PATH}} = [\text{COM/Route} - \text{INST/Source} - \text{DAT/Goal}] - \text{P}_{\text{dir}} - \text{DP} \]

   (25) \[ \text{CP}^{\text{PLACE}} - \text{Dx}^{\text{PLACE}} - \text{Asp}^{\text{PLACE}} = [\text{LOC} - \text{GEN/PART/Stative}] - \text{P}_{\text{loc}} - \text{DP} \]
3. Finally, Romeu's analysis of Spanish spatial P yield a simpler universal structure (before the modifiers he proposes):

\[(27) \quad \text{Relator} - \text{AxPart} - \text{Region} - \text{DP}\]

These hypotheses will be put to the test in Chapter 6 by means of an empirical procedure based on the distributional analysis of P usage in textual corpora. In the next chapters we examine theoretical and empirical arguments regarding the acquisition of P that illuminate the perennial question about the categorial status of P.
Chapter 2. Constructivism

Introduction

In this chapter and the two following we approach the question of the acquisition of functional categories, with specific emphasis on functional P. The discussion up to this point – from Leibniz through nanosyntax – has assumed an essentially rationalist perspective on language: the hypothesis of Universal Grammar. As we turn to questions of acquisition a rather different philosophical tradition comes into focus. Drawing on the empiricism of Locke and Hume and the psychology of Piaget, Constructivism in language research emerges as a fundamentally distinct interpretation of linguistic phenomena. I will argue that a central disagreement between constructivists and rationalists concerns the conditions of possibility of abstract representations in general. Constructivists believe that all abstractions emerge gradually as products of complex developmental processes; abstraction is hard. Rationalists, in contrast, hold that at least some complex representations are available to humans a priori, i.e. before experience itself; such representations are then by definition abstract since their content precedes the particularity of experience. If the category of P is at least partly functional, and if functional categories constitute abstract representations within the system of syntax, then constructivists would predict that functional P is acquired gradually, more slowly than lexical P.

In order to emphasize the magnitude of the disagreement over the nature of abstract representations, I begin with a brief consideration of the intellectual roots of constructivism in language research, which I locate in the research program Piaget called “genetic epistemology”.
Origins

Descartes’ radical skepticism argues that if in perception we never actually have (in hand, as it were) the qualities of the objects of perception, but only representations of these qualities – this is the “copy” theory of knowledge – then our representations of the world might be wrong. Under a correspondence notion of truth the copies must be verified. Yet knowledge refers representations to other representations (the copies), “models”, not the actual objects. Piaget begins here: knowledge is caught in vicious circle. The circle is vicious because as we go around it nothing ever changes, the correspondence between idea and reality is not strengthened – whereas in a virtuous circle, on every turn something is gained.

Kant solves this problem by consigning objects to an unknowable substrate that induces unstructured sensations in the subject. The subject organizes and classifies the input by applying its concepts and knowledge of space-time. The verification of models is then on the side of the subject and is transcendentally guaranteed by the affinity of our cognitive faculties to reality.

In sharp contrast, Piaget’s solution is essentially anti-Platonic, empiricist, and opposed to innatism. He wants to interrupt the vicious cycle of skepticism by historicizing the relation between self and world within the life of the individual and in that of the culture. Knowledge is primarily a matter of development.

Piaget’s action-oriented epistemology posits an early pre-representational stage of development of the human animal, the “sensorimotor” stage. The human subject does not create mental images right from the start. If there are no representations of the world, then the problem of circularity does not arise. The infant interacts with the world it inhabits, and learns via operant conditioning how to
do things that give it pleasure; but initially this learning is restricted to its actions in the world. Through its actions the child transforms the world and obtains its rewards.

I think that human knowledge is essentially active. To know is to assimilate reality into systems of transformations. To know is to transform reality in order to understand how a certain state is brought about (Piaget, 1970, p. 15).

By experimenting with reality the child develops schemes, which are “whatever is repeatable and generalizable in actions” (p. 42). Action schemes specify the recognition markers of applicable situations, the activities demanded by the situation, and the expectation of the reward (von Glaserfeld, 2002). And since actions typically compose simpler actions, schemes have structure. Having granted the child, at some point in its growth, a representational capacity and memory – perhaps because these capacities mature in the biological individual – Piaget says: the child not only learns about specific actions, but also about the relations between multiple actions as exercised in the world. The objective structures of actions are reflected back into the mind, to form the bases of logical thinking. That is, the part-whole structure of actions is abstracted as classification – and thus categorization – and the sequencing of actions is reflected as the mental process of ordering. So by doing, by acting on the world and transforming it, and subsequently by abstracting, the child learns about kinds of logical structures. The child’s understanding of the world is thus mediated by generalizing from the structures of actions. The crucial developmental notion is that logic itself emerges, every time again in the life of each individual, as an effect of experience.

Remarkably, not even time is an innate notion. A purely topographic, relational spatial knowledge (i.e. non quantitative, so no vectors) is available early since, Piaget says, space is conceptually simpler, being reversible and perceivable without reconstruction from memory (p. 60). Basic spatial
awareness allows the child to determine the *speeds* of objects relative to each other. Time is then abstracted from relations of speed:

> It is very easy to show that in the development of the notion of time in small children this relationship is not a primitive intuition. Judgments of time are based on how much has been accomplished or on how fast an action has taken place, without the two necessarily having been put into a relationship with one another (p. 70).

The distance between Piaget's epistemology and those of the rationalists and of Kant cannot be exaggerated. True to his empiricist roots, Piaget thinks of the initial state of the child as an almost perfectly clean slate. The child begins as a little behaviorist learner, needing only the tools required for operant conditioning: recall of experience, the ability to compare conditions in experience, and the capacity for preferential judgments. But the child has neither logic nor language, both of which arise much later – even at six years old Piaget consigns the child's reason to a "semilogic" – and indeed Piaget likely underestimated the syntactic sophistication of younger children.

> Chomsky goes so far as to say that the kernel of reason on which the grammar of language is constructed is innate […] I think that this hypothesis is unnecessary […] I agree that the structures that are available to a child at the age of fourteen to sixteen months are the intellectual basis upon which language can develop, but I deny that these structures are innate (p. 47).

In the constructivist perspective, learning is tied to a notion of systemic equilibrium. As long as the knowledge structures created by the subject work well in the subject's interactions with the world, experience is *assimilated* under the subject's existing schemes – the "percepts" of experience (themselves constructs, see von Glaserfeld (1974)) are modulated to fit their concept. Biological development and changing environments inevitably result in the inviability of earlier schemes. By
means of the mechanism of accommodation the subject reconfigures its conceptual structures to fit the precepts, giving rise to more complex knowledge. Later structures incorporate those built earlier:

> [I]t will be necessary, at least with regard to the building up of new structures, to distinguish two levels of regulation. On the one level the regulation remains internal to the already formed or nearly completed structure and, thus, constitutes its self-regulation, leading to a state of equilibrium when this self-regulation is achieved. On the other level, the regulation plays a part in the building up of new structures, by incorporating one or more previously built-up structures and integrating them as sub-structures into larger ones (Piaget, 1968 translated by von Glaserfeld).

We see then how constructivism views language, logic, time (and possibly three-dimensional space) as by-products of ontogenetic processes of assimilation and accommodation driven by the functional needs of the individual. In viewing knowledge in general as the sum of functional adaptations, constructivism in effect psychologizes epistemology, per Piaget's stated principle of “taking psychology seriously” (Piaget, 1970, p. 7). This psychologizing tendency emerges also in Tomasello’s work in language acquisition:

> [The] usage-based principle – that people form constructions at different levels of abstraction and use them at different levels of abstraction as well – underlines the fact that *what we are dealing with here is not formal linguistics but rather a kind of psycholinguistics* (Tomasello, 2003, p. 108 emphasis mine).

Just as with Piaget epistemology becomes developmental psychology, by which newer concepts are gradually built up by extending yet preserving older concepts, Tomasello would replace linguistics with “a kind of” psychology of language, whose only formal notion is that of the “construction”. His approach is psychological because, unlike theory-oriented linguists, “we”, i.e. today’s empiricists,
“are not concerned with providing a set of algebraic rules that covers the data in the most elegant manner possible; we are concerned with how people use a natural language” (p. 109). Parsimony, on this view, is merely an abstract criterion of the mature intellect, of interest perhaps to the theorist but not necessarily a general principle of psychology. It may therefore well turn out that the practical ways in which humans learn and use language are marked by systematic inefficiency. The incremental development of linguistic knowledge and the persistence of earlier representations in fact necessarily entail the suspension of economic constraints on the overall grammar, at least with respect to its representations, i.e. its long-term memory load. If, as suggested by constructivist linguists, a mature speaker is in principle able to generate multiple analyses for a single expression-meaning pair – some based on earlier, more concrete constructs, others on more abstract rules – then we must set aside the assumption that the syntactic system is representationally optimal. And if the grammar amounts to just what people say, without underlying representations, then Saussure’s distinction between *langue* and *parole* – arguably one of the foundational principles of modern linguistics – is lost.

It is on this thoroughly Piagetian foundation that the theoretical confluence of empiricist notions of language acquisition and construction grammar build their argument against generative linguistics. In the next section I will develop an exposition of the theoretical ensemble Tomasello offers in (2003), a work that presents itself as a *summa* of constructivism under the banner of usage-based linguistics (UBL). The key move is to recast the *final state* in a constructivist mold. Rather than thinking of adult grammar as a set of theoretically efficient, maximally general abstract rules whose acquisition must then be explained, the UBL viewpoint thinks of adult grammar as the accretion of heterogeneous constructs induced over the history of the individual. This argument, I suggest, establishes a necessary correlation between the anti-nativist stance of empiricism and the
constructivist model of the final state. Consider the following argument, where x scopes over linguistic notions in general:

(1)  \( \forall x[\text{innate}(x) \rightarrow \text{abstract}(x)] \)

Uncontroversially, a priori (innate) notions are by definition abstract, since they are not based in experience (Hume).

(2)  \( \forall x[\neg\text{innate}(x) \rightarrow \text{acquired}(x)] \)

A notion that is not innate must of course be acquired by experience.

(3)  \( \forall x[\text{acquired}(x) \rightarrow \neg\text{abstract}(x)] \)

This is the key proposition asserted by constructivism: acquired knowledge is never fully abstract; it lacks the character of a priori knowledge. Note this does not follow from (1) and (2) alone. It is for this reason that it is definitional for the notion of construction (described later) that the history of its development – the extended process of abstraction – be packaged with the construction itself. The idiosyncratic, lexical traces of that history remain active and may apply in appropriate contexts: Piagetian accommodation applied to language acquisition.

If a notion is abstract, then by (3) it is not acquired, and is thus innate (2). So one proposition that fits (1)-(3) is:

(4)  \( \text{innate}(x) \land \text{abstract}(x) \)

If a notion is not abstract, then by (1) it is not innate. By (2) it must be acquired. Thus a second solution is:
That is, either \( x \) is both innate and abstract, or it is a posteriori and constructed. Since these are the only two possibilities, in establishing the status of a linguistic concept, the researcher can attack the issue either in terms of its status with respect to a priori knowledge, or in terms of the type of generalization it constitutes with respect to mature knowledge. Thus if it proves difficult to determine whether some aspect of language is available very early in acquisition (because the evidence from child behavior is ambiguous), one might instead investigate whether the competency involved is idiosyncratic and constructed. Because constructivists deny (4) is ever true, only (5) is left. The project of UBL can thus be summarized as the following universal claim:

(5) \( \neg\text{innate}(x) \land \neg\text{abstract}(x) \)

There are no innate notions and no fully abstract notions. The rest of the present chapter amounts to an exploration of this claim, mostly vectored through Tomasello (2003) but also making reference to the work of Adele Goldberg in Construction Grammar.

**Constructivism in Language Acquisition**

> Language is nothing more than another type – albeit a very special type – of joint attentional skill; people use language to influence and manipulate one another’s attention (p.21).

Tomasello collects a number of research strands across disciplines (psychology, linguistics, anthropology, and others) under the banner of usage-based linguistics, to piece together a broad hypothesis about children’s acquisition of language. The theory is composed of: (i) a use-theoretic notion of meaning, the “social-pragmatic theory”, which provides a starting point for acquisition; (ii)
a model of syntax that views adult grammar as a collection of “child-friendly” constructions, the endpoint on a developmental path that is continuous with child language; (iii) a constructivist psychology of learning, whereby children engage in gradual, conservative, functionally-driven adaptations that minimize over-generation. I will summarize each sub-theory in turn.

**The Social-Pragmatic Theory of Language**

The essence of language is its symbolic function. A symbol is a social convention whose purpose is to share and direct the **attention** of others. Among the animals, only humans use language, for only humans understand that conspecifics have mental states that are susceptible to manipulation by others. Symbols are therefore illocutionary in character – meaning is social, because primitively intersubjective, and pragmatic, because rooted in its situational effects. Structurally, a symbol is triadic, incorporating the domains of self, other and world: the self invokes social conventions to direct the other to grasp the world in some specific manner: “Linguistic symbols are fundamentally perspectival” (p.12). Since language is primarily a means for functionally-oriented intersubjective communication, it is not in principle distinct from other semiotic modes.

Linguistic symbols are social conventions that may be used to manipulate the attentional and mental states of other people in a way that is different from, but still similar to, the way this is done with other joint attentional behaviors (such as nonlinguistic gesturing) (p.91).

Knowledge of other minds and the ability to read the intentions of others are the key biological adaptation that brought language to humans “fairly recently”. Grammar, as the art of piecing together symbols, instead emerges gradually as a purely cultural artifact, the product of historical processes of grammaticalization of linguistic symbols. Grammar presents as phylogenetically and
ontogenetically derivative of, and epiphenomenal to, abstractions and generalizations that emerge under communicative pressure – in the name of efficiency. The mental faculties at work in grammaticalization are nevertheless domain-general: schematization, analogy, comparison, statistics. The emergence of grammar itself is thus not directly dependent on a genetic event.

Following Wittgenstein, Piaget and Vygotsky, Tomasello centers the meaning of linguistic acts on their pragmatic effects, their use:

Human linguistic symbols are socially learned, mainly by cultural (imitative) learning in which the learner acquires not just the conventional form of the symbol but also its conventional use in acts of communication [...] linguistic symbols are understood by their users intersubjectively in the sense that users know their interlocutors share the convention (p.12).

The conventional use of symbols has biological, chronological and logical precedence over the existence of narrowly linguistic structures, exerting top-down semantic effects on meaning. Prior to expression, uses are unitary intuitions (“Gestalts”, per Lakoff) that assign meaning to expressions in toto; the meaning of the parts is then dependent on the prior use-theoretic whole. Crucially for Tomasello’s model of language acquisition, the priority of social-pragmatic meaning over linguistic articulation applies ontogenetically as well: the child can start learning language once she understands, in some pre-linguistic fashion, her specific interactions with adults in social contexts. This approach gets around the famous “referential indeterminacy” problem described by Wittgenstein and Quine: given some situation and a statement in an unknown language, how do we determine what aspect of the scene is being commented on?

Absent from Tomasello’s social-pragmatic theory of language is any mention of formal semantics. Since meaning is pragmatic and top-down, compositional semantics does not enter the program. A
reason for this, I contend, is that compositional semantics depends on a hierarchical notion of structure, i.e. the syntax of formal linguistics. To allow compositional meaning would open the door to a priori knowledge of grammatical structure, which of course is precisely what Tomasello is arguing against. This brings us to the usage-based theory of grammar.

**Construction Grammar**

A construction is prototypically a unit of language that comprises multiple linguistic elements used together for a relatively coherent communicative function, with sub-functions being performed by the elements as well (p. 100).

Proponents of construction grammar reason that if heterogeneity is so very prevalent in language – in the lexicon, idioms, frozen expressions and idiosyncratic word combinations – then perhaps there are no fully abstract rules at all, i.e. Montegovian procedures for compositional meaning or the syntactic principles of generative grammar are merely artifacts of the theory. Instead, a construction pairs a linguistic complex, which might be a single morpheme or a completely abstract template over categories, to a particular meaning (Goldberg, 2009). The traditional lexicon is thus generalized to encompass the whole grammar. And since, as we saw, meaning is ultimately a question of social-pragmatic use, Tomasello can write: “Constructions serve as a ‘zoom lens’ which the speaker uses to direct the listener’s attention to a particular perspective on a scene” (p. 146)

Constructions are conventional and ontogenetic in character. Within the life of the individual a construction is born fully concrete as one or more morphemes paired with a meaning, thus learning words and learning constructions are fundamentally the same process. Experience teaches how, by convention, elements of the initial form may be abstracted, as slots or category labels. The construction then represents a pattern of usage. But, echoing Piaget, the most concrete form is
Importantly, and radically, in usage-based approaches a given linguistic structure may exist psychologically for the speaker both as a concrete expression on its own—at the bottom of the structural hierarchy, as it were—and, at the same time, as an exemplar of some more abstract construction or constructions (p. 106).

This stipulation prevents the construction from becoming a mere rule-like template, as its more concrete forms will apply when possible, and the history of its development is retained—constructions that converge via abstraction retain their concrete character in their “native” contexts, overturning the “rule versus list” fallacy (attributed by Goldberg to Langacker). At the same time, the abstract forms determine the generative and recursive character of language:

The observation that language has an infinitely creative potential (Chomsky 1957, 1965) is accounted for by the free combination of constructions, where constructions can have open slots and underspecified aspects of their overt realization (Goldberg, 2009, p. 4).

Since a non-nativist theory of the initial state must rigorously exclude any language-specific a priori knowledge, the categories over which abstract constructions are formulated must themselves not be given in advance. In particular, functional elements pose a problem for usage-based accounts because of their very general, relational meaning and wide distribution. Tomasello grants that the traditional notions of noun, verb and adjective are psychologically real, possibly universal, and learnable a posteriori, as also all studies of distributional learning have found. For the rest he argues against their cross-linguistic generality, suggesting that the difficulties encountered by distributional techniques in categorizing functional elements are symptomatic of the fact that functional categories do not in fact exist.
A number of theorists are even moving toward the idea that so-called closed classes of grammatical words and morphemes (auxiliaries, prepositions, determiners, complementizers, and so on) do not really form coherent linguistic classes at all. People just learn to use the particular lexical items and grammatical morphemes individually (must, the, that), and these typically number in the dozens for each so-called closed grammatical class of words/morphemes (p. 105).

We might deny that functional elements bear category features, but we must still explain the vast distributions of items such as English of. On the usage-based account, humans learn separate partitive, genitive, and even more general relational constructions, each of which incorporates the morpheme of. UG proponents instead hold that sentences merge, say, a feature PARTITIVE and that the exponent of is listed with that same feature. Though the latter approach seems more efficient, we have seen that constructivists are willing to trade representational parsimony for fewer non-general psychological faculties. “In usage-based approaches, contentless rules, principles, parameters, constraints, features, and so forth are the formal devices of professional linguists; they simply do not exist in the minds of speakers of a natural language” (p. 100). Tomasello is not, in any case, too troubled by the acquisition of closed-class elements, as these are merely “parasitic” (his term) to nouns and verbs. As he also states that closed-class elements mark relations between substantives, it seems that these relations too are of a parasitic quality. This position marks a sharp contrast with generativists for whom functional morphemes express the spine of the syntax and compositional semantics of sentences.

The notion of construction is clearly close – perhaps identical – to that of schema in constructivism and cognitive psychology. Constructions and schemas frame and control possible meanings of complex expressions, beyond the meanings of lexemes. Thus argument structure, writes Goldberg, is underdetermined by the lexical semantics of the verb – the construction as a whole determines its
arguments, and that only loosely. As violable meaning boundaries, constructions are the device through which creativity is expressed in language. Metaphor, for example, amounts to the substitution of elements in the construction by novel content, triggering effects of semantic coercion brought on by the top-down semantic directive of the construction’s unitary meaning.

Creativity of language comes from fitting specific words into linguistic constructions that are non-prototypical for that word on a specific occasion of use, with no implication that this requires a corresponding permanent lexical entry for the verb involved (p. 161)

The point about the lexicon Tomasello intends as polemic against lexically-minded generativists, though I am not aware of any theory of metaphor that calls for listing every possible metaphoric sense of words. As for creativity, the echo of constructivist adaptation here is not accidental.

Rumelhart (1979) views schema-stretching metaphorizing as the primary acquisition mechanism:

The child’s acquisition process should not be construed, as it often seems to be, as a process of first learning literal language and then, after that is thoroughly mastered, moving on into nonliteral language. Rather, it would appear that the child’s early comprehension and production processes involve the production and comprehension of what is for the child nonconventional (and probably) nonliteral language (p. 73).

We will see below how Tomasello captures this theme via the concept of analogy in acquisition.

*Constructivist learning*

**Basic Mechanisms**

If meaning is primarily a matter of intention-reading and manipulation, and if the fundamental relation of language is the symbolic pairing of constructions with social-pragmatic meanings, then the task for the language learner is to learn the correspondences between constructions and their
intersubjective effects. It is essential for the empiricist venture that the child have pre-linguistic access
to the contextual intentions of others in social interactions; this is the understanding that initiates,
directs and supports the learning process. Tomasello’s acquisition model is thus a broad form of
semantic bootstrapping, where the child’s theory of mind, maturing naturally, determines her
subsequent entry into language. Acquisition cannot begin until the child is able to “read” and direct
the attention of others:

[T]he social-pragmatic theory has a clear answer to the question of why language
acquisition begins when it does. Language acquisition begins when it does because it
depends on the ability to share attention with other human beings communicatively and
so to form symbols, an ability that emerges near the end of the first year of life […]
word learning awaits the emergence of children’s more fundamental social-cognitive
skills of joint attention and intention-reading, on which it depends fundamentally (p.
90–91).

Between 12 and 24 months the child is able to “read” the communicative intent of adults against
complex situations. This understanding is “adult-like”, is logically and chronologically prior to
language, and is thus dependent on non-linguistic forms of communication, such as gestures, the
reading of facial expressions and so on. At this stage of development, on Tomasello’s assumptions,
the child must be able to:

• read, follow, direct, and share the attention of self and others;

• associate single or multiple word expressions with communicated attention;

• conceptualize space, time, objects, and causality, via reflective abstraction from Piagetian
  sensory-motor action schemata;
• conceptualize the social-pragmatic situation in order to make comparisons between situations;

• carry out distribution-based statistical analyses;

• create analogies (one-to-one mappings) between linguistic expressions.

These amount to “powerful” learning processes that enable the child to acquire language without prior knowledge of grammar. Given this initial state, the subsequent path of acquisition extends roughly over four phases:

1. The production of holophrases, in which the child employs single words (phonologically defined, presumably) to holistically express his intention.

2. The production of pivot schemas, composed of multiple words that indicate some structural decomposition of the child’s conceptualization of the situation, but exhibit no syntactic markers.

3. The production of item-based constructions, formulaic lexical sequences that show some productivity, such as slots for arguments, and some morphosyntax and function words for marking relations between content words, but no syntactic knowledge uniform across constructions. These are Tomasello’s verb islands and the limited scope formulae of Pine and Martindale (1996).

4. The abstraction across item-based constructions to yield adult-like structures: constructions over empirical categories.

The major prediction about children’s language made by the theory – a discriminating prediction that
does not also follow from nativist principles – is that children’s syntactic **productivity** increases gradually over time. It is not crisply clear on how this prediction can be tested in practice. What constitutes “gradual” learning? How do we disentangle the linguistic data from cognitive and physical limitations? There is no disagreement that the mapping of lexemes to meanings takes time – time to carry out the distributional analyses that are in any case required, for example – and that the process is likely to involve some experimentation on the part of the child and even exhibit effects of memorization. Low productivity early in children’s multi-word phase fails to deliver clear, unambiguous evidence of item-based constructions. Nevertheless, the empiricist claim here is this: multiword utterances before children’s third birthday are demonstrably formulaic in character.

One prediction *not* made by the social-pragmatic theory, contra earlier work such as Nelson’s (1977) acquisitional sequence of nouns, concerns the categories of words produced. Namely, early in development children’s speech is not restricted to concrete substantives only. If the child’s communicative intentions are primarily directed at intersubjective effects, then the words she utters will depend on the pragmatic situation and her prior exposure to adult constructions in analogous situations. In particular, the theory would expect that early on children will employ commands, introjections, deictics and other non-referential items in performative utterances. According to Tomasello this is the case. “Most children learn many different kinds of words early in development -- regardless of relative frequencies -- thus demonstrating that they can, in the appropriate conditions, individuate many different kinds of referents in the world” (p. 47) Word learning is primarily driven neither by frequency in the input (else function words would prevail early) nor simply by referential concreteness, but by the pragmatic goals and practical concerns of the child. Therefore whatever words the child needs to achieve her communicative goals she will be learn and produce:
The first words that children learn and use include exemplars from almost all of the major parts of speech from adult language: proper nouns, common nouns, pronouns, verbs, adjectives, adverbs, prepositions, and so forth [...] The early uses of these words thus serve specific discourse functions only, and a fully adult-like understanding awaits children's encounters with these words in a fuller range of functional contexts (p. 45).

Specifically with regard to Ps, Tomasello notes that while his daughter employed many spatial prepositions in her early language, she also produced functional prepositions such as of, in partitive and relational contexts, e.g. piece of bread, scared of that. He also notes that “most” English-speaking children employ both the double-object and the prepositional forms of dative constructions from the start, which again shows that children certainly can acquire closed-class morphemes early in acquisition. Again, the issue is not one of the presence or even the relative frequency of functors in early speech. The empiricist argument from heterogeneity, whereby anything goes as long as the context demands it, will not differentiate word learning based on grammatical type; rather, because the full adult meanings of functional elements are, to the empiricist, so abstract, thus requiring more experience and type-variety for learning, the only dimension in which the two word classes will differ is that of productivity over time.

In UBL frequency is the primary driver of and constraint on learning. In particular, for learning grammar, the important statistic is type frequency. It counts “the number of distinct items that can occur in the open slot of a construction or the number of items that exemplify a pattern” (Bybee & Beckner, 2009, p. p.841), and is the main determinant of productivity. Schematicity refers to the degree of abstractness of the category marking the open slot = X, representing, say, semantic restrictions on X. A construction might show high type frequency but low schematicity if it is instantiated with many distinct concrete types from a restricted semantic domain. Goldberg suggests
that schematicity, which she terms “degree of openness”, is in truth what causes children to creatively extend their constructions to new situations.

As for the standard poverty-of-stimulus argument of nativists – that the input underdetermines the patterns of absence of errors in child language, e.g. the seemingly principled ways in which children over-generate – UBL researchers posit three developmental mechanisms that together are sufficient for learning: the conservatism of children, by which they take few risks in expression; entrenchment, whereby greater frequency of the use of a construction by adults in a specific context lowers the probability of the child testing the construction in a new context (lower type frequency leads to greater conservatism); and preemption, whereby the child abandons a candidate generalization of a construction on hearing an adult employ the construction in a specific context that conflicts with the generalization. Preemption is the mechanism by which positive adult data exerts its corrective effect on the child’s grammatical hypotheses.

Goldberg is confident these mechanisms explain the lack of over-generation by children. Tomasello too is at least cautiously optimistic, though he notes that the basic empirical situation has yet to be fully described: “we do not even have good descriptions of the nature of and frequency of children’s syntactic overgeneralization errors” (p. 194).

The Acquisition Process

Having defined these general learning processes, let us focus in greater detail on Tomasello’s account of the acquisition of syntax, taking the adult syntax to have the form described by construction grammar. How does the child learn linguistic structures? If the target is nothing more than an enormous set of more-or-less abstract, more-or-less grammaticalized idioms, then a single,
continuous, recapitulative developmental arc joins child to adult. But by what means is structure first represented? Or, since the examples of constructions provided by Tomasello and Goldberg are all strictly linear strings, not trees, perhaps the question ought to be, is structure ever represented?

If there is no clean break between the more rule-based and the more idiosyncratic items and structures of a language, then all constructions may be acquired with the same basic set of acquisitional processes—namely, those falling under the general headings of intention-reading and pattern-finding (p. 6).

Again, learning begins when the child gains entry into the symbolic realm of adults, by a form of top-down semantic bootstrapping:

"Human linguistic communication can take place only when there is some "common ground" (joint attentional frame) between speaker and listener, which sets the context for the reading of the specific communicative intentions behind a word or utterance […] In situations where there is no common ground—which happen all day every day in the lives of young children—children simply are not learning new words."

By about twelve months children can create their own conceptual representations of many “scenes” of daily life, such as kinematic effects and interactions of forces (pushing, pulling), animacy, figure/ground movement and position (containment and support), possession—broadly, spatiotemporal and causal relations—what Kant terms the categories and their schematization in intuitions, except that here the conceptualization is understood to emerge a posteriori via experience. Such concepts frame the child’s understanding of the external world. As the child hones her ability to read the intentions of mature speakers, she associates the latter’s complete utterances with her interpretation of intentions plus the external situation (the context), inducing the three-way relation

---

27 The claim that no learning occurs until the child understands the holistic meaning of utterances is dependent on what is meant to “learn” a word. The production of so-called “filler syllables” in early speech may indicate positional awareness of functional elements early on, see e.g. Pepinsky and Roark (2001)
of sound, intention and situation that Tomasello calls symbolic. A symbol in this sense is “semantic-pragmatic package”, a conventional whole that refers to (or attempts to affect) an attitude and a situation, before the form of the utterance is itself analyzed. For Wittgenstein and Tomasello, such symbolic use, before grammar, is quite sufficient for language – the components of an expression are only required in special cases:

[I]f when someone says “Bring me a slab”, he could mean this expression as one long word corresponding indeed to the single word “Slab!” […] we mean the sentence as one consisting of *four* words when we use it in contrast to other sentences such as “Hand me a slab”, “Bring him a slab” […] (Wittgenstein, Hacker, & Schulte, 2010, p. p.12)

At this stage the child has also learned to segment speech into phonological words. Word learning then proceeds through a top-down analysis of the meaning of whole utterances, by means an incremental “blame-assignment” procedure that matches components of the overall meaning to individual words. The child is therefore able to analyze adult expressions into speech components (presumably phonological segmentation is sufficient), then match those to the semantic components of the adult's communicative intention. “We are led to a picture of word learning in which the child is determining the meaning of the utterance as a whole and then partitioning out those parts due to particular lexical items” (p. 78). Word meanings are therefore not learned individually and later combined; instead, world learning proceeds through the decomposition of inherently meaningful *Gestalten*.

As children learn the uses of individual words, they also construct categories for their lexical items. Here Tomasello leans on the standard references for statistical distributional analysis, e.g. Redington, Chater & Finch (1998), but correctly notes that this analysis is not sufficient for successful categorization. One must also understand the common semantics of word categories. Again, the
assumption of the child’s social-pragmatic understanding saves the day: items are grouped not simply on the basis of surface positional similarity in overt strings, but also in terms of shared meaning. Functionally based distribution analysis is the process by which “the learner groups together into categories those linguistic items that function similarly—that is, consistently play similar communicative roles—in different utterances and constructions” (p. 145).

The child now understands the use of utterances, is able to destructure meaningful wholes into subcomponents to reveal the meaning contributions of words, and is actively grouping words into functionally-determined categories. He also has developed structured representations of the world. The creation of item-based constructions, and later fully abstract constructions, then proceeds by means of analogy. The basic intuition here is straightforward: if while kicking the cat the child hears the parent order “don’t kick the cat!”, then later again “don’t kick the dog!” in the parallel situation, the child will relate the two statements by analogy, mediated by his understanding of the parent’s intent, and give birth do the construction “don’t kick the X”. The X may itself represent a “functionally based” category, e.g. KICKABLE ANIMALS.

In general, in order to set up an analogy the child must be able to determine three sets of relations: the functional relations inherent within each of the two expressions, and the higher level bijective mapping of one expression to the other. “The essence of analogy is the focus on relations” (p. 164). This requires the child to control a fairly complete understanding of the structures being compared. In particular, “an important part of making analogies across linguistic constructions is the meaning of the relational words, especially the verbs, involved – particularly in terms of such things as the spatial, temporal, and causal relations they encode” (p. 165). Now, clearly not only verbs encode those relations, not even primarily, as it is the task also of functional elements to mark relations
within the sentence. So those “parasitic” morphemes now emerge as essential to the formation of abstract constructions.

Analogies, along with metaphor, similes and equivalent rhetorical structures, belong to the general family of operations that assert correspondences between structures belonging to distinct domains on the basis of structural comparisons (morphisms). This is the fundamental operation of language Jakobson termed selection, “substitution”, and eventually “metaphor” (Jakobson, 1971). But as “any linguistic sign involves two modes of arrangement” (p.119), analogy cannot function in language without that second operation, combination, which Jakobson ties to the figure of metonymy.

Any sign is made up of constituent signs and/or occurs only in combination with other signs. This means that any linguistic unit at one and the same time serves as a context for simpler units and/or finds its own context in a more complex linguistic unit. Hence any grouping of linguistic units binds them into a superior unit: combination and contexture are two faces of the same operation (p.119).

Analogy is felicitous when the individual structures of the relata correspond. Therefore the expressions must be analyzed into their constituent structures. Rumelhart and Tomasello both stress the work of analogical/metaphoric processes in acquisition, upon which the child constructs her knowledge of structure, a posteriori. On what basis might the child execute the prior analysis required to determine the structures of expressions? Tomasello has the child develop top-down structural analyses of statements, guided by functional meaning. But, as Kant teaches, analysis presupposes synthesis – how does the child know that sentences break down hierarchically in the first place?

The axis of combination cannot be induced from the orderings of words given by the child’s having
learned to segment speech, even when supplemented by the generous semantic knowledge afforded
the child by the theory. Tomasello makes a particular point of listing evidence against the “myth of
word order”, the notion that children learn word order faster than e.g. dependent case morphemes
in the expression of thematic relations. In the latter case thematic roles and other relations are
expressed relative to tight local domains requiring little more than morphological knowledge. Word
order, instead, tends to operate on the full sentence. Evidence that the syntactic understanding of
word order develops later is thus consistent with the UBL assumption that syntactic knowledge is
acquired only gradually.

Word order schematizes syntactic relations in time (and its dual space). But the conceptualizations
of time and space are themselves subject to contamination a posteriori by language. In an oddly
Whorfian moment Tomasello writes:

Young English-speaking and Korean-speaking children conceptualize differently such
basic spatial relations as containment (in) and support (on)-as evidenced by their
behavior in preferential-looking studies-because English encodes these concepts with
prepositions such as in and on, whereas Korean uses verbs that indicate such different
kinds of spatial relationships as “tight fitting” and “loose fitting.” Recently these
investigators have extended these findings to other concepts and languages (p.63).

The concepts of space and time, linguistic and thus unstable, cannot provide the conceptual forms
of language because they are already conditioned by language. This reminds us of Piaget's
hypothesis that time (but not space) is an acquired concept.

The core philosophical problem here is that phenomena do not come packaged with labels
describing their relations. This applies to linguistic expressions as well. So either we humans are the
source of the relations we encounter in phenomena – the rationalist and transcendentalist position –
or we somehow induce the structure of experience from experience itself, the empiricist and therefore constructivist position. Tomasello seems to play it both ways, on the one hand invoking the slippery category of the natural:

people, including young children, focus on relations quite naturally and so are able to make analogies quite readily [...] These correspondences between processes in the creation of nonlinguistic analogies and in the creation of abstract linguistic constructions constitute impressive evidence that the process is basically the same in the two cases (p. 164).

On the other hand, since grammatical relations are to emerge from such prior “natural” relations, grammars amount to unnatural, artificial, merely cultural phenomena; thus the child’s ability to recognize them must arise from elsewhere. Tomasello now finds his footing in Piaget’s solution: “[S]ome of the fundamental syntactic relations apparent in children’s early language correspond rather closely to some of the categories of sensory-motor cognition as outlined by Piaget (1952).” As we saw, Piaget held that children abstract the nested character of their action schemas into relations of parts and wholes and containment. It is here that Tomasello locates the ontogeny of the knowledge of structure that, in conjunction with social-pragmatic understanding, enables children to decompose the utterances of adults and eventually (re)discover adult grammar. The origin of language, for the individual, is located in the extra-linguistic activity of the child in the sensory-motor stage. The structural relations of language are not native to language itself.

