NORTH OF PHONOLOGY

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF LINGUISTICS
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Luc Vartan Baronian

December 2005

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Paul V. Kiparsky Principal Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

William R. Leben

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Thomas A. Wasow

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Arnold M. Zwicky

Approved for the University Committee on Graduate Studies.

Abstract

The author proposes the Theory of Connected Word Constructions (TCWC), a generative theory of morphology, focusing on phonic, rather than semantic, structure. It is unique by its reductionist nature and integration of the lexicon inside the morphological constraints. The constraints, or Connected Word Constructions (CWCs), are declarative statements on the structure of fully inflected words using LexiBlocs, a tool resembling distributed disjunctions, formalized within set theory and implemented with feature-structures. The set of CWCs form a compressed lexicon that is expanded into the words of a language by a formal algorithm. The LexiBlocs encode facts of suppletion and specific/general morphological strategies, while storing words in a maximally economical way. Because fully inflected words are obtained by expanding the CWCs, simple ease-of-processing assumptions correctly predict that it is more common for a morphological strategy to refer to a stem, rather than a word. The author proposes further a five-step acquisition procedure by which speakers acquire the CWCs of their language by the simple learning of fully inflected words, as well as three Lexical Insertion Conditions that constrain the ways in which speakers may insert words within the existing CWCs, in order to inflect or derive new words. Errors in the five steps correspond to cases of category merger, folk etymology, contamination, loss of suppletion and leveling, while Lexical Insertion Conditions make much more accurate predictions than traditional four-part analogy. The latter also serve to explain paradigm gaps of defective English, French, Spanish and Russian verbs, while the five steps account for Aronoff's two Laws of the Root. A TCWC account of the complex system of Western Armenian verbal morphology is provided, and neglected phenomena such as phonesthemes and pluralia tantum are explained within TCWC. TCWC represents rarer phenomena with more complex structure: a rare type of double morphology in Armenian, as well as the more common ordering of derivational affixes within inflectional ones are explained this way. TCWC shares the pattern-seeking goals of morpheme-based theories and the moderate view of morphophonology of lexicalist theories. It accounts for an impressive number of facts with a minimal set of assumptions.

Cette thèse est dédiée à ma mère, Jeannine Bouchard,
qui aurait bien aimé avoir la chance d'aller à l'université,
et à mon père, Tatoul, qui fut toujours son propre patron.

Acknowledgements

I thank, first and foremost, my adviser, Paul Kiparsky, for having listened to my good and bad ideas with great patience during my five years at Stanford, and for an invaluable learning experience in the fields of phonology, morphology and historical linguistics, all of which find an important place in this dissertation. I would like to particularly thank four other Stanford professors from whom I have learned a lot. Will Leben, whose knowledge of tone systems and melody-based proposals left a greater impression on me than he suspects, and I only wish I had had more time to integrate similar analyses in this dissertation. Ivan Sag, who helped me shape the formal aspects of my dissertation, either by direct advice, or by inspiration drawn from reading his books and papers. Elizabeth Traugott, thanks to whom I discovered grammaticalization, and who opened my eyes to several foundational linguistic works that I may have overlooked otherwise, to my great loss. Arnold Zwicky, whose rigorous attention to often neglected linguistic facts have greatly shaped my look on language, and whose understanding of morphology I admire. Although I have not become as familiar with the breadth of topics and subfields covered by my department's other faculty members, I also thank Eve Clark and Peter Sells, for often explaining the real-world workings of this field of ours, John Rickford, for a very thorough introduction to the world of creoles, where I plan to spend more time in the near future, and Tom Wasow for help with the formal aspects of this dissertation. Edward Flemming, now at MIT, was also an inspiration in defending original ideas. I consider all the Stanford linguistics graduate students my friends, but I would especially like to highlight the friendship and influence on my thoughts of Ash Asudeh, John Beavers, Lev Blumenfeld, Luis Casillas, Brady Clark, Ashwini Deo, Cathryn Donohue, Itamar Francez, Veronica Gerassimova, Philip Hofmeister, Florian Jaegger, Andrew Koontz-Garboden, Jean-Philippe Marcotte, Melanie Owens, Rob Podesva, Colleen Richey, Mary Rose, Devyani Sharma, Julie Sweetland, Ida Toivonen, Judith Tonhauser and Andrew Wong. Warm thanks also go to the wonderful departmental staff, whose help in urgent situations and daily attention to little details made my graduate career so enjoyable; in particular, Melanie

Levin, Socorro “Coco” Relova and Gina Wein. Responsibility for errors or inexactitudes is my own.

These memorable years would of course not have been possible without financial support. I thus acknowledge a Stanford University scholarship throughout my five years in the Linguistics Department, as well as a scholarship during the middle three years, granted by the Canadian Social Sciences and Humanities Research Council (SSHRC). Other supporters are the Linguistic Society of America, for a scholarship giving me the opportunity to attend their 2001 Summer Institute at the University of California Santa Barbara, where I met several professors working in many different frameworks, that broadened my horizons even more; Harvard University and its Linguistics Department, who hosted me during the Fall 2001 semester, through the Graduate Student Exchange Program, and where I deepened my understanding of Armenian, especially thanks to James Russell and Bert Vaux, who were both very generous with their time for me; a Stanford Graduate Research Opportunity grant used during the Summer of 2003 for fieldwork on French and Creole in Louisiana. During this latter experience, Tom Klingler (Tulane) and Bernie Ricard welcomed me with an unsolicited hospitality that will always be warm to my heart, and a help in the field that allowed me to double the work I would have otherwise been able to accomplish. Some of the data accumulated during that period made it in this dissertation, while other data made it in separate publications and conferences. I am also grateful to the University of New Brunswick Saint John, its Department of Humanities and Languages, and, in particular, Virginia Hill, who allowed me to gain a very valuable teaching experience during the academic year 2004-2005. Finally, I thank Beth Levin for tuition support during my filing semester and a pleasant introduction to lexical semantics.

In turn, I would never have been able to obtain any of these scholarships, if not for my first linguistics professors who turned the math major that I was into a linguistics masters student at the Université de Montréal: Richard Kittredge, Igor Mel’cuk, Jean-Yves Morin, Yves Charles Morin, John Reighard, Rajendra Singh and Daniel Valois, as well as Jacques Samson, who was my first linguistics instructor at Collège de Maisonneuve. And likewise, I would never have had any academic success, had it not been for the constant support of my friends Mark Sumbulian, Olivier Hébert, Jean-François Lépine, Geneviève Labine and Marc Fredette, my aunts Arménouhie Baronian and Gisèle Bouchard, who both inspired me to study, my sister Julie, my parents Jeannine and Tatoul, as well as my beloved grand-parents. I reserve my final thanks for Stephanie Michelet, without whose love, these last few years would have been incredibly lonely.

Luc Baronian

Vieux-Hull QC

Contents

Chapter 1

Introduction

1.1 Background

The Theory of Connected Word Constructions (TCWC) presented here is a theory of morphology. It is a generative theory of morphology in the sense that it aims to generate exclusively the words that are accepted by speakers as existent or possible words in their language. It also tries to capture generalizations that linguists have made about the morphological systems of the world's languages.

In some respects, the theory is a middle ground between morpheme-based theories, such as Distributed Morphology (Halle & Marantz 1993) or Meaning-Text Theory (Mel'cuk 1993-2001), and stem-based theories such as Paradigm Function Morphology (Stump 2001). Like the former, it tries to capture all morphological patterns observed, even the unproductive ones, and like the latter it is strongly lexicalist. The breed of lexicalism at the very foundations of TCWC is however very different from what is generally assumed: instead of morphology being realized in the lexicon, it is the lexicon that is realized in the morphology. In this sense, it is closest to Lexical Morphology (Kiparsky 1982b), where every lexical entry was replaced by a rule mapping the underlying form of a word onto itself. The means employed are inspired by those of Seamless Morphology (Ford et al. 1997, Singh & Starosta 2003) and classical word-and-paradigm theories, framed in a novel and unique way that avoids the pitfalls of the classical theories..

Linguists with a strong empirical grounding such as Chafe (1997) or theorists sympathetic to their empirically-grounded vision of things such as Blevins (2003) advance the argument that unproductive morphological strategies are only part of grammar in a linguistic sense, but in no real cognitive sense for the native speaker. Advocates of such a position argue that while the speaker may be aware

of the patterns, the said patterns are in no way part of his/her grammar, or simply that these words are just learned as such and remain unanalyzed. The answers to the question of how the "productive" strategies are accounted for (the ones with which speakers seem able to inflect or coin new words) range from proposals of vague analogical models to more elaborate theories such as Paradigm Function Morphology, where several look-alike but unidentical stems are recognized for each lexeme, instead of a unified root.

There are at least three problems with this way of approaching morphology. First, there is equating productivity with grammaticality. Though the strategy that relates *mouse* to *mice* or *louse* to *lice* in English is entirely unproductive, what does it mean to say that this strategy is not "part of the grammar"? A possible interpretation is that the four words are simply learned and speakers see no connection whatsoever between these words. This would be too strong a position and would require demonstration: how could it be that speakers see no connection between the singular and the plural of a same lexical meaning, whose forms share two segments out of three? A milder interpretation, where people *see* a connection, but don't encode it in their grammar, seems more plausible. For example, Blevins (2003:756), talking about Germanic ablaut, claims that "these patterns have not been encapsulated in a separate system of rules, templates or schemas", though, citing psycholinguistic evidence, he also believes that "it would be implausible to claim that speakers are unaware of these patterns." But then, what does it mean to be aware of a pattern that is not part of grammar? How and where is this awareness encoded? This weaker interpretation inevitably leads to the core/periphery debate: of all the knowledge speakers have about their language, which is internal and which is external to grammar? Typically, linguists who equate productivity with grammaticality do not accept the core/periphery distinction, which leads us to a nice paradox: if the core/periphery distinction is not justified, how could there be linguistic patterns handled by the grammar, and others handled by mechanisms outside the grammar?

A second problem is that unproductive strategies sometimes win over more productive ones. The Oxford English Dictionary attests the (admittedly less frequent) *mongeese* as a plural of *mongoose* instead of *mongooses*, generated with the more productive strategy. Many speakers of North American English have *I dove* instead of *I dived*, etc. How is this to be explained if the pattern is not "part of their grammar" or simply not recognized by speakers? How do we draw the line between the productive-enough-to-be-grammatical and the other strategies?

The answer cannot lie in the nature of the pattern, because one language's unproductive strategies are another's productive ones. Vowel changes are perhaps not productive in English, but this is

certainly not the case in Semitic languages like Arabic and Hebrew. Hence, this leads to a third problem: different mechanisms (e.g. rules vs. analogy) must be used by different languages to account for morphological patterns that are essentially the same. Therefore a theory of morphololgy such as TCWC that does not limit its object to productive strategies rests on simpler assumptions, while being more challenging.

At the other end of the spectrum, we find the generative morpheme-based theories such as Distributed Morphology. These theories have two main characteristics. In their strongest incarnations, they reject entirely the lexicalist hypothesis (which states that morphology is done in the lexicon), though different degrees of its acceptance may be found in practice. Their second characteristic is their inclusion of all morphophonological alternations in phonology. This divide and conquer strategy has as consequence the trivialization of morphology: half of it is handled by syntactic movement and half of it by phonological rules. What remains for morphology in, for example, Distributed Morphology is only to "spell-out" the clean features manipulated by syntax. The two dimensions of this family of theories each encounter their problems.

As mentioned above, the version of the lexicalist hypothesis assumed by TCWC is different than in most frameworks, in that it is the lexicon that is assumed to be part of morphology, rather than the opposite. Nevertheless, rejection of the lexicalist hypothesis, in the sense that morphology is handled with syntax, rather than with the lexicon, completely or even partially, reduces the intuitive difference felt by speakers between words and sentences, as argued by Mithun (2001). Blevins (2003) also points out that if features are equated with morphemes, this often leads to a mismatch where a given form does not correspond to a constant substantive feature, for example in the case of English alternative stems (the problem of morphomic stems—Aronoff 1994). Another problem is that if both word and morpheme order are the result of syntactic movement, then why is it that word order, but never morpheme order, is an issue in domains such as language acquisition and aphasia? Thus, the main argument for lexicalism is not that morphology is irregular, while syntax is regular; there are obviously such things as irregular syntax and regular morphology, as Marantz (1998) argues. However, since morphology is more closely tied to word-learning, one expects semantic and formal idiosyncrasies at more levels in morphology than in syntax.

As for the inclusion of all of morphophonological alternations into phonology, Chafe (1997) points out that, in many cases, this amounts to postulating an underlying form (or input) that corresponds to the reconstructed historical form of a previous stage of the language; in other words, it amounts to reconstructing the language's history in its synchronic system. This is problematic from the point

of view of the learner for whom it must be postulated that sophisticated techniques similar to those employed by the historical linguist are available (perhaps innately). This is a big assumption for a theory to rest on. Of course, not all of morphophonology amounts to reconstructing a languages diachrony in its synchrony, and in some cases, this may be justified, but to do so systematically leads to a proliferation of opacity cases in the phonology.

In order to analyze opacity cases in phonology, researchers in Optimality Theory (Prince & Smolensky 1993, McCarthy & Prince 1993), for example, have needed to introduce additional mechanisms (thus complicating the theory), such as Output-Output Correspondence constraints (Benua 1997), Sympathy Theory (McCarthy 1999) or levels (Kiparsky 2000). Although not all opacity cases are morphophonological, and not all morphophonological cases are opaque, more moderate theories of morphology, such as Paradigm Function Morphology, Seamless Morphology and TCWC, that provide analyses for at least part of morphonology lessen the burden on the shoulders of phonology. These latter theories, as well as Kiparsky's (2000) non-parallel OT, are also better suited to to reflect the different cognitive status of phonological and at least some morphophonological alternations revealed by language games, language acquisition (see Stampe 1987) and aphasia.

TCWC shares the more ambitious aims of pattern-seeking morpheme-based theories while retaining the advantages of a strictly lexicalist approach to morphology that also includes morphophonology. With respect to its interaction with phonology, it is a "horizontal" model as illustrated in (1), and as opposed to vertical models such as (2). It is important to recognize that many models of morphology are mixed, in that they incorporate both a horizontal and a vertical dimension. Paradigm Function Morphology is such a model, where lexeme-to-lexeme (derivational) morphology is assumed to be on the horizontal dimension, while lexeme-to-word (inflectional) morphology is treated vertically. The original parallel version of Optimality Theory was a vertical model, though it did not have intermediate representations. Output-Output constraints are an indirect way of incorporating a horizontal dimension in OT, since they allow one word to refer to the form of another word. Distributed Morphology would be an example of a truly vertical model of morphology. Traditional word-and-paradigm models are good examples of horizontal models. Blevins (2003) proposes a word-and-paradigm interpretation of Paradigm Function Morphology, but this interpretation still goes through a stage of postulating a higher up stem that connects the lower level words. The best examples of modern horizontal theories are found in some connectionist models or in Seamless Morphology.

(1) **Horizontal models of morphology**

	morphology	
Word	↔	Word
↓ *phonology*		*phonology* ↓
Phonetic Form		Phonetic Form

(2) **Vertical models of morphology**

Underlying Form		
↓ *morphology*		
Intermediate Form(s)	↖	cyclic morphophonology
↓ *phonology*	↗	
Phonetic Form		

1.2 Goals and assumptions

The goal of TCWC is to account for as many morphological patterns as possible, with as few assumptions as possible. These assumptions must be the strongest most falsifiable choices among the parameters that distinguish one theory of morphology from another. Empirical data can and will be used to warrant additional mechanisms or assumptions to the theory, or to weaken the assumptions made initially.

Once one admits that a theory of grammar must include a theory of morphology that is at least relatively independent of phonology and syntax, then one is confronted with several dimensions defining possible morphological theories. I can think of two conceptual dimensions and two formal ones. Conceptually, one must ask what distinctions are to be made between types of phenomena, e.g. morphophonology/morphology/morphosyntax, inflection/derivation, affixation/compounding/incorporation. Once distinctions are made, one must decide which phenomena the theory should account for: morphophonological alternations, compounding, incorporation, truncation, etc. Formally, some tools must be devised or borrowed from other theories and the unit(s) on which the tools operate must be chosen (theories can be morpheme-based, word-based, lexeme-based, stem-based).

(3) **Parameters distinguishing different theories of morphology**

a. **Conceptual dimensions**

1) Distinctions between types of phenomena (derivation/inflection/word-level/stem-level...)

2) The domain of morphology (morphophonology, compounding, truncation...)

b. **Formal dimensions**

3) Formal tools (word-formation rules, transformational rules, declarative constraints, ranked constraints, stochastic rules, finite-state grammar...)

4) The unit(s) (word, morpheme, lexeme, stem)

Let us start by positioning TCWC on the scales of these four dimensions. The basic philosophy here is to defend the strongest most falsifiable option on each dimension.[1] Theoretical decisions of this nature made for other frameworks do not necessarily carry over in TCWC. Therefore, at this point, I ask the reader to put aside any objections temporarily and see how far the strongest most falsifiable positions can take us. The idea is that, since TCWC is a new theory, we shall first make a clean slate, return to the default assumptions, and modify these assumptions as we go along only when needed. I firmly believe that every new theory should start out this way.

The program is a reductionist one. In order to have a simpler but falsifiable theory, we will start by assuming no distinctions between types of phenomena. For instance, we will not assume a distinction between derivational and inflectional morphology. The opposite assumption is held by Paradigm Function Morphology for example, which in its very design has a distinction between lexemes and words, such that the rules relating lexemes are derivational, while the ones relating lexemes to words are inflectional. If we started by assuming that distinction, it would be harder to prove that TCWC was wrong in doing so, since what one would have to do would be to show that it could have done without it, so one would need to redo all the work without the distinction. By not assuming the distinction from the beginning, one only needs to show that there is some data TCWC cannot account for without this distinction. I should also point out that though I may use the words inflection/derivation, affixation/compounding, etc. in the text of the dissertation, it should be clear that they have no theoretical value in the model. Until further notice, they are all morphological phenomena for which I use traditional names for reference purposes only. As we will see, this assumption will hold relatively well in TCWC.