I view this as essential to the empiricist research program: in the name of a certain psychological parsimony, the psyche is granted all sorts of a priori powers, including tremendous sensitivity to complex intersubjective phenomena -- except for the a priori knowledge of structure in language. This must be imported via reflection from the child’s activity in the world.
Critiques

Given the three legs of empiricist theory – meaning as use; grammar as ensemble of constructions; learning as domain-general memorization, categorization and analogy-making – critical engagements of the theory might develop over any or all of semantics, syntax, and acquisition. If it turns out that meaning in language is not exhausted by intention-reading and attention-setting, i.e. if logic\textsuperscript{28} and bottom-up composition are necessary, then the psycholinguistics of acquisition will need to account for semantic effects that are not simply given by the situational dynamics of discourse. If the descriptive adequacy of our grammatical analysis demands abstract rules and principles not easily generalizable from constructions, then the continuity question in acquisition must be addressed. And if the theory’s (few) predictions are not borne out in acquisition, then both the psychological model and the theory of grammar are at risk.

The semantic matter is of course a central question in the philosophy of language at least since Wittgenstein, so I have little to say here except to remark that in practice speakers must make fine distinctions that could only be explained by some form of bottom-up processing. Even in Wittgenstein’s famous example of an order, “Bring me a slab”, word-level distinctions of emphasis (“Bring me a slab”), which he views as inessential, are surely precisely the distinctions required for directing the attention of other minds. Also, the sorts of inferences required by usage-based acquisition, in creating categories and making analogies, are needed from the start, therefore a logical faculty is required and is presumably active in the child’s processing of adult speech.

As for syntax, here it seems one must distinguish the general notion of construction grammar from its adoption in UBL. Whether the former is even descriptively adequate will need to be settled by

\textsuperscript{28} See Crain (2012) for empirical arguments in support of the rationalist notion of \textit{logical nativism}, i.e. that humans are endowed a priori with linguistically-relevant knowledge of logical constraints.
syntacticians, though construction grammar strikes me as an unfalsifiable theory that restates the data and predicts nothing, as in Molière’s old joke about the “dormative virtue” of a sleeping potion\(^29\) – especially since the theory makes virtue of its representational inefficiency\(^30\). Anything can be freely stored and retrieved, including tokens, partial structures, meanings, even inferences (Bybee & Beckner, 2009 p.846), the complete contexts in which tokens occurs, and type and token frequencies. Once the criterion of representational parsimony has been jettisoned, it becomes harder to select the better theory.

Not all linguists working with constructions reject a priori linguistic structures. Jackendoff’s (2003) lexicalist model retains phrase structure rules, elements of UG, and indeed various forms of poverty of stimulus arguments\(^31\). Within the instantiation of construction grammar found in usage-based approaches, instead, UG is rejected in principle, while the explanatory power of frequency – what is said, parole – is vigorously affirmed. Yet frequency may not even succeed at adequately restating the data:

> Knowledge about the conventionality of [idiomatic and compositional word] sequences must be represented somehow in the grammar, since fluent speakers do not produce (or accept) the full range of utterances permitted by combinatoric syntactic rules (Bybee & Beckner, 2009, p. p.836)

Which syntactic rules do the authors have in mind in making this curious claim? A model of

\(^{29}\) “We likewise come to know that the caused-motion construction contributes the relational meaning that something causes something else to move, while the question constructions determines that a \textit{wh}-word appears sentence initially” (Goldberg, 2009, p. 4).

\(^{30}\) “Surely it is premature to give up hope that humans, with our rich cognitive abilities, complex social skills, predilection to imitate, and 100-billion-neuron brains, can learn language from the available input” (Goldberg 2006, p. 69).

\(^{31}\) “One would be laughed at for [Tomasello’s] complaint [that UG principles cannot be related to other cognitive domains] in the case of an undeniably specialized system, say visual stereopsis […] What makes a system specialized is in part that it performs processes not found elsewhere in the [mind]” (Jackendoff, 2003, p. 79)
grammar whose unfiltered rules systematically over-generate is simply a failed model; the solution cannot lie in patching the grammar with an output filter that judges acceptability by polling speakers’ store of past linguistic experience. There is just no way contingent experience can explain the interpretability of novel sentences, so if the grammar is adequate, there is no reason for it to incorporate frequencies. Normative grammar is a matter of world knowledge. As Newmeyer reminds us (2003) the basic observation that language users understand infinitely more than they produce must be explained. It is furthermore no explanation to assert that an infelicitous sentence is so judged only when unusual: “[T]here is nothing predictive about accounts that say that grammars encode whatever because that is what speakers need grammars to do” (Newmeyer, 2003, p. 14).

The notion that “combinatoric syntactic rules” must be somehow controlled (because they overgenerate) indicates a tension internal to the UBL program. In Bybee’s words, UBL takes language to be an “emergent” product of general cognitive systems. Precisely which systems, and how they interact to yield language, remains an open research question. The objection against UG at times appears to focus on the domain-specificity of the linguistic principles hypothesized by generativists, but not necessarily on a priori capacities of the mind in general, e.g. knowledge of conceptual structure. “A usage-based model thus takes as its null hypothesis the view that language is an extension of other cognitive domains” (Bybee & Beckner, 2009, p. 828), where, presumably, the faculties belonging to the other domains might exist a priori. The combinatorial faculty, for one, may originate in general cognition.

One might sympathize with a certain negative reception of the more baroque proposals in P&P literature (Tomasello points to Subjacency) -- the suspicion that such principles are unlikely to be psychologically real. It is of course a goal of the Minimalist Program (MP) to control the excessive
postulation of Principles\textsuperscript{32} and, yes, to shift, wherever possible, linguistic processes to general
cognition (whatever that may be). “How little can be attributed to UG while still accounting for the
variety of I-languages attained, relying on principles [of general cognition]? ” (Chomsky, 2007, p. 4) However the MP evolves, it seems that one operation will necessarily remain part of UG:
unbounded Merge. (See Krivochen, 2011 for an argument that Merge is also the only component of
UG.) In its simplest expression Merge groups a set of items into a unit, an ordered pair perhaps
(Langendoen, 2003), or synthesis, in Kant’s terminology; and because synthesis is necessary for
thought in general (Kant), Merge is perhaps an extension or borrowing of a fundamental
component of all thought. The key genetic mutation yielding language would then have linked the
human communicative system to the synthetic faculty of the mind: the Kantian imagination.

The conclusion that Merge falls within UG holds whether such recursive generation is
unique to FL or is appropriated from other systems. If the latter, there still must be a
 genetic instruction to use Merge to form structured linguistic expressions satisfying the
interface conditions. Nonetheless, it is interesting to ask whether this operation is
language-specific. We know that it is not (Noam Chomsky, 2007, p. 7)

Or, as Newmeyer puts it: “the roots of grammar lay in hominid conceptual representations and that
the shaping of grammar for communicative purposes was a later development” (Newmeyer, 2003, p. 21). Once syntax split off from conceptual structure, its combinatorial possibilities extended in
principle beyond possible meanings. The generative power of syntax must be regulated. Where
Bybee and Beckner would have frequency act as a filter on over-generation, for Chomsky it is the
primacy of the conceptual-intentional (CI) system in language that controls the shapes produced by
Merge. Linguistic structures must be interpretable or they “crash”. A more “radical” thesis, which

\textsuperscript{32} Setting aside the even more contentious issue of parameters. “Under no circumstances does [language] development
look like an instantaneous setting of parameters in which all verbs and other lexical items immediately participate in
a totally abstract construction” (p. 142).
Chomsky associates with Wolfram Hinzen, holds that not only is Merge in language associated with or borrowed by the composition of concepts in general, Merge just is that faculty, across cognitive modules. In Hinzen’s view all semantic structure simply is linguistic structure:

Where grammar virtually gives us the structures that we need for the logical analysis of our thoughts, it becomes very hard to see how we can maintain the view that ‘thought’ has some independent origin, and logical structure is unrelated to the advent of language on the evolutionary scene, viewed merely as a means of expression (Hinzen, 2011, p. p.516)

Hinzen’s radical view would “deflate” semantic notions like proposition and event to their syntactic forms, CP and vP. Chomsky notes the “correlation” between the two forms of Merge, internal and external, and the “duality of semantics”, argument structure vs discourse and scope, and suggests that meaning is likely the primary function. “The correlation is close, and might turn out to be perfect if enough were understood” (p. 10). Beyond perfect correlation, Hinzen hypothesizes the identity of the linguistic and semantic structure, leading to a “crash-free” syntax: if structures are by definition meaningful, then nonsense cannot be generated. Falling somewhere between these positions, and working from the somewhat distinct principles of biolinguistics, Cedric Boeckx also identifies Merge as the cross-modular creative principle, but retains the interface to CI in its role as regulator:

To be useful at all in thought and action, such a freely combining Merge must be regulated. […]W]e have linguistic reasons to believe that this regulation takes the form of integration/embedding: Merge is constrained in virtue of its interfacing with and being embedded inside cognitive systems responsible for interpretation and externalization. This regulation is what the formal linguistics literature refers to as “Spell Out” or “Unify” (Boeckx, 2014, p. 5)
These varying positions on the relationship between Merge and semantic structure are evidence of a certain instability in generative linguistics with respect to the syntax/semantics interface. The more conservative position views the generative principle as dependent on CI, phylogenetically and logically. Language is an outgrowth of Reason, and the synthetic faculty originates in Reason. Hinzen and Boeckx, though differing on the regulative function, view Merge as an independent function, a mental faculty distinct from the understanding. In metaphysical terms we might think of the former position as rationalist, while the latter is transcendental in the Kantian sense, to the extent that the independent generative principle matches the creative faculty Kant termed the imagination. And UBL-friendly models of cognition that posit a priori mental faculties (but none specific to language) might subscribe to aspects of the rationalist view.

Metaphysically opposed to all such speculation stands constructivism. Tomasello, nothing if not faithful to his constructivist roots, falls back onto Piaget to explain the origin of structure. As we saw, the moment of structure is unavoidable. The child’s pre-linguistic sensitivity to other minds may well be decisive for language acquisition, but at some point the child must understand and build structure in order for learning to advance, for analogies to be constructed. Should it turn out that the type of structure-making required by language is more sophisticated than the recursive nesting of binary sets (Merge), then the child must grasp this too. This knowledge is either hard-wired or learned. Per Piaget the structure of concepts emerges via reflective abstractions of action schemata developed in infancy; logic is ontogenetically (re)discovered a posteriori. Whether there is evidence to help decide the issue is the topic of the next section.
Evidence

Modeling Acquisition

If the UG polemic boils down to knowledge of structure, as it must for minimalist theories of UG, little surprise there was such intense debate around Perfors, Tenenbaum & Reiger (PTR, 2011) (PTR). PTR set out to show that a Bayesian learner will converge onto structure-dependent grammars based only on child-directed speech (CDS) input. Importantly, the position set forth by PTR belongs to the order of domain-general nativism that sticks to generative models of adult grammar: PTR are not constructivists. So the issue here is about the learnability of structure, granted that fully abstract grammars obtain as the end state. PTR's work is important because it constitutes a serious attempt at demonstrating (via simulation) that a UG notion of hierarchical structure can be induced from data.

Like all such simulations, however, PTR's study runs into a basic logical problem faced by empiricist arguments. As Virginia Valian has pointed out (p.c.), to assert that NO language-specific knowledge is innate amounts to the universal proposition that all linguistic competence is a posteriori. A simulation that models only a step in learning is therefore of limited value; the whole acquisition trajectory needs to be simulated – obviously an extremely challenging task! Otherwise one simply displaces the problem. In PTR's study this issue is visible in the input to their simulation, which incorporated part of speech (POS) tags. For the argument to go through we must assume that children are able to organize the lexicon into abstract categories without the help of the hierarchical analysis of linguistic structure, since it is precisely the latter that PTR want to show can be derived.

In computational models of acquisition the task of category induction is often treated as one of classification, akin to unsupervised part-of-speech (POS) tagging in computational linguistics. The
target of learning is a function that maps phonological words, plus some representation of context, to category feature values. A number of computational modeling studies over the past two decades have sought to demonstrate that domain-general statistical methods, applied to child-directed adult speech, are able to learn POS taggers (Mintz, Newport, & Bever, 2002; Redington et al., 1998) without other input. These models typically exploit collocation statistics extracted from strings of words (hence assume the segmentation of the input medium into words) to build vector representations of the distributions of words in samples of language — typically child-directed adult speech (CDS). Words are then grouped into categories by comparing their distributional vectors via metrics on the vector space.

PTR point to the work of Redington et al and Mintz as evidence that unsupervised POS tagging from only data is possible. While distribution-based statistical learners have been found to perform well with open-class categories when the problem is simply to bin items into a verb category, a noun category etc. — not a difficult statistical problem — their performance with respect to functional categories has been poor. Occurrences of the former are often framed by high-frequency closed-class items, allowing for classification via simple heuristics. Functional morphemes are not as easily framed. Thus, while nouns and verbs have been shown to group with reasonably high accuracy (the homogeneity of the induced classes) and completeness (the degree to which objects of the same true class cluster together), functional categories manifest sharp trade-offs between these measures: one obtains either many small, homogeneous clusters, or fewer, larger groupings that are less tidy.

If it is agreed that the traditional closed classes are mere descriptive devices, however, then POS learning problem can no longer be cast as one of classification. To ‘learn’ a functional element

---

33 Even for verbs and nouns the clustering is clearly insufficient for syntactic purposes, since the groupings typically fail to make distinctions of tense, aspect, phi-features, mood etc.
means to determine its precise function in syntax (and meaning) – so the target of learning already refers to and presupposes syntactic structure. For example, if learning functional elements entails associating the correct morphosyntactic features with functional heads, then clearly the features themselves must be given. A simulation that purports to show that structure is learnable from data where the data are POS-tagged has scuttled the research question from the start – twice – by supposing that functional morphemes are learned by grouping them into categories, and by assuming that categories are discoverable without prior syntactic knowledge.

The probabilistic and minimum description length methods detailed in Hsu and Chater (2010) and Hsu, Chater, & Vitányi (2011), and especially the frequentist experiments of Perl and Sprouse (2012), again appeal to Mintz and Redington et al in order to justify the assumption that a data-driven learner is able to derive categories and syntactic structure – that the module responsible for analyzing language as hierarchies of phrasal categories is derived from domain-general principles of statistical learning. If it is agreed, as I have argued, that this assumption is not firmly established, then PTR's and similar work form weak support for the empiricist position. At best they show that specific steps in the acquisition process, such as the induction of syntactic constraints, might not require UG; but only because UG-type knowledge is already incorporated into the input.

More promising for the UBL approach is the investigation by Bannard, Lieven and Tomasello reported in (2009). Here the authors recognize that the UBL position must confront the acquisition problem in full: “What any theory of language development needs is an evaluation that tests whether it can account for child speech in general” (p. 17294). Their model, crucially, is entirely data-driven, in that it is not dependent on part-of-speech tagging nor on any sort of syntactic analysis, although it assumes that the learner builds simple hierarchical representations (type-2,
context-free grammars) which, perhaps, belong to general cognition.

Bannard et al investigated whether the early multi-word utterances of two children were consistent with the model of language development hypothesized by UBL. Their approach consisted in automatically inducing construction grammars and “fully abstract” grammars trained on around 28 hours of each child's recorded speech at ages two and three. These grammars were then evaluated for coverage on two further hours of child data at the same age.

Concretely, the two types of grammar both belonged to the family of probabilistic context-free grammars (PCFGs), in which production rules, from single non-terminals to (sequences of) terminals or non-terminals, are weighted by probabilities. The distinction was that for the usage-based PCFG (UB-PCFG), words could appear on the right side of the rewrite rules (indeed words obligatorily appeared on the right, so that all rules were lexically-specific), while for the traditional PCFG words were inserted from the lexicon, independently of production rules between non-terminals. Categories in the UB-PCFG did not correspond to traditional notions: they merely labeled slots in more-or-less lexically explicit frames, and constrained the range of fillers. The rewrite rules of the UB-PCFGs were derived by a string-matching procedure over the training data. Unfortunately the authors do not state how the rule base of the PCFGs was constructed – by hand, one presumes, since in the experiment the PCFG represents a learner with innate knowledge of syntax. The probabilities for both grammars were induced using Monte Carlo sampling methods.

Bannard et al evaluated the induced UB-PCFGs by determining the proportion of utterances in the test set for which the grammars yielded a parse, and by comparing the perplexities of the UB-PCFGs and PCFGs. Perplexity is information-theoretic quantification of the degree of “surprise” triggered by data given a model, lower perplexity indicating greater model fit. They found that
grammar coverage ranged between around 70% to 80%, while the upper bound of the 95% confidence intervals for perplexity of UB-PCFGs was below the means of PCFG perplexity. Based on these results they conclude that the lexically-specific grammars “perspicuously account for their later productions” and perform better than the fully abstract grammars, thus validating the usage-based perspective.

Several problems with these claims emerge on closer examination. First, what of the 20% or so utterances the UB-PCFGs could not explain? It is precisely such edge cases that often are most informative about the human grammatical competency. To cite the classic example, parasitic gap constructions are extremely rare in spoken language; that humans nevertheless understand and produce them demands explanation. A lexically-specific, frequency-driven induction procedure is likely to represent such rare constructions simply as phrasal idioms, since the input cannot reliably supply the variety of distinct construction tokens required for their schematization. How then is the mere recognition of a (mostly or fully) concrete construction, as opposed to its structural analysis, a “perspicuous account” and not just a gesture of recognition married to an (otherwise unspecified) notion of social-pragmatic meaning? The productivity of rare syntactic phenomena thus remains unexplained, putting into question the capacity of the grammar to generalize the right way.

Related to this first point, it seems likely that Bannard et al's UB-PCFGs overfit the training data. The coverage results were obtained on test data consisting of child utterances recorded at about the same time as the training data. It is not surprising that a learning mechanism that allows strings of words to be encoded into the grammar as “signs”, and which imposes no economies on the numbers of such signs, would perform well on closely related test data – children (and their parents) are likely to talk about the same things in the same way. When the authors ran one child's grammar
against the speech of the other, however, the proportions of recognized sentences dropped precipitously: at age 2, on average only 15% of Annie's speech could be parsed by Brian's grammars, while grammars induced from Annie's training data parsed 36% of Brian's test data. It is hardly likely that at some hypothetical playdate the spoken interactions between Brian and Annie would have been handicapped by such mutual incomprehension. The authors view this particular result as a confirmation of the UBL approach, since at age 3 the two grammars cross-parsed at greater rates (59% and 63% respectively), suggesting that, as posited by the UBL hypothesis, initially idiosyncratic grammars converge over time onto a shared grammar. But such an argument assumes it is reasonable that children at age 2 misunderstand even the simplest utterances at alarmingly high rates – and perhaps makes virtue of a property of any statistically consistent learning procedures: more input improves performance.

As for the comparison to fully abstract PCFGs, here one must attend to (and partly guess at) the details, since surely the performance comparison would depend on the quality of the PGFG model. In supporting information to their article, Bannard et al point to work by Mark Johnson and colleagues on the application of Gibbs sampling to the estimation of PCFG parameters (Johnson, Griffiths, & Goldwater, 2007). But whereas Bannard et al seem confident of the effectiveness of such induction methods – “we can expect to infer excellent models while also having guarantees concerning coverage” (Bannard et al., 2009, pp. SI, p.2) – Johnson et al are less sanguine:

We believe that the primary reason why both [Inside-Outside, i.e. expectation maximization methods] and the Bayesian methods perform so poorly on [the] task [of describing the structure of natural language] is that simple PCFGs are not accurate models of English syntactic structure. We know that PCFGs that represent only major phrasal categories ignore a wide variety of lexical and syntactic dependencies in natural
As a very basic example, the case of a pronoun depends on whether or not the D non-terminal is governed by a verb, preposition, complementizer and so on. The relatively low performance of PCFG models in accounting for natural language would not affect the comparison between UB-PCFGs and standard PCFGs (the effect would cancel out) if the probabilistic models of the two devices were comparable. This was not the case in Bannard et al's study, however. The reason again comes down the lexicalization of rewrite rules. To see this, consider the Bayesian network representation of the generative procedure for the UB-PCFG (Figure 6).

![Bayesian network for Bannard et al's UB-PCFG](image)

*Figure 6: Bayesian network for Bannard et al's UB-PCFG (Bannard et al., 2009, p. 17285)*

The box labeled “Categories” shows an example of the derivation of the abstract components of a parse. Nodes labeled $z$ represent data-induced categories (types of slot fillers), and solid arrows indicate conditional dependence. The triangle subtended by $\{z_1, z_2, z_3\}$ describes a hidden Markov model (HMM); were $z_2$ or $z_3$ to produce more categories then these would give rise to further HMM structures, hierarchically arranged. Since PCFGs are probabilistically equivalent to hierarchical
HMMs, the generative process for categories is compatible with a standard PCFG\textsuperscript{34}. Differences emerge, however, when we consider lexical insertion. In a typical PCFG words are inserted at a last step in the derivation (conceptually, at least; a parser may operate bottom-up), through mappings from X-bar level-0 categories (e.g. N, V, D etc.) to vocabulary entries. Now, the probabilistic selection of lexical content for a given category leaf is conditionally independent of all other words given the categories. So the lower box in Figure 6, labeled “Signs”, would, for a standard PCFG, contain word nodes co-indexed to leaves in the parse tree, and display only arrows from category to word – just as in Figure 6. Notice, however, that in the sample derivation in the figure, the phrasal-like non-terminal $z_1$ also probabilistically generates the “sign” $x_1$. This possibility is consistent with – in fact defining of – construction grammar: that a category is rewritten as a string containing concrete material plus categorial slots. Again, Bannard et al make it a requirement that every production contain at least some lexical material. \textbf{Sign} is thus the generic term for chunks of language incorporating both lexical terminals and abstract non-terminals. So, though the graph somewhat masks this, the rewrite rule for $z_1$ is:

$$z_1 \rightarrow x_1$$

where in the example $x_1$ contains two categories, e.g. for the sign $x_1 = “X$ wants a $Y”$, in which $X$ and $Y$ are the categories rewriting $z_2$ and $z_3$.

Now, if the UB-PCFG model requires that productions contain words, then, in order to remain probabilistically equivalent to the PCFG (i.e. to make the same independence assumptions), the complete production rule, including all lexical material, must be modeled as an HMM. This in turn implies the existence of dependencies between words and between words and categories, just as

\textsuperscript{34} This follows easily if one thinks of each production rule as an HMM where the parent node is the state.
there are between categories (e.g. the arrow linking $z_2$ to $z_3$). The UB-PFCG is therefore less constrained than the abstract PCFG, as the latter fixes the weight, so to speak, of the arc between words to zero. The UB-PCFG is thus able to capture dependencies between words that the PCFG cannot represent. For example, the sign give me $X$ establishes a direct dependency between give and me that cannot be modeled by a simple PCGF. One could of course add features to the PCFG that would guide lexical selection, e.g. the production:

$$VP \rightarrow V \ D[+\text{accusative}]$$

But it seems unlikely the comparison PCFG in Bannard et al. (2009) was as sophisticated as that.

In short, instead of demonstrating that a data-driven, constructivist acquisition strategy is superior to one that incorporates a priori knowledge of syntax, the better performance of the UB-PCFG on the parsing task simply suggests that context-free grammars are not very good models of natural language – a long established notion (Chomsky, 1956).
Observing Acquisition

As previously noted, the main (and it seems only) prediction made by UBL concerning L1 acquisition is that children’s linguistic productivity develops “gradually”. Setting aside questions of what rate might be construed as “gradual”, according to this hypothesis early in acquisition children traverse a developmental stage in which their overt language is sufficiently explained by item-level constructions or limited-scope formulae. This ought to be particularly evident for children’s use of functional morphemes, whose abstract nature implies, on constructivist grounds, that the child must await exposure to a wide variety of contexts in order to induce fully abstract representations. So the issue is not one of presence or absence of f-morphs in child speech, as the child will use whatever she needs for her communicative purposes, but rather one of type frequency.

Naturalistic studies of child corpora face several conceptual and methodological hurdles. First, as is well known at least since Brown (1973), children tend to not produce f-morphs earlier in acquisition, but when they do, they commit very few grammatical errors. The debate then turns on whether omission is evidence of non-knowledge or is due to production limitations. Second, in recorded sessions of child-adult interactions, interlocutors often repeat each other’s utterances and are typically talking about the same things. It is therefore difficult to assess the degree to which children produce novel combinations. Third, the question of what counts as “gradual” learning is not easily avoided. Children are certainly exposed to the high type frequency of common functional morphemes from birth. Consider the case of determiners: if low initial productivity with, say, a and the is observed, is the implication that the input to date has been insufficient to demonstrate their abstract meaning? How many more input exemplars are needed?

Finally, there’s the matter of how precisely to quantify productivity. Pine and Martindale (1996)
developed a measure of determiner productivity with nouns they termed “overlap”, with which they sought to show that children’s use of determiners is less productive than that of adults. But their overlap computations are fatally correlated with sample size (Valian, Solt, & Stewart, 2009). Yang (2013) shows that overlap is also affected by the Zipf-like distribution of morpheme frequencies in the language. Once overlap is adjusted for sample size (Wang, 2012) and word frequency, children’s productivity approximates that of adults. Pine et al. (2013) accept these critiques of the overlap measure, but counter that one must also control for noun types when comparing overlap values for determiner+noun pairs between children and adults, since the presence of infrequent nouns in adult speech tends to lower adult overlap scores. By computing overlap only for nouns used by both adults and children in each recorded session, they again find child overlap to be lower than that of adults. This sort of back-and-forth debate on just the issue of determiner productivity suggests that quantitative approaches to productivity based on recorded language have yet to settle the matter.

Given the challenges encountered by corpus studies of child-adult interactions, perhaps experimental approaches stand a better chance of clearly evaluating constructivist predictions. Of particular interest are comprehension studies, since comprehension outstrips production from the start (Tomasello, 2003, p. 50) yet it is plausible that the same grammatical representations required in production are active in comprehension (Naigles, 2002). Also, comprehension studies are better able to isolate confounds resulting from cognitive and biomechanical limitations. A simple test of whether a child knows determiners independently of specific nouns might consist in familiarizing young subjects with a novel noun introduced by a frequent determiner, “a blick”, then having the subjects perform a task that requires understanding of that noun with a different determiner, “the blick”. Tomasello and his colleagues must predict that at some stage in acquisition the child subject will simply not understand “the blick”.
I have not found studies that test exactly this, though Shi and Melançon have shown that children at 30 months use the gender feature on determiners in French to classify nouns (Melançon & Shi, 2012) and already at 14 months seem to know the categorial distinction between determiners and pronouns (Shi & Melançon, 2010). In his (2003) book, written before these more recent investigations, Tomasello did not list any comprehension experiments that show children cannot understand a novel DET + NOUN combination. He cites Maratsos (1976) as support that children fail to map the semantics of English articles to relations of “givenness” between speakers/hearers and information, a complex intersubjective task. But that same study showed that children correctly map specificity to the English definite and indefinite articles:

> Young children understand quite early in development that the definite article indicates a specific referent whereas the indefinite article indicates a non-specific referent. But they do not master the subtleties of the use of articles depending on current listener knowledge and attention-givenness, the perspectival component-until much later. Indeed, it may be that mastery of the specificity function gets in the way of children’s discovery of the later-acquired perspectival function. (p. 210)

Maratsos’ data suggest that three is the age at which children already make the specificity distinction. These data also seem to suggest that a fundamental tenet of Tomasello’s model, the early availability of social-pragmatic understanding, has it backwards – that the sophisticated use of syntactic markers for intersubjective meanings develops more gradually than pure syntactic knowledge, perhaps because the necessary social understanding itself develops more slowly than Tomasello claims. Naigles (2002) makes a similar point: the fast and effortless analysis of speech data carried out by infants contrasts with the slower progress made by toddlers once further progress in acquisition involves the specific encyclopedic meanings of terms. In her words: “form is easy, meaning is hard.” Whatever the case, it seems that, as Tomasello repeatedly emphasizes, much research remains to be
done in this area.

Conclusion

This chapter opened with a review of core concepts of constructivism. The radicality of Piaget's empiricism emerged through a consideration of his ontogeny of structure:

1. From the start the child engages in unreflective but structured activity in the world. This activity is progressively reified in the mind through reflective abstraction, giving rise to structure in cognition. Temporal (and possibly spatial) relations are similarly products of experience.

2. Development proceeds through stages in which the later, more abstract understanding of phenomena supplements, but does not replace, earlier, more concrete knowledge. The drivers of development are repair processes of assimilation and accommodation by which the subject negotiates its interpretations and interactions of and with the world.

Applied to first language acquisition, the first point amounts to the claim that knowledge of structure in language is not innate but derived from experience, while the dual second asserts that mature grammatical knowledge is constituted by dynamic ensembles of constructions (of varying degrees of abstraction), rather than by purely abstract language-specific principles. The notion of analogical creativity in language, as understood by UBL theorist – amounting to the constructivist notions of accommodation and assimilation of representations – forms the link between the learning process and the origin of structure, since the child extends her understanding through analogies, and analogy itself presupposes structure for its felicity.

The usage-based theory of language acquisition was shown, through a reading of Tomasello (2003),
to articulate precisely these constructivist positions. Most broadly, the constructivist perspective argues that all specifically linguistic notions are: (i) derived, not innate; (ii) capable of being analyzed as constructions rather than principles of UG. This entails that the development of syntax is in some sense “gradual”:

Usage-based approaches expect children's learning to be more gradual, piecemeal, and lexically dependent—with the acquisition of particular linguistic structures depending heavily on the specific language to which a particular child is exposed, and with generalizations coming only after a fair amount of concrete linguistic material has been learned (Tomasello, 2003, p. 98).

We then considered landmark evidence for the constructivist claims. Various modeling studies that purport to demonstrate that principles of language—including knowledge of the hierarchical structure of expressions—are learnable from data were found to depend on the prior analysis of the input into (minimally) sequences of POS tags. While there is some evidence that unsupervised learning mechanisms are able to induce gross categories for nouns and verbs, the same cannot be claimed for functional elements. Indeed the problem may be fundamentally miscast, if the acquisition of functional elements is understood as the determination of their specific featural content—a position with which Tomasello himself seems to agree. The modeling literature therefore leaves open the very real possibility that a priori knowledge of language is required in order to acquire functional elements.

As for empirical studies of child production data, the question of how the productivity of specific constructs in children and adults is best compared remains subject to a lively methodological debate. Comprehension studies might in principle disentangle effects due to articulatory or independent cognitive development from those grounded in grammatical knowledge. Few have been completed,
however, presumably because they are challenging to design.

We conclude that the positive evidentiary foundation for constructivism in language acquisition remains weak. Part II of this dissertation focuses on the empirical aspect, through a series of studies of children's use of lexical and functional prepositions.
Part II. Empirical Studies of the Acquisition of P

Chapter 3. Introduction and Analysis of Errors

Introduction

The present chapter introduces three empirical investigations of child and adult data in CHILDES, for native English and Spanish speakers. The overall purpose of these studies was to determine whether records of child speech yield evidence that children's acquisition of functional Ps is meaningfully different than that of lexical Ps. First the data sources and the processing steps carried out to construct data sets for analysis are detailed. The three studies are then presented roughly in order of increasing statistical complexity: a mostly qualitative analysis of PP errors (this chapter), followed by two quantitative studies that (i) compared the rates of production of lexical and functional Ps, and (ii) compared the rates of production of novel P + complement complexes for the two types of Ps (Chapter 4).

Database

Corpora for the various studies were derived from the transcriptions of conversations between adults and children collected and distributed via the CHILDES project (MacWhinney, 1996).

Most of the English language (US and UK) corpora in CHILDES were incorporated into the database. Exceptions were those that focused on infant phonology or were too sparse to be useful. For the pilot stage of the investigation the session files came from the North American subset for which the CHILDES project has provided morphological analyses by means of its MOR analyzer.
It soon became clear by visual inspection that the accuracy of MOR’s POS tagging is unsatisfactory. For example, MOR marked too many instances of Ps as particles and vice versa. Switching to the Freeling suite of natural language processing tools (Padró & Stanilovsky, 2012) led to improved accuracy, the adoption of a single tool for handling both English and Spanish, and, crucially, access to sub-corpora beyond the MOR set, for both languages, since the raw transcriptions were parsed and tagged directly.

Summary corpus descriptives for the full data set are given in Tables 4 and 5 for English and Spanish respectively. For each transcript the utterances of speakers whose roles were in the set {“Mother”, “Father”, and “Investigator”} were merged and identified simply as a synthetic “Adult”. Child utterances were not merged (in some sessions there are multiple child participants), but only children identified as “Target_Child” were retained. Each child was associated with a name, either given in the file or automatically generated from file metadata. Child names thus delimited sets of transcripts as sequences of longitudinal data.
### Table 4: Description of English CHILDES Corpora

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Children Sessions</th>
<th>Child Utt</th>
<th>Adult Utt</th>
<th>Child Tokens</th>
<th>Adult Tokens</th>
<th>Min MLU</th>
<th>Max MLU</th>
</tr>
</thead>
<tbody>
<tr>
<td>bloom70</td>
<td>7</td>
<td>27</td>
<td>21,744</td>
<td>127,925</td>
<td>123,077</td>
<td>1.4985</td>
<td>5.681</td>
</tr>
<tr>
<td>bloom73</td>
<td>1</td>
<td>6</td>
<td>2,158</td>
<td>6,603</td>
<td>12,123</td>
<td>1.0932</td>
<td>5.0158</td>
</tr>
<tr>
<td>brown</td>
<td>4</td>
<td>214</td>
<td>92,177</td>
<td>414,124</td>
<td>439,518</td>
<td>1.87</td>
<td>5.256</td>
</tr>
<tr>
<td>braunwald</td>
<td>1</td>
<td>900</td>
<td>35,626</td>
<td>145,675</td>
<td>152,092</td>
<td>1.6086</td>
<td>7</td>
</tr>
<tr>
<td>clark</td>
<td>1</td>
<td>47</td>
<td>17,921</td>
<td>98,808</td>
<td>168,785</td>
<td>2.6818</td>
<td>8</td>
</tr>
<tr>
<td>demetras-trevor</td>
<td>1</td>
<td>26</td>
<td>6,704</td>
<td>30,455</td>
<td>44,728</td>
<td>2.67535</td>
<td>5.53125</td>
</tr>
<tr>
<td>demetras-working</td>
<td>3</td>
<td>38</td>
<td>8,973</td>
<td>35,745</td>
<td>59,836</td>
<td>2.0764</td>
<td>4.98115</td>
</tr>
<tr>
<td>feldman</td>
<td>1</td>
<td>21</td>
<td>1,685</td>
<td>5,677</td>
<td>35,823</td>
<td>1.1547</td>
<td>5.6908</td>
</tr>
<tr>
<td>gleason</td>
<td>25</td>
<td>70</td>
<td>18,492</td>
<td>81,570</td>
<td>221,818</td>
<td>2.3779</td>
<td>5.858</td>
</tr>
<tr>
<td>hall</td>
<td>37</td>
<td>37</td>
<td>70,936</td>
<td>396,796</td>
<td>473,330</td>
<td>1.8664</td>
<td>6.4956</td>
</tr>
<tr>
<td>kuczaj</td>
<td>1</td>
<td>210</td>
<td>22,860</td>
<td>193,439</td>
<td>148,207</td>
<td>3.8026</td>
<td>9.0532</td>
</tr>
<tr>
<td>macwhinney</td>
<td>6</td>
<td>273</td>
<td>49,253</td>
<td>303,201</td>
<td>346,093</td>
<td>2.43415</td>
<td>6.92055</td>
</tr>
<tr>
<td>macwhinney-boys</td>
<td>2</td>
<td>4</td>
<td>1,079</td>
<td>860</td>
<td>5,132</td>
<td>3</td>
<td>6.074</td>
</tr>
<tr>
<td>manchester</td>
<td>13</td>
<td>804</td>
<td>232,411</td>
<td>781,062</td>
<td>1,850,621</td>
<td>1.51375</td>
<td>4.9682</td>
</tr>
<tr>
<td>peters</td>
<td>1</td>
<td>78</td>
<td>19,982</td>
<td>51,873</td>
<td>132,251</td>
<td>1.01065</td>
<td>4.7044</td>
</tr>
<tr>
<td>post</td>
<td>3</td>
<td>30</td>
<td>7,909</td>
<td>25,351</td>
<td>107,256</td>
<td>1.47975</td>
<td>7.511</td>
</tr>
<tr>
<td>providence</td>
<td>6</td>
<td>364</td>
<td>123,708</td>
<td>530,853</td>
<td>1,633,930</td>
<td>1.2864</td>
<td>6.507</td>
</tr>
<tr>
<td>sachs</td>
<td>1</td>
<td>93</td>
<td>16,236</td>
<td>60,074</td>
<td>69,200</td>
<td>1.56935</td>
<td>5.9664</td>
</tr>
<tr>
<td>snow</td>
<td>1</td>
<td>42</td>
<td>13,164</td>
<td>52,658</td>
<td>115,581</td>
<td>2.40615</td>
<td>5.7031</td>
</tr>
<tr>
<td>suppes</td>
<td>1</td>
<td>52</td>
<td>32,134</td>
<td>137,333</td>
<td>221,303</td>
<td>2.1459</td>
<td>5.4998</td>
</tr>
<tr>
<td>valian</td>
<td>21</td>
<td>43</td>
<td>13,902</td>
<td>56,224</td>
<td>158,036</td>
<td>1.923</td>
<td>6.2952</td>
</tr>
<tr>
<td>weist</td>
<td>6</td>
<td>180</td>
<td>40,538</td>
<td>210,493</td>
<td>415,199</td>
<td>2.658</td>
<td>7.857</td>
</tr>
<tr>
<td><strong>Totals</strong></td>
<td><strong>143</strong></td>
<td><strong>3,559</strong></td>
<td><strong>855,449</strong></td>
<td><strong>1,201,267</strong></td>
<td><strong>3,752,825</strong></td>
<td><strong>6,933,939</strong></td>
<td></td>
</tr>
</tbody>
</table>

### Table 5: Description of Spanish CHILDES Corpora

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Children Sessions</th>
<th>Child Utt</th>
<th>Adult Utt</th>
<th>Child Tokens</th>
<th>Adult Tokens</th>
<th>Min MLU</th>
<th>Max MLU</th>
</tr>
</thead>
<tbody>
<tr>
<td>aguirre</td>
<td>1</td>
<td>29</td>
<td>7,482</td>
<td>25,116</td>
<td>60,653</td>
<td>1.7712</td>
<td>4.5352</td>
</tr>
<tr>
<td>becasesno</td>
<td>79</td>
<td>80</td>
<td>33,963</td>
<td>290,539</td>
<td>170,969</td>
<td>2.7396</td>
<td>10.9216</td>
</tr>
<tr>
<td>fernaguado</td>
<td>55</td>
<td>408</td>
<td>60,056</td>
<td>273,012</td>
<td>333,089</td>
<td>2.0757</td>
<td>5.42995</td>
</tr>
<tr>
<td>lgo</td>
<td>1</td>
<td>60</td>
<td>12,054</td>
<td>58,748</td>
<td>117,214</td>
<td>1.34725</td>
<td>6</td>
</tr>
<tr>
<td>linaza</td>
<td>3</td>
<td>25</td>
<td>2,591</td>
<td>10,689</td>
<td>16,286</td>
<td>1.773</td>
<td>5.79875</td>
</tr>
<tr>
<td>marrero</td>
<td>3</td>
<td>12</td>
<td>4,222</td>
<td>21,524</td>
<td>27,399</td>
<td>1.3082</td>
<td>6.4774</td>
</tr>
<tr>
<td>menda</td>
<td>3</td>
<td>32</td>
<td>15,109</td>
<td>37,610</td>
<td>83,561</td>
<td>1.9194</td>
<td>4.3279</td>
</tr>
<tr>
<td>OreaPine</td>
<td>2</td>
<td>127</td>
<td>26,784</td>
<td>87,804</td>
<td>212,765</td>
<td>1.5742</td>
<td>3.924</td>
</tr>
<tr>
<td>ornat</td>
<td>1</td>
<td>115</td>
<td>9,838</td>
<td>45,364</td>
<td>68,400</td>
<td>1.84325</td>
<td>6.13225</td>
</tr>
<tr>
<td>romero</td>
<td>1</td>
<td>2</td>
<td>411</td>
<td>1,384</td>
<td>8,162</td>
<td>1.95165</td>
<td>7.62195</td>
</tr>
<tr>
<td><strong>Totals</strong></td>
<td><strong>149</strong></td>
<td><strong>890</strong></td>
<td><strong>172,510</strong></td>
<td><strong>197,404</strong></td>
<td><strong>851,790</strong></td>
<td><strong>1,098,498</strong></td>
<td></td>
</tr>
</tbody>
</table>
Mean length of utterance (MLU) was computed for each speaker (synthetic adults and children) as the ratio of morpheme counts against utterance counts. The mlu program, part of the CLAN package published by CHILDES, was used to obtain the relevant counts. Kernel density plots and empirical CDF plots of child MLUs per session for the two languages appear in Figures 7 and 8.
For both languages a plurality of sessions (30%) cluster around MLU ~ 4. In over 80% of the sessions the MLU of the target child is less than 6.0.

Data Transformations

Stage I

The CHAT transcript files were fed to Freeling's morphological analyzer. The analyzer pipeline takes care of tokenization, lemmatization, sentence detection, and, for English text, Penn Treebank POS tags. In the Spanish case the analyzer outputs EAGLES tags ("The EAGLES Guidelines," n.d.)
Stage II

A filter script searched the tagged files for instances of Ps. In the English data these were tagged as prepositions proper (tag IN), except that all tokens of *to* were tagged with TO, a synthetic tag that merges prepositional *to* with the English infinitival marker. The script examined each occurrence of TO. Those that were followed by a bare infinitival verb were excluded; the others were taken to be dative/benefactive or directional Ps. A similar heuristic was applied to the Spanish tagged transcripts, for complements of the particle a “to”. The Freeling analyzer marks particles as adverbial elements, so these were masked out without extra steps. Unfortunately the analyzer is not always successful at distinguishing particle uses from Ps in construction, acutely so when the verb and the particle are separated by intervening material. I made no attempt to systematically handle these situations. Sentences ending with a P after a verb, likely particle uses, were however discarded.

Stage III

The output from stage I was sent to Freeling’s shallow parser (or “chunker”), a tool that groups string-adjacent elements into a nearly flat sequence of constituents (chunks). For example,

(1) The cats stole the food on the table.

might yield:

(2) [NP [D The][N cats]][VP stole [NP [D the][NP food]][PP on] [NP [D the][NP table]]]

A state machine walked through all occurrences of PPs in the chunked output, returning for each P its “complement”, defined as the chunk right adjacent to the P. For multiword complements a straightforward heuristic was applied to determine the heads of the complements: in both English
and Spanish PPs the rightmost noun or pronoun was taken as the head. Occurrences of multiword complements containing neither nouns nor pronouns were set aside for error analysis. PPs whose complements contained material marked unintelligible by the transcribers were discarded. English Ps followed by sentence-final punctuation (periods, interrogation and exclamation marks) were identified as possibly stranded and retained; these were instead removed in Spanish, since Romance disallows prepositional stranding. Totals for the number of adpositional phrases extracted are given in table 6, where counts for stranding candidates in English are in parentheses.

Table 6: PP Counts by language

<table>
<thead>
<tr>
<th>Language</th>
<th>Child PPs</th>
<th>Adult PPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>English (stranded)</td>
<td>95,732 (13,580 = 14%)</td>
<td>238,556 (20,709 = 9%)</td>
</tr>
<tr>
<td>Spanish</td>
<td>38,885</td>
<td>52,655</td>
</tr>
</tbody>
</table>

The next two tables show counts for the top ten P types in English and Spanish, by MLU group, where an MLU group is simply the truncated MLU of the child (so an MLU of 1.9 corresponds to group 1). Note that at this level of description the data are aggregated for all children. The nested character of the data – occurrences of Ps grouped by child – is addressed in the linear mixed effects model discussed later in the chapter, where we will also see that for the number of clusters in play (143 children), a mixed-effects model gives results very similar to a simpler model without random effects.

Table 7: P counts by MLU group (English)

| MLU group | P Type |
English Ps are fairly well represented throughout the MLU range. The frequencies of Spanish Ps instead drop dramatically when one consider types beyond the five most common (a, de, con, en, para, por). For example, there were no occurrences of desde for MLU group 1.

Table 8: P counts by MLU group (Spanish)

<table>
<thead>
<tr>
<th>MLU group</th>
<th>a</th>
<th>con</th>
<th>de</th>
<th>dentro</th>
<th>desde</th>
<th>en</th>
<th>hasta</th>
<th>para</th>
<th>por</th>
<th>sin</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1172</td>
<td>169</td>
<td>403</td>
<td>2</td>
<td>0</td>
<td>129</td>
<td>6</td>
<td>136</td>
<td>111</td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td>2243</td>
<td>683</td>
<td>1308</td>
<td>4</td>
<td>2</td>
<td>932</td>
<td>24</td>
<td>677</td>
<td>365</td>
<td>42</td>
</tr>
<tr>
<td>3</td>
<td>2698</td>
<td>862</td>
<td>1755</td>
<td>8</td>
<td>12</td>
<td>1202</td>
<td>60</td>
<td>967</td>
<td>592</td>
<td>45</td>
</tr>
<tr>
<td>4</td>
<td>1636</td>
<td>692</td>
<td>1530</td>
<td>3</td>
<td>9</td>
<td>853</td>
<td>60</td>
<td>573</td>
<td>570</td>
<td>36</td>
</tr>
<tr>
<td>5</td>
<td>1196</td>
<td>404</td>
<td>1176</td>
<td>2</td>
<td>17</td>
<td>587</td>
<td>48</td>
<td>293</td>
<td>302</td>
<td>25</td>
</tr>
<tr>
<td>6</td>
<td>781</td>
<td>383</td>
<td>922</td>
<td>1</td>
<td>9</td>
<td>413</td>
<td>22</td>
<td>189</td>
<td>190</td>
<td>15</td>
</tr>
</tbody>
</table>
Study 1: Error Analysis

The status of grammatical errors in child language

Two central syntactic facts of child production data demand explanation: (1) children's omission of functional elements and/or features (Bloom 1970, Brown 1973, Radford 1990 are loci classici); (2) their correct use of functional elements, modulo omitted features, when they do express them. In William Snyder's terms, errors of omission are frequent, those of co-mission rare (Snyder, 2011). The presence of seemingly ungrammatical (non-adult-like) exponents in child language is often (perhaps always) interpretable as the omission of structure and features. Consider the following sentence from the Post corpus:

(3) *Her in her pen. (Child, 2; 0.24, file tow06)

The accusative subject is of course not typical of adult speech, but before concluding that the child does not “have” English pronouns, or even Structural Case, we note that the case on the possessive determiner in the complement of the preposition is correct. The child here appears to overgeneralize object case to the subject. According to Schütze and Wexler (1996), this specific phenomenon is not uncommon among English children, while German children instead tend to overgeneralize the other way by extending nominative case to verbal objects. Whether or not we accept Schütze and Wexler's explanation – that children do not project Agr, so that default case is assigned, i.e. accusative in English and nominative in German – the systematic nature of the error, especially when viewed cross-linguistically, calls for a systematic explanation. By explaining the non-target use of functional elements in terms of omission (e.g. of Agr), a process is hypothesized for which the patterns of omission are seen as significant in their relation to syntactic knowledge. On
this view the systematicity of the errors – the constraints on types of errors – becomes evidence of a priori syntax.

So we observe a growing sense, e.g. arising from the experimental work of Shi and colleagues with infants (Shi & Melançon, 2010), that children can and do attend to f-morphs, and perhaps exploit the perceptual salience of f-morphs for basic linguistic processing such as word segmentation and the determination of phrase boundaries. The research on “filler syllables” (Veneziano & Sinclair, 2000), (Pepinsky & Roark, 2001) or monosyllabic place holders (MPH) (Tedeschi, 2009) suggests that, as evidenced by their production of schwa-like syllables in complementary distribution with full exponents, children know at least about the positions of f-morphs, and, given their early understanding of specific grammatical distinctions (e.g. definite vs indefinite, Shi, 2014), children understand their meaning. Why then the omissions in production?

The field has generated a wide range of proposals to explain the gap between children’s understanding and overt expression, from Rizzi’s (1994) and Wexler’s (1996) maturational models, to phonological constraints on speech (Gerken & McIntosh, 1993), to processing limitations (Valian, 1991) and various discourse, pragmatic and information-structural accounts, e.g. Serratrice, Sorace, & Paoli (2004) who argue children omit functional material that is redundant given the discursive context, thus explaining the asymmetry in elision between subjects and objects, since subjects often refer to given information. While deterministic accounts such as Radford’s small clauses hypothesis have been partially disconfirmed by child data showing that omissions are more a matter of tendency than clearly-marked developmental patterns in time, the union of the remaining viable accounts point to a complex of factors – processing, pragmatic, phonological – with respect to which children must optimize production (Eisenbeiss, n.d.). Remarkable here is that in “solving”
this web of constraints, children obey certain regular patterns which themselves constitute evidence of linguistic awareness. For example, determiners and the copula are very often absent at lower MLUs. In selecting candidates for deletion children therefore seem to grasp both the function and the entropy of grammatical units (Phillips, 1995).

Usage-based accounts explain observed errors in child speech simply as the conservative repetition of the input (for correct constructions), and otherwise as resulting from children's lack of structural knowledge (omissions). But this position runs into trouble when confronted by patterns of systematic deletion and overgeneralization. Why do children select just these items for deletion, and only sometimes actually delete? What knowledge of language do they draw upon in such computations? More urgently, what explanation can there be for the apparent effect of language-specific defaults? If children seem to know that in English the default structural case is accusative, how might this knowledge be represented if erroneous forms never occur in the input and if the “parameter” cannot be easily represented by item-specific constructions?

The urgent need for detailed and insightful analyses of errors in child production emerges strongly in Naigles (2002). Responding to experimental work reported in Olguin and Tomasello (1993) and Akhtar and Tomasello (1997), Naigles argues that the simplistic assignment of child speech to mutually exclusive “adult-like, correct” and “non-adult-like, incorrect” categories leads to questionable conclusions about the linguistic competence of toddlers – particularly if one fails to take into consideration the specific demands that characterize the experimental tasks.

For example, for a complex stimulus in which Ernie is shown jumping onto one end of a see-saw, resulting in Bert, at the other end, being projected through the air, Olguin and Tomasello and Akhtar and Tomasello train the child with the sentence *Ernie is chamming Bert*. Note that linguists typically...
decompose this sort of causative accomplishment into two events, e.g. the jumping event that causes the flying event. Which of these events (or both) is indicated by *chamming* is therefore not obvious. If a child later uttered *Bert is chamming*, the researchers logged an error because object and subject were switched. But is this truly an error? As Naigles points out, the very fact of this production shows that children are able to employ a verb in multiple syntactic configurations (transitive and intransitive in this case). More importantly, at the level of meaning, the child may have determined that *chamming* expresses the second event only, in which case the intransitive form is not unreasonable. Naigles offers the example *Adam drops the ball / The ball drops* to demonstrate this sort of alternation in the use of causative verbs. On her analysis, therefore, what at first appeared as an error – perhaps because the child merely failed to interiorize the meaning intended by the researchers – might in fact constitute evidence for the sophistication of the child's syntactic knowledge.

It therefore will not do, as Littlefield has done for prepositions (Littlefield, 2006), to total up all omissions and substitutions as “errors” (omissions were 83% of her count) and declare the category known (or not) based on a criterion of 90% error-free production. That functional prepositions fail the 90% criterion at lower MLU might, for all we know, indicate that children are aware of the difference between lexical and functional prepositions and choose to emit only the cognitively richer kind when pressed. When Morgensteren and Sekali (2009) compared the development of a single English-speaking child to that of a French child, they found – perhaps to their own surprise – that while the English learner used more spatial Ps than functional, the French child exhibited exactly the opposite preference. And Yáñez and Zúñiga (2009) found no significant difference (and almost zero usage errors) between P types for monolingual Spanish learners. These data are troubling for empiricists like Littlefield, since their claim of non-continuity is necessarily independent of language
A similar cross-linguistic contrast is visible in the CHILDES dataset constructed for the present study. English learners at low MLUs unquestionably express the basic spatial Ps *in* and *on* at higher proportions of their total P productions when compared to adult behavior. The following graph (Figure 9) shows the smoothed distributions (using locally weighted scatterplot smoothing [LOWESS]) across selected Ps for MLU from 1.5 to 5 in increments of 0.25, where the child relative frequency for each P is computed in proportion to the corresponding relative frequency for adults. Adult-like behavior is represented by the line $y = 1.0$. For example, for the MLU band (2.00 to 2.25) 36% of children’s prepositions are *in*; for adults, 20%. The ratio of relative frequencies is $36/20 = 1.8$ (the plotted value is smaller due to smoothing).

![Figure 9: Child emissions of selected Ps relative to adult (English)](image-url)
At low MLUs it is evident that children show a strong preference for high-frequency locative Ps over functional Ps. Figure 9 summarizes the essence of Littlefield's argument – the asymmetry in expression between lexical and functional Ps means the latter develop more slowly, and this is an essential fact about language. On that account, however, how would one explain the observation that in Spanish no such asymmetry is visible (Figure 10)?

![Graph](image)

*Figure 10: Child emissions of selected Ps relative to adult (Spanish)*

The descriptive problem is brought into sharper focus by comparing the relative frequencies over MLU bands of Ps partitioned into two sets, lexical and functional. A “long” data set was constructed by computing, for each individual target child in the corpus, the proportions of selected
(frequent) lexical and functional Ps relative to total morphemes uttered over each MLU band (in increments of 0.25), per the classification in Table 9 (see Chapter 1 for theoretical arguments supporting the classification).

**Table 9: Classification of English and Spanish Ps**

<table>
<thead>
<tr>
<th></th>
<th>Lexical</th>
<th>Functional</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td><em>in on onto</em></td>
<td><em>of for with to at about</em></td>
</tr>
<tr>
<td>Spanish</td>
<td><em>en hasta desde</em></td>
<td><em>a de con para por</em></td>
</tr>
</tbody>
</table>

Each row in the dataset specified the corpus, the child's name, the MLU band, the P classification (lexical or functional), and the relative frequency of the P class. The relative frequencies per child per MLU band were then averaged for each MLU band and plotted (Figures 11 and 12 for English and Spanish respectively). The plots clearly show that English learners produce lexical Ps at almost constant rates right from the start, while the relative frequencies of English functional Ps increase gradually over MLU. For the Spanish data, in contrast, we observe essentially flat growth for both orders, the base rate for functional Ps being higher since most highly frequent Spanish Ps are functional.
Figure 11: Relative frequencies of lexical and functional Ps by MLU band (English children)
Interestingly, if the same procedure is applied to adult data, we observe a perhaps surprising correlation between the relative frequencies of Ps for children and adults. A second long dataset was constructed that was identical to the previous one, except that the relative frequency computed was based on the adult speech in the transcripts for the given child in the given MLU band. The resulting plot for the Spanish parents is virtually identical to that of the Spanish children:
The plot for the English parents reveals slopes that are similar to those of the English children – positive and substantial for functional Ps, slightly negative for lexical Ps – but the locations (intercepts) are very different.

*Figure 13: Relative frequencies of lexical and functional Ps by MLU band
(Spanish adults)*
In Figure 14 the functional Ps are always more frequent than lexical Ps, while for the children plotted in Figure 11 the growth in frequency of functional Ps intersects the mostly horizontal line of the lexical Ps at around MLU = 3. In other words, it seems that with respect to the expression of Ps, Spanish children behave just like adults right from the start, while for English children the expression of functional Ps is relatively delayed until MLU is around 3.

Further confirmation of this difference was obtained by submitting the frequency data sets to statistical analysis using a linear mixed-effects model (LMM). The model attempted to account for the effect of language development (MLU band), the classification or “order” of Ps (lexical vs
functional), and adult behavior on the relative frequencies of Ps in child speech. A mixed effects
model in which child observations were nested in children was adopted in order to account for
longitudinal effects (non-independence of observations). Using R’s formula notation, extended in
R’s lme4 package for multilevel regression models, the model was specified as:

$$ rf_{child} \sim (mlu\_band + rf\_adult) \times order + (1 + mlu\_band + rf\_adult | child) $$

The formula says that child relative frequency is regressed onto MLU band, adult relative frequency,
the order of P, and the interactions of MLU band and adult relative frequency with order, as fixed
effects. The intercept of child relative frequency and the effects of MLU and adult frequency are
also random effects, i.e. are allowed to vary between children. The purpose of the random effects
was to consume the variation between subjects; interpretation focused on the fixed effects. The
standardized coefficients\(^{35}\) estimated by the \texttt{lmer} function for the English data were:

<table>
<thead>
<tr>
<th>Estimate</th>
<th>Std. Error</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-0.36870</td>
<td>0.06588</td>
</tr>
<tr>
<td>orderL</td>
<td>0.74633</td>
<td>0.11858</td>
</tr>
<tr>
<td>mlu_group</td>
<td>0.38994</td>
<td>0.04838</td>
</tr>
<tr>
<td>rf_adult</td>
<td>0.28629</td>
<td>0.06675</td>
</tr>
<tr>
<td>orderL:mlu_group</td>
<td>-0.33067</td>
<td>0.05414</td>
</tr>
<tr>
<td>orderL:rf_adult</td>
<td>0.22648</td>
<td>0.08539</td>
</tr>
</tbody>
</table>

All fixed effects were significant. There was a positive effect for MLU band that went to nearly zero
for lexical Ps (orderL = 1, where orderL denotes lexical Ps, interaction effect = -0.331), i.e. there was
a significant slope only for the functional Ps, as is also evident in Figure 11. The largest effect was
that of the order of Ps: the relative frequency of lexical Ps was almost \(\frac{3}{4}\) of a deviation (0.746)
higher. There was a visible effect for the adult frequency covariate, whose magnitude was higher for
lexical Ps, (interaction of order with rf\_adult = 0.226). We can therefore conclude that while there

\(^{35}\) Variables rf\_adult, rf\_child and mlu\_band were standardized (scaled and centered) prior to running the estimation,
such that the regression coefficients were standardized. As a consequence, the standard errors and t-values refer to
the standardized data.
was an interaction with respect to P usage between the speech patterns of the English-speaking adults and children (we cannot determine the causal relation), this was more true for lexical Ps than functional. The analysis thus provides a fairly robust confirmation of the pattern we see in Figure 11: at or near the initial stages of acquisition English-speaking children omit more functional Ps compared to their almost adult-like use of lexical Ps.

It is instructive to compare the estimates of the LMM to that of a simple model with fixed effects only, since it is sometimes suggested that as the number of clusters in a hierarchical model reaches 50 or so, the two methods yield similar results. There were 143 children in the English data set.

<table>
<thead>
<tr>
<th>Estimate</th>
<th>Std. Error</th>
<th>t value</th>
<th>df</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-0.37725</td>
<td>0.05296</td>
<td>-7.123</td>
<td>2.32e-12 ***</td>
</tr>
<tr>
<td>orderL</td>
<td>0.91977</td>
<td>0.07629</td>
<td>12.056</td>
<td>&lt; 2e-16 ***</td>
</tr>
<tr>
<td>mlu_group</td>
<td>0.37235</td>
<td>0.04485</td>
<td>8.303</td>
<td>4.19e-16 ***</td>
</tr>
<tr>
<td>rf_adult</td>
<td>0.33547</td>
<td>0.05232</td>
<td>6.411</td>
<td>2.44e-10 ***</td>
</tr>
<tr>
<td>orderL:mlu_group</td>
<td>-0.36323</td>
<td>0.06159</td>
<td>-5.897</td>
<td>5.40e-09 ***</td>
</tr>
<tr>
<td>orderL:rf_adult</td>
<td>0.28602</td>
<td>0.07834</td>
<td>3.651</td>
<td>0.000278 ***</td>
</tr>
</tbody>
</table>

Again all effects were significant, and most were rather close to the estimates for the LMM – the main effects for order and rf_adult were stronger in the fixed effect model. As a point of methodology, this result supports the rule-of-thumb that above 50 clusters LMMs and non-nested LMs return roughly equivalent results.

The Spanish data present a very different picture:

<table>
<thead>
<tr>
<th>Estimate</th>
<th>Std. Error</th>
<th>t value</th>
<th>df</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>0.52282</td>
<td>0.04855</td>
<td>10.769</td>
<td>&lt; 2e-16 ***</td>
</tr>
<tr>
<td>orderL</td>
<td>-1.29427</td>
<td>0.12668</td>
<td>-10.217</td>
<td>&lt; 2e-16 ***</td>
</tr>
<tr>
<td>mlu_group</td>
<td>0.05816</td>
<td>0.03506</td>
<td>1.659</td>
<td>0.0977 .</td>
</tr>
<tr>
<td>rf_adult</td>
<td>0.29037</td>
<td>0.04497</td>
<td>6.457</td>
<td>2.31e-10 ***</td>
</tr>
<tr>
<td>orderL:mlu_group</td>
<td>-0.04895</td>
<td>0.05081</td>
<td>-0.963</td>
<td>0.3358</td>
</tr>
<tr>
<td>orderL:rf_adult</td>
<td>-0.22357</td>
<td>0.13901</td>
<td>-1.608</td>
<td>0.1083</td>
</tr>
</tbody>
</table>
The effect of MLU band is slight, as is its interaction with P order, suggesting little change during language development. The effect of adult relative frequency is moderate, while order = lexical has a strongly negative effect. So when it comes to P usage, Spanish learners easily reflect the characteristics of the target from the earliest stages of language development; we see no sign of the delay in producing functional Ps that is observed in the English data.

How might the empirically observed differences between English and Spanish learners be explained? Or rather, what significance can be assigned to errors of omission in children's speech?

Morgensteren and Sekali view their findings as evidence that children’s “discourse” is influenced by the target language (no argument there), in line with Talmy's typological distinction between verb-framed languages (e.g. Romance) and satellite-framed languages (e.g. Germanic). They want to argue that the greater production of functional Ps in French merely reflects a language trait, and in any case children use these morphemes for “pragmatic” rather than syntactic reasons. Yet one could equally well argue that in making smart pragmatic decisions about which items to leave out of their expressions, children must already somehow know what roles the units in play in language.

Whatever the case, the point is that a perspicacious analysis of errors would establish the precise linguistic characteristics of each occurrence, the better to understand their linguistic motivation.

Since my aim here is to apply automated methods to larger corpora, the approach to error analysis is more superficial. It is mostly intended as a way into the database; a descriptive tool. But before proceeding I want to make clear that, based upon scanning a large number of child PPs from the CHILDES data, and setting aside omissions, children overall use preposition with remarkable accuracy from the very start. Valian (1986) already established this on a smaller sample of English learners. And because my dataset is constructed from overt expressions of prepositions I cannot
address patterns of omission. In the end, due to the scarcity of errors of co-mission, the error analysis will have yielded little.

**Empirical analysis of errors**

**Procedure**

Only the English corpus was considered for error analysis. Four computer programs looking for specific error patterns were run over the child data in the P database. The four criteria were:

1. Ps taking only a bare determiner as complement.

2. All cases in which the transcriber noted a correction, where either the original token or the correction was in the set \{*in, on, for, with, of, at, like*\}, representing the most frequent P types in correction annotations. In the CHAT format these are annotated as `TOKEN [: CORRECT TOKEN ]`.

3. One thousand PPs selected at random from the dataset of presumed correct constructs, limited to sessions in which the MLU of the target child was less than 2.5.

4. Child utterances containing sequences of two or more Ps, not including cases in which the same P was repeated, which is typically just the child stuttering or repeating himself.

We can immediately dispose of (1) as these were all cases of interrupted phrases or utterances. The rest I investigated by hand in order to isolate cases of true errors from other phenomena. Significant numbers of broken-off phrases and self-corrections were removed in the process. Also excluded

---

36As is well-known, the normative uses of Ps are difficult for L2 learners; I did not trust myself in Spanish.
were particle uses and instances of the infinitival marker to mis-tagged as P; adjectival senses of like; and instances where the omission of required material resulted in sequenced Ps, such as in (4) where the one is missing between with and on. (The parent immediately after clarifies by echoing the utterance but adding the missing element.)

(4) wrong with on the window sill (Peter, 2;6.16, file peter15).

Finally, correctly sequenced Ps like “from inside” were removed from the error dataset.

Results

Annotated Substitutions

My intent was to determine whether, as Littlefield found, there are measurable counts of substitutions in the data, cases where the child chose the wrong P. Transcribers seem to indicate substitutions only rarely, unfortunately, and even then the overwhelming majority of the entries correct the pronunciation, not choice of preposition. This is especially prominent in the Hall corpus where sessions are classified for race and class. So there are events such as:

(3.3) I on [: don’t] know. (Kag, 4;6. - 5;0., file Hall/WhiteWork/kag.cha)

All occurrences of spoken on were corrected as don’t.

(3.4) I on [: don’t] want my water (.). (Mis, 4;6. - 5;0., file Hall/BlackWork/mis.cha)

(3.5) cookie e [: in] pot. (Laura, 01;04.27, file Braunwald/diary/010427.cha)

(3.6) inna [: in] du [: the] woom [: room] . (Jimmy, 2;9.23, file Demetras2/j_mot16.cha)

Schwa-substitution for of was common in the Hall corpus, 126 out of 146 occurrences of the
pattern.

(3.7) you pack a [: of] raisins are empty? (Sat, 4;6, file Hall/WhiteWork/sat.cha)

Setting all such cases aside, all of nine occurrences (by token) of proper substitutions (P to P, out of over 95,000 events) were found. See Table 10.

Patterns of pronunciation might be of interest for a study of filler syllables but I did not pursue the issue. In the event, given my focus on what the children intended to say, I collapsed all substitutions in favor of the correct element.

Table 10: Transcriber-Marked P Substitutions

<table>
<thead>
<tr>
<th>Utterance</th>
<th>Child</th>
<th>Age</th>
<th>MLU</th>
<th>Corpus</th>
<th>File</th>
</tr>
</thead>
<tbody>
<tr>
<td>yeah , [/] they're something different in [: from] lobsters.</td>
<td>Rob</td>
<td>4;6 - 5;0</td>
<td>4.59</td>
<td>hall</td>
<td>rob</td>
</tr>
<tr>
<td>no they went in [: to]** Ashtabula.</td>
<td>Emi</td>
<td>2;6,20</td>
<td>3.5</td>
<td>weist</td>
<td>emi02</td>
</tr>
<tr>
<td>thank+you of [: for]** making oatmeal to day.</td>
<td>Laura</td>
<td>03;01.22</td>
<td>5.58</td>
<td>braunwald</td>
<td>030122</td>
</tr>
<tr>
<td>Thank+you of [: for]** buying this much strawberries.</td>
<td>Laura</td>
<td>03;01.22</td>
<td>5.58</td>
<td>braunwald</td>
<td>030122</td>
</tr>
<tr>
<td>Mama, thank+you of [: for]** making nice oatmeal for me.</td>
<td>Laura</td>
<td>03;01.22</td>
<td>5.58</td>
<td>braunwald</td>
<td>030206</td>
</tr>
<tr>
<td>tow [: two] minutes [: minutes] for [: before] my mommy gets here.</td>
<td>Jimmy</td>
<td>2;9,22</td>
<td>3.25</td>
<td>demetras</td>
<td>j_fat15</td>
</tr>
<tr>
<td>uhhuh we gonna go at [: like] dis [: this]</td>
<td>Kig</td>
<td>4;6 - 5;0</td>
<td>3.94</td>
<td>hall</td>
<td>kig</td>
</tr>
<tr>
<td>I just went at [: to]** s:r the doctor.</td>
<td>Emi</td>
<td>02;08.13</td>
<td>3.84</td>
<td>weist</td>
<td>emi06</td>
</tr>
<tr>
<td>and Santa come at [: to] my house.</td>
<td>Emi</td>
<td>02;10.28</td>
<td>4.04</td>
<td>weist</td>
<td>emi10</td>
</tr>
</tbody>
</table>

Other than Laura’s confusing of for for, these substitutions all involve spatial Ps, but no conclusions can be drawn from this minuscule dataset.
Analysis of sample PPs

Out of the 1000 PPs randomly sampled from children of MLU <= 2.5, I counted sixteen true errors. But seven of these were cases of omitted *of*. Removing these (for the reasons given above) left nine errors (Table 11). To put this number in context, consider that in her analysis of five sets of transcripts, Littlefield found that children committed 1,601 errors out of 25,697 PP events (Littlefield, 2006, p. 195). Since, according to her report, 83% of these were omissions (p. 210), the child error rate without counting omissions was \( \frac{0.17 \times 1601}{25697} \approx 0.01 \) (or ~ 1%). An exact binomial test of the current result, 9 errors for 1000 trials, gives a 95% confidence interval for the error rate of 0.41% to 1.70% (p < 0.001 that the probability is 0.5) and an estimate of 0.9%. The child error rates (excluding omissions) in the two studies are thus quite close.

<table>
<thead>
<tr>
<th>Corpus</th>
<th>File</th>
<th>P type</th>
<th>Complement</th>
<th>Error</th>
<th>Utterance</th>
</tr>
</thead>
<tbody>
<tr>
<td>providence</td>
<td>nai18</td>
<td>out</td>
<td>the/DT way/NN</td>
<td>missing &quot;of&quot;</td>
<td>yyy out the way, yyy out the way</td>
</tr>
<tr>
<td>manchester</td>
<td>liz03b</td>
<td>out</td>
<td>tunnel/NN</td>
<td>missing &quot;of&quot;</td>
<td>out tunnel .</td>
</tr>
<tr>
<td>brown</td>
<td>sarah020</td>
<td>out</td>
<td>step/NN</td>
<td>missing &quot;of&quot;</td>
<td>out step .</td>
</tr>
<tr>
<td>manchester</td>
<td>anne08a</td>
<td>in</td>
<td>table/NN</td>
<td>&quot;in&quot; for &quot;at&quot;</td>
<td>sit in table .</td>
</tr>
<tr>
<td>brown</td>
<td>adam05</td>
<td>from</td>
<td>for/IN</td>
<td>missing &quot;of&quot;</td>
<td>come from for ?</td>
</tr>
<tr>
<td>manchester</td>
<td>john22b</td>
<td>on</td>
<td>the/DT carriage/NN</td>
<td>&quot;on&quot; for &quot;of&quot;</td>
<td>top on the carriage .</td>
</tr>
<tr>
<td>suppes</td>
<td>nina10</td>
<td>in</td>
<td>food/NN</td>
<td>missing &quot;of&quot;</td>
<td>that top in food .</td>
</tr>
<tr>
<td>manchester</td>
<td>liz15a</td>
<td>in</td>
<td>my/PRP$ mat/NN</td>
<td>&quot;in&quot; for &quot;under&quot;</td>
<td>I want put my food in my mat .</td>
</tr>
<tr>
<td>manchester</td>
<td>carl13b</td>
<td>on</td>
<td>the/DT water/NN</td>
<td>&quot;on&quot; for &quot;in&quot;</td>
<td>fishie on the water .</td>
</tr>
<tr>
<td>manchester</td>
<td>john22b</td>
<td>on</td>
<td>the/DT carriage/NN</td>
<td>&quot;on&quot; for &quot;of&quot;</td>
<td>top on the carriage .</td>
</tr>
<tr>
<td>brown</td>
<td>sarah020</td>
<td>out</td>
<td>step/NN</td>
<td>missing &quot;of&quot;</td>
<td>out step .</td>
</tr>
<tr>
<td>manchester</td>
<td>gail20a</td>
<td>out</td>
<td>way/NN</td>
<td>missing &quot;of&quot;</td>
<td>don't move that out way .</td>
</tr>
<tr>
<td>bloom70</td>
<td>peter06</td>
<td>out</td>
<td>church/NN</td>
<td>missing &quot;of&quot;</td>
<td>people . out church .</td>
</tr>
<tr>
<td>sachs</td>
<td>n47</td>
<td>on</td>
<td>my/PRP$ mouth/NN</td>
<td>&quot;on&quot; for &quot;in&quot;</td>
<td>I put it on my mouth .</td>
</tr>
<tr>
<td>providence</td>
<td>wil26</td>
<td>in</td>
<td>the/DT floor/NN</td>
<td>&quot;in&quot; for &quot;on&quot;</td>
<td>they stay in the floor .</td>
</tr>
<tr>
<td>manchester</td>
<td>john03b</td>
<td>out</td>
<td>way/NN</td>
<td>missing &quot;of&quot;</td>
<td>out way .</td>
</tr>
</tbody>
</table>

Looking at the errors, a child substitutes a high frequency locative for *of* (three cases) or a less common locative (two cases), or swaps *in* for *on* or vice-versa (three cases). The last case, *come from*
for\?, is opaque. Keeping in mind that in all cases the children making these substitutions elsewhere demonstrate productive use of the correct P, the fact that typically a more frequent P is expressed in place of a less frequent one (not the case when on is substituted for in) suggests these are mere slips of the tongue. In any event there is no direct evidence from these sentences of differential acquisition of Ps – lexical before functional, as argued for by Littlefield. If anything, the slippage between spatial Ps (e.g. in for on or at) points to some instability around the descriptions of space, not syntax.