The theory tries to account for as many observable phenomena as possible with as few assumptions as possible. Hence, on the question of the domain of application of this theory of morphology, I adopt the strongest possible claim: every phenomena for which the theory can provide a morphological analysis should be included in morphology, be they morphophonological alternations, compounding, incorporation, morphological tone, truncation, etc. I thus adopt the strongest version

[1] I realize that other linguists have taken different positions in the past, but I feel that the theory I present is different enough from other frameworks to at least try to do things differently.

imaginable of the Lexicalist Hypothesis. While it would be impossible in this dissertation to try and account for all the phenomena loosely related to morphology, I believe that by the end of it, the reader will be impressed by the breadth of challenges that TCWC can handle. We will however encounter some fuzzy areas in the domain of morphophonology (Chapter 6), where we will see some alternations that can be treated in TCWC, but for which it seems to be undesirable to do so. Many theories exclude a priori some phenomena from their reach. For example, Seamless Morphology excludes truncation and phonesthemes from its conception of morphology (Neuvel & Singh 2002) and even perhaps from linguistics, although there exists (linguistic) prosodic constraints on truncation (Anderson 1975) and a cognitive reality to phonesthemes (Bergen 2004).

The main formal tool I adopt is a type of declarative constraint on words called the Connected Word Construction (CWC). In this theory of morphology, every description must be a description of a word. The tool is a simple one, which, I believe, correctly characterizes morphology. No roots, stems, affixes or lexemes are defined *independently* from words. Only fully inflected words have an independent status. The constraints use a special kind of distributed disjunction that I call Lexical Building Blocks (LexiBlock). The purpose of these disjunctions is one of economy: words share segments using LexiBlocks as much as possible. The set of all the constraints create a network through which words are stored in an economical fashion and generated in speech. These constraints on words are formed following an acquisition procedure, which I outline in a separate section on acquisition. The simplest assumption is then that LexiBlocks are sufficient analyze all of morphology. In Chapter 3 on acquisition and analogy, it will turn out that we need a few more assets in our toolbox (namely some constraints on the insertion of new words inside existing LexiBlocks), but in turn, these extra constraints will take us far in limiting the possible creations of new words, as well as historical morphological changes, and will provide an analysis for the neglected phenomenon of paradigm gaps (Chapter 5).

Words are almost universally recognized as a real linguistic object. In order to keep the formal apparatus to a minimum then, TCWC avoids giving morphemes or lexemes any primitive status. Of course, since we are trying to describe how words are formed, reference to parts of words such as roots, stems and affixes will be useful at some point, and I will shamelessly use this terminology to name parts of words. However, I do not take these names to give morphemes any status other than "parts of words", just like referring to cold is not a proof that physicists should consider it anything else than the absence of heat. I thus make the assumption that language learners learn words. They do not directly learn sub-units of the word such as roots, stems, suffixes, nor do they learn directly

higher organizations such as lexemes. These are all taken to be derived concepts that may or may not be useful in the system of the language once the words are encoded in the grammar. (Of course speakers may also learn some idiomatic phrases or constructions). The morphological component of Meaning-Text Theory is probably at the other end of the spectrum as far as this assumption is concerned; most if not all morphological objects that have ever been proposed can be found in this impressive theory.[2]

TCWC thus admits that any one of the assumptions on these four dimensions may be complicated to a certain degree. One way to falsify TCWC would then be to provide enough empirical data showing that each one of these theoretical choices is completely ill-founded. Another way would be to justify a complication of one of these dimensions that would lead to a contradiction elsewhere. For example, nothing in the formal tool allows it to "count" morphemes, and I don't see how that could be integrated without completely starting over. Imagine that it could be proven that in some language, an affix is systematically inserted three morphemes from the right of the word, and that the morphemes counted do not always represent the same features. Since morphemes are derived from words in TCWC, there is no way to represent them formally as a countable entity, unlike phonic material such as syllables.[3] Another example would be to show that the acquisition steps I propose in Chapter 3 do not correspond at all to the real-time acquisition of language, and that there logically could not exist an acquisition algorithm leading to the CWCs I propose that would correspond to those real-time acquisition facts.

1.3 Content and structure of the dissertation

The chapter following this introduction is a formalization of CWCs and LexiBlocks. It starts by an intuitive presentation of the theory. While this first presentation is informal, it is probably sufficient for most readers to understand how to read the formalism, convince them of its rigor and obtain a window on the multitude of advantages that TCWC holds in morphology, including the representation of phonesthemes and accounting for intrinsically plural nouns. The rest of Chapter 2, I'm afraid, is much dryer. It consists of a double formalization of LexiBlocks. First, LexiBlocks are formalized using familiar notions from set theory, including lists. This is necessary from the

[2]While TCWC defines a theory and lets empirical data justify complications, Meaning-Text Theory directly reflects the empirical phenomena with different theoretical statuses (prefixes, suffixes, circumfixes, etc.). Perhaps this due to the strong cross-linguistic grounding of Meaning-Text Theory, being built mainly from the ground up, but the downside is that it is harder to falsify it.

[3]Of course, syllables may be derived as well, but crucially, they are derived by another module, phonology, so that as far as morphology is concerned, they can be considered primitives.

point of view of mathematically-trained readers, in order to understand LexiBlocks and operations on LexiBlocks; while the notions involved do not go beyond what an undergraduate mathematics students can understand, this section is nevertheless necessary to rigorously show that the formalism is well-defined. Finally, a no less dry implementation using feature-structures is proposed. While feature-structures are not as common as set theory notions in general, they are the main tool used by computational linguists to implement syntactic theories such as Head-driven Phrase-Structure Grammar (Pollard & Sag 1994) or Lexical Functional Grammar (Kaplan & Bresnan 1982). Hence, this second formalization allows interested formal and computational linguists to judge TCWC on the standards established by other highly formal theories, and judge its compatibility with them. I would like to stress however that the two most formal sections are in no way necessary for the average reader to understand the theory or appreciate the next chapters; the intuitive presentation at the beginning of Chapter 2 suffices.

In Chapter 3, we are still in the process of defining the theory. First, I provide a basic background of the debates surrounding the question of the historical mechanism of analogy, assumed in the Neogrammarian tradition to account for most of what we consider today to be historical morphology, and the modern cognitive mechanism of analogy that is proposed by psychologically-grounded theories of linguistics (e.g. Ramscar 2002). Then, an acquisition procedure for the Connected Word Constructions (CWCs) is explained at length. The procedure consists of five steps inferring the CWCs used by the theory from the simple learning of fully inflected words, and three conditions constraining how speakers may insert new words in the existing CWCs. The force that drives the five acquisition steps is maximal economy of representation, and without them, we could posit any CWC we wanted, making the theory more difficult to falsify. The justification for the insertion conditions is to disallow ungrammatical word-formations or inflections, such as plural **ciche* of *couch* (on the model of *mouse/mice*), plural **deers*, or plural **preef* of *proof* (on the model of *goose/geese*). The pay off of the acquisition procedure and insertion conditions is then illustrated: it turns out that each step of the procedure coincides with a type of morphological change (category mergers, folk etymology, contamination, loss of suppletion and some types of leveling), while the insertion conditions highly constrain the TCWC equivalent of four-part analogy. As we will see in the following chapters, the explanatory power of the acquisition procedure does not end here.

Chapter 4, an account of the verbal morphological system of Western Armenian, was written with the preoccupation of avoiding ending up with a theory that accounts for isolated phenomena in a sporadic choice of languages, without being able to handle a system as a whole. The English

and French languages used to exemplify the principles of the two preceding chapters are also not known to be the morphologically richest languages. Hence, Armenian, a language with many more productive affixes, cases of suppletion, and other phenomena serves our purpose well. TCWC turns out to provide a rather economical account of the morphology of this language, the complexity of which culminates in the aorist tense. The acquisition procedure of Chapter 3 also proves to be well suited for this new language.

In Chapter 5, we cover cases of paradigm gaps in English, French, Spanish and Russian. As we will see, very few linguists have proposed serious accounts of this neglected phenomenon. In fact, the two main morphological frameworks, Distributed Morphology and Paradigm Function Morphology, completely fail to bring any insight on the phenomenon. I will present two different accounts in two related languages[4] that each help us gain some insight on the nature of gaps, but that each run into their set of problems. As we will see, TCWC avoids these problems and provides an account of the phenomenon with the very same insertion conditions from Chapter 3. The unifying characteristic of TCWC is thus once again illustrated.

Because this dissertation is mainly concerned with the phonic or formal side of morphology, as opposed to the syntax/semantics or content side, Chapter 6 is concerned with delimiting the border between phonology and morphology within TCWC. While several cases of morphophonological alternations unambiguously fall in phonology or morphology under TCWC assumptions, it turns out that many cases can be treated by either module of grammar, depending on the model of phonology used along with TCWC and evidence specific to each case. In the ambiguous cases, we need to assume that TCWC first accounts for the phenomena, and then, later, if there is enough evidence available for speakers to treat the alternations in phonology, yielding a simpler grammar, etc., then the CWCs are restructured to accommodate the new analysis. We can't exclude however that some alternations are treated both in phonology and morphology, representing a transitional stage of grammaticalization. The examples in this chapter are drawn from English, French, Armenian syllable-based phenomena and Margi tone-based phenomena.

In Chapter 7, we review several generalizations made by other linguists about morphology: Kiparsky's observation about stems in compounding, Ford & Singh's observation about the diachronic stability of morphophonology, Aronoff's two laws of the root, the Adjacency Condition, Greenberg's Universal 28 about derivation preceding inflection and Carstairs-McCarthy's Peripherality Constraint about the inside out nature of allomorphy. Considerations of simplicity/complexity within TCWC are sufficient to account for these generalizations, except Carstairs-McCarthy's. As

[4]Morin (1987) for French and Albright (2003) for Spanish.

with other generalizations made by Carstairs-McCarthy (1987), TCWC needs to adopt them as independent principles. Fundamentally though, TCWC is not incompatible with those principles.

Because different readers will read this dissertation with different interests, I will end this introduction with its "road map", indicating the order in which the reader may chose to read or skip chapters, according to his/her interest.

(4)

<table>
<tr><th colspan="2">The road map of this dissertation</th></tr>
<tr><td colspan="2">1: Introduction</td></tr>
<tr><td colspan="2">2.1: Intuitive presentation of the formalism</td></tr>
<tr><td colspan="2">2.2-2.4: Formalization
(optional, intended for formalists)</td></tr>
<tr><td colspan="2">3: Acquisition procedure and morphological change</td></tr>
<tr><td>Morphologists
4: Armenian
5: Paradigm gaps
7: Generalizations
6: Phonology</td><td>Phonologists
6: Phonology
5: Paradigm gaps
7: Generalizations
4: Armenian</td></tr>
<tr><td colspan="2">8: Conclusion</td></tr>
</table>

As mentioned earlier, the strict formalization of the LexiBlock tool is optional, and will probably not interest every reader. After reading the first three chapters, morphologists and phonologists, may want to start by reading chapters that are more relevant to their interest, hence I propose two reading profiles. Of course, readers who are equally interested in everything TCWC has to say will want to read the chapters in the normal order.

Chapter 2

Connected Word Constructions: A Formalization

This chapter has three parts: an introduction to the theory, a set-theoretic formalization and a feature-structure implementation. In §2.1, I introduce from an intuitive standpoint the formal tool called LexiBlock (for Lexical Building Block), which serves as the building block of the Theory of Connected Word Constructions (TCWC), and I illustrate some of its basic advantages for morphology. In §2.2-2.4, TCWC is formalized. After introducing some basic definitions in §2.2, §2.3 offers a set-theoretic formalization of the theory, while §2.4 offers a formal implementation using feature structures. The intent of the set theory formalization is to allow the mathematically oriented reader to conceptualize the theory with familiar notions such as sets and lists. The advantage of the feature-structure implementation is that feature-structures are already used in some linguistic frameworks (HPSG[1] among others), so it allows us to better understand the relationship between TCWC and this family of frameworks.

The first section of this chapter is essential and sufficient for the understanding of the rest of the dissertation. Readers may refer back to the other sections only as needed, especially those readers who are not formally oriented.

Before beginning, I want to stress out that *no example from this chapter should be taken as a claim on how the lexicon of a particular language, or the "architecture of language in general", is structured*. The points are essentially didactic, and are simply intended to help the reader understand

[1]Head-driven Phrase-Structure Grammar. See Pollard & Sag (1994).

the formalism. Also, in TCWC, languages obtain their Connected Word Constructions (CWCs) and the LexiBlocks that the CWCs are made of, by following an acquisition procedure that will be outlined in the next chapter.

2.1 The formal tool: an intuitive introduction

2.1.1 Properties of LexiBlocks

In this section, I introduce the formal tool I call LexiBlock and illustrate its linguistic use. The LexiBlock organizes the lexicon in a linguistically significant and economical way. We can think of a LexiBlock as an ordered set (a list) with three tags. Having several tags allows us, among other things, to reorder the linguistic forms or meanings from one linguistic description to another, as is often necessary in linguistics. To start with a *non linguistic* example, we know on one hand that cats and lions go together because they are both felines, while dogs and wolves are both canines, but on the other hand, we also know that cats and dogs are domestic animals, while lions and wolves are wild animals. Hence:

(5)

ANIMALS	X	LISTBYSPECIES

FELINES	CANINES
cats	dogs
lions	wolves
etc.	*etc.*

ANIMALS	Y	LISTBYHUMANCONTACTTYPE

DOMESTIC	WILD
cats	lions
dogs	wolves
etc.	*etc.*

There are two LexiBlocks in (5) above. They each consist of a grey label rectangle containing three tags and a wider rectangular block each containing two LexiBlocks. (To simplify, I only labeled the embedded LexiBlocks with one tag). Let us leave aside for now the tags [X] and [Y] whose purpose will become clear shortly. The two LexiBlocks in (5) both contain animals, hence their common tag ANIMALS. However, the animal names they contain are ordered and organized differently: one lists them by their belonging to the canine or feline groups, while the other lists them by their being domestic or wild animals. This illustrates the usefulness of the separate tags on each side of the tags labeled [X] and [Y]: the tag preceding [X] or [Y] is a name given to the unordered set of objects; the tag following [X] or [Y] is a name given to the particular arrangement of objects. However, because the elements inside the LexiBlocks are overtly displayed, it will typically be clear what the arrangement is, so we will often omit the last tag.

This system is very useful in linguistics, especially in morphology. Take for example the different grouping requirements of morphonology and morphosyntax. While the verbs *drink* and *sit* both form their past tense by vowel change, *drink* is a transitive verb like *keep* and *sit* is intransitive like *sleep*. Hence:

(6)

VERBS X	
VOWEL CHANGING	REGULAR
drink	sleep
sit	keep
etc.	*etc.*

VERBS Y	
TRANSITIVE	INTRANSITIVE
drink	sit
keep	sleep
etc.	*etc.*

I have omitted the tags following [X] and [Y], because it is pretty clear what the verb grouping difference between the two LexiBlocks is. The reason why we need to have yet another tag labeled [X] or [Y] is to establish connections *within* LexiBlocks. The tags ANIMAL and VERBS, respectively in (5) and (6), allow us to establish connections *between* LexiBlocks, by stating that, in each case, the two LexiBlocks in (5) or (6) contain the same elements (though they are arranged differently). We also establish a connection between LexiBlocks, for example, when two inflections use the same stems.

We establish a connection *within* a LexiBlock when, for example, we want to associate the form of a word with its meaning. For example, in the following LexiBlock, the forms and meanings of the words *bird*, *cat* and *dog* are stored in separate embedded LexiBlocks, but the form and meaning LexiBlocks are connected by the common tag [1]:

(7)

WORDS 0	
FORM 1	MEANING 1
bɜ˞d	'bird'
kæt	'cat'
dɔg	'dog'

Thus, the tag [1] gives us a handle to manipulate together the elements present is two or more different LexiBlocks. As we will see throughout the dissertation, the usefulness of this tag is at least double: we often need to manipulate form and meaning at the same time (e.g. suffixation associated with a meaning change); different classes of words may require to be associated with different allomorphs or entirely different suffixes.

Let us recapitulate what we have learned so far about LexiBlocks. In the example below, the set name (ANIMAL NAMES) is used to refer to animal names, while the list name (LIST-AN-1) refers to the particular ordering they have in this LexiBlock, and an arbitrary number 1 is assigned:

(8)

ANIMAL NAMES	1	LIST-AN-1
cat		
dog		
...		
bird		

We can formulate this more generally as so:

(9)

SETX	74	LISTX
object 1		
object 2		
...		
object n		

SetX = {object 1, object 2, ..., object n}
(unordered set)

ListX = <object 1, object 2, ..., object n>
(list, or ordered set)

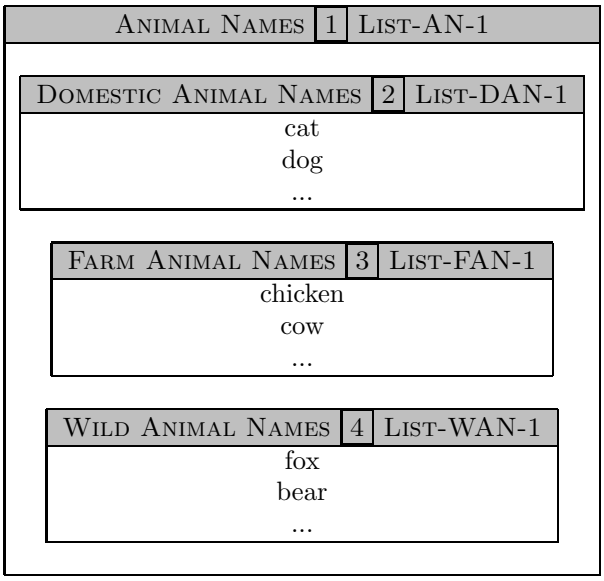
The LexiBlock above lists n objects, numbered from 1 to n. SETX in the label rectangle is the name of the unordered set, while LISTX is the name of the ordered list of objects in the box located below the gray label. If we think in general—i.e., not (necessarily) linguistic—terms, the objects can be anything conceivable (phonological forms, semantic information, apples, etc.), as long as the sets are well-formed according to the principles of set theory. We can embed LexiBlocks one into the other, in order to make subgroupings of the set members:

(10)

SETX	22	LISTX
SUBSETY 75 LISTY		
object 1		
object 2		
object 3		
...		
object n		

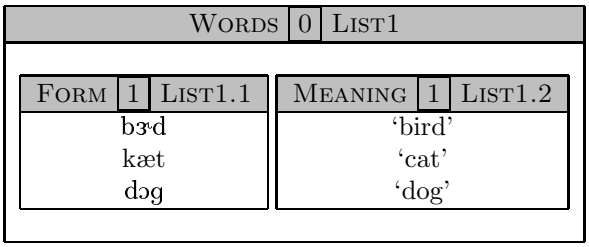
In this latter example, SETX is the set of objects 1 through n, but—and this is the crucial part—LISTX has only n - 1 elements, the LexiBlock SubsetY 75 ListY being the first, "object 3" being the second, etc. To keep with our linguistic example, we could distinguish between domestic, farm and wild animals, using the LexiBlock below:

(11)



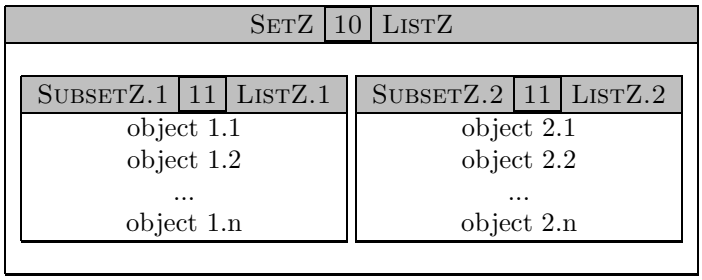
Two embedded LexiBlocks that share the same boxed number, even if they don't share the same set name, should be understood as associated. (It is a stipulated well-formedness condition for two embedded LexiBlocks that share a common number that they possess the same number of objects). For display purposes, the list elements of a LexiBlock can be listed vertically (top to bottom) or horizontally (left to right).