**Repeated Ps**

There were originally 711 cases of repeated, non-identical Ps in child utterances. Seventy-three of these were isolated as true errors.

Forty-one cases consisted of the functional P for taking a TO-infinitival complement. The phenomenon was relatively widespread: 25 children produced at least one such construction; Nina did so five times. Although not relevant to the present study, since the P piece of these construction was correct, these data might inform the long-standing debate on the structure of the clause in early child language. That children choose the infinitival form of the verb, instead of the present participle, might result from their not projecting a complete INFL (say, lacking Fin) that would host the ing morpheme. This would support a proposal in Rus and Chandra (2005), according to which (bare) participles in root position are finite clauses from which an auxiliary has been elided. In other words, for-to constructions obey a possibility in UG that is manifested cross-linguistically (e.g. in Dutch) and even in some varieties of English (Den Dikken, p.c.).

Setting aside these cases left 32 cases of non-target-like P uses, which mostly involved odd combinations of spatial Ps; see Table 14.
Table 12: Sequenced P errors

<table>
<thead>
<tr>
<th>Error Type</th>
<th>Utterance</th>
<th>Child</th>
<th>Age</th>
<th>MLU</th>
<th>Corpus</th>
<th>File</th>
</tr>
</thead>
<tbody>
<tr>
<td>K</td>
<td>look at in there.</td>
<td>Peter</td>
<td>2;0.10</td>
<td>2.286</td>
<td>bloom70</td>
<td>peter06</td>
</tr>
<tr>
<td>K</td>
<td>daddy sit on over there.</td>
<td>Peter</td>
<td>2;5.03</td>
<td>3.10</td>
<td>bloom70</td>
<td>peter13</td>
</tr>
<tr>
<td>K</td>
<td>on in there ?</td>
<td>Adam</td>
<td>3;0.11</td>
<td>3.33</td>
<td>brown</td>
<td>adam20</td>
</tr>
<tr>
<td>K</td>
<td>shine on in there !</td>
<td>Shem</td>
<td>2;3.16</td>
<td>3.41</td>
<td>clark</td>
<td>shem04</td>
</tr>
<tr>
<td>K</td>
<td>xxx if I put water in here , it goes down in into here .</td>
<td>Chj</td>
<td>4;6 – 5;0</td>
<td>3.84</td>
<td>hall</td>
<td>chj</td>
</tr>
<tr>
<td>K</td>
<td>Robbin could you put this inside of here ?</td>
<td>Ref</td>
<td>4;6 – 5;0</td>
<td>3.77</td>
<td>hall</td>
<td>ref</td>
</tr>
<tr>
<td>K</td>
<td>if you bust his butt going to fly up to into the the air haha .</td>
<td>Ref</td>
<td>4;6 – 5;0</td>
<td>3.77</td>
<td>hall</td>
<td>ref</td>
</tr>
<tr>
<td>K</td>
<td>what's the name of over here ?</td>
<td>Violet</td>
<td>2;6.16</td>
<td>3.68</td>
<td>providence</td>
<td>vio34</td>
</tr>
<tr>
<td>K</td>
<td>hey. it's in to here.</td>
<td>Sarah</td>
<td>5;1.06</td>
<td>4.84</td>
<td>brown</td>
<td>sarah139</td>
</tr>
<tr>
<td>K</td>
<td>waiting in to the bus.</td>
<td>Peter</td>
<td>2;3.24</td>
<td>2.78</td>
<td>bloom70</td>
<td>peter11</td>
</tr>
<tr>
<td>K</td>
<td>dad () &lt;in this&gt; [<del>/] in this little kid for over on () &lt;at t(v)&gt; [</del>/] on tv over there .</td>
<td>Ross</td>
<td>4;11.25</td>
<td>6.57</td>
<td>macwhinney</td>
<td>57b2</td>
</tr>
<tr>
<td>K</td>
<td>I going sit on in there.</td>
<td>Adam</td>
<td>3;8.01</td>
<td>4.09</td>
<td>brown</td>
<td>adam35</td>
</tr>
<tr>
<td>K or Frozen</td>
<td>why are the boys at to our house ?</td>
<td>Ross</td>
<td>3;4.26</td>
<td>4.33</td>
<td>macwhinney</td>
<td>40a2</td>
</tr>
<tr>
<td>K or Frozen</td>
<td>they're at to work.</td>
<td>Ross</td>
<td>3;4.26</td>
<td>4.90</td>
<td>macwhinney</td>
<td>39b2</td>
</tr>
<tr>
<td>K or Frozen</td>
<td>I've () breakfast () in at night I mean I have xxx xxx in in morning () morning () .</td>
<td>Bom</td>
<td>4;6 – 5;0</td>
<td>3.36</td>
<td>hall</td>
<td>bom</td>
</tr>
<tr>
<td>Part</td>
<td>these because they'll get . get caught in by the bad guys .</td>
<td>Trevor</td>
<td>2;10.04</td>
<td>3.29</td>
<td>demetras</td>
<td>tre13</td>
</tr>
<tr>
<td>Part</td>
<td>I'm going to dive my I'm going to dive my pole in into the ocean and I'm going to get some more fish .</td>
<td>Trevor</td>
<td>3.1.17</td>
<td>5.97</td>
<td>demetras</td>
<td>tre20</td>
</tr>
<tr>
<td>Part</td>
<td>the airplane just landed in for us .</td>
<td>Trevor</td>
<td>3;11.11</td>
<td>4.40</td>
<td>demetras</td>
<td>tre26</td>
</tr>
</tbody>
</table>
The first column gives a classification of the errors, based on my own interpretations of the structural characterization of the errors. The claim is that in most cases the nature of the error points to what children already know of the structure of PPs.

- Errors marked with K ("Kase") nest spatial Ps, so that the lower P occupies a position below the typical insertion point for locative Ps., but above the Ground. The infelicity of these errors is due to an unfilled argument structure slot: if the higher P introduces the subject of the overall PP, then the subject introduced by the lower P remains unfilled. We saw in
Chapter 1 that in Svenonius' model of the PP a functional case (“Kase”) node intervenes between the Ground DP and the rest of the extended projection, and is the insertion point for pure RELATORS such as *of*. The semantic function associated with Kase maps the ground object to a region. And in fact in expressions such as *look at in there*, the PP *in there* might be said to describe a region in space. According to this analysis, these errors seem to indicate that children are dropping locative Ps into Kase. The semantic oddity results from the locative *in* playing the semantic role of a function from entities to regions. On Den Dikken's analysis, instead, the lower P might merge into the aspectual position above the lexical P, the latter then raising over it as per the incorporated *into*. Regardless of which P theory we adopt, K errors suggest that children are aware of the structural complexity of the PP. An alternative interpretation is that the lower PPs are frozen expressions. This impression is especially strong in Ross's case: *why are the boys at to our house*, they're *at to work*, where *to + X* act as formulae denoting Grounds. The same might be said of the more common P + deictic constructions, but because these are not as lexically specific – *here* and *there* are likely functional – and therefore more likely to be constructed rather than frozen, I have preferred the syntactic interpretation of the errors. That a subset of cases might constitute frozen expressions does not in any case invalidate the analysis of the remaining cases.

- **Part(icle)** errors are non-target-like verb+particle uses, such as *dive in, land in, catch in* (vs *catch on*). Here the children appear to invent non-idiomatic particle uses of Ps. The linguistic creativity at work shows that children understand the grammatical mechanism at work: in a verb + particle construction, the P is de-transitivized. This suggests that children (at or above MLU of 4) are able to learn such constructions not merely (or only) as low-level
formulae, but in terms of the syntactic analysis of their behavior.

- **PaPl** (“Path over Place”) errors are characterized by a path or aspectual P taking a locative P as its complement. Most are of the form for/from + in/on + Ground, that is, a Goal or Source path over a Place. On Marina Pantcheva’s structure of the PP, these utterances—which, semantically speaking, are easily interpretable—realize a possibility of the functional sequence of P that is unusual (though not impossible) in English, e.g. the contrast in acceptability between *the smell comes from in here/there,* #*the smell comes from in my kitchen.* We saw in Chapter 1 that there is broad agreement about the nesting of Place within Path expressions, i.e. [Path [Place …]]. Languages in which the lower locative is realized make this structure evident, e.g. the sample of Arawakan language Yanesha in 2.8, where the locative case marker is closer to the Ground *canoe* than the ablative case marker (Pantcheva, 2010, p. p.11), literally meaning “from in/at the canoe”.

\begin{align*}
(3.8) \quad & \text{non’t}^\text{i}-\text{o-t}^\text{v} \\
& \text{canoe-LOC-ABL} \\
& \text{‘from the canoe’}
\end{align*}

In sum, while repeated P errors indicate a developing understanding of the precise behavior (feature content) of P$s in the target language, the patterns of the errors is at least suggestive of structural possibilities afforded by Universal Grammar (K and PP errors) or by the grammar of English (Part errors).
Discussion

After describing the construction of the CHILDES data set used in the empirical studies, this chapter considered the general question of the informativeness of children's non-target-like constructions involving functional categories with respect to language development. Drawing on Snyder's key distinction between errors of omission and those of co-mission, it was argued that, at least for functional categories, it is not immediately obvious that omissions indicate immature knowledge about these categories. Their selective omission may instead argue for an optimization strategy (of expression) that depends on the prior understanding of their functions in language.

The P category is useful for investigating differences in language development between functional and lexical elements, since, on traditional hypotheses at least, the category incorporates both types of elements; inter-categorial differences are therefore sidestepped. The statistical analysis (using robust mixed effects methods) of rates of production of lexical and functional Ps found that in English children omit functional Ps at greater rates than lexical Ps early in acquisition. This result confirmed Littlefield's findings for a smaller set of children (Littlefield, 2006). The difference remained even after controlling for adult behavior. Spanish children however demonstrated no such preference – they used functional Ps at near adult rates right from the start of multi-word speech. Littlefield's empiricist thesis that the greater linguistic generality of functional Ps implies they are harder to learn is therefore disconfirmed by the Spanish data. It seems omissions alone (86% of Littlefield's “errors”) tell us little about the learnability of functional elements.

Errors of commission were then investigated by searching the corpus of child speech for two error patterns in the English data (only): overt indications of substitutions in the annotations, and sequenced Ps. In addition, one thousand randomly selected child utterances were manually reviewed
for errors in usage. In all there were remarkably few errors in children’s use of Ps; the number found was consistent with the rate of non-omission errors counted by Littlefield. The same was found for Spanish in other studies, e.g. Yánez & Zúñiga, (2009)

The errors that were found in some cases support an interpretation according to which children exploit the structural possibilities of UG, perhaps when they have not yet fully solidified the feature assignments of prepositions in the language. This last point is surely debatable, however, as many of the repeated P errors, for example, might also be interpreted as frozen expressions.

English and Spanish learners both employ prepositions very accurately from the start. The former produce considerably fewer functional prepositions at low MLUs, though this is hardly unambiguous evidence that functional prepositions are harder to learn in general. It seems other tools are necessary to determine whether there are visible differences in the acquisition of the two orders of adpositions. The next chapter presents two studies aimed at determining whether there are relative differences in children’s grammatical productivity with Ps over the course of language development.
Chapter 4. PP Productivity Studies

Introduction

The two experiments described in this chapter sought to shed light on the productivity of Ps in early child speech. In usage-based approaches children will conservatively reproduce adult constructions, gradually abstracting from them on their way to adult competence. On this account the variety of complements taken by their Ps is narrowly limited to that of their adult interlocutors. Development is gradual.

How gradual? On any theory it takes time for children to learn the lexicon and specific grammatical rules of the target language, and inevitably the conceptual content of their discourse is closely matched to those of the parents – the parents “prime” the children. We would expect the degree to which children’s PPs are primed by the parents to drop as the children build fluency and vocabulary. Since constructivists believe children to be conservative in their use of grammar – and because they hold adult grammar to be based on constructions that differ from those of children only in their degree of abstraction – no continuity problem is seen, nor is children’s speech expected to differ significantly from that of adults. So if the (slow) pace of acquisition is the principal distinguishing characteristic of usage-based theories of learning, then in order to test this idea we would want to compare observed rates of acquisition against some criterion of velocity. Unfortunately it is far from obvious how such a developmental criterion might be established.

The idea underlying Studies 2 and 3 is that we might verify the predictions made by constructivism by comparing the rates of acquisition of functional and lexical elements. Recall Tomasello’s thought on the acquisition of function words:
The early uses of these words thus serve specific discourse functions only, and a fully adult-like understanding **awaits** children’s encounters with these words in a fuller range of functional contexts (Tomasello, 2003, p. 45 my emphasis)

Goldberg’s notion of schematicity and its relation to linguistic creativity also points in this direction: children need time to experience language in distinct semantic settings in order to abstract over schemas. To the usage-based theorist, children learn abstract syntax from data only; hence it takes time for the more abstract elements of language to emerge. Functional elements bear abstract (and arguably a priori) meanings. Therefore Tomasello and colleagues must predict it will take children longer to exhibit target-like facility with functional elements. UG-based models of acquisition, in contrast, maintain that there are (potentially highly abstract) aspects of grammar that emerge early in development because they are known a priori. The degree of abstraction of a grammatical notion is therefore not itself predictive of the course of acquisition of that notion – in fact, an abstract feature may even emerge faster than some highly idiosyncratic lexical notion, precisely because of its a priori availability, or at least because the search space is highly constrained. For example, in learning the specific differences between prepositions in a language, the learning problem is enormously simplified if there is a finite set of known prepositional features that must be distributed over a finite set of exponents.

The P category, because of its split nature, is useful for testing these prediction. Within the category we have elements sharing many properties – distributional, phonological – yet whose semantics assign them to one or both of the classes (which I sometimes refer to as “orders”) of functional and thematic elements. In other words, the category as a whole provides something like (sets of) minimal pairs. The contrasting hypotheses, expressed in terms of Ps, are:
• **Constructivism:** Children must learn all facets of Ps from data. The increase in children’s productivity with functional Ps is slower than that of lexical Ps, because the former are harder to learn.

• **UG:** Children know the grammatical features available to Ps from the start.

  Productivity of functional Ps increases at the same rate (or faster, given their simpler feature set) as lexical Ps.

How are we to measure the productivity of a grammatical category? This is a difficult conceptual and methodological problem. In the debate on the productivity of nouns with determiners (Pine & Martindale, 1996), (Valian et al., 2009), (Charles Yang, 2011), (Wang, 2012), (Pine et al., 2013) the focus has been on examining the degree to which nouns are fused with specific determiners. The statistical workhorse in the research has been **overlap**, essentially the probability of a given noun’s appearing with more than one determiner. In the present work we need to compare the productivity of (orders of) Ps, so we must proceed in the opposite direction, by investigating the **conditional distribution of adpositional complements given the P**. That is, for each candidate P in turn, for children at some developmental stage, we obtain the distribution over the P's complements (normalized counts). For semantic reasons each P type is likely to privilege certain complements, so the distributions of any two Ps are not directly comparable, but if we compare children's distributions to those of adults, we gain a sense of the similarity (or difference) in P usage patterns obtaining between children (at the relevant stage) and parents. For example, consider these artificial empirical distributions for individuals A, B, C, for the preposition *of*:
<table>
<thead>
<tr>
<th>Complement</th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>the book</td>
<td>0.2</td>
<td>0.05</td>
<td>0.05</td>
</tr>
<tr>
<td>the envelope</td>
<td>0.3</td>
<td>0.05</td>
<td>0.1</td>
</tr>
<tr>
<td>the barrel</td>
<td>0.4</td>
<td>0.2</td>
<td>0.3</td>
</tr>
<tr>
<td>all trades</td>
<td>0.05</td>
<td>0.5</td>
<td>0.3</td>
</tr>
<tr>
<td>the table</td>
<td>0.05</td>
<td>0.2</td>
<td>0.25</td>
</tr>
<tr>
<td>(total probability mass)</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
</tbody>
</table>

By visual inspection (see bar chart), the PPs consisting of of + complement produced by B are more similar to C's than A's are to B or C. This is the basic principle behind Study 2: we construct the distributions over complements for each of a set of Ps for children at various MLU intervals and for all the adults (once). We then employ a measure of the distance between distributions to investigate the relationships between patterns of P usage among children and adults.

It is not immediately obvious how these patterns are expected to unfold over the course of development. We might expect children to gradually become more adult-like in their discourse, hence for the differences in usage to diminish over time. But if that difference is small earlier in acquisition, we might instead conclude that children are then primed by the adults, formulaically parroting adult utterances. In this case differences between the conditional distributions would widen during development. In the present study we allow the data to tell the story: whichever way
the empirical results trend we take as the developmental trajectory. The crucial question is whether
the lexical Ps “get there” before the functional Ps. On constructivist principles lexical Ps ought to
approach the target of the trend faster than functional Ps, while a UG perspective suggests parallel
development or indeed for functional Ps to become productive faster (again, due to their simpler
feature constitution, where the space of features is constrained a priori).

So far we have discussed the analysis of P productivity during development by comparing the
relative frequencies of complements between children and adults, for complements that are shared
by both. A second (and complementary) indication of productive, non-formulaic P usage in
children is the rate at which children construct full PPs that are not found in their input. For
example, if a child utters to the ducky but never hears this expression, we have some evidence that to is
being used productively. (Constructivists might counter that the relevant frame is to the + X, but
then the arguments about the non-productivity of the determiner would be affected.) Of course we
have no complete record of the input to any child, so the question is how we might estimate from
the data we do have whether a given child-produced PP is novel with respect to adult language. In
Study 3 the approach taken consisted in merging all adult PPs in the CHILDES-derived database
into a single “virtual” adult, and taking the resulting sub-corpus as a proxy for the input to each
individual child in turn. We now turn to the details of methods and results for the two studies.

Study 2: Analysis of shared P complements

Method

Dataset

First, the adult PPs were scanned for eight very frequent P types: lexical \{in, on, over\} and functional
\{of, for, with, about, to\}. Since the intention was to compute statistics for each \( P \), only the most common \( P \)s were drawn. For each \( P \) type the distribution over the (pro)nominal *heads* of the object was tallied. For example, the top 10 most frequent complements for *in, on and of* in the adult data were:

**In:**

<table>
<thead>
<tr>
<th>Complement</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>there/EX</td>
<td>0.098</td>
</tr>
<tr>
<td>it/PRP</td>
<td>0.036</td>
</tr>
<tr>
<td>box/NN</td>
<td>0.031</td>
</tr>
<tr>
<td>car/NN</td>
<td>0.023</td>
</tr>
<tr>
<td>room/NN</td>
<td>0.021</td>
</tr>
<tr>
<td>house/N</td>
<td>0.019</td>
</tr>
<tr>
<td>mouth/N</td>
<td>0.019</td>
</tr>
<tr>
<td>bag/NN</td>
<td>0.017</td>
</tr>
<tr>
<td>bed/NN</td>
<td>0.015</td>
</tr>
<tr>
<td>water/NN</td>
<td>0.013</td>
</tr>
</tbody>
</table>

**On:**

<table>
<thead>
<tr>
<th>Complement</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>it/PRP</td>
<td>0.065</td>
</tr>
<tr>
<td>floor/NN</td>
<td>0.045</td>
</tr>
<tr>
<td>there/EX</td>
<td>0.03</td>
</tr>
<tr>
<td>table/NN</td>
<td>0.025</td>
</tr>
<tr>
<td>head/NN</td>
<td>0.022</td>
</tr>
<tr>
<td>top/NN</td>
<td>0.017</td>
</tr>
<tr>
<td>side/NN</td>
<td>0.017</td>
</tr>
<tr>
<td>chair/NN</td>
<td>0.013</td>
</tr>
<tr>
<td>you/PRP</td>
<td>0.013</td>
</tr>
<tr>
<td>back/NN</td>
<td>0.012</td>
</tr>
</tbody>
</table>

**Of:**

<table>
<thead>
<tr>
<th>Complement</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>them/PRP</td>
<td>0.038</td>
</tr>
<tr>
<td>it/PRP</td>
<td>0.036</td>
</tr>
<tr>
<td>you/PRP</td>
<td>0.02</td>
</tr>
<tr>
<td>things/N</td>
<td>0.016</td>
</tr>
<tr>
<td>these/DT</td>
<td>0.013</td>
</tr>
<tr>
<td>those/DT</td>
<td>0.013</td>
</tr>
<tr>
<td>way/NN</td>
<td>0.013</td>
</tr>
<tr>
<td>paper/NN</td>
<td>0.013</td>
</tr>
<tr>
<td>course/N</td>
<td>0.012</td>
</tr>
<tr>
<td>that/DT</td>
<td>0.012</td>
</tr>
</tbody>
</table>

Similar conditional distributions were computed for each child in the database, binned by MLU interval. The distributions over complement heads were computed for each of the seven \( P \)s in question for MLU from 2 to 5 in steps of 0.5 (Spanish) or 2 to 6 in steps of 0.25 (English).

**Comparing Distributions**

In order to compare the adult distributions of prepositional object heads conditioned on \( P \)s to those
of children, the event space (the complements) was limited to cases in which the complements appeared at least once in both distributions. So when comparing the relative frequencies of complement to in, for example, the set of objects was limited to the intersection of all adults’ complements and those of the children in the relevant MLU band. There were thus no zero frequencies. (As a reminder, Study 3 looked at the complementary situation by analyzing child PP objects not in the adult data set.)

The degree of dissimilarity of distributions was computed using relative entropy (or Kullback–Leibler (KL) divergence), a standard information-theoretic measure of the difference between two probability distributions. Relative entropy can be thought of as the degree of surprise at child behavior given our knowledge of adult behavior. For probability distributions $P$ and $Q$, the KL distance is given by:

$$D_{KL} = -\sum_x p(x) \log q(x) + \sum_x p(x) \log p(x)$$

It is the cross-entropy of $P$ and $Q$ less the entropy of $P$. During piloting it quickly became clear that the KL distances between child and adult data are heavily influenced by the gross number of child events considered and by the number of $P$ complement head types – yet more evidence that sample size is a powerful element in computing productivity, and necessarily so given the Zipf-like nature of word frequency in language (Yang, 2013). Intuitively, when comparing distributions for individuals $A$ and $B$, as the number of shared complement types increases, holding the number of events constant, the divergence between the distributions is likely to diminish. On the other hand, if the number of events for $A$ increases holding the number of types constant, then the divergence of $A$ from $B$ is likely to increase, because smaller differences will be amplified. We will see below these intuitions are borne out by the results, but the core point is that because the numbers of
complement head types and PP events vary by MLU, we cannot directly interpret the KL distances between children and adults for functional and lexical P between MLU ranges without taking into account the variation in the counts of PPs and complement types.

One approach to controlling for the variation in counts is to simply fix the numbers of PP and complement types across all MLU groups. This method, in the spirit of Pine et al (2013) entails either subsampling or oversampling the data, depending on the raw counts at each MLU group. Drawbacks of such sampling procedures are the introduction of sampling biases and the fact that when subsampling, not all available information is exploited (Charles Yang, Wadsworth, & Valian, 2014). It is arguably better to use all the data, making the necessary adjustments for dependencies between variables.

The approach taken here consisted in adjusting the empirical KL distances by the effects of covariates (variables not of interest) – in effect carrying out a form of residuals analysis.

The procedure was as follows:

For each 0.25 MLU interval in the range [2, 6] ([2, 5] for Spanish):

1. Select all child PPs whose Ps are in the set \{in, on, into, of, for, with, about\}.

2. For each P, repeat 1000 times:

   1. Draw a replicate of all PPs (P + head only) for the given P by first randomly selecting (with replacement) as many children as there are in the original set, then copying all PPs for each child into the replicate. This 2-step selection process was intended to moderate the longitudinal effects of the nested character of the data (recording sessions within
children).

2. Establish the event space as the intersection of the child and adult complements (size = N1). Compute the resulting pair of probability distributions.

3. Count the number of child PPs = N2.

4. Compute the KL-distance = KL.

3. Return the mean KL-distance from the 1000 replicates.

The entire procedure was then repeated on the Spanish data, for the top five most frequent Ps in the Spanish data set, \{a, de, en, con, para, por, hasta\}.

This process yielded for each language a data set consisting of the following variables: MLU group; FUNC, an indicator variable for whether a P is functional; KL, the empirical KL-distance; N1, the number of P complement types; and N2, the number of child PP events.

A stepwise linear regression of KL on the other variables was then performed, beginning with the base case incorporating the main effects only:

KL ~ MLU + FUNC + log(N1) + log(N2)

All models including two-way interactions were explored using the Akaike information criterion (AIC). Once the best model was found, the empirical KL scores were adjusted using the coefficients of all effects that included the covariates N1 and N2, to give the adjusted empirical KL score, AKL. This final quantity was then interpreted as a measure of the degree to which children’s PP differ from those of adults. The continuous variables, MLU and the log-transformed N1 and N2, were standardized (centered and scaled) prior to the fitting procedure.
Results

It was found that for these frequent Ps, the mean KL-distance was statistically consistent in the number of bootstrap replicates. At 1000 replications the confidence intervals were very small. I therefore will only report the means.

The best model returned by the stepwise regression was:

\[ KL \sim 0 + MLU + FUNC + \log(N1) + \log(N2) + MLU:FUNC + MLU:log(N1) + FUNC:log(N2) \]

There was no intercept, owing to the strong bimodality of the data (differential behavior of functional/lexical Ps). All main effects were retained, and two-way effects of FUNC and MLU with \log(N2) and \log(N1) respectively, and with each other, were added. The coefficients (standardized because the input variables were standardized) were:

|     | Estimate | Std. Error | t value | Pr(>|t|) |   |
|-----|----------|------------|---------|----------|---|
| mlu | 0.35104  | 0.07464    | 4.703   | 7.90e-06 | ***|
| func| 0.44828  | 0.06571    | 6.822   | 6.11e-10 | ***|
| ln1 | 4.24617  | 2.72488    | 1.558   | 0.1222   |   |
| ln2 | -4.72494 | 2.71815    | -1.738  | 0.0851   |   |
| mlu:ln1 | -0.26737 | 0.06086    | -4.393  | 2.70e-05 | ***|
| func:ln2 | -0.38285 | 0.07728    | -4.954  | 2.83e-06 | ***|
| mlu:func | -0.29424 | 0.06341    | -4.641  | 1.02e-05 | ***|

The model achieved a moderate fit (adjusted R-squared = 0.607). The coefficients tell us that KL distance increases with MLU (standardized beta = 0.351), and that the overall mean KL was considerably higher for the functional Ps (beta = 0.448). N1 (non-significant) and N2 affected KL in the expected directions: an increase in the number of child observations, holding the number of complement types constant, increased the divergence, while for constant observations an increase in
the count of complement types lowered the divergence. These effects were attenuated for increasing MLU (see the two-way interactions), but for functional Ps the lowering effect of the count of complement types was amplified. In other words, for functional more than lexical Ps, the greater the variety of complements, the lesser the divergence of child PPs with adult productions. Also, and importantly, the effect of FUNC was attenuated for increasing MLU.

The empirical KL scores were then adjusted by subtracting the effects of log(N2) and the two way interactions involving N1. Figure 15 displays a plot of the resulting adjusted KL scores for each preposition at each MLU interval.

![Figure 15: Adjusted KL distance by MLU group (English)](image)

There is a clear trend in the plot. First, the general tendency is for the adjusted KL divergence to
increase with MLU. That is, once we control for variation in the number of complement types and tokens, children's conditional distributions over complements tend to diverge from the adults, suggesting greater independence in the construction of their PPs. Second, for functional Ps the adjusted KL score is greater than that of the lexical Ps over the whole MLU range. A Kolmogorov-Smirnov test for the differences between the two classes of Ps indicates that they are statistically distinct (D = 0.6609, p < 0.0001; though note that the difference is present only based on the point estimates of the effects of sample size; were we to apply the confidence intervals of the effects, no difference would be found). By the logic of our interpretation, the result is evidence that children's use of functional Ps is more independent right from the start, confirming the UG hypothesis.

For the best model fit to the Spanish data there was again a positive and significant, effect for MLU and a small, almost significant, positive effect for FUNC. There was a significant interaction was that of MLU with FUNC, which was negative, as in the English data, and a positive interaction of FUNC with N1. Thus, as we saw previously, there appear to be few differences between P orders in the Spanish data. Figure 16 has the plot for the Spanish results. Note that because of the database of Spanish PPs was smaller, MLU intervals of 0.5 were employed.

| Estimate   | Std. Error | t value | Pr(|t|)   |
|------------|------------|---------|----------|
| mlu        | 0.20719    | 0.08209 | 2.524    | 0.0148 * |
| func       | 0.15662    | 0.09504 | 1.648    | 0.1057   |
| ln1        | 0.82547    | 0.11213 | 7.362    | 1.63e-09 *** |
| func:ln1   | 0.18217    | 0.07837 | 2.324    | 0.0242 * |
| mlu:func   | -0.20700   | 0.09431 | -2.195   | 0.0328 * |
| mlu:ln1    | 0.13834    | 0.09293 | 1.489    | 0.1429   |
The adjusted KL measure for the Spanish functional Ps is statistically flat over the MLU range (95% CI of Pearson correlation of adjusted KL with MLU = [-0.362, 0.259]), while for lexical P there was visible growth (95% CI of Pearson correlation of adjusted KL with MLU = [0.219, 0.700]). The lines in Figure 16 mark the trajectories for allegedly lexical *en* (solid line) and functional *de* (dotted line). The two lines are fairly close, suggesting that *en* locative may not be lexical at all. The Kolmogorov-Smirnov test for the differences between the two orders of Spanish Ps indicates, nevertheless, that they too are statistically distinct, and the value was similar to what was found for the English data (D = 0.600, p = 0.000249). It thus seems that while the overall pattern in Spanish is similar to English – development tends toward greater adjusted KL, and lexical P lags behind...
functional P – Spanish the children in the CHILDES corpus are fluent users of functional prepositions earlier than the English children.

Relating these findings to the straightforward production data detailed in Study 1 as measured by relative frequency of P types (see Figures 9 and 10 in Chapter 3), we have for English:

- Lexical Ps are produced at near-adult rates (initially greater) from the start, but their productivity with respect to P complements increases over development.

- Children under-produce functional Ps compared to adults, but exhibit greater independence in their construction of PPs with functional Ps compared to lexical Ps, though this too increases with development.

Whereas for Spanish:

- Children produce lexical Ps at less than adult rates initially, and the productivity of lexical PPs also increases over development.

- The rates of production of functional Ps are adult-like from the start, and their productivity is essentially constant over development.

Common to both languages is the observation that the productivity of Ps is asymmetrical between P orders – assuming productivity is proportional to, and thus measured by, adjusted KL – where the more syntactically abstract items exhibit greater independence from the start, while the semantically richer and more specific items remain more firmly bound to adult-type distributions. This result conflicts directly with the frequency-driven notions of constructivism: it seems children do not require a wide range of exemplary uses in order to acquire the syntax of abstract functional
elements.
Study 3: Novel PPs in child speech

Study 2 compared the distributions over PP objects in child speech to that of adults, for complements used both by children and adults. What about the complementary situation, PP objects that were not instantiated by one or the other of the two groups? Utterances in which adult objects do not appear in child PPs are ambiguous between the child not knowing the complement or not knowing that a particular P can select for that particular object; there is little to be said about this condition. (In principle one could search through the complete transcripts for such terms in non-PP child contexts – a sizable effort I did not undertake.) But we can, in theory at least, look at the other possibility – whether children build PPs by pairing Ps with prepositional objects in ways that adults do not – whether children produce PPs that are novel with respect to their input.

The issue is this: according to usage-based theorists the early grammar of children is lexically-conditioned and conservative with respect to the frequency of constructions in the input. On this account we would expect children to produce novel PPs at very low rates, if ever. Their linguistic creativity would lag that of adults. On a UG account, conversely, as soon as PPs emerge in child speech, children’s expression are in principle as creative as those of adults once one controls for their smaller vocabulary (and the associated Zipf effects).

Experimentally we want to ask whether, given a particular PP expressed by a child at time \( t \), the PP does not appear in the set of all adult utterances perceived by the child up to \( t \). Naturally there is no practical way of determining this from CHILDES records. The best we can do is estimate the probability of a given PP appearing in parent language from whatever samples CHILDES make available. One way of doing so, from an idea by Virginia Valian (p.c.), is to merge the records of all adults in the database, across all corpora, into a single file representing a “super-adult” of sorts; we
then test for the novelty of child PPs against the record of this merged corpus. We assume, in effect, that a single probability distribution underlies all adult speech, and that the union of all adult records in CHILDES gives us a good estimate of this distribution. While this procedure is the most conservative possible (given CHILDES), for the purpose of evaluating the creativity of children’s expressions, it is still likely to severely under-represent the input. As I describe below, the evident drawbacks of the method are partly addressed by estimating the creativity of adult PPs by the same procedure, enabling us to benchmark the child estimates with respect to those of the adults.

There are two questions in play. First, is it the case that children produce novel PPs? Second, do we observe developmental differences for lexical versus functional PPs? That is, are there differences in productivity between the two orders of PPs?

**Method**

Every child-produced PP in all corpora was marked for, separately, whether the pairs \([P, \text{Full complement}]\) and \([P, \text{Complement head}]\) appeared at least once anywhere in the set of all adult utterances. In this study it was possible to expand the set of PPs, since there was no need to compute the key statistic individually for each P. The sets of PPs were as follows:

- **Lexical**
  - English: *in, on, into, over, through, before, after, out, until*
  - Spanish: *en, hasta, desde, dentro_de, después_de, encima_de*

- **Functional**
Sampling two instances per corpus from the lowest MLU band [2, 2.5] (Table 13) shows a few interesting patterns:
We observe combinations of Ps with objects belonging to semantic classes that are not typically s-selected for, e.g. *in Mama, in kitty*, though one can certainly imagine contexts where those make sense. Some complements are spoken in “child-ese”, e.g. *doggy, kitty*. And some cases, such as *on*
message and for the microscope, clearly could appear in adult speech but just chanced not to in the
CHILDES corpus. The top ten most frequent novel PPs for children were:

Table 14: Top 10 novel child PPs

<table>
<thead>
<tr>
<th>English</th>
<th>P</th>
<th>Complement (head)</th>
<th>Tokens</th>
<th># Children</th>
<th>P</th>
<th>Complement (head)</th>
<th>Tokens</th>
<th># Children</th>
</tr>
</thead>
<tbody>
<tr>
<td>at doing</td>
<td>at</td>
<td>Texas</td>
<td>9</td>
<td>4</td>
<td>a</td>
<td>pintura</td>
<td>13</td>
<td>1</td>
</tr>
<tr>
<td>at Texas</td>
<td>in</td>
<td>Mom</td>
<td>9</td>
<td>1</td>
<td>a</td>
<td>coto</td>
<td>13</td>
<td>1</td>
</tr>
<tr>
<td>in Mom</td>
<td>of</td>
<td>Super_Guys</td>
<td>9</td>
<td>1</td>
<td>a</td>
<td>cuatro</td>
<td>12</td>
<td>2</td>
</tr>
<tr>
<td>by mosquitoes</td>
<td>of</td>
<td>loft</td>
<td>7</td>
<td>2</td>
<td>a</td>
<td>patatas</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>in Tesco</td>
<td>by</td>
<td>loft</td>
<td>7</td>
<td>1</td>
<td>sin</td>
<td>cabeza</td>
<td>10</td>
<td>3</td>
</tr>
<tr>
<td>through doing</td>
<td>by</td>
<td>Missouri</td>
<td>7</td>
<td>1</td>
<td>de</td>
<td>pitorreo</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>by Missouri</td>
<td>by</td>
<td>my</td>
<td>7</td>
<td>1</td>
<td>de</td>
<td>mayoguet</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>a</td>
<td>babie</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>a</td>
<td>zapatilla</td>
<td>9</td>
<td>2</td>
</tr>
</tbody>
</table>

A similar procedure was applied to the adult PPs, with one important difference: for each adult, all
utterances by that adult were first deleted from the “super-adult” comparison set. Also, to control
for sampling effects with respect to the underlying frequencies of words in the language, the size of
the bootstrap sample at each replication was fixed to equal the total number of child PPs across all
MLU bands. Finally, an additional adjustment was made to remove possible priming effects by the
child on the adult. As stated the computations must bias the adult level upward compared to the
child. Since all utterances of each adult in a child corpus were removed prior to determining that
adult's score, any complement that was unique to the dialogue between that child and adult would have counted towards the adult score, while by definition it would not for the child. So if the child and adult both utter *to Mickey*, but *Mickey* is never used by any other adult, then the adult's score would be incremented, but the child's would not. To account for this all PPs of the children in the corpus corresponding to each adult were added to the comparison set for that adult. Thus,

- a child PP was marked novel IFF it never appeared in a parent utterance anywhere in the corpora;

- an adult PP was marked novel IFF
  - it never appeared in any utterance by another adult anywhere in the corpora, and
  - it never appeared in any child utterance in the specific adult's corpus.