(12)



In (12), a LexiBlock of phonological forms of nouns are given a certain order and a label 1. The same label is given to a semantic LexiBlock. Therefore, the two LexiBlocks are to be understood together, pairing the first element of one set with the first element of the other, etc. This can again be stated more generally:

(13)

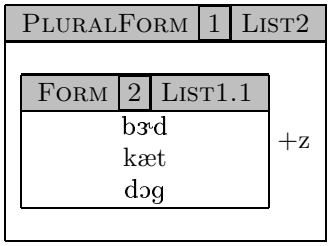


Well-formedness Condition: Co-indexed LexiBlocks that are embedded in a same LexiBlock must have the same number of elements.

In the example above, because the two LexiBlocks share the index number 11 the two subsets of objects (SubsetZ.1 and SubsetZ.2) are associated: object 1.1 with object 2.1 and so on.

If an operation is well defined on the set elements, it is possible to "factor out" common parts of the elements, using LexiBlocks. For example, if we wanted to describe the plurals of the nouns stored in (12), we could do this as follows:

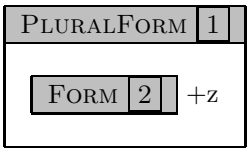
(14)



The embedded LexiBlock above is the same as the embeded LexiBlock in (12), but it is concatenated with the segment /z/. Thus the LexiBlock Plural Form 1 List2 above describes the word forms /bɜ˞dz/, /kætz/ and /dɔgz/ in an economical way by "factoring out" the common /z/ phoneme.[2] A further economy may be reached by not repeating the actual list of forms since it was already described in (12). From now on, I will also not give the name of the lists. Unless otherwise indicated, I will assume that the order is the same as stated elsewhere in the grammar/lexicon. So for example, in (15), the order of the embedded Singular forms is the same as the order they were given in (14).

[2] I assume that devoicing of plural z in *cats* can be handled automatically by phonology.

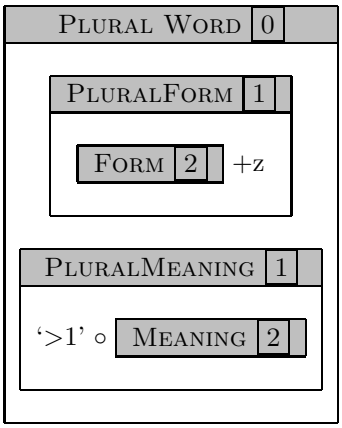
(15)



Abbreviation convention: The list name of a LexiBlock can be omitted. Unless otherwise indicated, the order is assumed to be the same as elsewhere in the grammar/lexicon.

In order now to describe the meaning of these forms when concatenated with the phoneme /z/, we need only describe the semantic function applied to the Meaning LexiBlock described in (12), and co-index the PluralForm and PluralMeaning within one LexiBlock.

(16)



The two LexiBlocks labeled 2 thus associate plural forms with plural meanings. The meanings of the singular words are defined elsewhere, in separate LexiBlocks, but are referred to here. Thus the meaning 'dog' and the form /dɔg/, though each is described only once, exist in two places at the same time and are always associated in the formalism.

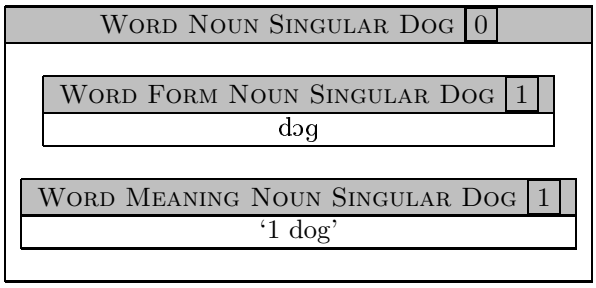
To summarize, our three tags have three separate functions. The index-number allows us to associate LexiBlocks embedded inside another LexiBlock, like the form and meaning LexiBlocks. The set and list names allow us to refer to a LexiBlock without repeating its content. The distinction between the set name and the list name is useful to allow for different groupings that serve different linguistic purposes. In practice though, we scarcely need to explicitly mark the difference between the set name and the list name.

2.1.2 Words and Connected Word Constructions

So far I have only described LexiBlocks as a formal tool that the theory uses. LexiBlocks are powerful and if we restrict their use in TCWC, we will be able to make stronger prediction, rendering the

theory more interesting. As mentioned in the introduction chapter, I assume that language learners first learn fully inflected words, not abstract lexemes or subparts of words such as stems and affixes. Hence, let's first focus on what the description of a word should be in this theory. For example, the word *dog* is represented in (17).

(17)



In TCWC, a word has four properties: 1) it has an overt segmental form; 2) it has an identifiable meaning; 3) it is associated to fully specified morphosyntactic categories; 4) it has an identifiable prosody that is different from phrase prosody. Thus, the following objects are not words: 1) an intonational pattern associated with questions is not a word, because it does not have segments; 2) some grammatical markers, such as classifiers, or morphemes, such as theme vowels, that do not have identifiable meanings; 3) English stems are not words, because they are used within both nouns and verbs and for many of their inflections;[3] 4) clitics are not objects that TCWC deals with, because they are not associated with a prosody (they are stressless or toneless). The reliance on word prosody to identify word boundaries is a strategy often used in language acquisition research—see the rich overview in Jusczyk (1997).

The several parts of the set names in (17) are understood to be independent sets brought together by intersection. Thus WORD NOUN SINGULAR DOG refers to both the FORM and MEANING of the SINGULAR NOUN *dog*. The set is considered to be the intersection of WORD, NOUN, SINGULAR and DOG or the union of WORD-FORM-NOUN-SINGULAR-DOG and WORD-MEANING-NOUN-SINGULAR-DOG. The label DOG is necessary, or else (17) would describe all singular nouns.[4]

I will adopt several abbreviation conventions to lighten the representation. First, the lexical categories such as DOG, because they are rather obvious, will often be omitted. The index numbers for WORD (0), FORM (1) and MEANING (1) being always the same, I will also omit them. Finally, since FORM and MEANING are subsets of the larger WORD LexiBlock, they should always

[3] Unless of course, a bare stem happens to be used to signify some particular inflection. For example, English SINGULAR NOUNS use their bare stems.

[4] This multi-part label system could be implemented with a type hierarchy system.

contain the same set names as the larger WORD LexiBlock. Therefore, I will omit these set names from the FORM and MEANING LexiBlocks. Hence, (17) may be rewritten as (18).

(18)

WORD NOUN SINGULAR	
FORM	MEANING
dɔg	'1 dog'

Abbreviation convention:
The lexical categories (here DOG) are omitted.
Index numbers are omitted for WORD, FORM and MEANING
FORM = WORD FORM NOUN SINGULAR DOG
MEANING = WORD MEANING NOUN SINGULAR DOG

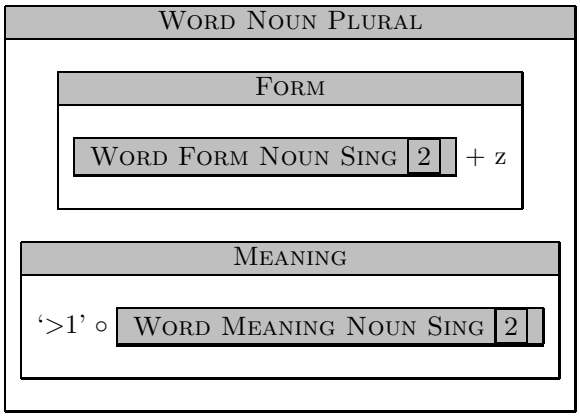
In the theory, all words have this shape. The whole is identified as a WORD, and it has two subparts: FORM and MEANING.[5] FORM is a list of phonological strings and MEANING is a list of meanings. As mentioned earlier, FORM and MEANING should be co-indexed (with $\boxed{1}$), indicating that they are associated, but this is omitted here by convention. It is possible to collapse lexical entries for words using embedded LexiBlocks as described earlier. Collapsed lexical entries are called Connected Word Constructions (CWC).

(19) a.

WORD NOUN SINGULAR	
FORM	MEANING
bɜ˞d	'1 bird'
kæt	'1 cat'
dɔg	'1 dog'
etc.	*etc.*

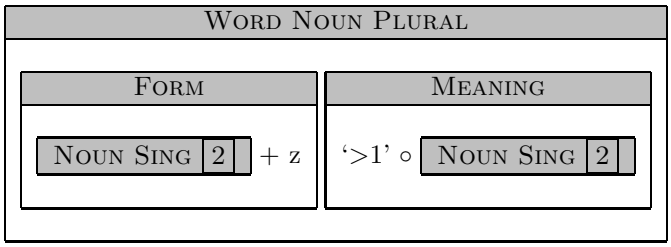
[5]Obviously, FORM should also include prosodic structure and MEANING (which should probably be called CONTENT) should also include argument structure and perhaps even pragmatic and sociolinguistic information. For the time being however, phonological form and (semantic) meaning will suffice.

b.



In order to lighten the formalization of CWCs such as the second one in (19), I stipulate that the LexiBlocks embedded in the FORM LexiBlock that refer to morphosyntactic categories only refer to WORD-FORMS and those embedded in the MEANING LexiBlock only refer to WORD-MEANINGS. Thus, in (19), in the second CWC, it is not necessary to repeat WORD FORM and WORD MEANING in the most embedded LexiBlocks. Hence, we can rewrite the second CWC in (19) as (20).

(20)



Abbreviation convention:
The LexiBlocks referring to morphosyntactic categories embedded in FORM are WORD FORM
The LexiBlocks referring to morphosyntactic categories embedded in MEANING are WORD MEANING

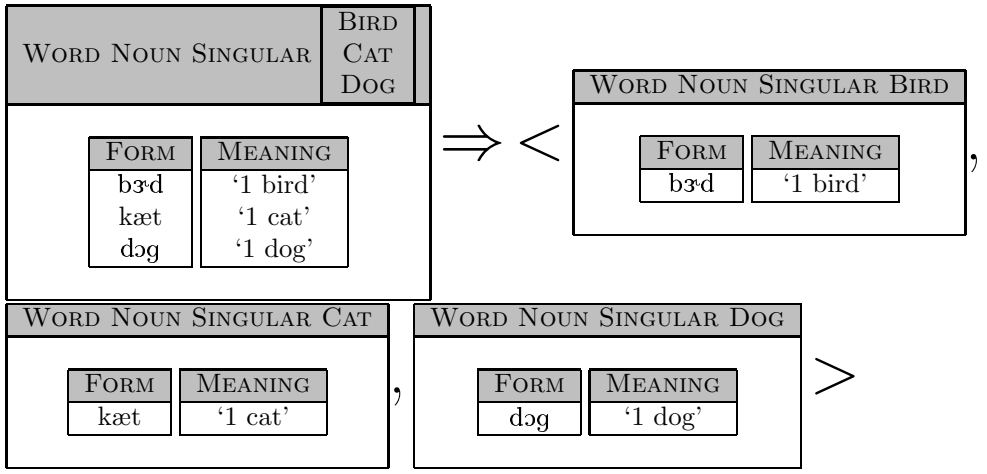
Because of the details of the formalization, it is necessary that embedded LexiBlocks have greater index numbers than the index numbers of the LexiBlocks in which they are embedded. This is a well-formedness condition on LexiBlocks.[6]

[6]Thus the index numbers increase inwards with the complexity of the embedding. There are two reasons for this choice. First, I wanted WORD, FORM and MEANING to be consistently indexed with the same numbers, so that I can omit them by convention. If the numeration increased outward, this wouldn't be possible, since they are the outermost layers—an insect skeleton if you will—of Connected Word Constructions. Second, it would have complicated the notation of the expansion algorithm (see the next sections) unnecessarily to have done things the opposite way.

(21) **Well-formedness Condition :** If a LexiBlock is embedded into another, its index number must be greater than that of the LexiBlock in which it is embedded.

In (18), by convention, I omitted the name Dog, that was present in (17), though I stated that this category is necessary to distinguish the word *dog* from *cat*. Thus the form /dɔg/, associated with the category Form, and the meaning 'dog', associated with the category Meaning, are associated with a category Dog (as well as the categories Word, Noun and Singular). It is assumed then that when we "expand" a CWC into individual words, category names are distributed to them. We can do this using a LexiBlock. For example, we could rewrite the first CWC in (19) as follows, with its expansion following it. However, for ease of reading, I will usually not include all the category-distributing LexiBlocks in CWCs.

(22)

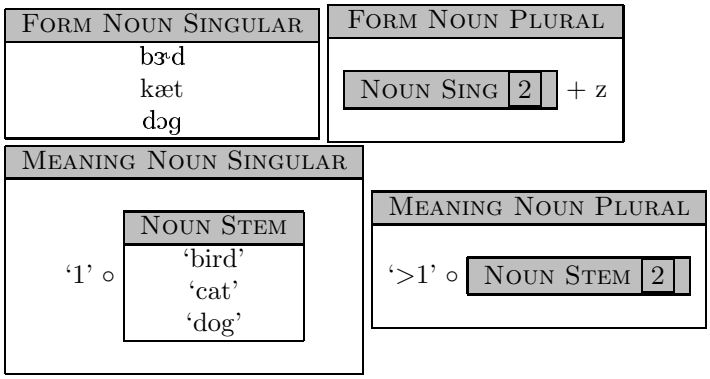


Abbreviation convention: The category-distributing LexiBlock is not always included in the representation.

Given then that the Form and Meaning parts of words and CWCs receive the same distributed category names, it is possible to describe them separately. The advantage of this is, as we know from Saussure, form and meaning enter different association patterns: while the word *cat*'s phonological form rhymes with *hat* and *mat*, its meaning may trigger associations with *dog* or *bird*. (I will illustrate this advantage further in §2.1.4). Our two CWCs in (19)-(20) can then each be split into two CWCs. The plural nouns are associated in form, because they share a phoneme (suffix) /-z/. They are associated in meaning, because of the meaning '>1' that they share. Because of formalization details, the LexiBlocks above should be part of a larger Word LexiBlock (labeled

[0]), but again, I omit this by convention.[7]

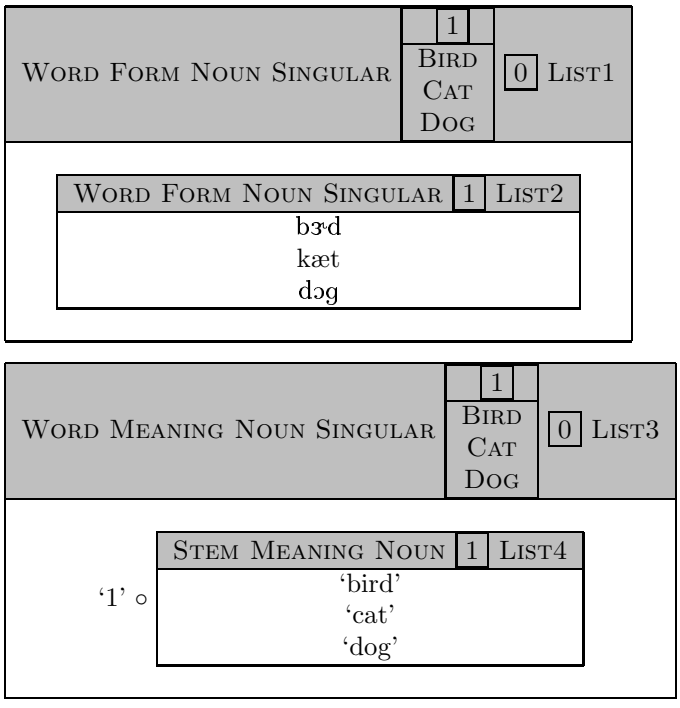
(23)



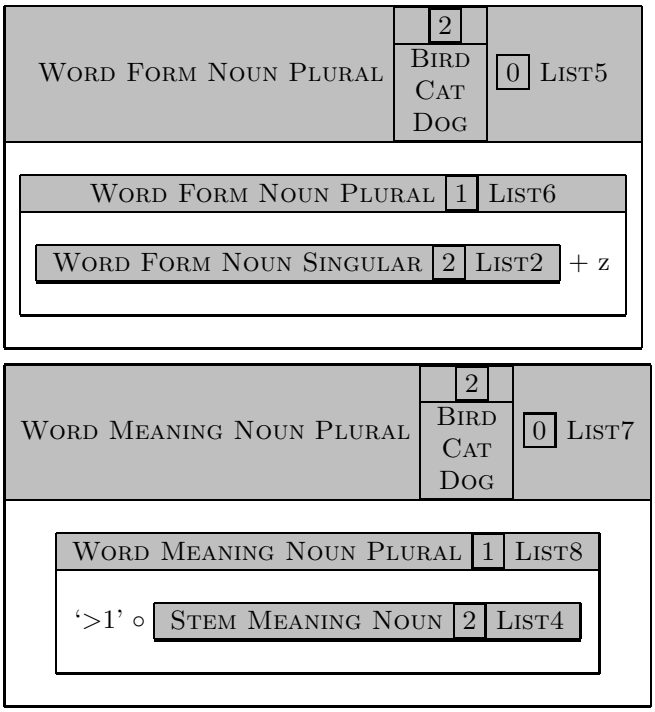
Abbreviation convention: When splitting CWCs into Form and Meaning Constructions, the outer Word LexiBlock labeled [0] is omitted.

Without the abbreviation conventions used so far, the four CWCs in (23) would be written up as follows:

(24)



[7]When we extract the meaning '1' from the singular meanings, we need to give the leftover lexical meaning LexiBlock a new name; I chose to call it Noun Stem (Meaning).



For convenience, I list below all the definitions, abbreviation conventions and well-formedness conditions discussed in this section.

(25) **Definitions**

a. LexiBlock: An ordered set of linguistic items, to which the unordered set and the ordered list are given names. In addition, an index number indicates that, when the LexiBlock is embedded into another LexiBlock, any other embedded LexiBlock with the same index number is associated with it.

b. Words are the basic unit of TCWC. They have four characteristics: 1) an overt form; 2) an identifiable meaning; 3) fully specified morphosyntactic categories; 4) an overt prosody that is distinct from phrasal prosody.

c. Connected Word Construction (CWC): Words collapsed as one using LexiBlocks.

(26) **Well-formedness conditions**

a. Two LexiBlocks coindexed with the same index number and embedded in a third LexiBlock must have the same number of elements.

b. A LexiBlock, if embedded into another, must have a greater index number than the LexiBlock in which it is embedded.

(27) **Abbreviation conventions**

a. The lexical categories, such as BIRD, CAT and DOG can be omitted.

b. Unless otherwise indicated, the ordering of elements in a LexiBlock is the same in all CWCs in which the LexiBlock is used and the list name may be omitted.

c. The index numbers of WORD (0), Form (1) and MEANING (1) are omitted.

d. FORM and MEANING are understood to be followed by the same categories as the larger WORD LexiBlock (including the category WORD).

e. The LexiBlocks referring to morphosyntactic categories embedded in FORM are WORD FORM, the ones embedded in MEANING are WORD MEANING.

f. The category-distributing LexiBlocks are sometimes omitted.

g. When FORM and MEANING are described separately, it is understood that they each belong to a larger WORD FORM or WORD MEANING LexiBlock (labeled 0).

2.1.3 Expanding the Compressed Lexicon

LexiBlocks and the way in which the connections are established both between and within Connected Word Constructions (CWCs) have now been defined. The next task is to explain how the actual words used by syntax are generated. The representations seen so far constitute a COMPRESSED LEXICON. Because the goal of morphology is not only to represent the structure of words, but to generate the actual words of a language, TCWC needs a method for decompressing the lexicon, or rather, expanding it into a list of fully inflected words usable by a (lexicalist) syntactic framework.[8]

In order to generate the EXPANDED LEXICON then, I propose a derivation in three phases. In the first phase, the operations that are stated on LexiBlocks in the COMPRESSED LEXICON are eliminated, while in the second phase, the CWCs themselves are gradually replaced by the fully inflected words. Finally, in the third phase, the forms and meanings on the lists generated by the Form and Meaning Constructions are unified to form the inflected words.

Phase 1

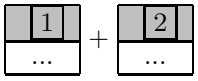
In the first phase, operations on LexiBlocks, like concatenation or semantic composition, are eliminated. I will illustrate this with concatenation, because this dissertation is more concerned about the formal side of morphology than the semantic side.[9] First, there are three different ways of applying an operation on LexiBlocks:

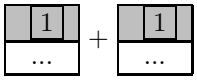
[8] Though perhaps some non lexicalist syntactic frameworks are somewhat compatible with the view of morphology I am proposing.

[9] For an explanation non-specific to concatenation, see the formalization section.

(28) **Three types of operations on LexiBlocks**

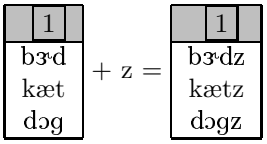
a) **Type A:** 

+ z b) **Type B:** 

c) **Type C:** 

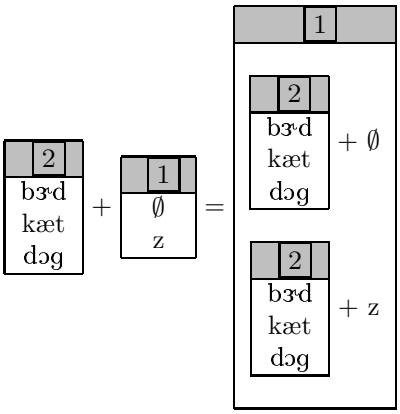
In (28), a LexiBlock is concatenated with (a) a phoneme; (b) a LexiBlock with a different index number; (c) a LexiBlock with the same index number. The first case is the simplest. When concatenating a LexiBlock with a phoneme, or a string of phonemes, the string is simply concatenated on every line of the LexiBlock, as follows:[10]

(29) **Type A:** when a phonemic string is concatenated to the left or right of a LexiBlock, concatenate the phonemic string respectively to the left or right of each line of the LexiBlock.



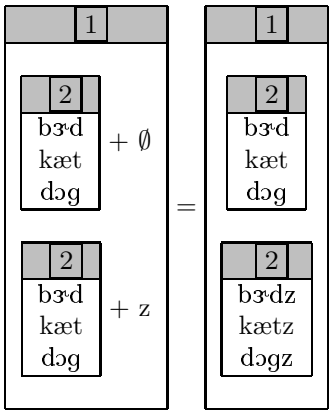
When a LexiBlock is concatenated with another LexiBlock bearing a different index number, the LexiBlock with the largest index number is concatenated on every line of the other LexiBlock, as follows:

(30) **Type B:** When two LexiBlocks with different index numbers are concatenated, concatenate the LexiBlock with the largest index number on each line of the other LexiBlock.

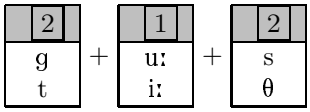


In this particular example, we are now left with two LexiBlocks each concatenated with a phoneme. This is a Type A operation, so as seen earlier, we now simply concatenate each phoneme on every line of the LexiBlock with which it is concatenated:

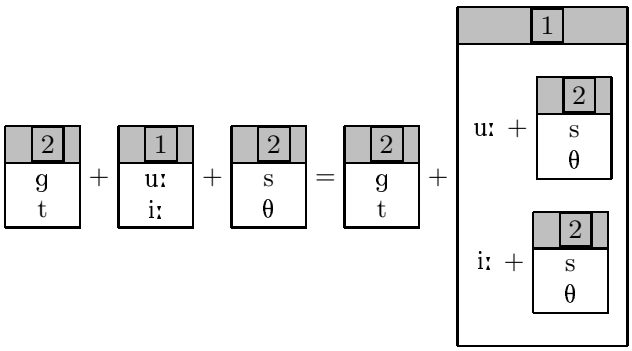
[10]Obviously, the order is important. If the phoneme or string of phonemes preceded the LexiBlock, then it would be concatenated to the beginning of each line.

(31) 

The third and final case, is the one where a LexiBlock is coindexed with another LexiBlock that shares the same index number. In order to illustrate this however, I will use the CWC for the singulars and plurals of *goose* and *tooth*, which uses all three types of LexiBlock concatenation.

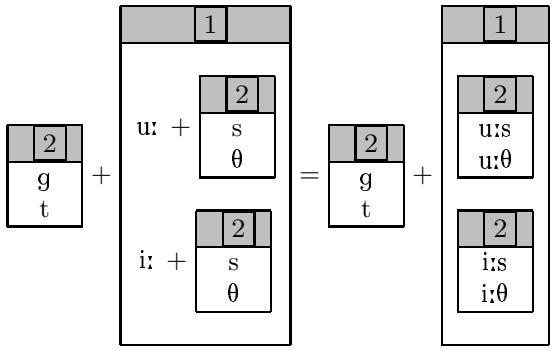
(32) 

Here we have the concatenation of three LexiBlocks. Because concatenation is defined on pairs, not on triplets, we must first concatenate two adjacent LexiBlocks in (32). I arbitrarily pick the second and third, but I could just as well have picked the first two.[11] Since the two LexiBlocks I picked do not share the same index number, the rightmost LexiBlock (with a greater index number) is concatenated on every line of the middle LexiBlock:

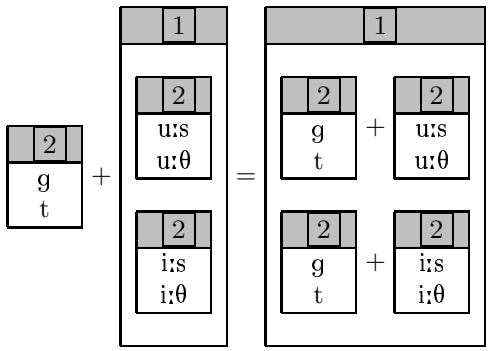
(33) 

Within the larger LexiBlock, we now have phonemes concatenated with LexiBlocks, and as we know from Type A operations, we simply concatenate these on each line of the LexiBlocks:

[11] This is ultimately due to the fact that concatenation is an associative operation: in concatenating a+b+c, it doesn't matter if I first calculate a+bc or ab+c, the end result will be abc.

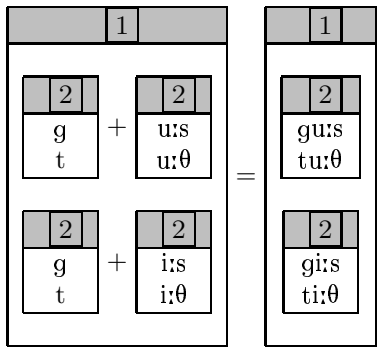
(34) 

At this point, we again have the concatenation of two LexiBlocks with different index numbers, so, following Type B operations, we take the LexiBlock with the largest index number and concatenate it on each line of the other LexiBlock:

(35) 

Finally, we are left with the third type of operation, the concatenation of two LexiBlocks bearing the same index number. Since it is a well-formedness condition that two such LexiBlocks have the same number of elements, we can define this third type of concatenation as a line-by-line concatenation:

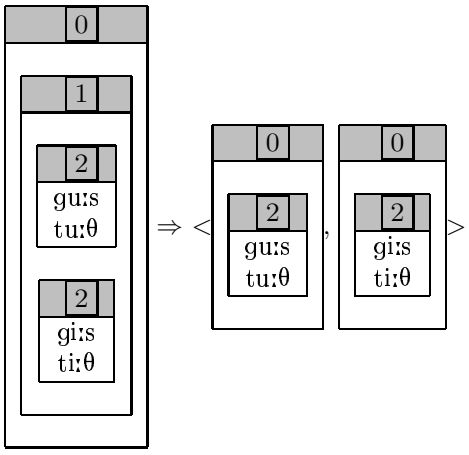
(36) **Type C:** When two LexiBlocks with the same index number are concatenated, do a line-by-line concatenation.



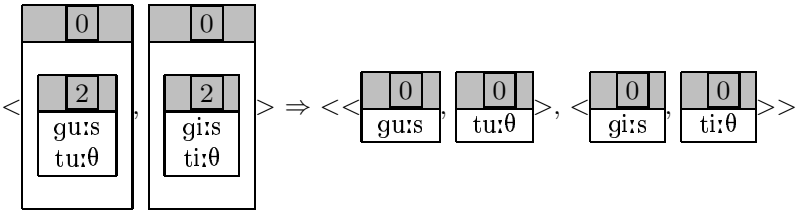
Phase 2

Once we have concatenated all that needs to be concatenated, we must "expand" the CWCs into a list of fully inflected words. I illustrate the systematic way in which this is done with the last CWC we have seen, after application of the concatenative operation. The general idea is that we successively replace each LexiBlock by its respective lines.

So first, we build a list of LexiBlocks where every LexiBlock with the index number [1] is replaced first by its first element, then by its second, etc. Remember that Form Constructions have an outer [0]-labeled layer that is usually omitted for ease of reading, but that I need to specify here.

(37) 

Then, in the list we have built, we replace every LexiBlock with the index number [2] with its first element, then its second, etc.:

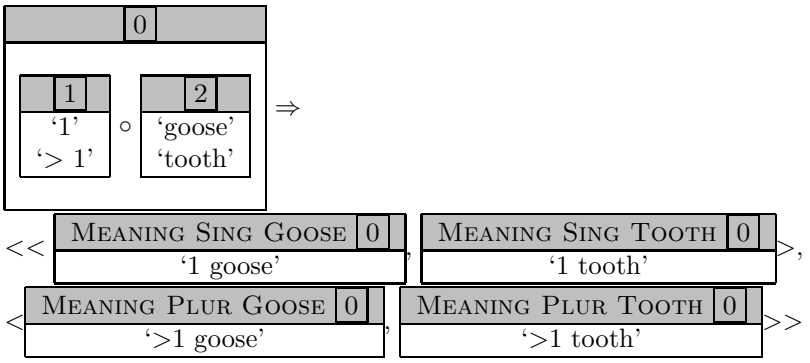
(38) 

If there were LexiBlocks with the index number [3], we would replace them successively by their lines, as we have done so far, and we would repeat this derivation until there are no more LexiBlocks left, except the outer [0]-labeled LexiBlock.

The fact that the COMPRESSED LEXICON yields a list (an ordered set) will become an advantage to account for suppletion and elsewhere effects. Indeed, some forms or constructions can be prioritized by putting them higher up in LexiBlocks. If we assume then that words generated higher up

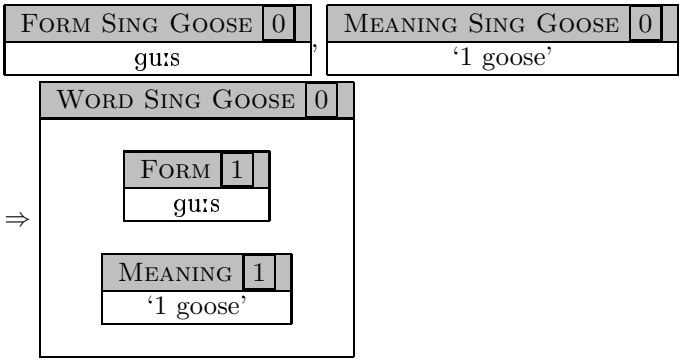
on the EXPANDED LEXICON list are used in preference to equivalent words generated lower down on the same list, then this provides a pretty straightforward handle on facts of suppletion and specific/general rules. This will be illustrated in detail in the chapter on Armenian.

By following the very same two phases as illustrated above, it is possible to derive the semantic side of the words derived above. I will not go into this in detail, but it should be pretty clear how by following the same principles for semantic composition as for concatenation, TCWC yields the following list:

(39)


Phase 3

The third and final phase is the simplest one. The FORM and MEANING of a word that are generated separately are unified thanks to the shared categories, and are ready for use by the syntax of the language:

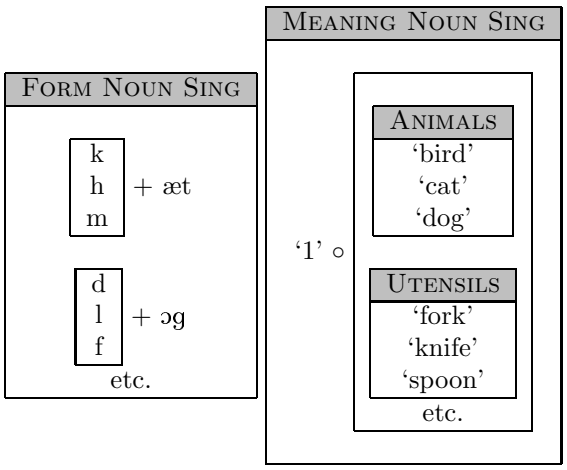
(40)


2.1.4 Form, Meaning and *pluralia tantum*

To conclude this illustration of how LexiBlocks and CWCs work, I will show how the separation of FORM and MEANING in different constructions can be an advantage. Separating the CWCs into a

Form Construction and a Meaning Construction allows us to group the nominal forms according to rhyme patterns, while grouping the meanings by semantic field, an advantage that will become clear by the end of this section. Let us take a look at the CWCs where the SINGULAR NOUNS of English are encoded.

(41)



As seen in the third phase of the expansion algorithm, the word *goose* (WORD NOUN SING GOOSE) can be generated by unifying its FORM and MEANING, since they each assign the category GOOSE respectively to /guːs/ and 'goose'. Division of CWCs into FORM and MEANING constructions has the advantage that lexical semantic categories with which one can group and subgroup meanings can be represented without complicating the phonological groupings. The advantage for nouns may not be as crucial as for verbs for syntactic purposes. As we know from its rich literature,[12] Lexical Semantic categories are very useful for syntactic purposes. For example, different syntactic constructions refer to stative, transitive or motion verbs. In terms of FORM though, a lexicon does not necessarily require the same groupings.

While the meaning groupings are ultimately the task of Lexical Semantics, the form groupings are of a greater importance for this dissertation concerned with the phonic side of morphology. There are three factors that influence the groupings: economy of representation, form-semantic correspondence and prosodic organization. A more detailed account of how these three factors are prioritized will be given in the next chapter; for now, the following observations will suffice.

For the economy factor, to give a straightforward example, in the case of classifier suffixes, when all nouns of a language end in one of three phonemes, it would be economical to group the three

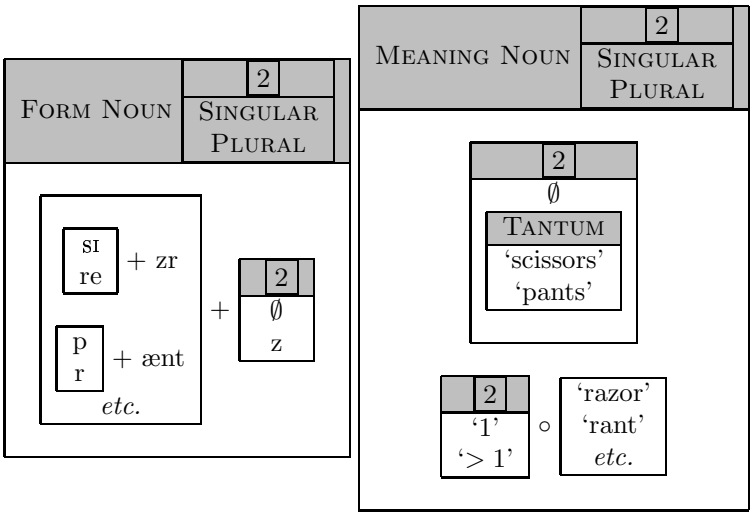
[12]See Levin (1985), Cruse (1986), Pustejovsky (1995) for general references.

classes separately and to factor out the three phonemes in question.[13]

Concerning semantic category, speakers seem to be aware of the presence of phonesthemes (see Bergen 2004). For example, many English nouns for round or circular objects begin with /b-/: ball, bowl, bulb, bearing, bubble, balloon. Factoring out the /b-/ of these nouns would then reflect speakers' knowledge.[14]

Besides the different general semantic and phonological motivations behind the organization of the lexicon, *pluralia tantum* nouns are a second illustration of the advantage of describing the form/meaning mismatches in the lexicon. *Pluralia tantum* nouns are plural nouns that refer to a singular meaning. For example in (42), we can see that the tantum meanings must be described independently from the other plurals, while their plural forms are described along with all other nouns.[15]

(42)



The FORM CWC in (42) generates singular forms for the *pluralia tantum*, but crucially, they do not generate meanings for them. I believe this is consistent with English speakers' judgements: if the *plurale tantum scissors* had a singular it is pretty obvious its form would be *scissor*, though it is not clear what its meaning would be. This "singular form" *scissor* however, by virtue of being grouped with the other noun stems, is available for compounding and derivational morphology, as is evidenced by formations such as *Scissorhands* (from the movie *Edward Scissorhands*) or the verb *scissor*. Thus, though English morphology generates a WORD FORM SINGULAR corresponding to each English noun

[13] A similar case is given on the chapter on Armenian, regarding theme vowels.