Two descriptive analyses were carried out. Analysis 1 simply compared the overall rate of novel PP production for all children at fixed MLU intervals to the adult rate. Child PPs were once again binned by MLU intervals of 0.25 (0.5 for Spanish) over the range [2, 6]. Then, for each MLU interval, an estimate of the proportion of novel PPs was computed as the mean of 1000 bootstrap replicates of the PPs in that MLU band. The bootstrap replicates for children used the same two-step procedure as in Study 2: first children were selected with replacement, then all records for each selected child were added to the sample. Similarly, the adult proportion of novel PPs was computed as the mean of 1000 bootstrap replicates over the whole adult dataset. As discussed in the results section below, adjustments were subsequently made to account for the effects of sample sizes.

Analysis 2 addressed the longitudinal nature of much of the child data. A mixed effects logistic
regression was fit to the child data (only), where the outcome variable was a binary indicator specifying whether a PP (P + head only) was novel, an ID for the child functioned as the grouping variable, and the within-subject random effects were MLU and MLU-squared. FUNC, an indicator for the class of preposition, and its interaction with MLU were entered as a fixed effects. The slopes and intercepts of the regression on MLU and MLU-squared were allowed to vary randomly; these were regressed onto the total count of PPs per child, since the likelihood of a novel PP occurring in a child's corpus is clearly related to the number of events in the corpus, and exploratory analyses revealed that this level two effect was significant. Motivating this particular model was the interest in isolating the effect (if any) of a P being functional from other drivers of novel-PP construction, such as developmental factors represented by MLU, and sampling effects. The path diagram for the model (Figure 17) breaks down the two levels of the analysis. The filled black circles in the “within” part denote latent random variables (slopes and intercept), which in the “between” part are regressed onto token counts.
The model was fit in MPlus by the following specification, where `o_head` is the binary outcome, `mlufunc` is the interaction of `FUNC` with `MLU`, `mlu` and `mlu2` are the developmental terms, `s` is a random variable that captures the per-child slope, and is `c` the log of the per-child count of PPs (sample sizes):

MODEL:

%WITHIN%
  o_head on func mlufunc; ! fixed effects, random intercepts
  s  | o_head on mlu;       ! random effects, random intercepts

%BETWEEN%
  s on c;          ! slope regressed onto count
  o_head on c;     ! intercept regressed onto count

Figure 17: Path diagram of mixed-effects logistic regression
Again, in the latter analysis only novel P + complement heads were explored. This outcome is more conservative than one using the full complement, i.e. in which any element of the complement NP may result in the uniqueness of the construct.

**Results**

Around 20% of all child PPs and 22% of all adult PPs were identified as novel according to the heuristics detailed above. Aggregate counts for PP tokens (P + head only) over the whole MLU child MLU range were:

*Table 15: Aggregate counts of novel PP tokens*

<table>
<thead>
<tr>
<th>MLU band</th>
<th>Novel P + head (%)</th>
<th>Not Novel P + head (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>Children: 15,836 (19.6)</td>
<td>Adults: 47,066 (22)</td>
</tr>
<tr>
<td></td>
<td>Children: 10,340 (27.0)</td>
<td>Adults: 9,297 (18.1)</td>
</tr>
<tr>
<td>Spanish</td>
<td>Children: 64,942 (80.4)</td>
<td>Adults: 166,837 (78)</td>
</tr>
<tr>
<td></td>
<td>Children: 27,959 (73.0)</td>
<td>Adults: 42,095 (81.9)</td>
</tr>
</tbody>
</table>

The higher rate for the Spanish children (27% vs 19.6%) is likely a consequence of the smaller adult corpus in Spanish – the ratio of child tokens to adult tokens overall was about 3:4 for the Spanish data, versus 1:2 for the English corpus (see tables 4 and 5 in Chapter 3). Tables 16 and 17 display the counts of original PPs by MLU band. The values in parentheses are the number of P + head combinations for which the head never appeared anywhere in the adult corpus, as determined by a search through the complete set of all tokens emitted by the adults.

*Table 16: Novel P+head tokens by MLU band (English)*

<table>
<thead>
<tr>
<th>MLU band</th>
<th>Novel P + head</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>92 (43)</td>
</tr>
<tr>
<td>2</td>
<td>861 (379)</td>
</tr>
<tr>
<td>3</td>
<td>1699 (695)</td>
</tr>
<tr>
<td>4</td>
<td>1193 (480)</td>
</tr>
<tr>
<td>5</td>
<td>566 (246)</td>
</tr>
<tr>
<td>6</td>
<td>288 (121)</td>
</tr>
<tr>
<td>7</td>
<td>188 (90)</td>
</tr>
<tr>
<td>8</td>
<td>155 (66)</td>
</tr>
</tbody>
</table>
in both languages the proportion of novel PPs increases with MLU, almost monotonically in English. The proportions of complements in novel PPs that never appear in adult utterances were higher in Spanish (from 59.7% to 77%) than in English (40.2% to 47.9%), again likely due to the relatively smaller Spanish adult data set.

**Analysis 1**

The results are summarized below by means of three panels of four graphs: one set each for English and Spanish, consisting of the proportion of novel PPs for complement heads only and for complete complement phrases, first by PP token, then by PP type. That is, in the first case, if a child repeatedly expresses a PP unseen in the adult data, each such instance will add to the overall proportion. In the types-only condition that particular pairing of P and its complement will count
once. A third panel of graphs gives the four proportions for the Manchester sub-corpus from CHILDES, consisting of longitudinal transcripts of equal length for thirteen British children all in the MLU range [1.01, 3.993]. This sub-corpus is of particular interest because its density (given the limited MLU range) and its balanced nature: we have approximately the same number of utterances from each child over the full range.

First a note on the confidence intervals of the bootstrapped estimates. Since the proportions computed were frequently small it was important to establish whether they were statistically different than zero. The smallest estimated adult proportion of novel PPs was for lexical PPs. Even at only 100 bootstrap replications the mean was 0.067 (i.e. 6.7%), and the 95% confidence interval was (0.064, 0.068). Fairly small, but clearly larger than zero. Similarly for the children: the lower bounds of all CIs were above zero.

Note that, per the legends, in the following graphs the dotted lines around the solid bars mark the 95% CIs for 1000 replicates. The vertical scales vary between graphs for better clarity.
Table 18: Proportions of Novel PPs (English)

Parental and child levels were everywhere larger than zero. Adults mostly produced higher proportions of novel PPs, though at higher MLUs (above 5) proportions by tokens for the children were essentially equal to those of the adults for both classes of Ps (top row).

The child proportions grew gradually (bottom row), but increased more quickly for the token
counts. In terms of absolute magnitudes, the type statistics were roughly twice as large as the proportions by token, though by the end of the MLU range the latter approached the former. For the token counts the functional units showed greater rates of novel PPs than lexical Ps, and at lower MLUs (up to 4.0) the difference was significant per the 95% CIs. For proportions based on full PP types there were no substantial differences between the two orders of prepositions.

There are two perplexing contrasts between the graphs for the tokens and for the types. One is that the children's type rates never intersect those of the adults, at least not in this MLU range, while the token counts do meet. Second, in the types/head-only condition (bottom left), the proportion of novel PPs for lexical Ps is significantly greater than that for functional Ps, while in the three other conditions the reverse is true. Both contrasts are explained by the effect of sample size on the proportion statistic for type (only). Figure 18 plots the values of this statistic, computed over the adult data, for random samples (with replacement) whose sizes range from 0 to 1,000,000. The proportion statistic for tokens is constant over the range, indicating that even small samples are very good estimates of the “population” value (population understood as the complete CHILDES database, which of course is itself a sample). For types, however, we see the law of large numbers at work: the value of the statistic increases as the probability of every novel PP type being included in the sample approaches its expected value of 1 once the size of the sample equals that of the complete dataset (and the probability of each type is approximately the reciprocal of the number of types, by Zipf's law).
Figure 18: Proportion of novel PPs for varying sample sizes
Since there are many more adult PPs than child PPs, it follows that we must make an adjustment for sample size in order to compare rates of novel PP types. The solution adopted here was to limit the size of the bootstrap replicate for adults to the number of PP types in the child data for the relevant MLU band, such that the adult proportion varied over the MLU range. Since the full adult data set was sampled at each replication, all the data still contributed in the overall analysis, but not for the statistics generated for each bootstrapped set. An analogous adjustment was made to the child data by updating the computed proportions by the effect of sample size on the proportion, as estimated by a linear model run on the child data. The revised plots appear in the bottom row of Table 19 (less the associated confidence intervals, for improved legibility).

*Table 19: Proportions of Novel PPs (English; Revised)*
After the adjustments the differences between P classes is greatly reduced for the parents in both types conditions, and the children proportions intersect those of the parents around MLU=5 (as per the tokens). More dramatically, for the types/full-PPs (bottom right) condition there is no visible difference at all between adults and children.

The growth pattern exhibited by the Spanish children with respect to the type conditions is broadly parallel to the English performance (again the confidence intervals are left out of the lower row). P orders are not differentiated in the adults (in any condition). Child proportions in the types/P+head condition (bottom left) lag those of the adults at first, reaching parity somewhat earlier than the English children, at around MLU= 4. For P+full complements (bottom right) children and adults overlap from the start – indeed child rates are higher than the adults later in development. As for the tokens, children’s rates for functional Ps overlap those of adults from the start; for lexical Ps they intersect the adults at around MLU=3.5 (top left).
These patterns are confirmed when the English data set is narrowed to the Manchester corpus only (Table 21). Note that the horizontal scale ends at MLU=4, the upper limit of the Manchester transcripts; we cannot compare the longer-term trends to those of the full corpus.
Table 21: Proportions of Novel PPs (English, Manchester corpus)

A few observations:

- Tokens/P+head (top left): functional Ps generate significantly more novel PPs at MLU < 3. Both rates are comparable to that of adult lexical Ps, but less than adult functional Ps. The behavior in this condition follows that of the full English dataset (see Table 19).
- 191 -

- Tokens/P+full complement (top right): no significant differences between lexical, functional Ps and adult lexical Ps. The contrast between the first two found in the full data set is lost.

- Types/P+head (bottom left): no significant differences between lexical, functional Ps. As we saw also with the full data, both sets of proportions are below adult levels, though the gap is almost closed by the top of the MLU range.

- Types/P+full complement (bottom right): no differences between adults, children, and P classes, again repeating the earlier findings.

In all plots the rate of novel child PPs for functional Ps trends upwards over the course of development. This is also the case for lexical Ps in all cases except for the types conditions in the Manchester plots, though one suspects there too the rates would increase for data at higher MLUs. The results must condition our understanding of the undeniable fact of novel PPs' appearing in child speech. While the observed rate for child PPs early in acquisition might be explained by patterns of infelicitous usage or indeed grammatical errors, this explanation will not convince for the rates we observe at higher MLUs. If it is granted likely that children put together PPs that are at least improbable in – and original with respect to – adult speech, and if there is evidence of growth from non-zero initial levels, we can reasonably assert that these English and Spanish children exhibit creativity in P usage beyond their default conservatism right from the start of acquisition, and throughout their language development. This is especially so for functional Ps, as the presentation of the second analysis will show.
Analysis 2

The two-level logistic model of Analysis 2 attempted to predict whether a given child PP (P + head only) appeared in the full adult corpus. MLU, which was mean-centered, was a random effect for clusters defined as the PPs for each child. FUNC, an indicator variable describing the categorial status of the P (functional or lexical) and its interaction with MLU were fixed effects. Slopes with MLU and intercepts per child were also regressed on the number of PPs available for each child. The 99% and 95% confidence intervals for the fixed within-level effects for the English data, as estimated by Mplus, were as follows, where the point estimates appear in bold type. Note that random and between-level effects are left out since these served merely as controls for sample sizes and individual variation.

<table>
<thead>
<tr>
<th>CONFFIDENCE INTERVALS OF MODEL RESULTS</th>
<th>Lower .5%</th>
<th>Lower 2.5%</th>
<th>Lower 5%</th>
<th>Estimate</th>
<th>Upper 5%</th>
<th>Upper 2.5%</th>
<th>Upper .5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Within Level</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>O_HEAD ON</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FUNC</td>
<td>0.189</td>
<td>0.220</td>
<td>0.236</td>
<td><strong>0.318</strong></td>
<td>0.399</td>
<td>0.415</td>
<td>0.446</td>
</tr>
<tr>
<td>MBYFUNC</td>
<td>-0.233</td>
<td>-0.209</td>
<td>-0.197</td>
<td><strong>-0.134</strong></td>
<td>-0.070</td>
<td>-0.058</td>
<td>-0.034</td>
</tr>
<tr>
<td>Thresholds</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>O_HEAD$1</td>
<td>1.913</td>
<td>1.943</td>
<td>1.959</td>
<td><strong>2.038</strong></td>
<td>2.118</td>
<td>2.133</td>
<td>2.163</td>
</tr>
</tbody>
</table>

All parameters were significant (the confidence intervals never cross zero). Inferring from the model estimates, for a child at mean MLU = 3.435, the probability of a novel PP for a non-functional P is 1 less the inverse logit of the threshold estimate, around 13%. For a functional P the probability increases to 17%. But as MLU rises the difference in probabilities drops, becoming zero at around MLU = 6. For the Spanish data FUNC was not a significant predictor. Keeping in mind that MLU was centered, this means that by the time the child is at mean MLU, FUNC is no longer significant. The significant slope of the MLU X FUNC interaction, twice as large as in the English model, shows that the effect of FUNC drops faster with MLU such that by MLU=4 (near the mean MLU) no difference is predicted.
This is shown by fitting to the Spanish data set a model in which MLU is not centered:

<table>
<thead>
<tr>
<th>Within Level</th>
<th>Lower .5%</th>
<th>Lower 2.5%</th>
<th>Lower 5%</th>
<th>Estimate</th>
<th>Upper 5%</th>
<th>Upper 2.5%</th>
<th>Upper .5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>O_HEAD</td>
<td>-0.236</td>
<td>-0.190</td>
<td>-0.167</td>
<td>-0.043</td>
<td>0.080</td>
<td>0.104</td>
<td>0.150</td>
</tr>
<tr>
<td>MBYFUNC</td>
<td>-0.318</td>
<td>-0.304</td>
<td>-0.297</td>
<td>-0.259</td>
<td>-0.221</td>
<td>-0.214</td>
<td>-0.199</td>
</tr>
<tr>
<td>Thresholds</td>
<td>O_HEAD$1</td>
<td>1.551</td>
<td>1.594</td>
<td>1.616</td>
<td>1.732</td>
<td>1.848</td>
<td>1.870</td>
</tr>
</tbody>
</table>

For a virtual child of MLU = 0, FUNC has a strong, positive effect on the likelihood of a novel PP: the probability of a novel PP with a functional P is 38% versus 14% for a lexical P. Thus in both languages functional prepositions were associated with higher probabilities of child-produced novel PPs at low MLU, after accounting for the effect of sample size.

Discussion

The empirical work presented in Chapter 3 agreed with Littlefield (2006) that English children – but not Spanish children – exhibit some delay in their production of functional Ps. This observation alone tells us little about how children represent P in their early grammars. We must establish whether a delay occurs because younger children have yet to fully acquire these grammatical elements, as Littlefield would have it, or if the delay is merely apparent, falling out of the interaction
of constraints on production and language-specific pragmatics. The question then becomes: given
the evidence we do have of early usage of functional and lexical Ps (as opposed to non-
occurring), are there comparative differences in the syntactic productivity for the two orders of P?

The two studies presented in this chapter attack the problem by comparing children's choice of
complements for the most frequent Ps to the corresponding patterns of co-occurrence found in
adult speech. The analyses uniformly indicated that the productivity of functional Ps in the speech
of English learners is not below that of lexical Ps, and is typically higher. These results held for
Spanish as well. Hence there is evidence that children are able to acquire the syntax of functional
prepositions early and fast.

The investigations described in this chapter are nevertheless vulnerable to serious criticisms. Study 3,
the novel PPs study, suffers from the obvious weakness that parental speech was assumed drawn
from a single probability distribution. For the English data this assumption is not as far-fetched as
one might think. Consider the token count for adults in the corpus, 7 million. If at three a child
will have perceived somewhere between 20 to 30 million words (Bart and Risley, 2003) then the
CHILDES-derived corpus is only three or four times smaller than the estimate of children's
experience. Since the token statistics in Study 3 stabilize quickly as sample size increases, for the
longer longitudinal sub-corpora at least the sample of adult speech we have may constitute a reliable
sample of the language of that adult. It thus seems plausible that for the token-based comparisons,
for many of the children we have a useful comparison set of adult PPs.

A more fundamental problem for Study 3, with respect to its relevance to the generativist-
constructivist debate, manifests when we recall that Tomasello and especially Goldberg have argued
linguistic creativity must be measured over schemata, not individual words. A child who utters “with
kitty” and “with ducky”, where the latter happens to not show up in adult speech, might be said to exhibit less creativity than the child who also says “with glue”. Computing an index of productivity over schemata would require a mechanism for semantically classifying complements – a task made additionally complex by the high rate of pronominal complements, as these would need to be resolved with respect to the discursive context.

Similarly, in Study 2 the method was again based on counts of word forms rather than (something approximating) schemata. Here the crucial issue of developmental velocity also comes into play. In motivating the study I indicated that while we lack an obvious measure of “gradualness” in linguistic development, we can test the constructivist claim of differential growth for functional vs lexical categories. But what if the constructivist side (Tomasello, in this instance) were to withdraw that prediction? After all, the very high frequency of essential closed-class words in child-directed language suggests that children should have plenty of data to work with. Tomasello almost hints at this possibility:

As soon as just a small amount of grammar begins, young children’s utterances are peppered with a relatively small number of high-frequency lexical items such as certain pronouns and function words with highly recurrent discourse functions—with the more well-known nouns and verbs, which are typically thought of as the prototypical items in young children’s vocabularies, used relatively infrequently as their specific referents occur in the child’s experience at only irregular intervals (Tomasello, 2003, p. 82).

Of course the appearance of such function words must not trick us into imputing adult-like knowledge to the child:

What if instead of leaping at every opportunity to attribute to children abstract linguistic categories and constructions, researchers simply acknowledged the empirical
fact that linguistic abstractions build up during ontogeny in a much slower and more piecemeal fashion than previously believed – and then adjusted their models accordingly (Tomasello, 2003, p. 97)?

“Much slower” is the key modifier here. Researchers working within the usage-based paradigm might concede the point: yes, children seem to build fluency with functional elements as quickly or faster than with thematic elements. But the graphs in Study 2 suggest that growth is nevertheless extended in time; the notion of across-the-board “gradualness” of learning for all categories is not necessarily disconfirmed if the order of acquisition is altered. Why posit more than one learning mechanism? If acquisition is primarily parametrized by the frequency of schemata, then perhaps it is simply the case that f-morphs are learned very quickly, too quickly to show up in Study 2, e.g. because MLU >= 2 is already too late. “Slower” would perhaps only apply to lexical Ps, and in so doing underline the experiential (versus innate) essence of learning, which can then be assumed to occur with functional categories as well (just faster).

I cannot directly falsify this argument. I will suggest, however, that the underlying factor dampening the rate of change of the measure of syntactic fluency for lexical Ps developed in Study 2 is the size of the children’s vocabulary, not the slow “build up” of syntactic knowledge. Lexical Ps are to a greater degree than the functional selected by their objects – in the table is conceptually unusual, on the hole too. Their distribution is therefore more tightly bound to the lexical and experiential knowledge of the speaker. If we posit that functional Ps express a limited, rarefied set of morphosyntactic features in UG, then it stands to reason that children acquire target-like facility with functional Ps earlier than the lexical.

Whether or not one accepts this argument, there appears to be converging evidence that, far from
being “much slower”, the acquisition of f-morphs occurs early and fast. This does not entail UG per se, but it means the focus of research must turn to the earliest stages of acquisition, before the two-word stage, when of course children produce very little language. We are thus faced with sharp limitations on the usefulness of corpus analytics. I suggest that for progress to occur it will be necessary to deploy experimental protocols focusing on the grammatical comprehension of infants and young toddlers. Until then, the results presented in this chapter point to nativist UG-oriented theories as the better explanation – precisely because such approaches offer concrete and specific hypotheses that explain the data, rather than mere research programs. The remainder of this thesis is dedicated to the exposition and implementation of one such proposal.
Part III. Models of Acquisition of Function Words

Chapter 5. Learning Subject Pronouns

Introduction

This chapter is concerned with developing an abstract model for the acquisition of \( f \)-morphemes that integrates various learning mechanisms to precisely assign feature content to closed-class forms. I say “abstract model” because, as is typical in the modeling literature, I make no direct claims about the actual psychological processes active in language learning. The focus is rather on developing and testing hypotheses about how data and knowledge that are plausibly available to children might be exploited for learning the feature specifications of \( f \)-morphemes. More specifically, I will outline a method for integrating distributional information, a priori knowledge of feature geometries (introduced in Chapter 1), and semantic-pragmatic understanding to learn the assignment of morphosyntactic features to \( f \)-morphs.

In this chapter we take a break from prepositions. As discussed in Chapter 1, the field’s broad effort to identify the set of universal morphosyntactic features of \( P \) is young yet, hence an established consensus on the morphosyntactic feature assignments to individual \( P \)s has yet to emerge. At present it is impossible to carry out linguistically-informed evaluations of learning models applied to \( P \)s. Current research programs investigating the fine structure of functional projections (cartography, nanosyntax) give us hope for significant advances in our understanding of \( P \), and in the Chapter 6 we will investigate competing proposals using the tools developed in the present chapter.

The general strategy in this thesis consists in first developing a hypothesis for the acquisition set of
function words whose features are better established, thanks especially to the work of Harley and Ritter on the morphosyntactic feature geometry of phi-features (Harley & Ritter, 2002): subject pronouns and their phi-features (Person, Number, Gender). That is the subject of this chapter. If the model is able to assign the correct bundles of phi-features to pronouns, then – on the rather strong assumption that it can be successfully extended to learning functional Ps – the output of the model when applied to P data ought to yield useful information about the features of Ps. In other words, if in this first act linguistic theory provides criteria to guide the development of the model, in the second the favor is returned, so to speak, to the extent that the model is able to empirically inform current theoretical debates on the nature of P.

**The evaluation problem**

A useful starting point is to consider the thorny methodological question of how a model of acquisition might be evaluated, whether qualitatively or quantitatively. The traditional approach in the literature on distributional learning falls out from the manner in which the problem itself has been posed – namely, as a categorization task. Cartwright and Brent (1997) introduced two evaluation measures that have become standard. Both are based on knowing a priori whether or not each pair of morphs in the dataset belong to the same category. When the model correctly places a pair in some category, it has generated a hit. If it failed to assign two elements that belong together to the same category we have a miss. And if the system mistakenly places two items that do not belong together into the same category, we have a false alarm. The evaluation criteria are then defined as:

\[
\text{Accuracy} \% = \frac{\text{Number of Hits}}{\text{Number of Hits} + \text{Number of False Alarms}} \times 100
\]

\[
\text{Completeness} \% = \frac{\text{Number of Hits}}{\text{Number of Hits} + \text{Number of Misses}} \times 100
\]
These statistics are roughly similar to measures of precision and recall in computational linguistics. Intuitively, accuracy measures the degree of homogeneity of the categories created by the system, while completeness measures the degree to which like items are grouped into fewer, larger classes. Ideally the learning model would generate one perfectly homogeneous class per gold-standard category, thus scoring 100% accuracy and completeness. In practice there is often a tradeoff: one can induce many high quality groupings (high accuracy, low completeness), or fewer, more heterogeneous groups (low accuracy, high completeness). The accuracy of Cartwright and Brent’s system, for example, came in at or above 70% but completeness never rose beyond 20%. Redington, Chater and Finch improved overall completeness to 47% (accuracy was 72%), but the accuracies for functional categories were extremely low: using their labels, 9% for the numerals, 10% for articles, 25% for pronouns, 6% for conjunctions, 33% for prepositions (Redington et al., 1998, p. 455). Interestingly, some functional categories broke out of the accuracy/completeness tradeoff downwards: completeness for pronouns was only 24%, 33% for conjunctions, 53% for prepositions – but articles were grouped at 100% completeness, numerals at 82%. The very low accuracy and high completeness for the article group is explained by the fact that in English it consists of exactly two morphemes and an allomorph, \{the, a, an\}. Unless the model is allowed to generate a great number of very small groups (dooming completeness), these three will inevitably be grouped together into a larger set with units of other categories (number, perhaps), leading to low accuracy.

Wang (2012) grouped f-morphemes into four categories: determiner, pronoun, preposition, and auxiliary. His scores varied by the number of clusters induced. As one would expect, accuracy rose and completeness dropped with cluster count. Mean accuracy across CHILDES corpora varied between 59% and 76%, while completeness went from 29% to 17%. Unfortunately Wang did not report scores for individual categories, but in any case we cannot compare his results to those of
Redington et al because the two studies did not evaluate against the same categorization scheme.

For example, Wang’s determiner category is split across at least articles and numerals in Redington et al. We might say each hypothesis is tied to its own evaluation method. So at the very least what is needed is an agreed-upon categorization scheme – one not necessarily based upon the traditional categories of grammar. In short: if the measure of success is based on how units are grouped, then naturally the field must agree on a gold standard for categorization.

I want to argue, however, that this methodological issue is itself a symptom of a more basic conceptual problem. When it comes to f-morphemes, the whole notion of grouping may be of limited value. Even assuming perfect accuracy and completeness, what is the value of the classification? Take the category of pronouns, about whose constitution presumably Wang and Redington et al would agree. Grouping pronouns perfectly would yield a category that current mainstream theory does not quite recognize (see for example Déchaine & Wiltshko, 2002, for whom English 1st and 2nd person pronouns are of category D along with determiners), and would give the learner little useful information with respect to the precise syntax of pronouns. Category is at best one morphosyntactic feature among many (see Chapter 1); PERSON, NUMBER and GENDER (phi-features) are also relevant, as is CASE. Indeed, Dechaine and Wiltshko propose that phi-features project their own pronominal category. The better test of distributional methods (or other learning procedures) is therefore whether they can yield clues about all relevant features.

In general, models of lexical acquisition of function words must be evaluated with respect to the feature bundles they assign to each morph, by comparing the computed assignments to our theory-based knowledge of the feature content of functional elements. On this view the learning model is a function from the input to feature bundles:
The grouping algorithms reviewed above always involve quantitative comparisons between the distributional behaviors of individual terms, typically by means of a metric on a vector space (i.e. a function from a pair of vectors to a scalar). Methods based purely on distributional analyses are strictly non-semantic; they cannot output features directly. Their direct evaluation against theoretical feature matrices is impossible. Instead we have a function:

Language input (e.g. raw text) $\rightarrow$ Model $\rightarrow$ relations of similarity and difference between words

Now, even though the output of the distribution-based learner cannot be evaluated directly against a feature-based gold standard, we would expect the patterns of similarities and differences between the theoretical feature bundles to be reproduced by those inherent in a correct clustering of distributional vectors. That is, we can evaluate the model by comparing the correlation matrix of the distributional vectors to the correlation matrix constructed from the theoretical feature assignments of the functional elements.

Consider this simplified example: there are six morphosyntactic features \{F1 \ldots F6\} to be assigned in sets of three to tree morphs \{A, B, C\}. Linguists have discovered the following mappings:

\[
\begin{align*}
A & \rightarrow \{F1, F2, F3\} \\
B & \rightarrow \{F1, F2, F4\} \\
C & \rightarrow \{F1, F5, F6\}
\end{align*}
\]

Clearly A and B are more similar than either is to C. These similarity relationships can be visualized by means of a tree diagram of clusters, a dendogram (Figure 19).
Imagine now that we have two methods, M1 and M2, for generating and comparing distribution vectors from speech data. The corresponding vector spaces can also be thought of as feature spaces, though they typically span hundreds of dimensions and of course have no direct relationship to morphosyntactic features. By comparing the distribution vectors of morphs \{A, B, C\} a dendogram is generated for each model (Figure 20). The dendogram for M1 is identical to the linguistically-informed clustering, whereas M2 determines that C and A are more alike than either is to B, contradicting the ground truth established by linguistic research. Model M2 is therefore judged unsatisfactory, while M1 is retained.

*Figure 19: Dendogram for theory-based assignments*
A model of acquisition

Moving from evaluation to the acquisition problem itself, assume that the learner is able to carry out an accurate distributional analysis, i.e. an analysis of input language that, for function words, yields relations of similarity and difference that are consistent with the true (but unknown at first) feature assignments of the units. How might the learner exploit this information?

The key point is that the correlations of distributional vectors yield a set of constraints on the true feature assignments, enabling the learner to discard all assignments that fail to match the pattern of observed correlations.

Returning to the example: having established that M1 (but not M2) yields linguistically correct correlations, we hypothesize that a learner might use a method comparable to M1 during the acquisition process. The goal of acquisition is to infer the correct mapping from exponents \{A, B,
C\} to their corresponding feature sets (i.e. entries in the Vocabulary in Distributed Morphology). If the learner happens to know that in the language features F1 through F6 are assigned to the three sets given above (DM: the Lexicon) – that is, the learner knows beforehand that there is a matrix \{F1, F2, F3\} but not that A maps to it – the distributional analysis immediately yields the correct mapping simply by associating corresponding nodes in the empirical (distributional) dendogram to the theoretical one. We assume, however, the Lexicon to be language-specific; it too must be acquired. Otherwise we have simply shifted the learning problem from the Vocabulary to the Lexicon.

Let us therefore not assume prior knowledge of the sets of abstract features that characterize the functional lexicon of the language being acquired. If the learner only knows from UG that features F1 through F6 exist and must be somehow assigned to the three forms \{A, B, C\}, the empirical dendogram can assist the process by excluding all non-conforming candidate assignments. Of the mappings below, those in the first column are conforming, those in the second are not and are therefore discarded as possibilities.

\textit{Table 22: Classification of candidate feature matrices}

<table>
<thead>
<tr>
<th>Conforming to distribution-based similarities</th>
<th>Not conforming to distribution-based similarities</th>
</tr>
</thead>
<tbody>
<tr>
<td>A → {F1}</td>
<td>A → {F1, F2, F3}</td>
</tr>
<tr>
<td>B → {F1, F2}</td>
<td>B → {F1, F4, F5}</td>
</tr>
<tr>
<td>C → {F3, F4, F5, F6}</td>
<td>C → {F1, F5, F6}</td>
</tr>
<tr>
<td></td>
<td>(B and C closer)</td>
</tr>
<tr>
<td>A → {F1, F2, F3}</td>
<td>A → {F1, F2}</td>
</tr>
<tr>
<td>B → {F1, F2, F4}</td>
<td>B → {F3, F4}</td>
</tr>
<tr>
<td>C → {F1, F5, F6}</td>
<td>C → {F5, F6}</td>
</tr>
<tr>
<td></td>
<td>(no overlap; all equally distant)</td>
</tr>
<tr>
<td>A → {F1, F2, F3}</td>
<td>A → {F1, F2, F3, F4}</td>
</tr>
<tr>
<td>B → {F3, F4, F5}</td>
<td>B → {F2, F3, F4, F5}</td>
</tr>
<tr>
<td>C → {F6}</td>
<td>C → {F3, F4, F5, F6}</td>
</tr>
<tr>
<td></td>
<td>(C closer to B than to A, but distance between A and B and B and</td>
</tr>
</tbody>
</table>
Clearly there are many possible solutions matching the constraints imposed by the distributional information, and exponentially more as the number of features to be assigned increases. It is therefore necessary for the learner to draw upon additional sources of information in order to reduce the ambiguity to the single correct solution. Feature geometries offer one such mechanism. We saw in Chapter 1 that a feature geometry specifies entailments between features, i.e. that the presence of some feature determines that of another, e.g. a NEUTER node implies a CLASS node in Harley and Ritter's feature geometry of phi-features. In our example, if F5 must always co-occur with F6, then the third and fourth entries in Table 22 are no longer viable.

Now, even though the distributional filter plus feature-geometric constraints may reduce the solution space, there are still several possible solutions left. We remain confronted by the “linking problem” that, according to Tomasello, bedevils UG-based theories of language acquisition (Tomasello, 2003, p. 7), namely that the terms of the concrete language must be linked to the abstract notions of UG. As we saw in Chapter 2, usage-based theorists root children's understanding of language in their “semantic-pragmatic” ability to “read” attention and intention in intersubjective situations. The existence of such learning processes does not by itself invalidate the UG hypothesis; on any theory the process of learning the meanings of words must call upon the concrete experience of the individual. Such experiential learning may well supply the additional constraints the learner needs to converge onto the correct Vocabulary. For example, if the child determines, through its interactions...
with adults, that C has the feature F1, the first and third entries in the first column of Table 22 must be wrong. Of the four initial candidates only the second remains.

So in this contrived example, by combining distribution-based queues, the a priori knowledge of the feature space and its geometry, and the human ability to extract form-meaning relations from experience – perhaps by relating intersubjective dynamics to linguistic expression -- it was possible to converge onto the linguistically correct solution. It is true that in this composite learning strategy, the third component, the experiential, functions as the non-explanatory *deus ex machina* that enables the model to succeed. One might indeed object that if the third component is capable of determining *any* morph-to-feature relations, then perhaps it alone can learn them *all*, making distributional analysis redundant, thus simplifying the learning hypothesis. How exactly the child might go about it becomes a question for psychological research that cannot be decided analytically. The proposal explored by this thesis is that, from the point of view of theory development, we want to minimize the range of linguistic facts whose acquisition depends on poorly understood mechanisms. For example, I have not found systematic studies of how children might learn functional prepositions by intention reading.

We can, on the other hand, develop concrete representations of the informativeness of distributional data; and for at least some domains linguists have constructed detailed feature geometries. These enable us to determine with some precision the ambiguity in feature assignments that remains for experiential learning to reduce. In the rest of this chapter we will investigate these ideas by carrying out a distributional analysis of English and Spanish subject pronouns, in order to determine whether the patterns of clustering predicted by their features, as determined by Harley and Ritter's feature geometry (Harley & Ritter, 2002) (hereafter H&R), are reproduced by the
distributional analysis (Study 4). An algorithm for learning the mappings from morphs to feature sets is then developed in some detail.

**Study 4: The distributional analysis of subject pronouns**

*Method*

At a high level, the models developed in the literature on distributional learning call for the same general sequence of processing steps. Again, the locus classicus is Redington et al (1998); other relevant works are Schütze (1995), Cartwright and Brent (1997), Mintz et al (2002), and Wang (2012).

1. Raw language data are collected into n-grams, where n is an odd number – typically 3 or 5 – though other windowing criteria, e.g. complete sentences, have been investigated. Smaller windows, such as 3-grams, imply fewer (implied) features – distinctions – are made available to the clustering procedure. Larger windows stretch the length of the syntactic dependencies, perhaps to the point where little information or no is added. Redington et al manipulated the windows size to empirically determine the window size that resulted in the best balance of accuracy and completeness, finding precisely that 5-grams were optimal.

2. The n-grams centered on some set of types of interest, perhaps the N most frequent terms, are selected. In Wang (2012), these were limited to function words. The assumption is that a learner might know to isolate function words by phonological and/or pragmatic queues,

---

37 Though I focus only on subject pronouns, there is evidence that the procedure described below works also for object pronouns. Wang (2012) induces a dendogram containing a large number of functional items, including nominative and accusative pronouns. In his dendogram, he and she form a cluster early, as do him and her; in parallel, I and we, and me and us, also form clusters. The pair of accusative clusters then merge, as do the nominative pair. So in both cases we have masculine+feminine and first person pronouns clustering first. This same pattern was found in the present study. The isomorphism between the two pronoun cases suggests that the following analysis applies to objective pronouns.
perhaps on the basis of the Phonological-Distributional Coherence Hypothesis (Monaghan, Christiansen, & Chater, 2007). In the present study the target set was further limited to just subject pronouns.

3. Frequency counts of word types in the remaining n-gram slots are used to form high-dimensional vector representations of the contexts of the target types. The resulting vectors are the unreduced distribution vectors for the target types.

4. The dimensions of the distribution vectors are reduced, either simply by truncation to the M most frequent types in each n-gram slot (Redington et al., 1998), or by more elaborate matrix factorization, e.g. the Singular Value Decomposition (Schütze, 1995). The purpose of this step is to smooth out the relationships between vectors by considering only the most relevant dimensions.

5. The reduced distribution vectors are entered into a clustering algorithm and the resulting clusters are evaluated.

The next section details how these steps were concretized in Study 4.

**General Procedure for Distribution Modeling**

The data sources for the distributional analyses were adult child-directed speech (CDS) in CHILDES (English only) and the Spanish data files from the Google Books n-gram corpus.

For the English CHILDES data, a computer program searched the tagged utterances – the output of Stage I of Chapter 3 – for instances of the target tag, either pronouns or (for Chapter 6) prepositions (the precise tag depended on the language and tagset). Only adult utterances were
considered. Each instance of a pronoun was then output with its two neighbors to its left and right, resulting in a 5-gram per instance. Non-word neighbors, such as punctuation, were interpreted as sentence-start or sentence-final markers and replaced with the tags <START> or <END>.

The Google 5-grams were filtered for Spanish subject pronouns in the 3d (middle) position. The target pronouns were \{yo, tú, él, ella, nosotros, nosotras, vosotros, vosotras, ellas, ellos\}. The resulting n-grams were then marked up with and start and end tags. At this stage in the analysis pipeline English and Spanish data had the same structure.

A computer program counted up the distributions over vocabulary items in each of the four positions around the target category, resulting in four distinct unigram distributions, two places before and two after the target. A 1200-dimension distributional vector space was then constructed by taking the 300 most frequent types from each of the four contextual distributions. Finally, each of the target types, from the set of pronouns or prepositions, was assigned a vector in the distributional vector space by extracting its four 300-sized unigram distribution over contextual types. Note that the counts were not normalized. The output of this stage consisted in an N x 1200 matrix, where N was the cardinality of the category of interest.

The distance between each pair of distribution vectors was computed as their rank (Spearman) correlation. Many other distance measures could have been employed – a whole cottage industry of such measures has emerged in the statistics community. Cosine similarity, Pearson correlation and Euclidean distance are also popular, but rank correlation has been found to be particularly effective (Redington et al., 1998) for distributional analysis. It seems that paying attention only to the relative ordering of components in distribution vectors, not their magnitudes, leads to greater accuracy for the sort of linguistic application under consideration here. One might even speculate that rank
correlations are less implausible as models of actual human learning processes, since relative ordering is less memory-intensive than keeping specific, error-free counts of co-occurrences.38

These distances were then input to R’s `hclust` function for hierarchical cluster analysis (HCA) (R Core Team, 2015), using the “ward” agglomeration method. The final product was the visual representation of the hierarchy of clusters, the dendogram.

Results

English Subject Pronouns

There were 609,516 5-grams centered on pronouns. Most were nominative; counts are given below:

Table 23: Counts of English subject pronouns

<table>
<thead>
<tr>
<th>Count</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>213,934</td>
<td>you</td>
</tr>
<tr>
<td>73,869</td>
<td>it</td>
</tr>
<tr>
<td>59,145</td>
<td>I</td>
</tr>
<tr>
<td>31,734</td>
<td>we</td>
</tr>
<tr>
<td>31,520</td>
<td>he</td>
</tr>
<tr>
<td>20,730</td>
<td>she</td>
</tr>
<tr>
<td>23,578</td>
<td>they</td>
</tr>
</tbody>
</table>
Figure 21: Dendogram of English subject pronoun distributions

Figure 21 shows the resulting dendogram. Unfortunately the Freeling tagger does not distinguish between pronouns in subject position versus those having other grammatical functions, so such distinctions are only possible on the basis of surface form (i.e. case). As a result, the distribution vectors for case-invariant it and you incorporated both subject and other uses. In order to counter this, the only occurrences of it and you in the original data that were incorporated into its distribution vector were those where the pronoun was followed by a verbal or modal tag (respectively, Penn Treebank tags beginning with “V”, or equal to “MD”).

We note that the third person pronouns are grouped into one branch, first and second on the other.

She and he cluster more tightly than either with they, understandably as they share GENDER and

39 Den Dikken (p.c.) has pointed out that this strategy will mistakenly interpret it and you in complex subjects as subject pronouns, e.g. in “this picture of it is nice”.
NUMBER features, while first person *we* and *I* merge before *you*, the latter underspecified for NUMBER but sharing with the other two a PARTICIPANT feature, as discussed below. The high position of *it* indicates its vector is fairly different from those of the pronouns it dominates. This is likely a side-effect of its selection criteria.

**Spanish Subject Pronouns**

The Google Books corpus is far larger than CHILDES. Not surprisingly, despite the fact that Spanish is a pro-drop language, the total number of subject pronouns in the Spanish corpus was far larger than the English CHILDES data set. Over nine million pronoun 5-gram tokens were extracted, distributed over 81,518 5-gram types. Because more Spanish pronouns were input than for English (ten versus seven), the dendogram was somewhat more complex.
Here we see contrasts in GENDER, PERSON, and NUMBER represented in a plausible manner, with GENDER relations dominating at the terminals of the dendogram. This is an important clue about the sensitivity of the method to the underlying feature assignments, as I discuss next. In contrast with the English graph, third person pronouns (él, ella, ellas, ellos), while grouped together into a sub-branch, cluster with the plural first and second person units, which are morphologically richer than in English because gendered.
Evaluation

In order to evaluate the clusterings found in the two languages, the H&R feature geometry was applied to English and Spanish pronouns to assign their theoretical feature assignments. The feature assignments were coded as vectors of binary feature values (1 or 0) that were then submitted to hierarchical clustering. Note that the coding shown in Tables 24 and 25 abandons the purely privative feature scheme employed by H&R, since absence of a feature is now represented (by a zero bit). As this change is merely an artifact of the clustering process it has no theoretical impact. Also, features that H&R take as default, such as INANIMATE under CLASS, were made explicit in the coding. Note that in H&R's scheme, because first and second person pronouns are discourse-specific indexicals, they are not ANIMATE – only pronouns denoting gendered entities (including neuters) are also ANIMATE.

Table 24: Features of English subject pronouns

<table>
<thead>
<tr>
<th></th>
<th>I</th>
<th>you</th>
<th>she</th>
<th>he</th>
<th>it</th>
<th>we</th>
<th>they</th>
</tr>
</thead>
<tbody>
<tr>
<td>PARTICIPANT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>speaker</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>addressee</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INDIVIDUATION</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>minimal</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>group</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>CLASS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>animate</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>neuter</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>masculine</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>feminine</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
**Table 25: Table of Spanish subject pronouns**

<table>
<thead>
<tr>
<th>PARTICIPANT</th>
<th>yo</th>
<th>tú</th>
<th>él</th>
<th>ella</th>
<th>nosotros</th>
<th>nosotras</th>
<th>vosotros</th>
<th>vosotras</th>
<th>ellos</th>
<th>ellas</th>
</tr>
</thead>
<tbody>
<tr>
<td>speaker</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>addressee</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>INDIVIDUATION</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>minimal</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>group</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>CLASS</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>animate</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>neuter</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>masculine</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>feminine</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
Distances between all pairs of theoretical feature vectors were measured by their dot products; the latter were then input to R’s `hclust` procedure and graphed in the usual manner (Figures 23 and 24).

The theoretical dendogram for English pronouns is remarkably similar to its distributional counterpart. The only difference is that *they* and *it* have swapped positions. If the learner knew beforehand the feature constitution of English nominative pronouns, but not their exponents, she could read off the mapping of the latter to the former directly off the distributional dendogram; pragmatically-derived information would be needed only to exchange the bundles associated with *they* and *it* and to disambiguate between morph pairs in terminal nodes, e.g. *I* vs *we*. But here we will take the more conservative position that the feature bundles of the language (the Lexicon) are not known.

*Figure 23: Dendogram of English pronoun features (H&E)***
Essentially the same considerations apply to the Spanish result. The distributional dendogram differs from the theoretical one only in that the cluster of masculine plural \{nosotros, vosotros\} merges with the cluster consisting of the third person pronouns \{ella, el, ellas, ellos\}, then with the feminine plural \{nosotras, vosotras\}, while in the theoretical clustering the plural first and second pronouns join first.

The idealized learner that already has the Lexicon would in this case need some extra information to work out the correct correspondences, e.g. that ellos or ellas is plural and that, say, nosotras is first person and feminine.

Clearly, because the similarity between any two items depends on the number of shared features,
pairs of items that have more features in common will merge first in HCA. This means that the more marked units – those bearing larger numbers of features – will tend to merge first, lower in the dendogram. Now, according to H&R’s model, the presence of the FEMININE or MASCULINE feature implies three others: INDIVIDUATION, CLASS and ANIMATE. If we set aside defaults and enforce NUMBER features when INDIVIDUATION is present, then a gendered lexical item will have at least five features: the four to specify the gender and one at least for number. For this reason in the theoretical dendograms for both English and Spanish, it is the gendered pronouns that merge lowest in the tree, and, for pronouns marked masculine or feminine, gender establishes the contrasts lowest in the hierarchies: he/she, ella/el, ellas/ell, and so on. The remarkable fact is that the empirical dendograms reflect this pattern as well. I take this result as double evidence: at once support of the H&R model, and support for the hypothesis that distributional data can directly inform the acquisition of functional elements. The next section presents a hypothesis of how the empirical dendograms, in conjunction with the morphosyntactic feature geometry and assumed semantic-pragmatic learning, can be put to work to derive the mapping between morphs and abstract feature bundles.

Learning Functional Morphemes

The Learning Task

The feature-centered perspective discussed in the previous sections yields a clear characterization of the learning problem: the learner must acquire the mapping of morphs to feature sets, as in the feature assignment matrices of Tables 24 and 25. Simply grouping elements, justified perhaps by their sharing features, is insufficient; the complete feature bundle assigned to each morph is unique.
Considering pronouns, one way of thinking of the problem is by analogy to the sort of popular logic puzzles in which, given some clues, one seeks to match sets of attributes to subjects. As published these puzzles are of course completely deterministic given the clues as constraints, and are easily solved by mechanical means. For the pronoun acquisition puzzle we have three sets of constraints on possible solutions: the distributional dendogram, representing patterns of differences and similarities; the morphosyntactic feature geometry, which limits the possible combinations of features; and, unknown a priori, the information the learner is able to extract from experience through “semantic-pragmatic” means – information that Tomasello posits as the dominant source of linguistic knowledge. Again, since at the current state of acquisition research we cannot be sure what, in the acquisition of subject pronouns, is learned by experience, the approach taken here is to determine the minimal contextual clues required to solve the learning puzzle given the other two information sources, i.e. a priori feature dependency relations and distributional data. Concretely: given a distributional dendogram and the feature geometry, what explicit feature assignments to f-morphs must the learner determine in order to converge onto the correct mapping? The goal, then, is to specify the minimal mappings the learner must acquire via experience.

To gain a sense of the scale of the learning problem, consider, for the seven English pronouns, the following brute-force search strategy:

1. Select seven bundles (= the number of pronouns) from the complete set of possible bundles as determined by the feature geometry;

2. compute the matrix of distances between all pairs of pronoun assignments (21 values);

3. compute the distance between the hypothesized distance matrix and that derived from
distributional vectors.

4. Repeat steps (1) through (3) for all possible feature assignments. The assignment that minimizes the scalar in (3) is the best solution.

Setting aside the possibility of default values, i.e. explicitly stating all features and disallowing bundles containing bare parent nodes that in the H&R scheme take default children if not otherwise specified – for example, disallowing bundles containing a CLASS feature that also do not contain ANIMATE and/or INANIMATE – the total number of possible feature assignments is 99. There are over fourteen billion ways to choose seven out of 99 – a huge solution space. One might employ better search methods, such as evolutionary algorithms, to home in on possible solutions faster, but a second difficulty is that because the function from the feature assignment matrix to the distance matrix is many-to-one (surjective), search methods are likely to identify multiple best fits.

One way to shrink the solution space is to hypothesize feature assignments that are learned by experience. The child might, for example, work out that they is always plural, thus its feature bundle must contain \{\text{INDIVIDUATION}, \text{GROUP}\}. But as yet we have no principled way to decide which experientially-determined assignments would be most useful – clearly, the more the better for learning. One possibility is that by simulating the brute-force search procedure we might determine the set of best solutions, from which it would be possible to determine the experiential knowledge required to pick out the single correct solution.

We can also exploit the distributional dendogram to further limit the solution space. For example, given some feature set assigned to she, it is very unlikely that the set for he would differ significantly, given their relative positions as sisters in the distributional dendogram. This constraint is expressed
when comparing distances (step 3 above), but it seems desirable to push the constraint into the
generation procedure so that every hypothesis tested already captures the relative similarity of *he* and
*she*. The algorithm described in the next section builds on these considerations to define a purely
*logical* (i.e. non-statistical) procedure that identifies the correct solutions for both English and
Spanish in finite (computational) time.

**A Learning Algorithm**

I will describe and illustrate the procedure mostly with reference to the English data set, which also
constituted the development set for the implementation; the latter was then tested on the Spanish
data. The Spanish case is understandably more involved due to the larger number of pronouns (ten
versus seven). The exposition is informal, and I stress that in the absence of significant, principled
constraints on learning algorithms, the method I offer stands as merely one solution among many
one might construct.

1. The first step consists in generating all possible combinations of features from the feature
genometry (the powerset of the feature space filtered by the feature geometry). Call this set
of sets $F$.

2. Next, the abstract features for each terminal morph in the distributional dendogram (seven
for English and ten for Spanish) are generated. This is accomplished by associating an
abstract feature with each node in the empirical dendogram. Each feature bundle consists of
all features on the path from a terminal to the root (Figure 25, where RE = Referring
Expression). Sister terminals will therefore share all features but one, and the resulting
assignment matrix will induce a dendogram identical to the distributional clustering.
3. Crucially, even though the initial hypothesis recreates the similarity relationships falling out of the distributional analysis of the English pronouns, it cannot, without alteration, represent a template of English feature assignments. The reason is that for some nodes the number of features available is incorrect. English *I* requires four features: \{PARTICIPANT, SPEAKER, INDIVIDUATION, MINIMAL\}, but the hypothesis in Figure 25 only has three, \{B, D, E\}. The latter bundle must be stretched to incorporate a fourth feature (abstract, for now) so that it can eventually match the correct feature specification. The proposal is that a function of experience is to adjust the dimensions of the hypothesized bundles. This occurs by its imposing constraints on hypothetical bundles that force expansion (or contraction), to enable unification with bundles produced by the feature geometry, i.e. sets in

![Figure 25: Assignment of abstract features to morphs](image_url)
F. For example, if the learner knows (by whatever means) that I has features \{SPEAKER, MINIMAL\}, since no set in F incorporates those two but not their two parents, PARTICIPANT and INDIVIDUATION, it follows that the initial assignment of \{B, D, E\} must be expanded with a fourth element, \{B, D, E, X1\}. Note that according to the H&R feature geometry 1SG might be larger than it is in English, e.g. a language might assign gender, as Spanish does for 1PL. The conservative approach described here will require positive evidence for this in order to consider such possibilities.

For English, the necessary and sufficient node-altering information the learner needs are:

FEMININE for she, (or MASCULINE for he), SPEAKER and MINIMAL for I (or SPEAKER and GROUP for we), and INANIMATE and MINIMAL for it.

4. Sister terminals are balanced (upwards) in their cardinality. That is, wherever two terminals are merged into a cluster, if one member of the cluster is smaller than the other, then the smaller set is expanded by copying features from the larger set, such that at the end of this step both have the same number of features. The two are also constrained to share all but one feature. Continuing with I, which now has features \{B, D, E, X1\}, because we is assigned \{B, D, F\}, it is adjusted by copying over feature E, giving \{B, D, E, F\}. I and we now share three features. Since I's bundle is also constrained to contain \{SPEAKER, MINIMAL\}, one or both of these must also be assigned to we – one only if the feature discriminating I from we, X1, is also one of \{SPEAKER, MINIMAL\}, as indeed is the case (since the two pronouns contrast in number).

At the conclusion of step 3 the situation will be as in Figure 26.
5. The feature templates and constraints from Step (3) are unified with all compatible subsets of F. If there is only one match, then that is the solution; if several, then additional considerations (e.g. experiential data) are required to further prune the results. In English there are two solutions, the only difference being the PARTICIPANT specification for you, either SPEAKER (incorrect) or ADDRESSEE (correct). To delete the former, either the learner must work out that you refers to second person, or we might stipulate an additional structural constraint whereby the lowest feature in each bundle, then one representing the terminal node (C for you in Figure 26) must never share with any other c-commanded terminal features. This must be true in the case of you: if the PARTICIPANT type of you were SPEAKER then you and I would be very similar and ought to merge first.
A last adjustment for English concerns *they*. Its bundle begins with three features. On unification it is assigned \{INDIVIDUATION, MINIMAL, GROUP\}, which H&R interpret as the specification for dual. The learner will need to rely on experience to delete the MINIMAL feature from the bundle.

This procedure yields the correct solution for English. A possible set of extrinsic feature assignments required by the process are summarized in Table 26. For Spanish seemingly more the data are required to restructure the initial bundle hypotheses, as exemplified in Table 27.

**Table 26: English experientially-derived feature assignments**

<table>
<thead>
<tr>
<th>Pronoun</th>
<th>Required Assignments</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>SPEAKER, MINIMAL</td>
</tr>
<tr>
<td>it</td>
<td>INANIMATE, MINIMAL</td>
</tr>
<tr>
<td>she</td>
<td>FEMININE</td>
</tr>
<tr>
<td>they</td>
<td>delete MINIMAL</td>
</tr>
</tbody>
</table>

**Table 27: Spanish experientially-derived feature assignments**

<table>
<thead>
<tr>
<th>Pronoun</th>
<th>Required Assignments</th>
</tr>
</thead>
<tbody>
<tr>
<td>yo</td>
<td>SPEAKER, MINIMAL</td>
</tr>
<tr>
<td>ella</td>
<td>FEMININE</td>
</tr>
<tr>
<td>ellas</td>
<td>FEMININE, GROUP</td>
</tr>
<tr>
<td>nosotros</td>
<td>SPEAKER, FEMININE, GROUP</td>
</tr>
<tr>
<td>nosotros</td>
<td>SPEAKER, MASCULINE, GROUP</td>
</tr>
</tbody>
</table>

Now, while there are certainly more extrinsic assignments for Spanish (eleven versus five plus a deletion for English), these may well fall out from morphological generalizations the learner may
induce. Specifically, the assignments to FEMININE might all fall out from associating the word-final vowel “a” with feminine grammatical gender, and GROUP from associating plurality word-final with “s”, which in general spells out plurality in Spanish and is therefore easily learned. Thus the “s” in nos-, vos- and ellos provides a strong clue that these have plural reference.

There is some evidence from acquisition in support of these conjectures. While the expression of grammatical gender in English is limited to 3SG pronouns, the richer gender agreement system of Romance provides interesting data about children’s acquisition of DP syntax. Socarrás (2011) reviews a series of previous studies, including Aguirre (1995), Hernández Pina (1984), López Ornat, (2003), Schnell de Acedo (1994), and presents her own analyses based on records of speech from three children in Puerto Rico. She shows that while the empirical picture is not perfectly clear (likely due to tiny samples and other methodological issues), children seem to be working out gender agreement from very early on, even at MLU ~ 1.6. Once again, observed errors are rather revealing, because of regularities not present in the input. One child uttered the DP “una papá”, where morphological agreement trumped semantic gender. Rafael, a child studied by Hernández Pina, at 21 months produced non-target-like word endings in order to regularize the agreement patterns (2011, p. 39)

<table>
<thead>
<tr>
<th>Child Production</th>
<th>Correct Form</th>
<th>Gloss</th>
</tr>
</thead>
<tbody>
<tr>
<td>moto roto</td>
<td>moto rota</td>
<td>motorcycle (fem) broken (fem)</td>
</tr>
<tr>
<td>un llave</td>
<td>una llave</td>
<td>a (masc) key (fem)</td>
</tr>
<tr>
<td>tierra azula</td>
<td>tierra azul</td>
<td>earth (fem) blue (fem)</td>
</tr>
</tbody>
</table>

I will return to the third case below in the discussion section. In the first two the child seems to take word-final -a as the marked gender, in the first case giving it priority over the masculine ending -o
despite its being suffixed onto the adjective, so that instead of moto roto we get mota roto; while in the second case the child constructs the non-existent adjective azula in order to agree with the noun. Whether or not he links the suffix -a to semantic gender, it is clear that Rafael knows it plays a morphosyntactic role in the language – that it expresses a feature. I return to the topic of gender in child language below in the discussion section.

**Implementation Notes**

The learning algorithm described above was implemented in Prolog; see the Appendix for a listing of the program. The program takes as input a Prolog description of an empirical dendogram and a set of assignments of features to exponents, returning complete feature bundles for the exponents. Because, as detailed below, the implementation is very free in how concrete features are assigned, it generates huge numbers of solutions, most – or all, when the program fully “converges” – amounting to the same logical solution. A completely successful run will result in exactly one logical solution, therefore this must also be the first one returned.

The empirical dendograms are materialized as nested lists:

```
dendogram(en, pron, [[it, [they, [she, he]]], [you, [we, i]]]).
dendogram(es, pron, [[[ella, el], [ellas, ellos]], [[nosotras, vosotras], [[yo, tu], [nosotros, vosotros]]]]).
```

The complete set of possible bundles from the H&R feature geometry are constructed by recursive descent initiated by the predicate re. The resulting structure is constrained to have at least one of a PARTICIPANT and/or INDIVIDUATION node. Resatisfaction of the predicate re will produce all possible combinations of features.
re(RE) :-
    participant([], RE1),
    individuation(RE1, RE),
    RE \= [].

A PARTICIPANT node, if present at all, must contain one of SPEAKER or ADDRESSEE. That is, if a PARTICIPANT node is added to the bundle under construction, then also either a SPEAKER or ADDRESSEE node must be added. Note the second, body-less statement of the predicate participant, which introduces a choice-point that does not add a PARTICIPANT node.

participant(RE1, RE) :-
    append(RE1, [participant], RE2),
    speaker(RE2, RE3),
    addressee(RE3, RE),
    not(reverse(RE, [participant | _])).
participant(RE, RE).

The rest of the features are (optionally) appended in similar fashion. The initial hypothesis for the feature bundles is built off the empirical dendogram by the create_vocab predicate. Its key clause employs the dif/2 mechanism to ensure that contrasting features never receive the same value:

create_vocab(F, [X, Y], Vocab) :-
    var(A), var(B),
    dif(A, B),
    append(F, [A], FX),
    append(F, [B], FY),
    create_vocab(FX, X, X1),
    create_vocab(FY, Y, Y1),
    Vocab = [X1, Y1, F].

This clause fires for non-terminals, taking the abstract features F from the root to the current non-terminal node as input. It creates two fresh variables which, via dif/2, are guaranteed to not have the same value (distinctive features) and appends each in turn to F and recursively invokes create_vocab. In the following query, variable V contains the bundles of abstract features for she and he.
?- tree(en, _, T), create_vocab(T, V).

... [she, _G1907, _G1925, _G1946, _G1949], [he, _G1907, _G1925, _G1946, _G1967], ...
dif(_G1949, _G1967), ...

The first three variables for he and she are identical, while the last pair, _G1967 and _G1949, are constrained to be not-equal (per the dif attribute). A crucial issue for this structure concerns the method by which features are assigned to abstract bundles. For example, if FEMININE is assigned to she, it matters whether the assignment sets the variable that is not shared with he, _G1949, or one of the other shared variables in the bundle. In the latter case, FEMININE would also be assigned to he. The algorithm must be allowed to consider both cases – it does not yet know whether gender is the distinguishing feature for he and she. Thus the predicate assign_features creates choice points for all possible unifications of input features to free variables in the bundle (a demonstration of the beauty of Prolog):

assign_features(Bundle, [FeatureValue | Rest]) :-
   include(var, Bundle, FreeVars),
   reverse(FreeVars, RevFreeVars),
   nth1(_, RevFreeVars, FeatureValue),
   assign_features(Bundle, Rest).

In general the algorithm converges faster if feature assignment starts with the “lowest” (rightmost) feature first, hence the reverse/2 call in the body of assign_features. This optimization does not further limit the solution set.

The set of (partially assigned) feature bundles is then unified with all alternative subsets (of equal cardinality) of the theoretical bundles. It would be a mistake, however, to simply rely on Prolog’s built-in unification mechanism, because it is sensitive to position. That is, a bundle [X, Y, Z, speaker] would not unify with [participant, speaker, individuation, minimal] because the
feature SPEAKER is not in the same position. We therefore require an operation \texttt{set unify}, that matches elements independently of position. The core predicate is the recursive \texttt{set unify elements}, which exploits Prolog’s \texttt{select/3} to test for identity at all positions in the target set:

\begin{verbatim}
set_unify_elements([], _).
set_unify_elements([H | T], Y) :-
    select(H, Y, Y1),
    set_unify_elements(T, Y1).
\end{verbatim}

These last two operations, feature assignment and set unification, together conspire to yield an enormous number of alternative solutions. Nevertheless, with the right starting values for the initial feature assignment, the program converges immediately onto the correct pronoun vocabularies in both languages.

\section*{Discussion}

The learning algorithm presented in this chapter assigns specific functions to the three sources of information that enable the acquisition of subject pronouns. On this hypothesis, the universal geometry of phi-features determines the set of features available and constrains their co-occurrence. Distributional analysis supplies a hierarchical model of similarities and differences based on usage that induces a first hypothesis of the feature structure of pronouns in the language. Experience supplies subsets of concrete UG features that force the restructuring (expansion or contraction) of the feature templates constructed by distributional analysis. As demonstrated by computer simulation, the procedure is able to make explicit predictions about what specific information is required from pragmatic experience to correctly learn the subject pronoun vocabulary of English and Spanish.
The obvious critique of this scheme from the usage-based perspective is this: if some of the features assigned to pronouns are determined by experience, why not all of them? The distributional analysis could still be used in learning, but economy of theory argues for scrapping the feature geometry in toto on the assumption that children are able to work out all relevant contrasts – and their links to conceptual meaning – through the sort of interactive, contextual and pragmatic processes Tomasello proposes.

A usage-based proposal might run as follows. Harley and Ritter's feature geometry is at best only a descriptive tool, an attempt to factor phi-features cross-linguistically; it bears no psychological reality. If there are no universals then what is learned need not refer to universal grammatical primitives. The inventory of phi-features can then be relativized to the language. In English we traditionally observe surface distinctions of PERSON (three values), NUMBER (two values) and (third person singular only) GENDER (three values). We might construct an assignment matrix as in Table 28.

Table 28: Surface feature assignments to English pronouns

<table>
<thead>
<tr>
<th></th>
<th>I</th>
<th>you</th>
<th>he</th>
<th>she</th>
<th>it</th>
<th>we</th>
<th>they</th>
</tr>
</thead>
<tbody>
<tr>
<td>First</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Second</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Third</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Singular</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Plural</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Feminine</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Neuter</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Masculine</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

We assume that all assignments of features to morphs are learned by constructivist means, aided by
distributional information. If *her* is correctly determined to be feminine, then, because of their shared position in the dendogram, the child may well hypothesize that *he* is masculine. And so on. A priori knowledge of morphosyntactic features is thus redundant.

One difficulty faced by this account is that if the child learns by semantic-pragmatic means, presumably he will initially develop a semantic understanding of features, and only later map the semantics to grammatical representations. Yet, as Valian (1986) has pointed out for English, and Socarrás for Spanish, for at least some feature classes children appear to develop the relevant syntactic distinctions before associating those with meaning. In particular, Valian cites an early study by Levy (1983) that “ruled out recognition of the sex of an object as a basis for learning linguistic gender in several languages” (p. 572). As previously noted, the issue turns on the reflex in language acquisition of the relation between grammatical gender and *natural* or semantic gender – the sex of (usually) animate entities. The essential question is this: do children, in their linguistic expressions, *initially* select gender morphemes based on the sex of referents (where this is overt), and only later work out the grammatical rules for gender, or instead do they attend to the formal (syntactic and morphophonological) nature of the gender system in their language right from the start? More generally, is the acquisition process for children at its root a matter of establishing form-meaning correspondences, or all along are children also intent on working out the structural aspects of their language? This latter possibility,

[C]oincides with an idea […] according to which there is no privileged region of experience in which the child forges her development, as would be derived by Piagetian theory, where the role of physical experience is indeed privileged. Physical experience as well as social and linguistic experience demand work on the part of the child, and will trigger the child's interest. As (Karmiloff-Smith, 1981) has expressed, language
constitutes a formal space that demands solutions to specific problems. (Pérez, 1990, p. 75 my translation)

That children and adults show a general preference for the formal character of linguistic gender over extra-linguistic gender has been shown for a variety of languages, e.g. Spanish (Pérez, 1990), Greek (Koromvokis & Kalaitzidis, 2013), French (Karmiloff-Smith, 1981) and Hebrew (Levy, 1983). In some of this work (the first two in the previous list, for example), the children are already too mature (four and six years old respectively in the Spanish and Greek studies) to help decide the question of what comes first, natural or grammatical gender. As Levy underlines, we must go back to the very beginning of multi-word speech by toddlers to examine the path of acquisition, taking into account both production and comprehension data.

One such study is Arias-Trejo and Alva (2013). Thirty-seven Mexican-Spanish toddlers were tested at 30 months using a preferential looking design. During the training phase a pair of unfamiliar inanimate objects were described, but not named, using gender-inflected adjectives (the suffixes -a and -o respectively for feminine and masculine terms). The children were then tested by again presenting them the objects, but this time the objects were referred to directly using gendered nonce terms: Mira, una betusa (“Look, a betusa”) and Mira, un pileco (“Look, a pileco”). The prediction was that toddlers would prefer the object that during the training phase was described using an adjective of the same ending as the noun in the test phrase (the properties of the objects to which the adjectives referred were scrambled during testing). The results bore out the hypothesis, indicating that at 30 months the children were able to infer the names of objects by means of morphosyntactic agreement between adjectives and nouns. Children at this age appear, minimally, to have the capacity to perform such linguistic analyses. What this study does not tell us, however, is how such learning interacts with natural gender.
Levy (1983) addresses this question head on. The hypothesis that around age 2 children's grammar is based on word-meaning correspondences, and that the formal system of language is learned later, leads to certain predictions about the types of erroneous generalizations children might make in the early stages of acquisition. Children might be expected to more accurately express the gender of animate entities of known sex, but mark gender at chance accuracy when referring to inanimate objects. Or they might extend the gender marking of known animate, sexed entities to associated objects. By contrast, a theory that lays stress on the formal aspects of acquisition will predict that at least some usage errors will occur as by-products of linguistic generalizations, e.g. *la mapa* “the-FEM map” instead of *el mapa* “the-MASC map”.

Levy marshals data from prior work in German, Russian, Polish and French that, in her telling, together suggest children's use of gender inflection reflects the morphological features of words primarily, and only secondarily relates to the objective sex of entities. Her own work with children acquiring Hebrew turned on the regularity of gender in the nominal system of the language. In Hebrew every noun (whether animate or inanimate) is assigned gender via the phonological features of the word's ending, where feminine is marked: nouns ending in /a/ or /t/ are feminine and take the suffix /ot/ in the plural, while all other nouns are masculine and take /im/ in the plural. If the early grammars of children are semantic in character, we would expect errors in the pluralization of non-sexed (inanimate) referents. Yet in a longitudinal study of her own son, and in a cross-sectional study of 32 Israeli children at 30 months, no reflex of the animate/inanimate distinction was observed in the formation of plurals:

[N]o evidence for a semantic notion of gender that would facilitate the acquisition of the plurals of animate nouns was found nor was there in fact any indication of a binary classification of nouns that would group inanimates and animates together in ways that
could result from a partial misunderstanding of natural gender distinctions. (p. 85)

It seems, then, that children are sensitive to syntactic and phonological features even in the first stages of acquisition. Once again, “form is easy, meaning is hard” (Naigles, 2002). The order of acquisition appears to be at odds with the purely constructivist model.

A second issue is that the hierarchical clustering of the feature assignments in Table 28 looks rather different from the distributional dendogram (Figure 27). For example, whereas distributional analysis suggests we and I share many features, as they do per H&R, in Figure 27 these two pronouns are only distantly related. We can look at the matter quantitively as well. If we take the distance matrix feeding the clustering and subtract from it the distance matrix based on a theoretical assignment, then square those differences and sum the elements of the resulting matrix, we have a rough scalar measure of the deviance of the empirical matrix from a theoretical one. For any two sets of hypothesized feature assignments to f-morphs, the one whose sum of squared differences is lower is the better fit to the empirically-determined distribution patters. As it happens, the sum of squared differences between the distance matrix for the distributional clustering and those of the “traditional” feature assignments versus those based on the H&R geometry are 0.062 and 0.033 respectively (again, lower is better). That is, the H&R model fits the distributional statistics better. This is not fatal for the empiricist position but it does suggest that, under its likely assumptions, distributional analysis is less useful to acquisition than it is to the UG-based approach. The mathematics are such that the only way to better reconcile the distributional facts with the theoretical target is to add more features to the latter. But since these could not be based on surface characteristics or meanings (we already represent those), the learner must construct several purely abstract features, like H&R’s organizing nodes. The account must then explicate the processes by
which such abstract features are hypothesized and (dis)confirmed – and how such theorizing on the child’s part might be constrained.

Finally, both accounts must contend with the issue of the availability of defaults in syntax, a topic that has been extensively investigated in the acquisition literature (Guasti, 2004). If features characterize the functional sequences of the grammar, then the issue of defaults is core to the UG debate, weighing on key issues such as minimal structure, bare noun phrases, and root non-finite verbs (Rizzi, 2005). Should defaults be found to be universal – if, for example, before its language-specific default is set, all children default some feature to the same value – then usage-based researchers must explain their cognitive basis – likely in terms of universals of the human experience and culture. But if defaults are always language-specific, and no language-independent prototypical settings are visible, then the UG position is weakened because there is less need for abstract, language-independent representations of grammatical features to act as hosts for the
specification of defaults. Specifically with respect to pronouns, Harley and Ritter (2002), building on earlier work by Rebecca Hanson (Hanson, 2000; Hanson, Harley, & Ritter, 2000) reviewed prior acquisition studies involving several language families to determine whether there is evidence of language-independent orderings in the acquisition of pronouns. They divided the findings into three developmental sequences: children whose initial pronouns are first person singular, those who first produce third person neuter pronouns, and a third group of subjects who were already producing multiple pronouns when the associated studies began. Setting the third group aside, Hanson et al interpret the alternation between 1SG and 3SGN as evidence of defaults for the PARTICIPANT and INDIVIDUATION organizing nodes. On a structure-building hypothesis, a child might begin with either the one or the other. If just mapping PARTICIPANT, the child defaults to SPEAKER, giving the first person. INDIVIDUATION defaults to MINIMAL for number and, if a CLASS node is also present, to INANIMATE, giving 3SGN. If both PARTICIPANT and INDIVIDUATION are in the bundle, then, given their respective defaults, we have 1SG.