[14] This does not make a morpheme of /b-/, it is simply a phoneme that all these nouns share. There is no systematic pattern that allows one to isolate /b-/ as a root or a prefix.

[15] Of course the semantic subcategories of (41) within the non-tantum nouns can be incorporated as well.

stem, it does not necessarily generate a corresponding WORD MEANING SINGULAR. Phase 3 of the expansion algorithm then fails to apply and a singular noun *scissor* is **not** generated.

2.2 Definitions used throughout the formalization

In this section, I introduce formal notions that will be used both in the set theory formalization and the feature structure implementation.

2.2.1 Indices

Throughout the formalization, I use the following indices that are all natural numbers:[16]

(43) g, h, i, j, k, m, n, p, s, t $\in \mathbb{N}\backslash\{0\}$

The indices *i* and *p* are both between *1* and *n*, while *j* is between *1* and *m*:

(44) $1 \leq \mathrm{i}, \mathrm{p} \leq \mathrm{n}$
$1 \leq \mathrm{j} \leq \mathrm{m}$

Finally, F stands for any feature and f for any feature-value. When f is the feature-value of a feature F, I will write this as F:f.

(45) F is a feature
f is a feature-value
F:f means that f is the feature-value of the feature F.

2.2.2 Feature elements ($\in$)

If a feature-value is a feature-structure, then all the pairs of features with their respective feature-values are its elements:

(46) $\forall$ i, $\mathrm{F}_i{:}\mathrm{f}_i \in \mathrm{f}_t \Leftrightarrow \mathrm{f}_t = [\mathrm{F}_i{:}\mathrm{f}_i]_{i=1}^{n}$

[16] In some notation systems, $\mathbb{N}$ includes {0} and $\mathbb{N}^*$ doesn't, while in other notations, it is exactly the opposite. To avoid ambiguity, I will use $\mathbb{N} \cup \{0\}$ and $\mathbb{N}\backslash\{0\}$.

I illustrate the notion with the feature-structure below. The syllable feature-structure has only two elements: Onset:l and Rhyme:rhyme.[17]

(47) **Feature-structure example**

$$_{syllable}\left[\begin{array}{ll} \textsc{Onset} & \text{l} \\ \textsc{Rhyme}_{rhyme} & \left[\begin{array}{ll} \textsc{Nucleus} & \text{u} \\ \textsc{Coda} & \text{c} \end{array}\right] \end{array}\right]$$

2.2.3 Feature members (⫛)

The members of a feature-structure are its elements and "everything inside them", that is, their own members. I define this recursively as follows:[18]

$$\text{(48)}\quad \mathrm{F}_i{:}\mathrm{f}_i \Cap \mathrm{f}_t \Leftrightarrow \begin{array}{l} \mathrm{F}_i{:}\mathrm{f}_i \in \mathrm{f}_t \\ \vee\ \exists\ \mathrm{F}_j{:}\mathrm{f}_j\ \ni\ \mathrm{F}_j{:}\mathrm{f}_j \in \mathrm{f}_t \wedge \mathrm{F}_i{:}\mathrm{f}_i \Cap \mathrm{f}_j \end{array}$$

Thus, the feature-structure in (47) has four members. In addition to the two elements already mentioned, Onset:l and Rhyme:rhyme, which are also members, it has the two members Nucleus:u and Coda:c.

2.2.4 Lists

I define lists in feature-structure terms. A list is a feature-structure with a feature FIRST that can have any feature-value (including another list) and a feature REST that may have as value either the empty list (elist) or another list:

(49) **List definition**

$$_{list}\left[\begin{array}{ll} \text{FIRST} & \text{f} \\ \text{REST} & \text{elist} \vee \text{list} \end{array}\right]$$

[17] The reader should keep in mind that this is a formalization section. Thus this example does not indicate that I am claiming that this is the way to handle syllable structure. The example is only meant to illustrate with a simple case how to read feature-structures, for readers who aren't familiar with them.

[18] I follow the convention that uses ∋ to mean 'such that'.

Since lists are feature-structures, they have elements and members as defined previously. However, it will be necessary to also define first order list members. These will be defined recursively below.

So for example, the list <a, <b, c>, d> can be represented with a feature-structure as follows:

(50) **List example**

$$\begin{bmatrix} \text{FIRST} & \text{a} \\ \text{REST} & \begin{bmatrix} \textsc{First} & \begin{bmatrix} \textsc{First} & \text{b} \\ \textsc{Rest} & \begin{bmatrix} \textsc{First} & \text{c} \\ \textsc{Rest} & \text{elist} \end{bmatrix}_{list} \end{bmatrix}_{list} \\ \textsc{Rest} & \begin{bmatrix} \textsc{First} & \text{d} \\ \textsc{Rest} & \text{elist} \end{bmatrix}_{list} \end{bmatrix}_{list} \end{bmatrix}_{list}$$

(51) **First order list members ($\underline{\pitchfork}$)**

$$\begin{aligned} \text{f}_i \mathbin{\underline{\pitchfork}} \text{list}_t \Leftrightarrow \quad & \text{FIRST:f}_i \pitchfork \text{list}_t \\ & \wedge \nexists\, \text{f}_j \ni [\text{FIRST}|\text{FIRST:f}_j \pitchfork \text{list}_t \wedge (\text{FIRST:f}_i \pitchfork \text{f}_j \vee \text{f}_i = \text{f}_j)] \end{aligned}$$

In the list $< a, < b, c >, d >$ illustrated in (50), the elements of the list are FIRST:a and REST:$<< b, c >, d >$. Its members are its elements, plus FIRST:$< b, c >$, REST:$< d >$, FIRST:$< b >$, etc. Its first order members are: a, $< b, c >$ and d.

The cardinality of a list is equal to the cardinality of the set of its first-order members. Thus, the cardinality of $< a, < b, c >, d >$ is 3:

(52) **List cardinality**

$$| \text{list}_t | = | \{\text{f} \mid \text{f} \mathbin{\underline{\pitchfork}} \text{list}_t \} |$$

It is also necessary to define list precedence. I define it such that for the list $< a, < b, c >, d >$, we have $a \prec b \prec c \prec d$:

(53) **List precedence (f_i precedes f_j on $list_t$)**

$$\frac{f_i \prec f_j}{list_t} \Leftrightarrow \exists\ \text{list}_k\ \ni \quad \text{list}_k = \text{list}_t \vee \text{FIRST:list}_k \Cap \text{list}_t$$

$$\wedge \quad \text{FIRST:f}_i \in \text{list}_k \vee \exists\ \text{list}_i\ \ni \text{FIRST:f}_i \Cap \text{list}_i \wedge \text{FIRST:list}_i \in \text{list}_k$$

$$\wedge \quad \text{REST|FIRST:f}_j \in \text{list}_k \vee \exists\ \text{list}_j\ \ni \text{FIRST:f}_j \Cap \text{list}_j \wedge \text{REST|FIRST:list}_j \in \text{list}_k$$

With these definitions in mind, we can now formalize the shorthand for lists that I have been using informally so far ($< a, < b, c >, d >$ is a list containing a, followed by the list $< b, c >$, followed by d):

(54) **Shorthand for lists**

$$<\text{f}_i>_{i=1}^{n} \equiv \text{list}_t\ \ni \forall\ \text{i, FIRST:f}_i \underline{\Cap}\ \text{list}_t \wedge \frac{f_i \prec f_p}{list_t} \leftrightarrow \text{i} < \text{p}$$

2.3 Set Theory Formalization

The goal of this section is to formalize the tool used by TCWC and allow one to conceptualize it in relation to already familiar notions from set theory. The section is divided in the following subsections: first, some definitions are introduced; then, the three phases required for generating the word list as illustrated in the first section of this chapter are formalized.

2.3.1 Definitions

In this formalization, a LexiBlock is a triplet (W, X, L), where W is a set, X is a natural number (including 0) and L is a list. I will however sometimes omit either the set name or the list name.

(55) **LexiBlock Definition**

W$\boxed{\text{X}}$L $\equiv$ (W, X, L) where W is a set, X $\in \mathbb{N} \cup \{0\}$ and L is a list.

(56) **LexiBlock Abbreviation**

LexiBlocks such as W$\boxed{\text{X}}$L may be abbreviated as either W$\boxed{\text{X}}$ or $\boxed{\text{X}}$L as needed in the context.

For example, the LexiBlock Word∩Noun∩Singular 0 ListX[19] can be written up as follows, omitting the list's name (ListX in this case):

(57)

WORD ∩ NOUN ∩ SINGULAR 0	
WORD ∩ FORM ∩ NOUN ∩ SING 1	WORD ∩ MEANING ∩ NOUN ∩ SING 1
bɜ˞d	'bird'
kæt	'cat'
dɔg	'dog'

In (57), ListX is the list that contains the two embedded LexiBlocks. LexiBlocks have members that are defined recursively. A member is either a first-order member of its list or the member of a LexiBlock that is a member of this LexiBlock:

(58) **LexiBlock Members (⫛)**

$w \Cap W[X]L \leftrightarrow (w \underline{\Cap} L) \vee (\exists M[Y]K \ni w \Cap M[Y]K \wedge M[Y]K \Cap W[X]L)$

Thus the LexiBlock in (57) has eight members, two of which are LexiBlocks themselves. The elements of a LexiBlock form a subset of its members. The elements are those members that are not themselves LexiBlocks. This way, though the LexiBlock in (57) has eight members, it has only six elements (the three forms and three meanings).

(59) **LexiBlock Elements (∈)**

$w \in W[X]L \leftrightarrow (w \Cap W[X]L) \wedge (\nexists M[Y]K \ni w = M[Y]K)$

LexiBlocks need to satisfy four (4) conditions in order to be well-defined. a) the set of elements of a LexiBlock W[X]L must be the same as the set of elements of the set W. b) If two LexiBlocks that share a same index number are both members of W[X]L, then their respective lists must be of the same length. c) All LexiBlocks that are members of W[X]L must have an index-number greater than W[X]L. d) All LexiBlocks member of W[X]L must be well-defined as well.

[19] Word∩Noun∩Singular is the intersection of the sets Word, Noun and Singular, as shown in (57). In the rest of this chapter, I will simply notate the intersection of sets by a space between them, as I do in the rest of the dissertation.

(60) **Well-defined LexiBlocks**

W$\boxed{\text{X}}$L is well-defined $\leftrightarrow$ W = {w | w $\in$ W$\boxed{\text{X}}$L }

$\wedge \forall$ $\boxed{\text{Y}}$K, $\boxed{\text{Y}}$M $\pitchfork$ W$\boxed{\text{X}}$L, |K|=|M|

$\wedge \forall$ M$\boxed{\text{Y}}$K $\pitchfork$ W$\boxed{\text{X}}$L, X < Y

$\wedge \forall$ M$\boxed{\text{Y}}$K $\pitchfork$ W$\boxed{\text{X}}$L, M$\boxed{\text{Y}}$K is well-defined

In (57), if we consider the set WORD to be the union of the sets WORD FORM and WORD MEANING,[20] then the elements of the main LexiBlock are the same as the elements of the set WORD-NOUN-SING.[21] The two LexiBlocks inside the main LexiBlock are attributed the index number $\boxed{1}$, which is a bigger number than the $\boxed{0}$ attributed to the main LexiBlock. The FORM and MEANING LexiBlocks coindexed with $\boxed{1}$ have the same number of elements on their respective lists (that number is three; recall the feature-structure definition of a list and its cardinality defined previously). The FORM and MEANING LexiBlocks are also well-defined. Our LexiBlock in (57) is thus well-defined.

Now that I have defined LexiBlocks in set theory terms, we need to formalize the three phases of LexiBlock expansion.

2.3.2 Phase 1

If an operation is well-defined on the elements of a LexiBlock, then it is possible to define these operations on LexiBlocks themselves as follows:

(61) **Operations on LexiBlocks**

Let X<Y, and let A, B and C be three sets, such that:

$A = \{a_i\}_{i=1}^{m}$ $B = \{b_j\}_{j=1}^{n}$ $C = \{c_k\}_{k=1}^{p}$

A$\boxed{\text{X}}$ = $\boxed{\text{X}}$$< u_g >_{g=1}^{s}$ B$\boxed{\text{X}}$ = $\boxed{\text{X}}$$< v_g >_{g=1}^{s}$ C$\boxed{\text{Y}}$ = $\boxed{\text{Y}}$$< w_h >_{h=1}^{t}$

Let $\circ$ be an arbitrary operation defined on E = (A $\cup$ B) $\cup$ C

[20] Some FORM or MEANING could be phrases.

[21] This conception sounds a little strange from a linguistic point of view. However, in set theory, if we define the set WORD as the union of all word forms and all word meanings, it does not mean that an individual form or meaning is a word. Individual words are referred to by the intersection of WORD with say the sets NOUN, SINGULAR and BIRD, which gives us the set (or the word) {/bɜ˞d/, 'bird'}.

a. $\forall$ e $\in$ E, A$\boxed{\text{X}}$ $\circ$ e $\equiv$ $\boxed{\text{X}}$$< u_g \circ e >_{g=1}^{s}$

b. $\forall$ e $\in$ E, e $\circ$ A$\boxed{\text{X}}$ $\equiv$ $\boxed{\text{X}}$$< e \circ u_g >_{g=1}^{s}$

c. A$\boxed{\text{X}}$ $\circ$ B$\boxed{\text{X}}$ $\equiv$ $\boxed{\text{X}}$$< u_g \circ v_g >_{g=1}^{s}$

d. C$\boxed{\text{Y}}$ $\circ$ A$\boxed{\text{X}}$ $\equiv$ $\boxed{\text{X}}$$<$$\boxed{\text{Y}}$$< w_h \circ u_g >_{h=1}^{t}>_{g=1}^{s}$

e. A$\boxed{\text{X}}$ $\circ$ C$\boxed{\text{Y}}$ $\equiv$ $\boxed{\text{X}}$$<$$\boxed{\text{Y}}$$< u_g \circ w_h >_{h=1}^{t}>_{g=1}^{s}$

I illustrate these operations on LexiBlocks below with the phonological list concatenation and semantic composition operations in (62).

In (a) and (b), I illustrate the same thing with an order difference. In (a), a LexiBlock is concatenated with a phoneme, which results in concatenating this phoneme on each line of the LexiBlock. Similarly, in (b), a LexiBlock is semantically composed with '>1', which results again in distributing the meaning '>1' on each line.

In (c), once the string /uː/ is concatenated on each line of the second LexiBlock, as in (a), the remaining LexiBlocks, which happen to share the same index number, are concatenated, which means that their corresponding lines are concatenated together.

Finally, in (d) and (e), we deal with the case where an operation is applied to two LexiBlocks with different index numbers. In (d), two such LexiBlocks are concatenated, which this time results in concatenating the LexiBlock with the greater index number on every line of the other LexiBlock. Likewise, in (e), semantic composition is applied to two LexiBlocks with different index numbers.

(62) **LexiBlock operation examples**

a.

Form Noun Singular 1
bɜːd
kæt
dɔg

+ z =

Form Noun Plural 1
bɜːdz
kætz
dɔgz

b. '>1' $\circ$

Meaning Noun Singular 1
'bird'
'cat'
'dog'

=

Meaning Noun Singular 1
'>1' $\circ$ 'bird'
'>1' $\circ$ 'cat'
'>1' $\circ$ 'dog'

c.

ONSET 1
g
t

+ (uː +

CODA 1
s
θ

)

=

ONSET 1
g
t

+

RHYME 1
uːs
uːθ

=

SYLLABLE 1
guːs
tuːθ

d.

FORM NOUN SINGULAR 3
bɜˑd
kæt
dɔg

+

SUFFIX 2
∅
z

=

FORM NOUN 2	
FORM NOUN SINGULAR 3	
bɜˑd	
kæt	
dɔg	
FORM NOUN PLURAL 3	
bɜˑdz	
kætz	
dɔgz	

e.

MEANING NUMBER 2
∅
'>1'

∘

MEANING NOUN SINGULAR 3
'bird'
'cat'
'dog'

=

MEANING NOUN 2	
MEANING NOUN SINGULAR 3	
'bird'	
'cat'	
'dog'	
MEANING NOUN PLURAL 3	
'>1' ∘ 'bird'	
'>1' ∘ 'cat'	
'>1' ∘ 'dog'	

2.3.3 Phase 2

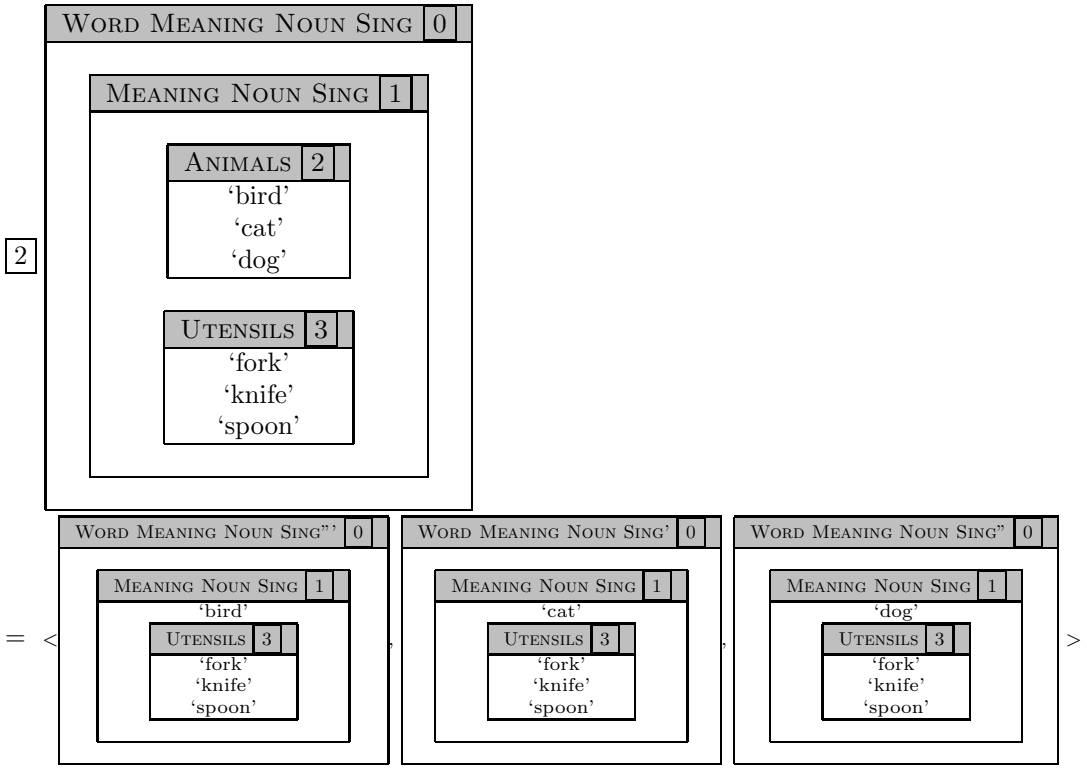
Now that we have defined what the operations on LexiBlocks mean, we need to expand the COMPRESSED LEXICON, which consists of the CWCs of the language.