It is hard to say whether Spanish too exhibits these defaults, since the language lacks neuter forms of its subject pronouns. Socarrás (2011) does conclude that singular is the default for number, but also that masculine is the default gender, by which she explains the majority of agreement errors found in the data she analyzed. For example, Rafael’s substitution of un llave ‘a (masc) key’ for una llave ‘a (fem) key’ can be explained by noting that llave ‘key’ is unmarked for gender (the final e can go either way); Rafael opts for a default, producing the male indefinite article. Assuming Socarrás is right in her conjecture, it would seem that Rafael has already learned the language-specific default, so his data fail to discriminate between nativist and usage-based positions. Indeed the example itself may not be helpful since nouns in Spanish (along with their determiners) are necessarily either
masculine or feminine. The language does have a neuter definite determiner *lo*, homophonous with its neuter object pronoun/clitic (e.g. *damelo*, ‘give-me-it’, ‘give it to me’). Its distribution is limited to nominalizing abstract adjectives: *lo importante es que estés bien* ‘the important is that you’re well’; *lo peor es caer en la calle* ‘the worst is to fall in the street’. Yet intriguingly in the appendix to Socorrás’ monograph there is a record of a child, Alonso, employing *lo* in front of nouns not overtly marked for gender (p. 188): *lo pece* ‘the fish’, *lo juete* ‘the toy’, *lo pie* ‘the foot’. Socorrás interprets these as reduced forms of plural articles and nouns, noting also the rarity of *lo* + noun constructions. But *lo* as object pronoun and clitic is very common, so it’s not impossible that a child would draw conclusions as to its gender feature.

As always, it is very difficult to derive general conclusions from such minuscule non-experimental samples. These scattered speculations merely point to the field’s dire need for data at scale, and especially for the insightful and statistically robust analysis of children’s non-target-like productions.

**Conclusion**

This chapter elaborated an algorithm that addresses the “linking problem” (Tomasello, 2003) as it relates to the acquisition of functional elements: how does a child come to associate universal morphosyntactic features with sounds specific to her language? The procedure was developed on English subject pronouns and tested on the equivalent units of Spanish. Based on the observation that the correlations between distributional vectors for pronouns are similar to the correlations between feature bundles for the same units – where features are assigned according to the H&R geometry of phi-features – I described a logic-driven search strategy over the space of phi-features that, constrained by the H&R geometry, associates sets of features to pronominal f-morph. The model examines the patterns of similarity and difference between the distributions of the pronouns,
and, comparing those to the comparable patterns implied by the UG-given feature geometry, makes
inferences about possible feature assignments to the pronoun types.

The procedure, as demonstrated by its computer implementation, is neither sufficient nor necessary
for acquisition. The search yields multiple possible sets of assignments which the learner must
somehow disambiguate. I have made no claims about the form of learning required to identify the
one correct solution, instead taking it to be “experiential” in character – some form of semantic
bootstrapping, perhaps by means of the social-pragmatic learning advanced by constructivists. The
value of the demonstration lies in its ability to make explicit the information the learner requires
from this other “experiential” source, assuming the learner is granted the a priori knowledge of the
feature hierarchy and the ability to gather, store and analyze lexical co-occurrence statistics.

Specifically, the learner requires:

1. extra information to determine the exact feature values of the contrast at nodes in the
distributional dendogram containing only a pair of terminals;

2. if the rank of the feature matrix is larger than that of the distributional correlation matrix –
   if more true features underlie are active in the syntax than there are contrasts in the
distributional data – then those extra features too must be determined by independent
   means.

An important avenue of future work in this area consists in investigating the learning strategies a
child might use to extract feature-specific knowledge of function words from observing and
participating in communicative situations. Again, I have seen little in the literature on the topic,
whether by constructivists or nativists. Another important question concerns the amount of
distributional information required by the proposal described in this chapter. As Virginia Valian (p.c.) put the question, does the model's accuracy increase as the input cumulates? How much input is required to generate the dendogram in Figure 21 – and at what developmental stage will that data have been accumulated by the child? In this connection it is noteworthy that Wang (2012) worked with a far smaller data set consisting of eight sets of CHILDES transcripts, but nevertheless obtained plausible structures for the organization of pronouns, e.g. *I* and *we* form a cluster in parallel with *he* and *she*, as in Figure 23.

In the next chapter we return to prepositions. The distributional analysis and acquisition models developed in this chapter will be applied to adult preposition data. As we will see, the learning problem is considerably simplified if feature dependencies are constrained to a linear hierarchy – a functional sequence.
Chapter 6. Learning Functional Prepositions

Introduction

In this final chapter the distributional analysis and learning mechanisms from the previous chapter are extended to English and Spanish prepositions. Because our knowledge about the features of Ps is less mature than that of the phi-features of pronouns, here the exercise must proceed in several steps. It is necessary to first establish a hypothesis about the nature and structure of features targeted by the acquisition process – for pronouns these came packaged as the feature geometry of Harley and Ritter. Only then is it possible to specify an algorithm for matching sets of features to exponents of P.

Recall from the latter portion of Chapter 1 that a hypothesis emerged about the f-seq of functional P, essentially a merger of Caha’s case sequence, Pantcheva’s decomposition of Path, and the overall extended projection of N posited by Svenoniuss:

\[(5) \quad \text{COM/Route} - \text{INST/Source} - \text{DAT/Goal} - \text{LOC} - \text{GEN/PART} - p - \text{AxPart} - \text{DP} \]

An alternative model is obtained if we inflate the Asp^{SPACE} nodes in Den Dikken’s extended projections of P_{dir} and P_{loc} into steps along Caha’s f-seq, as in the parentheses below:

\[(6) \quad \text{CP}_{\text{PATH}} - \text{Dx}_{\text{PATH}} - \text{Asp}_{\text{PATH}} = [\text{COM/Route} - \text{INST/Source} - \text{DAT/Goal}] - P_{\text{dir}} - \text{DP} \]
\[(7) \quad \text{CP}_{\text{PLACE}} - \text{Dx}_{\text{PLACE}} - \text{Asp}_{\text{PLACE}} = [\text{LOC} - \text{GEN/PART/Stative}] - P_{\text{loc}} - \text{DP} \]

If it is agreed that the hierarchical clustering method applied to distributional vectors is effective in recreating the relations of similarity between pronouns subtended by the morphosyntactic features of pronouns, it is then plausible that the same method, applied to Ps, might help inform a decision
about (i) whether the Pantcheva/Caha case sequence is correct; (ii) how spatial P relates to the functional sequence, and indeed whether relations of similarity among spatial prepositions appear to be primarily determined by interface features or by purely semantic-internal features.

In Chapter 1 we also took note of Romeu’s analysis of spatial P is Spanish. Though it shares core principles with Pantcheva and Caha, in practice Romeu’s approach yields fairly different conclusions about the structures he studies. In what follows the distributional method is applied to a large sample of Spanish PPs in order to evaluate Romeu’s proposals.

Finally, the learning algorithm described in Chapter 5 is adapted to the (far simpler) task of acquiring the functional vocabulary associated with cumulative functional sequences of the sort pioneered by Pantcheva and Caha. A Prolog implementation is able to easily learn the functional Ps of English.

Methods

The procedures used for the distributional analysis of English and Spanish Ps were exactly the same as those employed in Chapter 5, except of course the units extracted were prepositional. Again, because of the relatively low counts of PPs in the Spanish CHILDES data, Google Books data constituted the source of Spanish 5-grams centered on Ps.

For the English dataset, it was found that Freeling codes many instances of complementizers as prepositions; these were ignored by explicitly filtering on P types. Prepositions that are often used intransitively, i.e. particles, were also ignored by filtering out Freeling’s POS tag for particles. Ps whose adult frequency was at least 500 were retained. The item *like* was also eliminated, as it often heads adjectival phrases\(^{40}\).

\(^{40}\) Den Dikken (p.c.)
For Spanish the Ps selected for analyses were instead limited to the intersection of P types in the Google corpus and the Ps to which Romeu (2014) assigns structures. Because of the very large frequencies in the Spanish data there was no need for a lower bound criterion for inclusion.

Analysis

*English Prepositions*

A total of 222,543 adult English P 5-grams were entered into the distributional algorithm. The top ten most frequent P heads were distributed as follows:
Table 29: P Counts in CHILDES child-directed speech (English)

<table>
<thead>
<tr>
<th>Count</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>44067</td>
<td>in</td>
</tr>
<tr>
<td>40303</td>
<td>on</td>
</tr>
<tr>
<td>25799</td>
<td>to</td>
</tr>
<tr>
<td>24643</td>
<td>of</td>
</tr>
<tr>
<td>22987</td>
<td>with</td>
</tr>
<tr>
<td>19969</td>
<td>for</td>
</tr>
<tr>
<td>15248</td>
<td>at</td>
</tr>
<tr>
<td>7944</td>
<td>about</td>
</tr>
<tr>
<td>5723</td>
<td>over</td>
</tr>
<tr>
<td>4504</td>
<td>from</td>
</tr>
<tr>
<td>2165</td>
<td>into</td>
</tr>
<tr>
<td>1901</td>
<td>after</td>
</tr>
<tr>
<td>1611</td>
<td>under</td>
</tr>
<tr>
<td>1353</td>
<td>before</td>
</tr>
<tr>
<td>1212</td>
<td>through</td>
</tr>
<tr>
<td>992</td>
<td>around</td>
</tr>
<tr>
<td>827</td>
<td>behind</td>
</tr>
<tr>
<td>650</td>
<td>inside</td>
</tr>
<tr>
<td>645</td>
<td>until</td>
</tr>
</tbody>
</table>

The resulting distributional dendogram (Figure 28) might serve as the principal result of this dissertation.
The clustering procedure neatly organizes the Ps into three subtrees, marked by the rectangles in the figure. To the right (blue rectangle) are *all and only* functional prepositions, on the assumption that *from* lexicalizes Source (= Instrumental case) without descriptive content, and that *at* is functional (as in *look at me*) and a static “path” expression of sorts that creates locative expressions as proposed by Svenonius and others (Svenonius, 2006c). On the left (red rectangle) are spatial Ps, and in the middle are Ps that usually contribute temporal meanings (green rectangle), though the distinction here is hardly crisp (cf. Kant's notion of the dual character of time and space). Possibly, we ought to view the Ps in the green rectangle as the lexical Ps that are somehow closer to the functional units – perhaps on account of their broader semantic range.

Consider now the following two assumptions:
A) The method of distributional clustering is sensitive to the feature makeup of target lexical items, whether the features in question are of the synsem or a posteriori (encyclopedic) variety. This has been amply demonstrated in the literature as well as in Chapter 5, where it was demonstrated that the method organizes subject pronouns in a manner consistent with their phi-features.

B) Because, in constructing the clustering, the procedure first merges items that are most similar, and because similarity is greater where the total number of features shared is greater, there is a tendency for clusters lower in the dendogram to group items that are associated with larger feature bundles. It is a tendency because, as shown in the dendogram for Spanish pronouns (Figure 5.2), it is always possible for pairs of very similar elements to merge into the overall tree late since together they are sufficiently dissimilar from the rest. For example, in Figure 29 the plural first and second person pronouns merge above the third person pronouns, even though they have more features.

If we accept these assumptions, the dendogram of English Ps suggests that if we also take the Grimshavian view of PP as the extended projection of N (a single f-seq rooted in N), such that all Ps are functional, then the fact that the set of Ps that are traditionally (and here confusingly) referred to as functional (of, for, with etc., exponents of K in Svenonius’ proposal) merge with the temporal Ps before the spatial Ps suggests that the first and/or second group bear more features than the third – that is, one or the other of the first two groups lies above spatial P in the f-seq. This observation lends support to the notion that \( K \) lies above (little) \( p \), in agreement with Caha (2011) and contra Svenonius (2010).

Within the group of functional Ps, postponing discussion of about and at until we model the
acquisition of functional Ps below, we observe the following sequence, where “>” indicates “contains more features”:

(8) \[ \text{with} > \text{for} > \text{from} > \text{to} > \text{of} \]

That is, the graphs suggests that \textit{with} has most features, followed by \textit{for} with which it shares a subset of features with \textit{with}, and so on. This is almost exactly the like the Caha’s f-seq of the oblique cases:

(9) \[ \text{COM} > \text{INST} > \text{DAT} > \text{GEN/PART} \]

The one non-congruence is the presence of benefactive \textit{for} higher than the dative “zone”. Further work is needed to determine why \textit{for} is attracted to its higher position. That aside, I take Figure 28 as intriguing evidence for Caha’s analysis, which he arrived that by wholly different – non-distributional – means, namely his cross-linguistic study of the patterns of syncretism of case affixes. That a distributional analysis of functional adpositions lines up this well with a paradigmatic analysis of case affixes should give us confidence that the underlying theoretical model is well-supported empirically.

The fact that the distributional algorithm makes a distinction between temporal and spatial meanings of \( P \) is at odds with the single f-seq hypothesis. None of the proposals reviewed in chapter two designate a region of the f-seq for temporal Ps distinct from the fine structure of spatial \( P \); and certainly all temporal Ps in English can signify spatial relations (if somewhat oddly for \textit{until} in American English but a standard L2 normative error). The distinction between spatial and temporal meanings is semantic only. Therefore, in partitioning \( P \) into three broad sets, the algorithm is at best making a semantic distinction.

Again, within the cluster of functional prepositions we have evidence that further distinctions are
determined by the synsem features of the individual units, in accord with Caha’s f-seq for case. It is hard to determine what sort of distinctions the algorithm makes within the subtree in the green rectangle, as it incorporates just three items. But for the large cluster of spatial Ps, the pattern of sub-clustering suggests that the algorithm is more sensitive to semantic-internal features than it is to synsem features. Three sub-clusters are visible: roughly, we have the directional Ps \{into, through\}, the AxParts \{inside, under, behind\}, and the statives \{around, in, over, on\}, though the status of over and around is complex since both may act as AxParts (see below) and can also express path senses. In light of assumption A, the clustering patterns suggests that since the directionals merge in later, they have fewer features than statives and AxParts – the opposite of every syntactic theory of PP. Within the AxPart sub-cluster the organization of the items follows no obvious pattern, unsurprisingly since we have no proposal in hand that subdivides AxParts by synsem features. Rather, while the overt role of the AxPart category in Svenonius’ and Romeu’s models is to identify subregions of projective Grounds (Svenonius, 2006b), which is why in some contexts over and around behave like AxParts, in effect the category collects the many lexical-seeming components of adpositional phrases. As for the statives, that on and over bind more tightly than either to in again suggests a semantic criterion. On and over share the semanteme I have called [superiority] but are distinguished by [contact], or perhaps Romeu’s Con-junto, which in also has. For the pattern of grouping to be explained in terms of features we would need on and over to share one extra feature beyond what each shares with in. Translating directly from Romeu’s structures for Spanish Ps, we have:

\[
\begin{align*}
\text{in (from en)} & : \text{[RelP Con-junto [Rel’ Rel [DP … ]]]} \\
\text{on and over (from sobre)} & : \text{[RelP Rel [DP …]]}
\end{align*}
\]

Clearly more features are needed, also because the meaning of English \textit{in} is more constrained. Over implies two points at least, the Ground and the position of the Figure, thus it bears a Dis-junto
modifier. Adding *Con-junto* to *on*, and modifiers [interiority] and [superiority] yields:

\[
\text{in} : \text{[RelP [Con-juntoP [interiority][Con-junto]] [Rel' Rel [DP ... ]]]}
\]

\[
\text{on} : \text{[RelP [Con-juntoP [superiority][Con-junto]] [Rel' Rel [DP ... ]]]}
\]

\[
\text{over} : \text{[RelP [Dis-juntoP [superiority][Con-junto]] [Rel' Rel [DP ... ]]]}
\]

In this scheme there is a one feature difference between all three pairs of items; *on* and *over* must somehow together be further distinguished from *in*, another modifier X somewhere, either in [Spec, Rel] or [Spec, Con-junto] (the first option is shown next):

\[
\text{on} : \text{[RelP X [Rel' [Con-juntoP [superiority][Con-junto]] [Rel' Rel [DP ... ]]]]}
\]

\[
\text{over} : \text{[RelP X [Rel' [Dis-juntoP [superiority][Con-junto]] [Rel' Rel [DP ... ]]]]}
\]

Once the distinctions reach this fine level of granularity it is quite difficult to argue that they are syntactically relevant. Whatever X is, it is probably semantic-internal in nature.

Additional evidence for this supposition emerges when we consider the quantitative bases for the clustering pattern generated by the algorithm. The statistic for the similarity between any two distributional vectors is their Spearman's rank correlation, which is just the Pearson correlation of the relative rankings of the values in each vector. By looking at the smallest *absolute differences in ranks* for any two vectors we obtain a sense of the dimensions that mostly determine their similarity. For example, here are the 10 closest values (by rank) for *over* and *on*:
Table 30: Differences in ranks of vector dimensions for "over" and "on"

<table>
<thead>
<tr>
<th>Difference in Rank</th>
<th>Dimension</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>&lt;START&gt; at Pos(–2)</td>
</tr>
<tr>
<td>0</td>
<td>they at Pos(–2)</td>
</tr>
<tr>
<td>0</td>
<td>come at Pos(–1)</td>
</tr>
<tr>
<td>0.5</td>
<td>be at Pos(–2)</td>
</tr>
<tr>
<td>0.5</td>
<td>now at Pos(–2)</td>
</tr>
<tr>
<td>0.5</td>
<td>little at Pos(–1)</td>
</tr>
<tr>
<td>0.5</td>
<td>tonight at Pos(+1)</td>
</tr>
<tr>
<td>0.5</td>
<td>us at Pos(+1)</td>
</tr>
<tr>
<td>0.5</td>
<td>camera at Pos(+1)</td>
</tr>
<tr>
<td>1</td>
<td>you at Pos(–2)</td>
</tr>
</tbody>
</table>

Fractional values are due to tie-breaking scores. The first entry says that on and over assigned the same rank to the co-occurrence of each with an utterance start two positions to the left of the prepositions (the difference in ranks is zero). The last entry means that whatever their absolute ranks, the rank score of you (at two positions to the left) for on was off by one relative to the rank for over.

The key statistic I will call LEX-RATE is the proportion of context items in the top N rank differences that are lexical. So in Table 30, assuming that {they, us, you, be, <START> } are functional, the degree to which the similarity between over and on is due to semantic relations is 5/10. Now, if LEX-RATE for in with on or in with over is significantly different – lower – than the value for the pair (on, over), then there is evidence that the algorithm's preference for clustering on and over before merging in is due to the relative strength of their semantic relations. If on the other hand there is little change in the statistic when comparing (in, over) and (on, over), then the algorithm’s grouping criterion is not primarily determined by semantic relations.
The analysis was carried out for the four instances of “triangular” structures in the dendogram, where two singletons (terminals) are grouped, then another is merged in. The four cases were:

[in [on over]]
[inside [under behind]]
[until [after before]]
[for [with at]]

The differences in observed LEX-RATE values were tested by computing exact 95% confidence intervals using the binomial test, where the expected distribution was the LEX-RATE value for, in each case, the pair of most similar prepositions. For example, taking the first triangle [in [on over]], the confidence intervals around the LEX-RATE values for the pairs (in, on) and (in, over) were computed in terms of the corresponding value for (on, over), since the distribution vectors for the latter pair were more similar than the others. The top 30 rank differences were considered in each case.

*Table 31: LEX-RATE scores for pairs of P vectors*

<table>
<thead>
<tr>
<th>Item 1</th>
<th>Item 2</th>
<th>LEX-RATE</th>
<th>95% CI low</th>
<th>95% CI high</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>on</td>
<td>over</td>
<td>0.5</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>in</td>
<td>over</td>
<td>11/30</td>
<td>0.199</td>
<td>0.561</td>
<td>0.2005</td>
</tr>
<tr>
<td>in</td>
<td>on</td>
<td>12/30</td>
<td>0.226</td>
<td>0.594</td>
<td>0.3616</td>
</tr>
<tr>
<td>under</td>
<td>behind</td>
<td>0.267</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>inside</td>
<td>behind</td>
<td>5/30</td>
<td>0.056</td>
<td>0.347</td>
<td>0.301</td>
</tr>
<tr>
<td>inside</td>
<td>under</td>
<td>4/30</td>
<td>0.037</td>
<td>0.307</td>
<td>0.1457</td>
</tr>
<tr>
<td>with</td>
<td>at</td>
<td>0.5</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>for</td>
<td>with</td>
<td>15/30</td>
<td>0.313</td>
<td>0.687</td>
<td>1</td>
</tr>
<tr>
<td>for</td>
<td>at</td>
<td>14/30</td>
<td>0.283</td>
<td>0.657</td>
<td>0.855</td>
</tr>
<tr>
<td>after</td>
<td>before</td>
<td>9/30</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>until</td>
<td>before</td>
<td>9/30</td>
<td>0.147</td>
<td>0.494</td>
<td>1</td>
</tr>
<tr>
<td>until</td>
<td>after</td>
<td>9/30</td>
<td>0.147</td>
<td>0.494</td>
<td>1</td>
</tr>
</tbody>
</table>
While none of the differences were significant – the (exact) confidence intervals are wide, symptomatic of the relatively short sequences considered – there is a striking contrast between the spatial Ps on the one hand, and the “temporal” and functional Ps on the other. Namely, while we observe a drop-off in the degree to which similarity is determined by lexical factors for the spatial Ps, almost significantly so for the pairs \( (\text{in, over}) \) and \( (\text{inside, under}) \), the functional (and temporal) groups show essentially no change. Though I have as yet no explanation for the behavior of the temporal items\(^{41}\), the contrast between functional and spatial Ps in terms of how context determines the relative similarity of items is consistent with the proposal that for spatial Ps, the properties that drive their distributional analysis are semantic-internal, while for functional Ps they are syntactic-semantic. The evidence presented here therefore points to the lexical nature of spatial P, in line with Den Dikken’s proposals.

**Spanish Prepositions**

A total of 5,724,501 P contexts were extracted from the Spanish Google n-gram corpus. The majority were the generic relator \( \text{de} \).

---

\(^{41}\) Kant’s teaching that abstract relations (the Categories) are made manifest in thought, i.e. language, by their transposition into temporal intuitions may be relevant here, as it suggests a certain affinity between the way language expresses time and functional meanings in general.
Table 32: Frequency of Ps in Spanish Google Books corpus

<table>
<thead>
<tr>
<th>Count</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>3,675,573</td>
<td>de</td>
</tr>
<tr>
<td>928,245</td>
<td>en</td>
</tr>
<tr>
<td>604,578</td>
<td>a</td>
</tr>
<tr>
<td>234,969</td>
<td>por</td>
</tr>
<tr>
<td>193,392</td>
<td>con</td>
</tr>
<tr>
<td>131,535</td>
<td>para</td>
</tr>
<tr>
<td>47,393</td>
<td>sobre</td>
</tr>
<tr>
<td>44,148</td>
<td>entre</td>
</tr>
<tr>
<td>22,273</td>
<td>hasta</td>
</tr>
<tr>
<td>15,462</td>
<td>desde</td>
</tr>
<tr>
<td>9,449</td>
<td>hacia</td>
</tr>
<tr>
<td>8,535</td>
<td>bajo</td>
</tr>
<tr>
<td>1,741</td>
<td>alrededor</td>
</tr>
<tr>
<td>600</td>
<td>tras</td>
</tr>
</tbody>
</table>

The feature assignments developed by Romeu are summarized in Table 33:
Since Romeu does not treat temporal Ps, e.g. *antes* “before”, *después* “after”, *durante* “during”, the distributional dendogram breaks down into two major sub-trees rather than three (though many spatial Ps, especially *tras*, also have temporal readings). I have also included the comitative *con* in the clustering.

**Table 33: Features of Spanish spatial P (compiled from Romeu, 2014)**

<table>
<thead>
<tr>
<th>P</th>
<th>Reg</th>
<th>AsPart</th>
<th>Rel</th>
<th>Con-Junto</th>
<th>Dis-Junto</th>
<th>PuntoEscalar</th>
<th>Dispersion</th>
<th>[inicial]</th>
<th>[final]</th>
<th>[faz]</th>
</tr>
</thead>
<tbody>
<tr>
<td>de</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>en</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>a</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>por</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>tras</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>sobre</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ante</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>bajo</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>tras</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>alrededor</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>hasta</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>desde</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>hacia</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>para</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Rooting the whole structure, *de* “of” occupies the position we might expect for the most abstract relator, the least specified element, expressing only relation as such, and parallel to the position of *of* for the English data. That *de* appears even higher in the dendogram than *of* points to the fact that distribution of *de* is even wider than that of its English equivalent.

On the left (red rectangle), *bajo* “under”, *tras* “after, behind” and *alrededor* “around” are the clearly lexical units (AxParts in Romeu’s terms). Somewhat surprisingly the directional *desde* “from” and *basta* “to” are grouped with the AxParts. This is not expected on Romeu’s interpretation of the structures of *desde* and *basta*, to which he assigns the PuntoEscalar modifier (making them path-oriented) but not the AxPart feature. The fact that both are multi-morphemic and contain lexical material, at least in their etymologies (e.g. *desde* derives from the latin *de ex de* “from out of”, wherein
the ex is lexical), supports their membership in the order of lexical Ps.

The right cluster (blue rectangle) contains the dative/benefactive element a “to”, and por “by”, which like the others is highly polysemic but in spatial terms has a Route-type meaning, while in other domains it yields causative and instrumental readings. Recall that according to Romeu para “to, towards, for” is composed of the meanings of por and a, expressing the path to a determinate place. Romeu explains that PPs headed by para are true if the Figure traverses any length of the path to the destination, so in Pantcheva's model the meaning is approximately that of non-transitional routes, i.e. prolative case, lexicalizing Scale over a Route phrase (Pantcheva, 2011, p. 38). The appearance of con “with” at the bottom of the left branch in the functional P cluster indicates its feature matrix is at least as complex as that of para. Finally, a and en merge as a cluster slightly closer to de. We thus observe the sequence

\[(10) \{con, para\} > por > \{a, en\} > de\]

The stark difference between this ordering and the corresponding ordering of English prepositions is the presence of en “in” immediately to the left of de. As Figure (28) shows, the English locative in clusters with the other lexical Ps, whereas the Spanish en would appear to fall into the sequence of functional Ps. Indeed Romeu assigns en a markedly reduced structure containing no AxPart features, reflecting its broader distribution in Spanish compared to English. En is just Rel with a Con-junto modifier, perhaps lexicalizing only Pantcheva's Place, i.e. Asp\text{PLACE} or LOC in Caha's f-seq.

The notion that en is functional is supported by the acquisition findings in Chapter 4, e.g. Figure 12. There we noted the virtual absence of differences in the rates of acquisition of functional Ps versus the nominally lexical en. Cross-linguistic evidence for Romance further supports this conclusion.
While a in Italian has both stative and goal meaning, in Spanish a is limited to the goal sense (and its metaphoric transpositions), while en takes over as the exponent of locative case, parallel to English at, which it resembles (Torrego, 2002). Theory and multiple sources of empirical evidence thus converge onto this interesting contrast: Spanish en is functional, English in is lexical.

So if en expresses a locative meaning, a is an exponent for Goal, por is a Route, para a non-transitional Route, and con is a comitative/directional, the ordering in (10) corresponds to the case/directional sequence:

(11) \{COM, Scale\} > INST/Route > \{DAT/Goal, LOC/Place\} > GEN/PART

Other than the gap left by the missing Source exponent (desde, grouped with the lexical Ps), the ordering of Ps induced from the Spanish data fits the Pantcheva/Caha functional sequence very well.

Let us now compare the empirical dendogram with one projected from Romeu’s feature assignments (Figure 30).
Here too the AxPart ghetto is evident on the right. The major differences are that the pair \{hasta, desde\} group with the functional units, de and sobre merge early (their structures are identical), por is clustered with en (and others) because, on Romeu's account, it is essentially stative, and hacia is clustered with the directional Ps. Unsurprisingly, given that Romeu's account yields an f-seq that is quite different from Caha's, there is no reflex here of Caha's case hierarchy. So in broad terms, it seems that while agreeing with Romeu in some particulars – the AxParts, the closeness of de, en, and that between para, por – the output of the distributional algorithm points to Caha’s sequence as the more correct generalization. At the same time, by cleanly partitioning Ps into two broad groups, the distributional analysis suggests that the distinction between lexical and functional Ps is real, and that
for Spanish it is indeed the case, per Romeu's proposals, that \( \epsilon \) is functional.

The Acquisition of Functional Ps

The strictly linear implicational hierarchy of features typical of nanosyntactic approaches greatly simplifies the acquisition task compared to feature geometries (of phi-features) of the sort we saw in Chapter 5. This is true of f-seqs of the Caha/Pantcheva variety, but not necessarily of Romeu's. Strictly speaking, Romeu's model does not yield an implicational hierarchy, since exponents are allowed to lexicalize nodes higher in the f-seq by automatically incorporating all lower nodes (including modified nodes), and because modifiers in principle can attach anywhere. All that is banned is discontinuous lexicalization, e.g. of \( \text{Rel} \) and \( \text{Reg} \) but not \( \text{AxPart} \), which is equivalent to the ban on *A-B-A syncretism. The relationships between higher and lower nodes are therefore not fixed in advance by syntactic features, because, says Romeu, the constraints on modification are semantic: as far as the syntactic computation is concerned there are no restrictions on modification.

To see how cumulative, nested f-seqs simplify learning, consider an idealized sequence:

\[
[A [B [C [D [E [F]]]]]]
\]

So that the possible feature bundles listed in the vocabulary are:

\[
[F] \\
[E \ F] \\
[D \ E \ F] \\
[C \ D \ E \ F] \\
[B \ C \ D \ E \ F] \\
[A \ B \ C \ D \ E \ F]
\]

Imagine now that distributional analysis of four functional items in the language yields an ordering such as:
The latter ordering severely constrains how a learner might select four bundles from the six made available by UG. Namely, X must always have fewer features than W but more than Y, though jumps in the sequence are possible, i.e. it is not necessarily the case that each next exponent adds a single feature. Continuing with the example, the correct mapping mostly comes down to how W is mapped. If W corresponds to \([C \ D \ E \ F]\), then only one assignment to the remaining units is possible, namely X to \([D \ E \ F]\), Y to \([E \ F]\) and Z to \([F]\). If W instead lexicalizes \([B \ C \ D \ E \ F]\), one of the units will differ from the node it contains by two features; the learner must determine which one, a step that requires extra information. There are only four possibilities, however:

\[
W = [B \ C \ D \ E \ F] \quad X = [C \ D \ E \ F] \quad Y = [D \ E \ F] \quad Z = [E \ F] \\
W = [B \ C \ D \ E \ F] \quad X = [C \ D \ E \ F] \quad Y = [D \ E \ F] \quad Z = [F] \\
W = [B \ C \ D \ E \ F] \quad X = [C \ D \ E \ F] \quad Y = [E \ F] \quad Z = [F] \\
W = [B \ C \ D \ E \ F] \quad X = [D \ E \ F] \quad Y = [E \ F] \quad Z = [F]
\]

Only one piece of information must be learned from experience: the unit that adds two features to the bundle of its predecessor must be identified. The information need escalates as W moves up the sequence, but in principle we see how the linear implicational hierarchy and distributional learning work together to keep the problem tractable.

The linearization of the dendogram feeds feature assignment. The linearization operation itself parallels Kayne's Linear Correspondence Axiom (LCA, Kayne, 1994). It proceeds by a depth-first, left-to-right traversal of a transformation of the distributional dendogram. The transformation consists in reordering the non-terminal children of nodes such that the number of terminals spanned by the right child is greater or equal to that spanned by the left. This learning method has, compared to the more involved approach in described in Chapter 5, the advantage that the structure-building phase reuses a core element of linguistic competence, the LCA – if, that is, we
accept Kayne's antisymmetry as a fact of language.

An additional complication occurs when the distributional structure (the induced dendogram) is not purely right-branching, as we see in the functional branch of the English dendogram:

\[ \text{[of [to about] [from [for [at with]]]]} \]

Here the relative linear ordering of \textit{to} and \textit{about} cannot be decided; the same is true for \textit{at} and \textit{with}.

This corresponds to ambiguities in LCA-based linearization introduced by right branching and other structural symmetries. The problem is that the tree structure of the dendogram must be linearized for unification with the cumulative f-seq, but left branches introduce options in the linearization. The assignments to the nodes \([\textit{to/about}]\) and \([\textit{with/at}]\) therefore require disambiguation; that is, the learner must determine the one distinguishing feature for each pair. This issue obtains in the Spanish data for the pairs \{\textit{con, para}\} and \{\textit{a, en}\}. Again, the learner would have to determine from experience how to resolve such ambiguities.

The Prolog program to simulate the learning process is simpler compared to the one developed in Chapter 5. The input dendogram for the seven functional Ps in Figure 28 is declared in Prolog as:

\[ \text{dendogram(en, p, [of, [[about, to], [from, [for, [at, with]]]])}. \]

The dendogram is linearized by reading the terminals left to right. The two ambiguities in linearization noted earlier are modeled as Prolog choice points; thus four possible sequences are generated:

1. \text{of about to from for at with}

2. \text{of to about from for at with}
3. of about to from for with at

4. of to about from for with at

For each sequence, place-holder features were then assigned by scanning the sequence left-to-right, each time adding one fresh variable. Thus for the first sequence the Vocabulary entries (before assignment of concrete features) were:

[of F1] [about F2 F1] [to F3 F2 F1] [from F4 ... F1] [for F5 ... F1] [at F6 ... F1] [with F7 ... F1]

The features were then made concrete via set-based unification with feature bundles as determined by the functional sequence. The Caha f-seq posits five slots, which again are:

(12) \( \text{COM} > \text{INST} > \text{DAT} > \text{LOC} > \text{GEN} \)

where each position adds a discriminating feature. Five slots for seven exponents; two more positions/features are required to define the English functional Ps. \emph{About} might map to LOC, in a manner similar to \emph{sobre} in Spanish, and indeed there are contexts in which the one translates the other, e.g. \textit{el libro sobre la nada} and \textit{the book about nothing}\textsuperscript{42}. The distinction between \textit{to} and \textit{for} might be captured by a benefactive case, in the dative “zone”:

(13) \( \text{COM} > \text{INST} > \text{BEN} > \text{DAT} > \text{LOC} > \text{GEN} \)

The remaining problem is the presence of \textit{at} as a sister of \textit{with} in the dendogram. This seems to be an artifact of the dataset. By far the most common frame for \textit{at}, accounting for 40\% of occurrences of \textit{at}, is \textit{look at the-X/that/this}. This relational use of the preposition distorts its  

---

\textsuperscript{42} Den Dikken (p.c.) has pointed out that the bimorphemic nature of \textit{about} is composed of the historically locative \textit{a-} (from \textit{at} or \textit{on}) plus other material that might be lexical. Indeed, the form seems to derive from the Old English \textit{onbutan} “on the outside of, where “outside” suggests an AxPart. It is possible then that \textit{about} ought to be classified as a lexical P.
distributional vector away from its locative meaning, for English *at*, like German *auf* is the basic exponent of LOC, denoting static aspect (what Svenonius (2006c) calls “AT-paths” (see also a suggestion in Riemsdijk and Huybregts (2007) where “at” aspect (LOC for us) is characterized in terms of the features [-TELIC] and [-INCHOATIVE]). Rather than attempt an algorithmic repair of the mis-association of *at* and *with* I will drop *at* from the analysis, by allowing it to (mistakenly) map to COM – noting however that given the observed preponderance of the *look at X* frame in child-directed speech, the child will need to resort to other means for mapping *at* to LOC.

The target f-seq is thus:

(14) COM > INST > BEN > DAT > LOC > GEN

Unification of a candidate vocabulary projected from the dendogram with the expanded f-seq returns two (when *at* is excluded) possibilities, because of the complex branch [*to about*]. Pragmatically-derived information is required to disambiguate these. Specifically, the learner must determine that *to* has the feature associated with DAT. When given this information the simulator converges onto a single set of assignments, though one error remains: as a result of the dendogram’s nesting *from* under *for*, the latter is assigned Source meaning (INST/ablative) and the former Goal (BEN, in the dative zone). The learner would therefore need to swap these assignments.

The upshot is that given a slightly expanded form of Caha’s f-seq for case, the analysis of distributional information of Chapter 5, and the two gestures of disambiguation and the post-algorithmic repair explained above, we can straightforwardly describe a procedure that learns the functional prepositions of English. The Spanish units can mapped by a broadly similar procedure:
the learner would need to determine the synsem features that distinguish \textit{con} from \textit{para} and \textit{en} from \textit{a}, and that there is no single exponent for Source (\textit{desde} being categorized with the lexical Ps).