First, I will define the expansion of a single LexiBlock for an arbitrary number Y. The Y-expansion of a LexiBlock W[X]L for index Y is a certain list where all the LexiBlocks that share the index Y are successively replaced by the first-order members of their lists.

(63) **Y-Expansion**

$\exists!$ n M_i[Y]K_i=[Y]$<a_{ij}>_{j=1}^{m}$ $\Cap$ W[X]L
$\Rightarrow$ [Y](W[X]L)
$=<W'_j$[X]$L'_j \mid \forall i,\ M_i$[Y]$K_i \rightarrow a_{ij} \wedge W'_j$[X]$L'_j$ is well-defined$>_{j=1}^{m}$

(64) **Y-Expansion example**

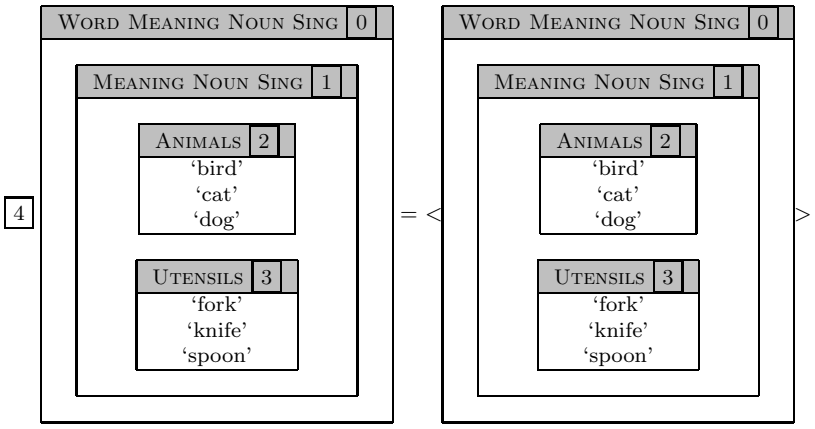


However, if there does not exist a LexiBlock with the index Y inside W[X]L, then its Y-expansion is the singleton list that contains W[X]L.

(65) **Trivial Expansion**

$\nexists$ M[Y]K $\Cap$ W[X]L $\Rightarrow$ [Y](W[X]L)=<W[X]L>

(66) **Trivial Expansion example**

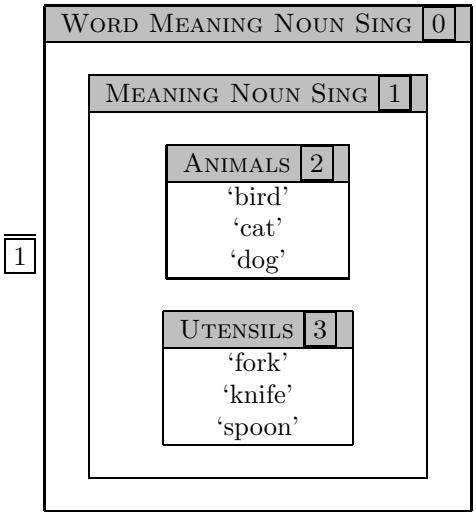


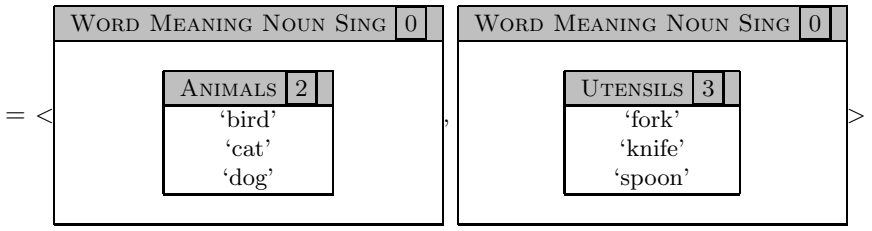
The 1!-expansion (one-factorial expansion) of W[X]L is the same as its 1-expansion.

(67) **1!-Expansion**

$\overline{[1]}$(W[X]L)=[1](W[X]L)

(68) **1!-Expansion example**



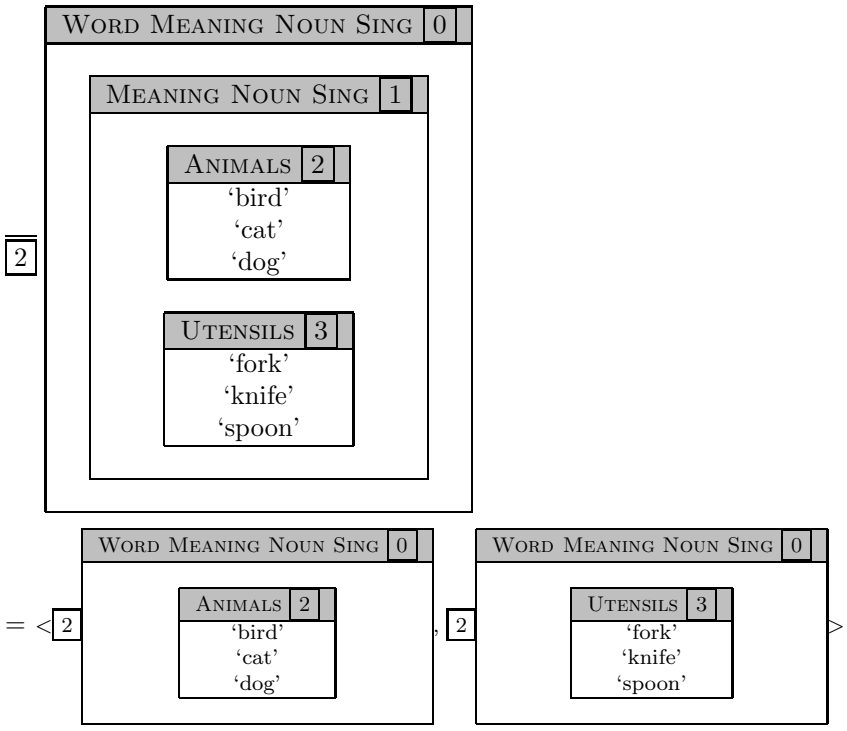


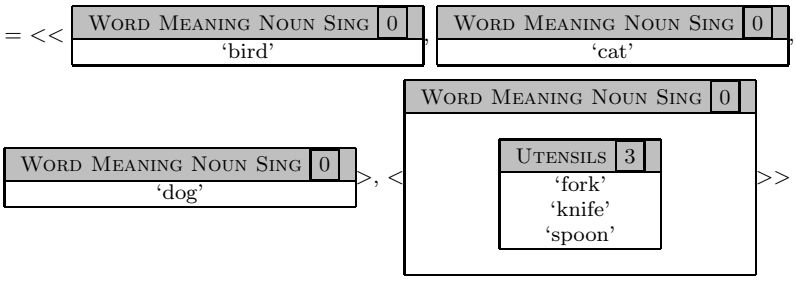
However, the Y!-expansion for any number Y greater than 1 is a series of successive expansions of expansions. This way, for example, the 5!-expansion of W X L is the 5-expansions of the 4-expansions of the 3-expansions of the 2-expansions of its 1-expansions.

(69) **Factorial Expansion**

$$\overline{\boxed{Y{+}1}}(W\,\boxed{X}\,L)=<\boxed{Y{+}1}(W'_i\,\boxed{X}\,L'_i)\mid W'_i\,\boxed{X}\,L'_i \Uparrow \overline{\boxed{Y}}(W\,\boxed{X}\,L)\ \wedge\ \frac{\boxed{Y{+}1}(W'_i\,\boxed{X}\,L'_i)\prec\boxed{Y{+}1}(W'_p\,\boxed{X}\,L'_p)}{\overline{\boxed{Y{+}1}}(W\,\boxed{X}\,L)}\ \leftrightarrow\ \frac{W'_i\,\boxed{X}\,L'_i\prec W'_p\,\boxed{X}\,L'_p}{\overline{\boxed{Y}}(W\,\boxed{X}\,L)}>$$

(70) **Factorial Expansion Example**





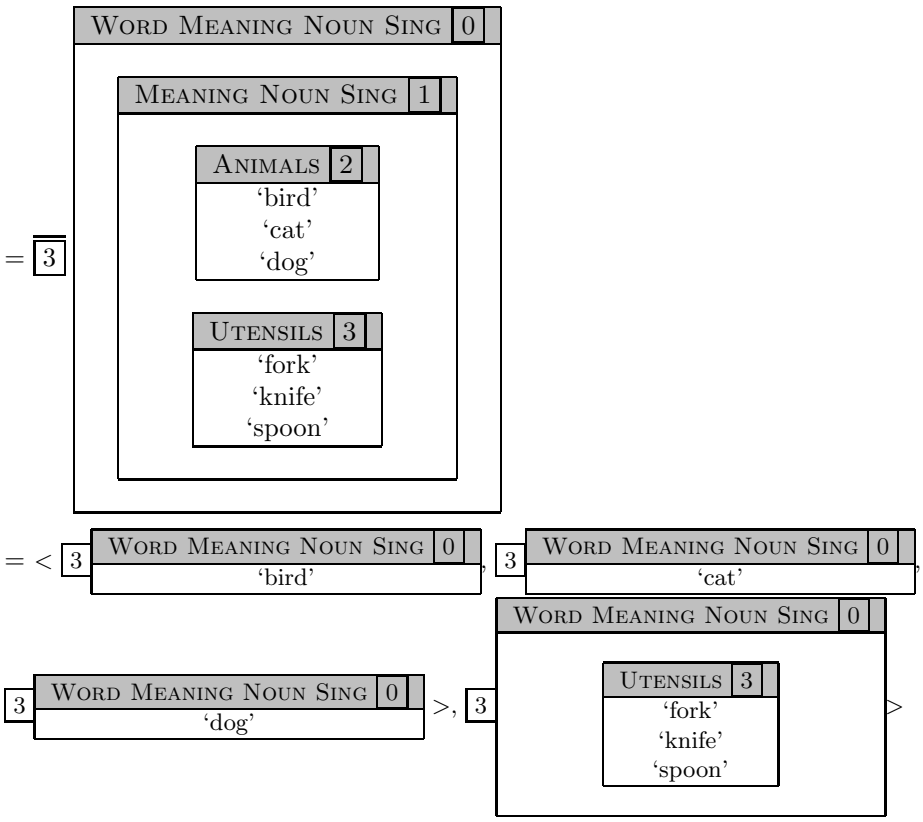
The total expansion (or simply the expansion) of W X L is its maximal factorial expansion:

(71) **Total Expansion**

N=max{Y | ∃ M Y K ⫛ W X L} ⇔ $\overline{W\,\boxed{X}\,L} = \overline{\boxed{N}}$(W X L)

(72) **Total Expansion Example**

$\overline{(Word \cap Meaning \cap Noun \cap Sing)\boxed{0}}$

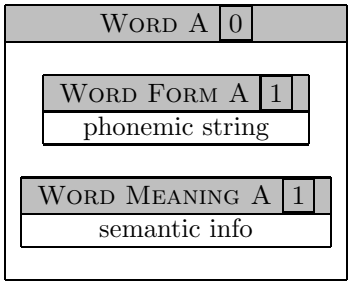


= << WORD MEANING NOUN SING 0 'bird' >, < WORD MEANING NOUN SING 0 'cat' >,
< WORD MEANING NOUN SING 0 'dog' >, < WORD MEANING NOUN SING 0 'fork' ,
WORD MEANING NOUN SING 0 'knife' , WORD MEANING NOUN SING 0 'spoon' >>

2.3.4 Phase 3

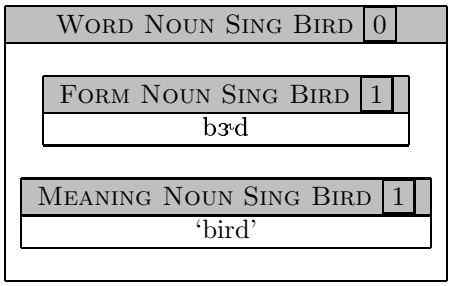
A *word*, in Set-theory LexiBlock terms, is defined in the following way. It is a LexiBlock with index [0]. Its set is *Word A* (the intersection of the sets Word and A). Its list contains two LexiBlocks, Word Form A and Word Meaning A, both of index [1] (so they are expanded together).

(73)



Thus, the word *bird* takes the form and meaning generated by CWCs such as (72) and unifies them into the mold of (73), yielding (74):

(74)



2.4 Feature-Structure implementation

In this section, I propose an implementation of TCWC using feature-structures. This is a useful thing to do, because many linguists are familiar with feature-structures and some syntactic frameworks

(among which is HPSG[22]) are formalized using them. Hence, this section will facilitate future comparisons of TCWC with such theories of syntax to see how they might be compatible.

In terms of feature-structures, there already exists a mechanism called DISTRIBUTED DISJUNCTIONS, that is similar to LexiBlocks. Distributed Disjunctions are used to merge descriptions in feature-based formalisms, allowing more efficient processing of disjunctions, while adding no expressive power to a formalism that already admits disjunctions.

(75) **Distributed Disjunctions**

$$\begin{bmatrix} \textsc{Pers-Num} & \left\{ \$1 \quad \text{1Sing, 2Sing, 3Sing} \right\} \\ \textsc{Phon} & \text{stem} \oplus \left\{ \$1 \left\langle \quad \text{e} \right\rangle \vee \left\langle \quad \text{s,t} \right\rangle \vee \left\langle \quad \text{t} \right\rangle \right\} \end{bmatrix}$$

$$= \begin{bmatrix} \textsc{Pers-Num} & \text{1Sing} \\ \textsc{Phon} & \text{stem} \oplus \left\langle \quad \text{e} \right\rangle \end{bmatrix} \vee \begin{bmatrix} \textsc{Pers-Num} & \text{2Sing} \\ \textsc{Phon} & \text{stem} \oplus \left\langle \quad \text{s,t} \right\rangle \end{bmatrix} \vee \begin{bmatrix} \textsc{Pers-Num} & \text{3Sing} \\ \textsc{Phon} & \text{stem} \oplus \left\langle \quad \text{t} \right\rangle \end{bmatrix}$$

In the example above, adapted from Krieger et al. (1993), a stem is concatenated with a Distributed Disjunction (DD) labeled \$1 and the Person-Number is again a DD labeled \$1. This is equivalent to a disjunction of three feature-structures, one for each person-number, where the stem is concatenated with the suffix that is paired to this person-number, via the DD labeled \$1. The label \$1 is the equivalent of our boxed index number from the previous sections.

LexiBlocks differ from DDs in two respects: 1) LexiBlocks are expanded into a list, whereas DDs correspond to a disjunction; 2) DDs do not have a separate tag for the set of objects. This latter difference is due to the difficulty of implementing set theoretic notions in feature-structures. In the implementation of LexiBlocks I am proposing, I will have to compensate this difficulty with a FORM and MEANING function at the end of this section. To learn more about DDs, see Dörre & Eisele (1989), Backofen et al. (1990), (1991).

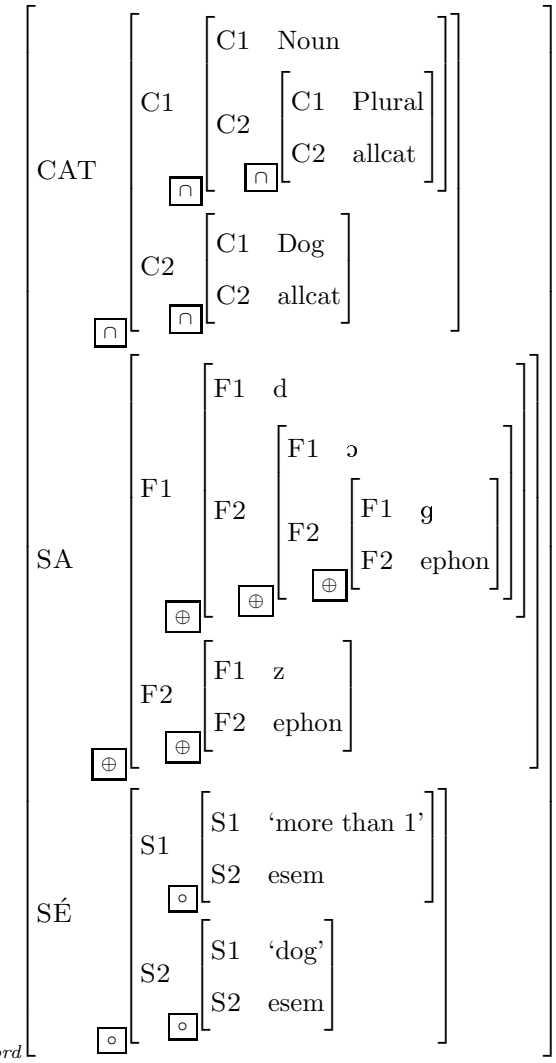
This section follows the same structure as the previous on a set theory formalization of TCWC. First, some definitions are introduced, then the phases of the expansion algorithm are described. In the feature-structure implementation, there is no Phase 1, but there are two additional functions that need to be defined to compensate the absence of set names.

[22] Head-driven Phrase-Structure Grammar. See Pollard & Sag (1994).