One might ask how the learner would know to enter only the right set of functional Ps into the algorithm. As far as pronouns and other closed-class items are concerned, it seems that heuristics based on frequency and phonological characteristics might be marshaled to isolate the set of functional morphemes in the language (Wang, 2012, p. makes a similar point). The hybrid nature of the P category, however, implies such procedures are less reliable. A frequency criterion will work in Spanish, since all most common Ps (including \textit{en}) are functional; but for English frequency fails to make the right cut, given that \textit{in} and \textit{on} are extremely common. Instead, the distribution-based clustering in Figure 28 can again be put to work. Since all non-functional Ps are grouped into two large clusters, a single clue – one induced feature – is enough to identify the set of all functional Ps. Knowing that \textit{to} maps to DAT, information that is required in any case, would be sufficient.

**Conclusion**

The previous chapter explored a method by which a learner might leverage distributional information and a priori knowledge of feature geometries to acquire the Vocabulary (the mapping of phonological forms to feature bundles) of a language. While neither sufficient nor perhaps necessary, the analysis of distributional data seems to be informative with respect to the featural content of function morphemes. The demonstration took as given the discipline's relatively mature view on the nature of phi-features and showed that the hierarchical clustering of distribution vectors of subject pronouns is able to partly reconstruct the structure of similarities and differences implied by the featural content of pronouns.
The present chapter applied the same procedure to the distributions of prepositions in English and Spanish. Here there is less consensus on the features in play. In Chapter 1 I compared the accounts of spatial P developed by Koopman and Den Dikken to those of Svenonius (and others), concluding that while these accounts coincide in what we might term the aspectual layers of the PP – Pantcheva and Caha's f-sequences for directional P and case – a fundamental theoretical point distinguishes these two approaches. The first takes spatial P to be lexical and to root its own extended projection (two extended projections, for Den Dikken), while the second views all Ps as functional, occupying positions in the extended projection of the noun. A third model of the (spatial) PP was presented in Chapter 1, Romeu's analysis of spatial P in Spanish. Romeu builds on elements from Svenonius and Pantcheva but introduces the notion of modifiers in the f-seq, leading to a reduced “spine” and ultimately a fairly divergent theory of the PP.

The hierarchical structures for English and Spanish prepositions developed in this chapter group functional Ps in broad clusters separate from (by hypothesis) lexical Ps. Within the functional branches, the ordering of the clustering supports Pantcheva's and Caha's functional sequences. Romeu's model finds less support. The empirical evidence marshaled here therefore suggests:

1) The category of P is indeed split: there are lexical and functional Ps.

2) The features of functional P are the same as those of case and directional P – i.e. there is a single aspectual layer in the extended projection of P that, in a spatial context, has directional meaning, and in other dimensions describes changes over time (e.g. ownership).

3) As a minor conclusion, en in Spanish is a functional preposition.

Naturally these inferences fall apart if the distributional analysis fails to correlate with the true
feature content of Ps.

Prior investigations of distribution-based categorization (i.e. unsupervised POS tagging) have previously established that lexical co-occurrence data are informative of the syntactic categories of words. Redington et al (1998) characterized the basic characteristics and parameters of the method: the construction of collocational vectors from language corpora and their grouping by means of hierarchical cluster analysis. Hao Wang’s dissertation (2012) applied the method to function words, finding that the resulting structures, visualized as dendograms, seem to match linguists' intuition about the organization of grammatical morphemes – at least for English and German, the languages Wang investigated.

It should hardly surprise us that the statistical analysis of word patterns is able to recover linguistically significant relationships between words. After all, distribution criteria have long informed linguists' classification of words into parts of speech. A contribution of this dissertation is to demonstrate that, beyond mere categorization, the method is sensitive to the similarities and differences of the feature bundles of individual f-morphs, even for items that are traditionally grouped into a single category. That is, patterns of similarities and differences of word collocation vectors are consistent with corresponding patterns constructed from the features associated with individual f-morphs.

The observed correlation between distributional vectors and feature vectors is just that – an observation, not itself a result, and again should not surprise us, as it merely generalizes from syntactic category to all synsem features. The correlation follows if, as the cartographers propose, the feature content of grammatical morphemes maps to their syntactic position. But we can exploit the apparent correlation to help decide theoretical questions about the features (and thus syntax) of
Vocabulary items.

When applied to adpositions in English and Spanish, distributional analysis yields supporting evidence for the functional sequence of case/directional P advanced by Caha and Pantcheva. Adopting Den Dikken's analysis of spatial P, we may term this the *aspectual* subsequence of P, one each for path and place P:

(1) \[ \text{Asp}^\text{PATH} = [\text{COM/Route} – \text{INST/Source} – \text{DAT/Goal}] \]

(2) \[ \text{Asp}^\text{PLACE} = [\text{LOC} – \text{GEN/PART}/\text{Stative}] \]

Romeu's (2014) modifier-based analysis of Spanish spatial prepositions instead found less support. Additionally, the distributional analysis of Ps in English and Spanish broadly organizes Ps into two groups, recognizably distinguished by the lexical vs functional status of the prepositions. While the functional group is structured in a manner consistent with (1) and (2), the relationships of similarity and difference among the lexical Ps appear to be conditioned by encyclopedic factors inherent in their semantics and their complements. While hardly conclusive, this is suggestive evidence that P is indeed a hybrid category, and that lexical Ps therefore root their own extended projections.

One might object that the distributional analysis's partitioning of P into two large groupings does not automatically entail a lexical/functional split of the category. The spatiotemporal Ps in the “lexical” category might constitute a different set of equally functional Ps that do not belong to the aspectual sub-sequence of the PP. One might imagine a line of reasoning based on Leibniz's metaphoric relation between spatiotemporal and more abstract Ps. If the semantics of the latter sort of P are rooted in the spatiotemporal meanings of the former, then we lose a principled basis for the lexical/functional distinction. Perhaps we simply lack the very fine structure of stative
spatiotemporal P.

The trouble with such speculations is that it is rather difficult to see how meaning contrasts among spatiotemporal Ps could be modeled using syntactically relevant features. Even the simple distinction between *in* and *on* is not likely to have anything to do with relations of agreement or government; *in* and *on* are in complementary distribution. For such spatial distinctions to be modeled by means of synsem features, we would be compelled to have the syntax manipulate semantic-internal features. This violates the hypothesis of modularity and amounts to generative semantics.
Chapter 7. Conclusion

The principal contributions of this thesis are:

1. A method for relating distributional information from textual corpora to the feature content of functional elements.
2. A proposal for the location of functional Ps in the functional sequence of P.
3. The application of (1) to English and Spanish Ps, yielding evidence in support of (2).
4. Two novel methods for comparing the productivity of functional elements in child speech to that of adults.
5. Applying (4) to functional and lexical Ps, evidence that children’s use of functional Ps is at least as productive as their use of lexical Ps right from the start of multi-word speech.
6. A model of acquisition that combines the distributional procedure in (1) with UG-given functional sequences to yield predictions about what children must learn from experience in order to map f-morphs to their morpho-syntactic-semantic feature content.

Items 1-3 are concerned with linguistic theory, while 4-6 are about language acquisition. The present chapter treats these two areas in turn.

The functional sequence of P

The question of the categorial status of P – lexical, functional, or both? – matters to contemporary linguistic theory in that it decides whether or not P hosts its own functional sequence. If (at least some) Ps are lexical, then in general the PP is a complete functional sequence in its own right, along with the verb and noun phrases. Indeed, Den Dikken (2010) has proposed that all three sequences embody the same abstract structure, which, crucially for the theoretical proposal of this thesis,
incorporates an aspectual element above the relevant XP. If instead all Ps are functional, the PP must belong to the upper regions of the noun phrase. The latter possibility blocks Den Dikken's elegant speculation, since, assuming that the mass/count distinction for nominal meanings is represented aspectually, there would be two aspectual regions in the noun phrase to the single one in VP.

The discipline has yet to define a set of criteria for distinguishing lexical elements from functional elements that decides the issue for P. Some Ps will match more lexical criteria than functional, others might show the opposite behavior, all of which leads the despairing linguist to declare P a “mixed” or “hybrid” category. In this dissertation we have explored an empirical, corpus-based approach to the matter. Earlier work employing distribution-based analyses of CHILDES data has shown that hierarchical clustering applied to collocation vectors for individual terms yields sets of terms that correspond, to varying degrees, to groupings based on traditional grammatical categories. As discussed in Chapter 5, the correspondence is often poor for smaller functional categories – there is a sharp tradeoff between the degree to which like items are joined into single clusters (completeness) and the internal consistencies of the resulting clusters (accuracy): completeness is positively correlated with cluster size while accuracy is negatively so correlated.

Now, if functional elements are in essence unique – if at non-Root nodes in a syntactic derivation there is only one “solution” to the feature-driven competition for insertion – then it stands to reason that an induction procedure for learning the morphosyntax of function words in a given language must aim at associating with each f-morph the feature bundle that contrastively identifies it. A grouping criterion – declaring that some set of items are all Xs by virtue of similar distributional characteristics – is at best descriptive: it fails to equip the language user with the knowledge required
for correctly spelling out functional nodes in the tree.

In the context of evaluating the results of hierarchical clustering, i.e. the dendogram, I take as a symptom of a certain impedance between the dendogram itself and the grouping criterion – do the clusters correspond to the researcher’s intuitions about grammatical categories? – the problem of determining the appropriate “attachment height” of the dendogram at which groups (and thus categories) are declared. For example, Mintz et al (2002) and Redington et al (1998) must confront the problem, since cluster size increases with attachment height, triggering the above-noted tension between accuracy and completeness. A feature-driven perspective can instead exploit the very structure of the dendogram for purposes of acquisition. The key move is to ask whether, given a known set of feature assignments to f-morphs, a hierarchical clustering based on similarities between feature bundles yields correspondences to the dendogram induced from distributional vectors for the same f-morphs. Chapter 5 showed that such is indeed the case for subject pronouns in English and Spanish: the correlation matrix of distribution vectors appears to reproduce the correlation matrix of feature bundles determined by theory. This result should not surprise us, if, as posited by the theory, the surface distribution of f-morphs is controlled by their associated feature content.

The very fact of the correspondence between the correlation matrix of features and that of distribution vectors, shown to obtain for subject pronouns, can then be exploited to help decide the question of P – on the assumption that the correspondence holds for Ps as well. Empirical distributional analysis is thus put to work in the service of theory. Grimshaw, Svenonius and others characterize all Ps as functional, with those units that are traditionally defined as functional in the narrow sense (e.g. English of, Spanish de) merging below spatiotemporal Ps in Kase, and directional
Ps and directional Ps merging above. Based on a monotonic assignment of features over the functional sequence, we would expect a distributional dendogram in which stative and directional spatial Ps and functional Ps are interleaved, with the items bearing the greater number of features, directional Ps, merging first. In contrast, an account in which spatial P is lexical instead predicts that lexical and functional Ps would constitute separate subtrees in the dendogram, and that the functional Ps would be arranged according to some hypothesis of their morphosyntactic content, while lexical Ps would be arranged according to their encyclopedic content.

What are the features of functional P? This was the guiding question through Chapter 1. A proposal emerged that assembled from much recent work on the syntax of adpositions and case (including contributions by Den Dikken, Pantcheva, Asbury, Zwarts, Caha, Svenonius and others), to suggest that functional P, directional P and the non-Structural Cases all express the region of the PP Den Dikken has identified as aspectual – the relationship between directional and non-directional senses being mediated by Lebniz’s doctrine of metaphoric transposition. The proposal simply explodes Den Dikken’s aspect node into Pantcheva’s and Caha’s nanosyntactic analyses of directional P and oblique case, distributing the positions of the latter between the aspectual positions for the two types of lexical Ps, directional and locative, in Den Dikken’s model.

Concretely, the proposed f-seqs for PP are, repeated here from Chapter 1, are:

(24) \[ \text{CP}_{\text{PATH}} \cdot \text{Dx}_{\text{PATH}} \cdot \text{Asp}_{\text{PATH}} = [\text{COM/Route} – \text{INST/Source} – \text{DAT/Goal}] - P_{\text{dir}} – DP \]

(25) \[ \text{CP}_{\text{PLACE}} \cdot \text{Dx}_{\text{PLACE}} \cdot \text{Asp}_{\text{PLACE}} = [\text{LOC} – \text{GEN/PART/Stative}] - P_{\text{loc}} – DP \]

The distributional dendogram of English Ps (reproduced below) supports this proposal and suggests that spatiotemporal Ps are indeed lexical in nature – that P is a mixed category.
The functional Ps appear in a distinct subtree (blue rectangle), and the distance relationships implied by the order of the hierarchical clustering is consistent with the proposed functional sequences. For example, the larger feature set associated with the comitative with (per Caha's analysis) clusters with the somewhat longer dative/benefactive/goal for, while the rarefied – almost featureless – of lies at the root of the functional subtree. The hierarchical structure of spatial Ps (red rectangle) instead show little relation to the analysis of their syntax developed by e.g. Svenonius. Path heads into and through merge with the other spatial Ps last, suggesting they are either lacking in features, or they share few features with the others, including over. Neither possibility is suggested by the theory. The organization within the red rectangle is thus more likely driven by the referential meaning of the complements of P.

An important theoretical question not treated by this dissertation relates to the syntactic structure of
PP in which no lexical P appears, e.g. purely functional uses such as piece of cake. If P roots its own extended projection, then what is the structure of functional PPs (Den Dikken, p.c.)? Is there a null lexical P, or none at all – i.e. can aspectual Ps project? Parallel questions obtain in the nominal domain, with respect to pronouns and determiners, so possibly solutions developed there are portable to the adpositional domain – see e.g. Déchaine & Wiltschko (2002) who allow phi-features to project, just as aspectual Ps might.

The acquisition of functional categories

If P is in truth mixed, then the category itself constitutes a useful tool for testing a core hypothesis of empiricist, usage-based theories of language and its acquisition. Advocates of usage-based approaches hold that adult grammar is composed of more-or-less abstract constructions that, even in the mature state, retain the full history of their development and thus are never fully rule-like and abstract as conceived in the generative tradition. In denying the rationalist hypothesis of UG – that essential elements of syntax are innate – empiricism necessarily views abstraction as the hard-fought achievement of extended, gradual developmental processes. This hypothesis has the interesting consequence that, granting the mixed status of P, children's facility with prepositions ought to develop asymmetrically, the use of the more concrete lexical P becoming productive before that of the more abstract functional P. Thus closed-class units that share common phonological, syntactic and (in some cases) semantic features are predicted to develop on distinct timelines. UG-based approaches would instead predict that, because the abstract features of functional P are already available at birth, the acquisition of functional P is rapid, possibly faster than that of lexical P.

These competing hypotheses were tested by means of a pair of novel measures of the productivity of child language. Adjusted KL measures the degree to which the conditional distributions over P
complements in child-produced PPs matches that of the parents, adjusting for sample size. The second measure, from an idea by Virginia Valian, computes the rates at which children produce PPs that are not seen in the adult data. On both measures it was found that the productivity of functional P grew faster than the productivity of lexical P.

One objection to the empirical contributions of this thesis concerns the treatment of errors, specifically errors of omission. I argued that children's non-production of function words in required contexts should not be taken as symptomatic of undeveloped syntactic knowledge of the category in question. Indeed, patterns of non-production may indicate knowledge, however latent, of the syntax and semantics of functional elements in the language. The contrast between the meanings of *in* and *on* may be more salient and urgent than those between *of* and *about*, so if the child is managing limited resources in production, she may opt to express the former and delete the latter; yet in so doing she also reveals implicit knowledge of the items in play. The point is debatable and I have not provided data to support it, at least not in English, where the rate of production of locative Ps in required contexts is greater than comparable rates for functional Ps at low MLU. But just the fact that in Spanish we do not observe this asymmetry suggests that children are inherently capable of mastering functional P early in acquisition, and that pragmatic considerations specific to English might enable the lag observed in their early production.

The results of the these empirical studies favor the UG hypothesis, though we should be clear this is because of the extended gradualism that infects the constructivist account. The hypothesis that e.g. the features of P are part of the genetic endowment of our species was never tested directly.

Nothing in this dissertation blocks an account of language acquisition that, while committed to constructivist principles, greatly accelerates the pace at which children work out syntax from only
their interactions with others. What the empirical results contributed here do achieve is to significantly restrict the time period over which this development occurs – perhaps to the point of absurdity, since adult-like syntactic behavior is already detectable at MLUs near 2. It is for this reason that Tomasello wants so badly to \textit{gain time} by insisting that development is slower than we think:

\begin{quote}
What if instead of leaping at every opportunity to attribute to children abstract linguistic categories and constructions, researchers simply acknowledged the empirical fact that linguistic abstractions build up during ontogeny in a much slower and more piecemeal fashion than previously believed – and then adjusted their models accordingly? (Tomasello, 2003, p. 97)
\end{quote}

When it comes to functional prepositions, the evidence presented in this dissertation argues for precisely the opposite stance: children acquire functional prepositions fast and with little apparent effort, and commit virtually no errors. So in the absence of a concrete theory about how “social-pragmatic” learning enables children within the first two years of life to acquire the relatively subtle distribution of labor over functional P$s$, I believe the UG hypothesis remains the more convincing theory.

A second way of putting to work the correlation between distributional vectors and feature assignments is as a constraint in acquisition processes. If, per the UG hypothesis, features are held to be psychologically real, then the learning problem with respect to the functional Vocabulary of a language consists in mapping phonological forms to sets of features. Distributional analysis helps constrain the solution space of the learning problem, and allows us to delimit what must be learned by other means. The relations of similarity and difference between distribution vectors enable an efficient search through the solution space. The algorithm presented in Chapter 5 computes over a
tree-like representation of these relations to define a set of possible word-to-feature mappings that must then be disambiguated by other means. If the target feature hierarchy is simple, as in a functional sequence, the process is particularly straightforward: the learner only needs to determine, in some cases, whether a morpheme lexicalizes more than one additional feature compared to the Vocabulary item immediately lower in the hierarchy.

A weakness of this proposal is that if the learning procedure still needs other sources of information, then perhaps those other sources are sufficient for the acquisition process. This is certainly true if acquisition targets UG features. On nativist assumptions the learning procedure contributed by this thesis amounts to just a proposal for how collocation statistics might accelerate the learning process – and even then, it is not clear that the time required to accumulate distributional data is shorter than alternative modes of evidence-gathering. In any case the procedure is neither sufficient and possibly not necessary. For constructivist approaches to acquisition the utility of distributional analysis is less clear. If the cardinality and structure of the target feature space are not pre-given, then the targets must be determined by the specific language being learned. The trouble is that for English, the syntactically active phi-features generate a correlation structure that matches the corresponding distributional vectors less well than one created from abstract UG features. In order to account for the distributional facts the learner would need to hypothesize abstract, unrealized features. But if the cognitive processes responsible for acquisition focus on attention-reading and the use of linguistic expressions, it is not obvious what, other than explaining the data, would motivate the learner to induce abstract syntactic features. It seems a different, perhaps simpler, type of analysis of collocation patterns in the input would be more useful to the constructivist learner.
Speculative Postscript

I end with an admittedly wild speculation on the subject of aspect in grammar. From its mild start in Slavic linguistics, as the contrast between perfective and imperfective expressions, the notion of aspect has grown in importance to the point that, in Tortora (2008)'s and Den Dikken (2010)'s proposals, every major extended projection incorporates an aspectual subsequence. The analogy between the mass/count distinction in the nominal domain and the aspectual analysis of event structure in the verbal domain had already been established when Bach sought to “elucidate this proportion: events : processes :: things : stuff” (Bach, 1986, p. 5). His algebra of events was inspired directly by Link's (1983) algebraic analysis of plurals and count/mass nouns. Krifka (1998)'s semantics of part structures, extended to paths and events, sought a unified algebraic analysis of space and time expressions. Krifka even speculated about extensions of the model to non-spatiotemporal domains, such as ownership. Zwarts developed an explicitly aspectual semantics of directional expressions based on their analogy to event structure, thus adopting terms such as “stative”, “dynamic”, “bounded”, “telic” in the metalanguage of movement in space

At a minimum, what these aspectual algebras seem to share are a pair of basic oppositions: cumulative vs quantized and bounded vs unbounded; and a function, the “sum” operation that builds larger units. (Other elements of aspectual expressions, such as point vs extension or stative vs dynamic, may be derived.) Very abstractly, whether applied to stuff or space or time or desire or property, the two basic oppositions and the sum operation appear to constitute the building blocks of how language, in the very grammar, describes structure in general. (See also Kracht, 2008, for an account of aspects (plural) as functions that re-interpret spaces as vectors, coordinate systems, or paths.)
The speculation is this: one could draw analogies between the terms by which aspect represents structure and the manner in which generative linguistics has characterized the operations of syntax. We might then relate the bounded/unbounded distinction to Chomsky's notion of phases in derivation, on the analogy of a spelled out phase to a completed action; the cumulative/quantized opposition to notions of projection – a “cumulative” structure is a line of projection of an element; and the sum operation would map, naturally enough, to Merge. On this view, aspectual notions in UG would derive from the fundamental structure-dependence in language that Chomsky has identified as essential to the faculty. If the various manifestations of aspect in syntax were reduced to a common set of semantic universals of the computational module itself, then the footprint of universal grammar itself would diminish. The distance between empiricists and rationalists would be thereby reduced.
Appendix

Source Code

Prolog program for learning English and Spanish pronouns

re(RE) :-
  participant([], RE1),
  individuation(RE1, RE),
  RE \= [].

participant(RE1, RE) :-
  append(RE1, [participant], RE2),
  speaker(RE2, RE3),
  addressee(RE3, RE),
  not(reverse(RE, [participant | _])).
participant(RE, RE).

speaker(RE1, RE) :-
  append(RE1, [speaker], RE).
speaker(RE, RE).

addressee(RE1, RE) :-
  append(RE1, [addressee], RE).
addressee(RE, RE).

individuation(RE1, RE) :-
  append(RE1, [individuation], RE2),
  minimal(RE2, RE3),
  group(RE3, RE4),
  not(reverse(RE4, [individuation | _])),
  class(RE4, RE).
individuation(RE, RE).

minimal(RE1, RE) :-
  append(RE1, [minimal], RE).
minimal(RE, RE).

group(RE1, RE) :-
  append(RE1, [group], RE).
group(RE, RE).

class(RE1, RE) :-
  append(RE1, [class], RE2),
  animate(RE2, RE3),
  neuter(RE3, RE),
  not(reverse(RE, [class | _])).
class(RE, RE).
animate(RE1, RE) :-
    append(RE1, [animate], RE2),
    masculine(RE2, RE3),
    feminine(RE3, RE),
    not(reverse(RE, [animate | _])).
animate(RE, RE).

neuter(RE1, RE) :-
    append(RE1, [neuter], RE).
neuter(RE, RE).

masculine(RE1, RE) :-
    append(RE1, [masculine], RE).
masculine(RE, RE).

feminine(RE1, RE) :-
    append(RE1, [feminine], RE).
feminine(RE, RE).

set_unify_elements([], _).
set_unify_elements([H | T], Y) :-
    select(H, Y, Y1),
    set_unify_elements(T, Y1).

set_unify(X, Y) :-
    list_to_set(X, X),
    list_to_set(Y, Y),
    length(X, L),
    length(Y, L),
    set_unify_elements(X, Y).

set_select(H, S, S1) :-
    select(X, S, S1),
    set_unify(H, X).

sub_set([], S, S).
sub_set([H | T], S, Ret) :-
    set_select(H, S, S1),
    sub_set(T, S1, Ret).

dendogram(en, pron, [[it, [they, [she, he]]], [you, [we, i]])].
dendogram(es, pron, [[[ella, el], [ellas, ellos]], [[nosotras, vosotras],
           [[yo, tu], [nosotros, vosotros]]]])].

create_vocab(Tree, Vocab) :-
    create_vocab([], Tree, Vocab).
create_vocab(F, [X, Y], Vocab) :-
    var(A), var(B),
    dif(A, B),
    append(F, [A], FX),
    append(F, [B], FY),
    create_vocab(FX, X, X1),
    create_vocab(FY, Y, Y1),
Vocab = [X1, Y1, F].
create_vocab(F, X, Vocab) :-
   not(is_list(X)),
   append([X], F, Vocab).

get_terminals(Tree, Terminals) :-
   get_terminals(Tree, [], Terminals).
get_terminals(Tree, Terminals, Terminals1) :-
   Tree = [H | _],
   not(is_list(H)), !,
   nonvar(H),
   append(Terminals, [Tree], Terminals1).
get_terminals([X, Y, _], Terminals, Terminals1) :-
   get_terminals(X, Terminals, T1),
   get_terminals(Y, T1, Terminals1).

get_morphs(Vocab, Morphs) :-
   maplist(nth1(1), Vocab, Morphs).
sisters([X, Y], X, Y).
sisters([X, Y], Y, X).
sisters([X, _], A, B) :-
   sisters(X, A, B).
sisters([_, Y], A, B) :-
   sisters(Y, A, B).
sister_atoms(Tree, X, Y) :-
   sisters(Tree, X, Y),
   not(is_list(X)),
   not(is_list(Y)).
bundle_lengths(Features, BundleLengths) :-
   lexicon(Bundles),
   include(subset(Features), Bundles, MatchingBundles),
   maplist(length, MatchingBundles, BundleLengths).

change_vars(L, 0, L) :- !.
change_vars(L, N, L1) :-
   N > 0,
   append([X], L),
   append(H, [X], L),
   append([_Y, X], L2),
   N1 is N - 1,
   change_vars(L2, N1, L1).
change_vars(L, N, L1) :-
   N < 0,
   append(L2, [X], L),
   append([_Y, X], L),
   N1 is N + 1,
   change_vars(L2, N1, L1).
alter_bundle(Bundle, NewLength, Bundle1) :-
   length(Bundle, L),
   Diff = NewLength - L,
   change_vars(Bundle, Diff, Bundle1).
assign_features(_, []).  
assign_features(Bundle, [FeatureValue | Rest]) :-  
  include(var, Bundle, FreeVars),  
  reverse(FreeVars, RevFreeVars),  
  nth1(_, RevFreeVars, FeatureValue),  
  assign_features(Bundle, Rest).

value_feature(Vocab, Morph, FeatureValues, Vocab) :-  
  member([Morph | Bundle], Vocab),  
  include(nonvar, Bundle, ValuedFeatures),  
  union(FeatureValues, ValuedFeatures, ValuedFeatures), !.

value_feature(Vocab, Morph, FeatureValues, Vocab1) :-  
  member([Morph | Bundle], Vocab),  
  include(nonvar, Bundle, ValuedFeatures),  
  union(FeatureValues, ValuedFeatures, AllFeatures),  
  % expand bundle size?  
  bundle_lengths(AllFeatures, Lengths),  
  min_list(Lengths, MinLen),  
  length(Bundle, L),  
  ( MinLen > L ->  
    alter_bundle(Bundle, MinLen, Bundle1)  
  ; Bundle1 = Bundle  
  ),  
  select([Morph | Bundle], Vocab, Vocab2),  
  Vocab1 = [Morph | Bundle1] | Vocab2,  
  % set values  
  assign_features(Bundle1, FeatureValues).

get_bundle(Vocab, Morph, Bundle) :-  
  member([Morph | Bundle], Vocab).

equate_bundles(_, _, 0).
equate_bundles(To, From, N) :-  
  N > 0,  
  nth1(N, From, X),  
  nth1(N, To, X),  
  N1 is N - 1,  
  equate_bundles(To, From, N1).

duplicate_merge_bundle(B1, B2, Bundle) :-  
  length(B2, L2),  
  N is L2 - 1,  
  copy_term(B2, B3),  
  equate_bundles(B3, B2, N),  
  append(_, [End], B1),  
  append(Head, [__], B3),  
  append(Head, [End], Bundle).

balance_terminal_pair(Tree, Vocab, Morph, [Morph | Bundle]) :-  
  not(sister_atoms(Tree, Morph, _)), !,
get_bundle(Vocab, Morph, Bundle).
balance_terminal_pair(Tree, Vocab, Morph, [Morph | Bundle]) :-
    sister_atoms(Tree, Morph, Morph1),
    get_bundle(Vocab, Morph, B1),
    get_bundle(Vocab, Morph1, B2),
    length(B1, L1),
    length(B2, L2),
    Diff is L2 - L1,
    ( Diff > 0 ->
        duplicate_merge_bundle(B1, B2, Bundle)
    ; Bundle = B1 )
).

balance_terminals(Tree, Vocab, Vocab1) :-
    get_morphs(Vocab, Morphs),
    maplist(balance_terminal_pair(Tree, Vocab), Morphs, Vocab1).

rest([_|Rest], Rest).

learn_vocab(Features, TheoreticalBundles, Vocab) :-
    maplist(rest, Vocab, EmpiricalBundles),
    term_variables(EmpiricalBundles, Features),
    sub_set(EmpiricalBundles, TheoreticalBundles, _).

vocab_subset(Vocab, Morphs, Vocab1) :-
    findall([Morph | Bundle], (member(Morph, Morphs), member([Morph | Bundle], Vocab)), Vocab1).

init_features(en, Vocab, Vocab1) :-
    value_feature(Vocab, she, [feminine], V1),
    value_feature(V1, i, [minimal, speaker], V2),
    value_feature(V2, it, [neuter, minimal], V3),
    value_feature(V3, you, [addressee], Vocab1).

init_features(es, Vocab, Vocab1) :-
    value_feature(Vocab, nosotras, [speaker, feminine, group], V1),
    value_feature(V1, nosotros, [speaker, masculine, group], V2),
    value_feature(V2, ella, [feminine, minimal], V3),
    value_feature(V3, ellas, [feminine], V4),
    value_feature(V4, yo, [speaker, minimal], Vocab1).

learn(Language, V) :-
    findall(X, re(X), TheoreticalBundles),
    assert(lexicon(TheoreticalBundles)),
    dendogram(Language, _, T),
    create_vocab(T, V1),
    get_terminals(V1, V2),
    init_features(Language, V2, V3),
    balance_terminals(T, V3, V),
    learn_vocab(_, TheoreticalBundles, V).
Prolog program for learning English and Spanish functional Ps

pp(PP) :-
    PP1 = [],
    partitive(PP1, PP).

partitive(PP1, PP) :-
    locative([part | PP1], PP).

locative(PP1, PP) :-
    dative([loc | PP1], PP).
locative(PP, PP).

dative(PP1, PP) :-
    benefactive([dat | PP1], PP).
dative(PP, PP).

benefactive(PP1, PP) :-
    instrumental([ben | PP1], PP).
benefactive(PP, PP).

instrumental(PP1, PP) :-
    comitative([ins | PP1], PP).
instrumental(PP, PP).

comitative(PP1, [com | PP1]).
comitative(PP, PP).

set_unify_elements([], _).
set_unify_elements([H | T], Y) :-
    select(H, Y, Y1),
    set_unify_elements(T, Y1).

set_unify(X, Y) :-
    list_to_set(X, X),
    list_to_set(Y, Y),
    length(X, L),
    length(Y, L),
    set_unify_elements(X, Y).

set_select(H, S, S1) :-
    select(X, S, S1),
    set_unify(H, X).

sub_set([], S, S).
sub_set([H | T], S, Ret) :-
set_select(H, S, S1),
sub_set(T, S1, Ret).

dendogram(en1, p, [of, [[to, about], [from, [for, [with, at]]]]]).
dendogram(en2, p, [of, [[to, about], [from, [for, [with]]]]]).
dendogram(es, p, [de, [[por, [para, con]], [a, en]]]).

right_branching(X, X) :-
    not(is_list(X)).
right_branching([X], [X]).
right_branching([X, Y], RightTree) :-
    right_branching(X, RightX),
    right_branching(Y, RightY),
    flatten(X, FlatX),
    length(FlatX, LX),
    flatten(Y, FlatY),
    length(FlatY, LY),
    ( LX > LY ->
        RightTree = [RightY, RightX]
    ; RightTree = [RightX, RightY]
    ).

linearize(Tree, Line) :-
    linearize(Tree, [], Line).

% terminal
linearize(X, Linel, Line) :-
    not(is_list(X)),
    append(Linel, [X], Line).

% vacuous projection implies c-command
linearize([X], Linel, Line) :-
    append(Linel, [X], Line).

% pair of terminals, no c-command, order not decidable.
linearize([X, Y], Linel, Line) :-
    not(is_list(X)),
    not(is_list(Y)),
    !,
    ( append(Linel, [X, Y], Line); append(Linel, [Y, X], Line) )
.

% higher structure, some c-command, proceed left -> right
linearize([X, Y], Linel, Line) :-
    linearize(X, Linel, Line2),
    linearize(Y, Line2, Line).

create_vocab(Tree, Vocab) :-
    right_branching(Tree, RightTree),
    linearize(RightTree, [Morph | Morphs]),
    var(Feature),
    create_vocab(Morphs, [[Morph, Feature]], Vocab).
create_vocab([], Vocab1, Vocab) :-
    reverse(Vocab1, Vocab).
create_vocab([Morph | Morphs], Vocab1, Vocab) :-
    Vocab1 = [[ _ | Features] | _],
    var(Feature),
    create_vocab(Morphs, [[Morph | [Feature | Features]] | Vocab1], Vocab).

bundle_lengths(Features, BundleLengths) :-
    lexicon(Bundle),
    include(subset(Features), Bundles, MatchingBundles),
    maplist(length, MatchingBundles, BundleLengths).

change_vars(L, 0, L) :- !.
change_vars(L, N, L1) :-
    N > 0,
    append(H, [X], L),
    append(H, [_Y, X], L2),
    N1 is N - 1,
    change_vars(L2, N1, L1).
change_vars(L, N, L1) :-
    N < 0,
    append(L2, [X], L),
    N1 is N + 1,
    change_vars(L2, N1, L1).

alter_bundle(Bundle, NewLength, Bundle1) :-
    length(Bundle, L),
    Diff = NewLength - L,
    change_vars(Bundle, Diff, Bundle1).

assign_features(_, []).
assign_features(Bundle, [FeatureValue | Rest]) :-
    include(var, Bundle, FreeVars),
    reverse(FreeVars, RevFreeVars),
    nth1(_, RevFreeVars, FeatureValue),
    assign_features(Bundle, Rest).

value_feature(Vocab, Morph, FeatureValues, Vocab) :-
    member([Morph | Bundle], Vocab),
    include(nonvar, Bundle, ValuedFeatures),
    union(FeatureValues, ValuedFeatures, ValuedFeatures), !.
value_feature(Vocab, Morph, FeatureValues, Vocab1) :-
    member([Morph | Bundle], Vocab),
    include(nonvar, Bundle, ValuedFeatures),
    union(FeatureValues, ValuedFeatures, AllFeatures),

    % expand bundle size?
    bundle_lengths(AllFeatures, Lengths),
min_list(Lengths, MinLen),
length(Bundle, L),
( MinLen > L ->
    alter_bundle(Bundle, MinLen, Bundle1)
; Bundle1 = Bundle )

select([Morph | Bundle], Vocab, Vocab2),
Vocab1 = [[Morph | Bundle1] | Vocab2],

% set values
assign_features(Bundle1, FeatureValues).

rest([_ | Rest], Rest).

learn_vocab(Features, TheoreticalBundles, Vocab) :-
    maplist(rest, Vocab, EmpiricalBundles),
    term_variables(EmpiricalBundles, Features),
    set_unify(EmpiricalBundles, TheoreticalBundles).

vocab_subset(Vocab, Morphs, Vocab1) :-
    findall([Morph | Bundle], (member(Morph, Morphs), member([Morph | Bundle], Vocab)), Vocab1).

init_features(en1, Vocab, Vocab).
init_features(es, Vocab, Vocab).
init_features(en2, Vocab1, Vocab) :-
    value_feature(Vocab1, to, [dat], Vocab).

learn(Language, V) :-
    findall(X, pp(X), TheoreticalBundles),
    assert(lexicon(TheoreticalBundles)),

    dendogram(Language, _, T),
    create_vocab(T, V1),

    init_features(Language, V1, V),

    learn_vocab(_, TheoreticalBundles, V),
    retractall(lexicon(_)).
Bibliography


Romeu, J. (2012). *Cartography and polysemy of spatial Ps. The case of in in English*. Presented at the The