2.4.1 Definitions

In the feature-structure implementation I propose for LexiBlocks, Words and Connected Word Constructions are represented as in (77).[23] The word [Dog Plural]={dɔgz, '>1 dog'} is represented in (76).

(76) **The word Dog**



[23] CAT stands for CATEGORY, SA for Saussure's SIGNIFIANT, SÉ for his SIGNIFIÉ. Note also that ∩, ⊕ and ○ are all types of lists. The names are mnemonic for the category intersection, phoneme concatenation and semantic composition operations.

(77) **Words and CWCs with feature-structures**

$$word\left[\begin{array}{ll} \text{CAT} & \boxed{\cap}\left[\begin{array}{ll} \text{C1} & \boxed{\cap} \vee \text{category} \\ \text{C2} & \text{allcat} \vee \boxed{\cap} \end{array}\right] \\ \text{SA} & \boxed{\oplus}\left[\begin{array}{ll} \text{F1} & \boxed{\oplus} \vee \text{phoneme} \\ \text{F2} & \text{ephon} \vee \boxed{\oplus} \end{array}\right] \\ \text{SÉ} & \boxed{\circ}\left[\begin{array}{ll} \text{S1} & \boxed{\circ} \vee \text{meaning} \\ \text{S2} & \text{esem} \vee \boxed{\circ} \end{array}\right] \end{array}\right]$$

Since $\boxed{\oplus}$, $\boxed{\circ}$ and $\boxed{\cap}$ are specific kinds of lists, we can use a shorthand for them, similar to the shorthand I introduced for lists back in (54) in §2.2.4. The shorthands I will use are quite straightforward: instead of the angled brackets, I will use slashes for phonological forms, single quotes for semantic information and nothing for the categories; instead of commas I will use adjacency; the presence of the operation symbols between two lists indicates that they are elements of a larger list.

(78) **Shorthand for $\boxed{\cap}$, $\boxed{\oplus}$ and $\boxed{\circ}$**
CAT X Y ≡ CAT <X, Y>
CAT X $\boxed{\cap}$ Y ≡ CAT <<X>, <Y>>
SA /XY/ ≡ SA <X, Y>
SA /X/ $\boxed{\oplus}$ /Y/ ≡ SA <<X>, <Y>>
SÉ 'XY' ≡ SÉ <X, Y>
SÉ 'X' $\boxed{\circ}$ 'Y' ≡ SÉ <<X>, <Y>>

We can then abbreviate the lengthy feature-structure in (76) by introducing appropriate symbols between the two members of the three features CAT, SA, SÉ, and parenthesizing appropriately as follows:

(79) **The word Dog abbreviated**

$$word\left[\begin{array}{ll} \text{CAT} & \Big(\textsc{Noun} \boxed{\cap} \textsc{Plural}\Big)\boxed{\cap} \text{Dog} \\ \text{SA} & /\text{dɔg}/ \boxed{\oplus} /\text{z}/ \\ \text{SÉ} & \text{'more than 1'} \boxed{\circ} \text{'dog'} \end{array}\right]$$

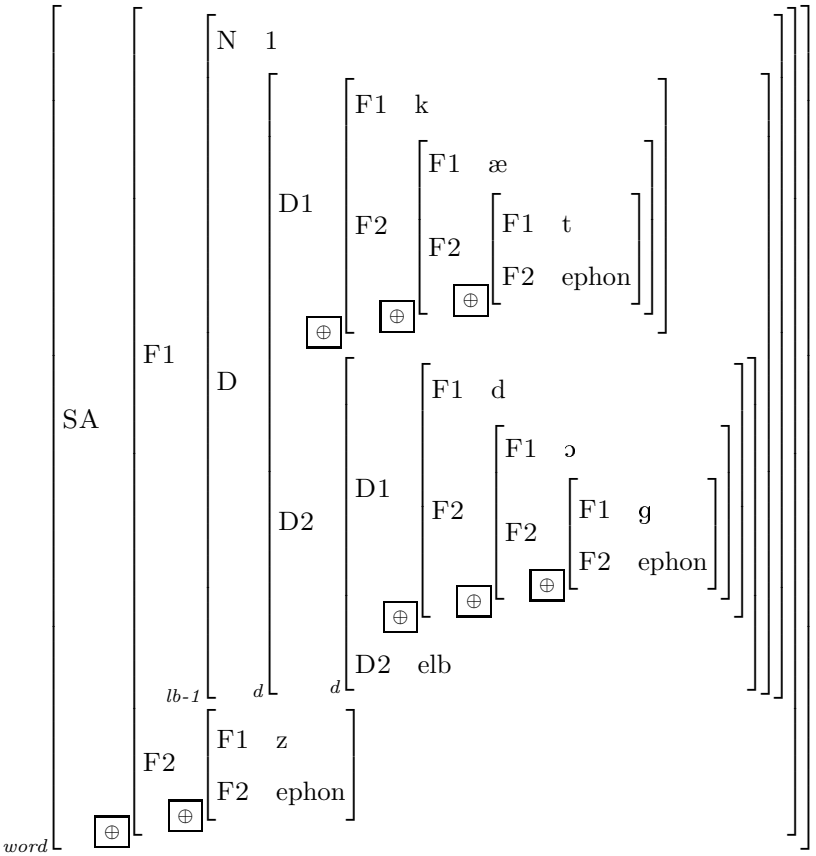
In this feature-structure implementation then, WORDS are not LexiBlocks themselves as I imply throughout the dissertation (although that is the case in the set theory formalization). A LexiBlock can be defined on any feature. It is defined recursively like a list, except that it has an index number and LexiBlocks may only appear on D1 (whereas lists may appear on REST).

(80) **LexiBlock Definition**

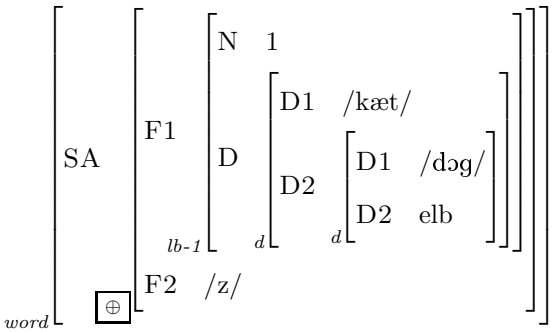
$$\text{F:f} \Leftrightarrow \text{F:} \quad \underset{lb\text{-}n}{\left[\begin{array}{ll} \text{N} & \text{n} \\ \text{D} & \underset{d}{\left[\begin{array}{ll} \text{D1} & \text{lb} \vee \text{f} \\ \text{D2} & \text{elb} \vee \text{d} \end{array}\right]} \end{array}\right]}$$

(81) contains the forms /kæt/ and /dog/ in a LexiBlock concatenated with /z/. Given the abbreviations introduced above, (81) may be rewritten as (82).

(81) **Cat and Dog in a Feature LexiBlock**



(82) **Cat and Dog abbreviated**



Because LexiBlocks aren't lists, we must define first order members and precedence for them, in order to adopt a shorthand for them as well. These definitions however, have essentially the same structure as they have for lists.

Thus, the first-order LexiBlock members of a lb-k are those members that aren't members of other LexiBlocks inside lb_k.

(83) **First-order LexiBlock Members ($\underline{\Cap}$)**

$\text{f}_i \mathbin{\underline{\Cap}} \text{lb}_k \leftrightarrow \text{D1:f}_i \Cap \text{lb}_k \wedge \nexists\ \text{f}_j\ \ni \text{D1|D|D1:f}_j \Cap \text{lb}_k \wedge (\text{D1:f}_i \Cap \text{f}_j \vee \text{f}_i{=}\text{f}_j)$

(84) **LexiBlock Precedence (f_i precedes f_j on lb_k)**

$$\frac{f_i \prec f_j}{lb_k} \Leftrightarrow \exists\ \text{d}_t\ \ni \quad \begin{array}{ll} & \text{d}_t \in \text{lb}_k \\ & \vee\ \text{D1:d}_t \Cap \text{lb}_k \\ \\ \wedge & \text{D1:f}_i \in \text{d}_t \\ & \vee\ \exists\ \text{d}_i\ \ni \text{D1:f}_i \Cap \text{d}_i \wedge \text{D1:d}_i \in \text{d}_t \\ \\ \wedge & \text{D2|D1:f}_j \in \text{d}_t \\ & \vee\ \exists\ \text{d}_j\ \ni \text{D1:f}_j \Cap \text{d}_j \wedge \text{D2|D1:d}_j \in \text{d}_t \end{array}$$

(85) **Shorthand for lb-k**

$$<\text{k f}_i >_{i=1}^{n} \equiv \text{lb-k} \ni \forall\ \text{i, f}_i \mathbin{\underline{\Cap}} \text{lb-k} \wedge \frac{f_i \prec f_p}{lb-k} \leftrightarrow \text{i} < \text{p}$$

Thus we can rewrite (81)-(82) as (86), which looks a lot like the DISTRIBUTED DISJUNCTIONS we saw at the beginning of §2.4:

(86) **Cat and Dog fully abbreviated**

$$_{word}\left[\text{SA} \quad \left\langle 1 \quad /\text{kæt}/, /\text{dɔg}/ \right\rangle \boxed{\oplus} /\text{z}/\right]$$

The cardinality of a LexiBlock counts all of its first-order members:

(87) **LexiBlock Cardinality**
$|\text{lb}_i| = |\ \{\text{f}_k \mid \text{D1:f}_k \underline{\pitchfork}\ \text{lb}_i\}|$

LexiBlocks are well-defined if they meet three conditions. a) The first condition is that all member LexiBlocks must have a greater index number. b) All members LexiBlocks that share an index number must have the same cardinality. c) All member LexiBlocks must be well-defined.

(88) **Well-defined LexiBlocks**

(a) $_{lb\text{-}i}\begin{bmatrix} \text{N} & \text{i} \\ \text{D} & \text{d}_i \end{bmatrix} \Rightarrow \forall\ \text{N:j} \ni \text{N:j} \pitchfork \text{d}_i,\ \text{j>i}$

(b) $_{lb\text{-}k_i}\begin{bmatrix} \text{N} & \text{k} \\ \text{D} & \text{d}_i \end{bmatrix}\ _{lb\text{-}k_j}\begin{bmatrix} \text{N} & \text{k} \\ \text{D} & \text{d}_j \end{bmatrix} \pitchfork \text{lb}_p \Rightarrow |\text{lb-k}_i| = |\text{lb-k}_j|$

(c) $\forall \text{i},\ \text{lb}_i \pitchfork \text{lb}_j \Rightarrow \text{lb}_i$ is well-defined

2.4.2 Phase 2

As I mentioned earlier, there is no need for an equivalent of the set-theoretic Phase 1 of the LexiBlock expansion algorithm in the feature-structure implementation. Because both LexiBlocks and lists are features, we can start the replacement of LexiBlocks by their members directly. In the set-theory formalization, LexiBlocks are different objects than the form and meaning lists that they are made of, so it was necessary to define concatenation and other operations on LexiBlocks. The feature-structure implementation allow us to do things in a more straightforward manner.

Thus, the k-expansion of a feature-structure is a list where the feature-structure in question is successively modified by replacing each lb-k by its first-order members.

(89) **k-Expansion**

$\exists!$ n $F_i{:}<_{k_i}\ f_{ij}\ >_{j=1}^{m} \Cap f_t$
$\Leftrightarrow$ lb-k(f_t)=$<(f_t)'_j \mid \forall i, F_i{:}<_{k_i} f_{ij} >_{j=1}^{m} \rightarrow F_i{:}f_{ij} >_{j=1}^{m}$

(90) **k-Expansion example**

$$\text{LB-1}_{word}\left(\left[\begin{array}{lll} \text{CAT} & \text{Noun} \boxed{\cap} \left\langle 1 \quad \text{Sing, Plur} \right\rangle \boxed{\cap} \left\langle 2 \quad \text{Bird, Cat, Dog} \right\rangle \\ \text{SA} & \left\langle 2 \quad \text{/bɔ:d/, /kæt/, /dɔg/} \right\rangle \boxed{\oplus} \left\langle 1 \quad \emptyset, \text{/z/} \right\rangle \end{array}\right]\right)$$
$$= \left\langle \left[\begin{array}{lll} \text{CAT} & \text{Noun} \boxed{\cap} \text{Sing,} \boxed{\cap} \left\langle 2 \quad \text{Bird, Cat, Dog} \right\rangle \\ \text{SA} & \left\langle 2 \quad \text{/bɔ:d/, /kæt/, /dɔg/} \right\rangle \boxed{\oplus} \emptyset \end{array}\right]_{word}, \left[\begin{array}{lll} \text{CAT} & \text{Noun} \boxed{\cap} \text{Plur} \boxed{\cap} \left\langle 2 \quad \text{Bird, Cat, Dog} \right\rangle \\ \text{SA} & \left\langle 2 \quad \text{/bɔ:d/, /kæt/, /dɔg/} \right\rangle \boxed{\oplus} \text{/z/} \end{array}\right]_{word} \right\rangle$$

If there are no lb-k inside a feature-structure, then its k-expansion is a list containing just that feature-structure.

(91) **Trivial Expansion**

$\nexists$ F:$<_k\ f_{ij}\ >_{j=1}^{m} \Cap f_t \Rightarrow$ lb-k(f_t)=$<f_t>$

The 1!-expansion (one-factorial expansion) of a feature-structure is its 1-expansion.

(92) **1!-Expansion**

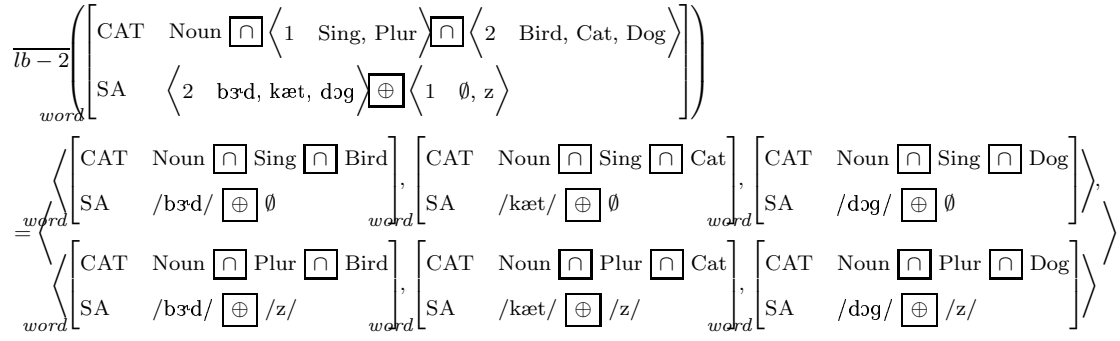
$\overline{lb-1}(f_t)$ = lb-1(f_t)

For the numbers greater than 1, the factorial expansion of a feature structure is define recursively as expansions of expansions. So for example, the 5!-expansion of a feature-structure is the 5-expansions of the 4-expansions of the 3-expansions of the 2-expansions of its 1-expansions.

(93) **Factorial Expansion**

$\forall$k, $\overline{lb-k+1}(f_t)$=<lb-k+1($(f_t)'_i$) |
$(f_t)'_i \Cap \overline{lb-k}(f_t) \wedge \frac{(f_t)'_i \prec (f_t)'_p}{\overline{lb-k}(f_t)} \leftrightarrow \frac{lb-k+1((f_t)'_i) \prec lb-k+1((f_t)'_p)}{\overline{lb-k+1}(f_t)} >$

(94) **Factorial Expansion Example**



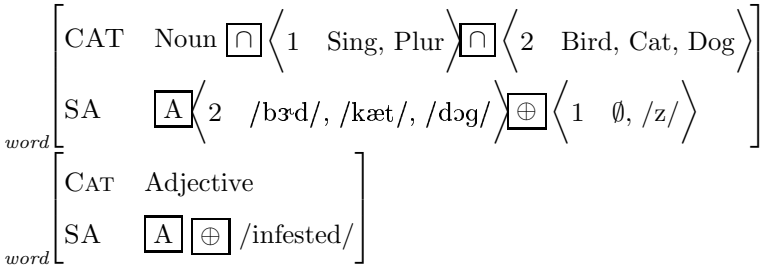
The total expansion (or simply the expansion) of a feature-structure is the factorial expansion of its maximal LexiBlock. (In the previous example, since the maximal LexiBlock is indexed with the number 2, its total expansion is its 2!-expansion).

(95) **Total Expansion**

$n=\max\{i \mid \exists\ F{:}<_{lb-i} f_j >_{j=1}^{m} \Cap f_t\}$
$\Leftrightarrow \overline{lb}(f_t)=\overline{f_t} = \overline{lb-n}(f_t)$

Feature-structures can be tagged in order to be referred to in other constraints. Tagging LexiBlocks is like tagging any other feature-structures. Therefore, it is possible to refer to parts of one CWC in another. For example, in the CWC of (94), the stems may be tagged and used in a CWC on compounding. Below, after tagging the noun stems, we use the tagged LexiBlock in a CWC that compounds them with the word *infested*.

(96) **Tagging LexiBlocks**



However, because there is no set-component in the feature-structure implementation, it is sometimes necessary to be able to refer to some CWCs that belong to an expansion in a different way (since we cannot tag LexiBlocks that appear only in the expansion). I can illustrate this problem with the example above. Because the word *infested* belongs to the EXPANDED LEXICON, we cannot tag it (or any other past participle) like we tagged the noun stems. What we need is a function that will pick the signifiers and signified present in the EXPANDED LEXICON associated with certain categories.

First, let's define the Lexicon LEX of a given language as a list consisting of the expansions of all the CWCs in a language. Since lists are feature structures, we should be able to define a function that will "dig into" LEX using feature membership to fetch the words we need. The functions for FORM and MEANING are defined separately.

(97) $\text{LEX} \equiv < \overline{Word_j} >_{j=1}^{m}$

(98) $\exists!$ n Word_i ɔ $\forall$i FIRST:Word_i ⫛ LEX
$\wedge$ C1:$\boxed{\cap}_k$ ⫛ Word_i, $\forall$ C1:$\boxed{\cap}_k$ ⫛ $\overline{\boxed{\cap}_t}$
$\wedge$ $\frac{Word_i \prec Word_p}{LEX} \leftrightarrow$ i < p

$\Rightarrow$ $\text{Form}_k(\boxed{\cap}_t) = <_k \boxed{\oplus}_i \mid \text{SA:}\boxed{\oplus}_i \in \text{Word}_i >_{i=1}^{n}$
$\wedge$ $\text{Meaning}_k(\boxed{\cap}_t) = <_k \boxed{\circ}_i \mid \text{SE:}\boxed{\circ}_i \in \text{Word}_i >_{i=1}^{n}$

Now if we want to expand the compound construction defined above to include all past participles, all we need to do is use the FORM function:[24]

(99) **Using the Form Function**

$$_{word}\left[\begin{array}{ll}\text{CAT} & \text{Adjective} \\ \text{SA} & \boxed{\text{A}}\,\boxed{\oplus}\,\text{Form}\Big(\text{PASTPARTICIPLE}\Big)\end{array}\right] = \quad _{word}\left[\begin{array}{ll}\text{CAT} & \text{Adjective} \\ \text{SA} & \boxed{\text{A}}\,\boxed{\oplus}\Big\langle \quad \text{infested, polluted, ...}\Big\rangle\end{array}\right]$$

[24]Once again, keep in mind that this is a didactic chapter, and I do not wish to imply that the example below captures all the subtleties of English ADJECTIVE compounding.

2.5 Conclusion

In this chapter, I have accomplished three things. First, I have introduced from an intuitive point of view the formal tool called the LexiBlock and its use in the Connected Word Constructions (CWCs) of the theory. As we have seen, LexiBlocks are relatively flexible, allowing one to classify word forms and meanings in cross-cutting groups required for various linguistic purposes, such as allomorphy, form/meaning associations, pluralia tantum, and so on. As we will see throughout the dissertation, they provide a way of organizing the lexicon that is very useful in analyzing various morphological phenomena.

After the intuitive introduction of TCWC and LexiBlocks, I formalized LexiBlocks, CWCs and their expansion using set theory. This allowed us to conceptualize TCWC using simple mathematical tools whose behaviors are well known. Finally, I provided a feature-structure implementation, which should facilitate the study of the extent to which TCWC is compatible with feature-based theories of syntax. Such an implementation seems like the natural thing to do because feature-based theories of syntax are usually lexicalist, just like TCWC.

Perhaps not every detail of this chapter was accessible to all readers, but it was a necessary one if we want to have a formally sound theory. As mentioned at the beginning of the chapter, it is not necessary to understand the set theoretic and feature-structure formalizations to appreciate the rest of the dissertation. The first section of this chapter should be sufficient to allow most readers to understand the theory and appreciate its insights.

Chapter 3

Analogy and Acquisition

This chapter proposes a learning procedure for the Theory of Connected Word Constructions (TCWC) and ties this procedure to diachronic change in morphology. Several phenomena in diachronic linguistics, and particularly in diachronic morphology, have been grouped under the label "analogy" since Neogrammarian times. It would be an anachronism to attribute a synchronic/diachronic distinction to the Neogrammarian school, but they also considered that analogy explained the synchronic inflection and derivation of new words, although the question of whether speakers generate "online" words they have heard before, or whether they simply repeat them from memory, was at least not a central concern to them. More recently, and more closely related to the synchronic use of the word, analogy has also been used to describe a postulated cognitive mechanism used by a family of linguistic and cognitive theories to generate the set of words of a language, in part or in whole. In these theories (e.g. Ramscar 2002, Pinker 1999), analogy is opposed to rules.

Therefore, in §3.1-3.2, I explicitly state which diachronic analogical phenomena this chapter deals with, and in which ways TCWC is related to analogical linguistic-cognitive models. In §3.3, I introduce a five-step acquisition procedure for Word Constructions, as well as three Lexical Insertion Conditions that allow speakers to generate and inflect new words.

In the following sections, I examine how TCWC, with its acquisition procedure, accounts for some diachronic changes in morphology. The examples are mainly taken from North American dialects of French. I should point out that no claim is intended about whether the changes examined were "imported" from Europe or whether they were independent developments; we will strictly be interested in the structural qualities of the changes. More precisely, the proposal is that various

analogical changes that have been labeled traditionally PROPORTIONAL ANALOGY, LEVELING, CONTAMINATION, etc. correlate with the five acquisition steps proposed and are constrained by the Lexical Insertion Conditions discussed in this chapter.

Finally, in §3.12, I propose some slight adjustments to TCWC that could help it account for the kinds of facts with which cognitive analogical models of morphology are concerned. These suggestions should not be taken as definitive, as much more research will be necessary to confirm their relevance. The adjustments fit naturally within the framework, and are simply fine tuning, rather than radical modifications.

3.1 Analogical change and sound change

Much attention has been given to analogy, both as a diachronic linguistic phenomenon and as a synchronic cognitive mechanism. Neogrammarians reserved the term analogy for changes that regularize languages, as opposed to the regular sound laws or sound changes that—Neogrammarians claimed—apply blindly to words and often yield irregularities in language.

Schuchardt (1885), a contemporary critic of the Neogrammarians, considered their hypothesis that sound laws are regular/exceptionless contradicted by his detailed studies of languages and language change (among which the first creole studies). Schuchardt's writings were important in the foundations of the American Variationist sociolinguistic school (that also has diachronic pretensions). Weinreich et al. (1968:139-140) criticize *the vacuities of the Neogrammarian doctrine of analogy*, and agree with the critics who point out that *"analogy" as an alternative to exceptionless sound laws not only was an* ad hoc *explanation, but also converted the sound law itself into an* ad hoc *concept* (Weinreich et al. 1968:139). Weinreich et al. (1968) further criticize the Neogrammarians, and especially Paul (1888), about language/dialect mixing.[1] The critique, then, was that if everything that wasn't regular sound change was analogy, then the hypothesis could not be disproved, making it a tautology.

Decades of variationist research later though, Labov (1994:v1:471) recognized that *if we were to decide the issue by counting cases, there appear to be to be far more substantially documented cases of Neogrammarian sound change than of lexical diffusion.*[2] This, however, could be a bias of the number of researchers who have worked with the assumption that sound change is regular. Labov (1994:v1:543) proposes that regular sound change and lexical diffusion may be in complementary

[1] In spite of these criticisms, the authors recognize Hermann Paul's contribution to historical linguistics, which go far beyond theory-internal concerns.

[2] Lexical diffusion being irregular sound change that spreads word by word. For a discussion, see Kiparsky (1995).

distribution. For example, consonant manner of articulation changes tend to be regular in the documented cases, while consonant place of articulation ones tend to be irregular. While analogy is not limited to lexical diffusion, Labov's position recognizes the empirical validity of both regular and irregular changes.

It is certainly true that Neogrammarians were primarily interested in regular sound laws. The reason for this is simple: it was—and still is—the most convincing tool to prove genetic relationships between languages and to reconstruct proto-languages, and that was the most exciting task for a linguist in those days. As a consequence of this focus on sound laws (today, we would say diachronic phonology), the theoretical definition of the complementary phenomenon that is analogy suffered. Another reason for the discredit given to analogy was the preponderance of four-part or proportional analogy. First, it was not until Kurylowicz (1949) that it was realized that proportional analogy was much too powerful for the facts observed; it needed to be constrained or else one could describe proportional changes that had never been observed (and by 1949, the number of languages studied was now significant). Second, because proportional analogy was so dominant, it made other types of analogy seem *ad hoc*. As recently as Kiparsky (1995) and Kiparsky (2005), linguists still need to argue for the unity of proportional and non proportional analogy. Hence, in modern terms, if one equates analogy with proportional analogy, then it both undergenerates and overgenerates.

These problems could all have been solved if someone had proposed a single coherent definition of analogy on which all agreed. Of course, the problem is that, unlike regular sound change, it is not straightforward what unifies all the instances of analogy, and it is thus not as easily formalizable as sound change. Kiparsky (1995, 2005) does a convincing job of unifying lexical diffusion and grammaticalization with analogy, two phenomena that seem to fall between cracks. According to Kiparsky, *analogical change is grammar optimization, the elimination of unmotivated grammatical complexity or idiosyncrasy.* Although nothing in this dissertation depends crucially on a single definition of analogical change that would be set in stone, Kiparsky's is a very convenient definition for our purposes.

The reason I like Kiparsky's definition is that it is well motivated. Sound change did not originally need such a well-motivated definition, because it was such a compelling and easily observable phenomenon to whoever had access to Latin, Greek, Sanskrit, the numerous German dialects or the evolution of the various written Romance languages. The motivation(s) behind sound change are only beginning to be seriously considered, with a better understanding of our perceptual system and physiology.[3] If all of analogy can indeed fall under a definition concerning grammar optimization,

[3]One could perhaps add identity construction as a motivation behind the disappearance or acceptance of a sound

then this is a great step forward. We all know from learning foreign languages how tempting it is to replace the idiosyncrasies and complexities of a language by using more regular and simpler patterns. Even in our native languages, we observe this tendency on many occasions.

Analogy then is truly in an opposite situation from sound change: sound change is readily identifiable, but obscurely motivated; analogical change is difficult to identify, but easily motivated. Astronomy and meteorology represent a similar pair. Several ancient cultures were able to discover extremely precise rules governing the movement of the celestial objects: the sun, the moon, the stars, comets, etc. The motivation behind these movements however was not at all understood; at least not until we gained a better understanding of the Solar System, galaxies, etc. Conversely, long-term weather predictions are extremely difficult to make. Yet, it is fairly easy to understand the basic motivations behind meteorological phenomena: water evaporates, when it's cold, water freezes, hence snow. In sum, the basics of meteorology are much more accessible than star movements, although it is easier to write formulas predicting the position of the Sun, rather than predicting the next thunderstorm.

3.2 Cognitive analogy, rules and constraints

As we have just seen, analogical change, which covered most of, but not exclusively, what we would consider today to be diachronic morphology, was the *parent pauvre* of the Neogrammarians' conception of linguistics. Ironically, but perhaps not accidentally, when Generative Grammar came about (Chomsky 1957 for syntax, Chomsky & Halle 1968 for phonology), it was again mainly morphology who suffered. In fact, it was not until Aronoff (1976) that morphology was given serious consideration as a legitimate part of the Generative program. From diachronic phonology, it was an easy step to formalize synchronic phonology, a step in fact that had already been taken (to a lesser extent) by the Structuralists. Formalizing syntax also came naturally. It was thought that the apparent exceptions were regimented by identifiable structural (or later, semantic) constraints. True, there were always here and there the annoying idiosyncrasy, but those could be dealt with convenient diacritics, and there was an agreement that there existed a core set of syntactic phenomena that could be handled by rigid rules.

This did not prevent parallel traditions from emerging. For example, variationists claimed that language was not to be studied with absolute rules, but with variable rules, that attributed a probability of realization to an outcome, based on various sociological factors (gender, social class,

change within a speech community.

age, ethnic group, etc.). In a similar vein, Bybee (1985, 2001), Rumelhart & McClelland (1986), Skousen (1989), Ramscar (2002) proposed that really, grammar is best studied with an analogical model of cognition that is coherent with our understanding of the human mind. Some of these researchers, e.g. Ford et al. (1997), restrict their more general cognitive approach to morphology and wonder (Singh 1990) why syntax would be so different.

Prince & Pinker (1988), Pinker (1999) adopt a more moderate view, the DUAL-ROUTE MODEL, where rules account for the productive cases, while unproductive and irregular morphology is accounted for by storing lexical items in an associative (analogical) memory.

The relevant question is: just what is the difference between rules and analogy? The cognitive concept of analogy takes into account factors like real-world meaning, similarity and frequency in order to decide how to inflect or derive a new word. The concept of a rule differs in that rules are taken to be categorical statements that apply to inputs in order to generate the inflected and derived words of a language, with no consideration for factors such as similarity, frequency, etc.

Ramscar (2002) has challenged the weaker dual-route model of Pinker (1999) by showing how at least semantic similarity has a crucial influence on the choice between regular-productive and irregular-"unproductive"[4] morphological strategies, and concludes that an analogical model is desirable to deal with all of morphology.

Albright & Hayes (2002) also conclude that there is no evidence for the rule/analogy distinction in morphology, though they conclude that all of morphology is to be dealt with stochastic rules. However, the difference they make between rules and analogy is only a matter of degree: a rule is an explicitly restricted form of analogy. Given this definition of a rule and given that their rules are stochastic and hence sensitive to factors such as frequency, we can take the difference between Ramscar and Albright & Hayes not to be important for our purposes.

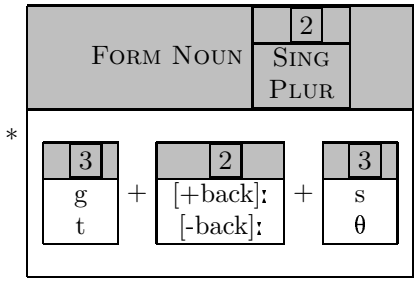
Because TCWC is not concerned with syntax, I will only express a morphological opinion on analogy and rules. TCWC is fully compatible with the results of Ramscar (2002) and Albright & Hayes (2002). No distinction is made between morphological rules and analogy. TCWC is "analogical" in the sense that it is a model that posits competing morphological strategies and no priority based on abstract structural properties is given to regular-productive strategies. Similarity of form is definitely a factor in deciding between competing morphological strategies, and, though proposals to this effect are not definitive at this point, frequency can and should be considered as one of the factors that influence the choice a speaker makes between competing morphological

[4] I write the word *unproductive* in quotes, because what is usually termed "unproductive" morphology is best characterized as *less* productive morphology.

strategies. I haven't paid much attention to semantic similarity, but as we will see in this chapter, this is certainly a kind of explanation that is welcome to account at least for folk etymology and contamination.

Crucially though, and as we will see throughout this chapter, TCWC is a restricted analogical model. Learners must follow five acquisition steps in order to obtain the CWCs of their language, they cannot construct them in any which way they want. For example, the CWC for the plurals of *goose* and *tooth* establishes an alternation between the *vowels* /uː/ and /iː/, not between any front and back vowel, which would imply that the plural of *bus* could be /bœs/.

(100) **Not the CWC for the plurals of *goose* and *tooth***



Further, the three Lexical Insertion Conditions that we will see impose restrictions on which kinds of words may or may not be inserted in a given CWC. It is within these parameters, when there is still a choice left between morphological strategies, that frequency, similarity of form and meaning may be invoked to choose one over the other.

Finally, a word on constraints. Since the introduction of Optimality Theory (OT) in linguistics—Prince & Smolensky (1993), McCarthy & Prince (1993)—it has been increasingly popular in phonology in particular and in linguistics in general, to talk in terms of constraints rather than rules. Constraints were already known in linguistics, especially in declarative models of grammar—see e.g. Stanley (1967), Ross (1967), Perlmutter (1971), Gazdar et al. (1985) for syntax, Kisseberth (1970), Bird (1992) for phonology. While again, there is some variation as to what is to be understood as a constraint,[5] in OT terms, a constraint may be either positive[6] (as in "syllables have onsets" or "sonorants are [+voiced]") or negative, in which case, we call them filters (as in "syllables don't have codas" or "obstruents are not [+voiced]"). While a given rule determines a single output, a filter only prevents certain outputs from surfacing, and another mechanism is required, e.g. constraint ranking or repair strategies, to determine the correct output.

[5]For example, Singh (1985) equates constraints with what I call filters below.
[6]Rules are an example of positive constraints.

CWCs are positive constraints, though they are not "rules" in the sense that they do not determine a single output, but allow for several options. A choice between the available options is made by the speaker and this choice can be influenced by several factors, such as frequency, similarity with other stored forms, etc.

There are also (negative) filters: those are the Lexical Insertion Conditions we will see at the end of this chapter. However, in TCWC, the positive constraints are all violable by definition, since they allow for several options, while the filters are non-violable. By contrast, OT constraints are all violable.[7]

The diachronic meaning of analogy is thus not unrelated to the one that describes a family of cognitive linguistic frameworks. Indeed, the factors that analogical models claim to be at the source of choice between morphological strategies are the same as the ones Neogrammarians claimed to be responsible for historical analogical change. In fact, researchers who promote analogical models of morphology often cite experiments involving the inflection of nonce-words (wug tests) to show how speakers extend existing patterns to new words. Analogy in its diachronic sense may be concerned with the extension of existing strategies to new words, but also to already existing words which over time move from using one strategy to using another. Diachronic analogy may also include how new strategies come along in a language.

In this chapter, I will use the study of analogical change as a testing ground for TCWC, rather than the study of the synchronic extension of an existing strategy to new (or nonce) words. The advantage of studying language change to uncover how analogy (the phenomenon) works is that the facts are easily available. The study of analogy and sound change also form the oldest scientifically studied area in Western linguistics. The phenomena are thus well known.

3.3 Connected Word Construction acquisition

3.3.1 Acquisition steps

I assume that learners are equipped with the capacity to store words in the way described in the previous chapter: forms and meanings of words are stored separately, so that we may reflect the different groupings that form and meaning require, though they are coindexed in a way that allows them to be associated. Further, words may be collapsed into CWCs in order to factor out identical

[7] Although there are attempts to do morphology within OT, e.g. Russel (1997), there is no accepted canon that morphologists work with, as of yet. Therefore, I will generally not take on the difficult task of comparing TCWC with potential OT accounts of morphological phenomena.

parts of their form or meaning in an economical way. The relevant question now is: under what circumstances can word descriptions be collapsed? I propose a five-step procedure for the acquisition of CWCs (collapsed lexical entries), followed by three conditions on insertion of newly learned words in the existing CWCs.

I present the learning procedure as if it were monotonic. This is obviously an abstraction, what I present is an idealized model of acquisition. I do not intend to imply that learning the morphology of a language consists solely of these five steps applied a single time. If the steps are on the right track, I imagine learners can backtrack and loop back to previous steps in the course of acquisition. But the goal of this section is first to show that the theory is learnable. Of course, the procedure will be used throughout the dissertation, and so eventually, it should be more finely tuned to match the facts observed by linguists working on language acquisition.

I The Word Step

The first step, called the Word Step, is that the form and meaning of all words that share the same morphosyntactic categories are stored together in two CWCs. So for example *cat*, *dog*, *mouse*, *louse* and all the other singular nouns of English are stored in two constraints, one for their forms, one for their meanings, while *cats*, *dogs*, *mice* and *lice* are stored in two other constraints, along with the other plural nouns of English. This step relies on the assumption that language learners are equipped with a way of learning the forms and meanings of fully inflected words, as well as the relevant morphosyntactic categories of their language. Perhaps some categories are also universally available, or there are good cognitive arguments for why the same categories show up again and again in one language after another. I will leave these questions open, though I acknowledge that TCWC should not be indifferent to the answers.

(101)

Form Noun Singular	Form Noun Plural
kæt	kætz
dɔg	dɔgz
maws	majs
laws	lajs
etc.	etc.

Meaning Noun Singular	Meaning Noun Plural
‘cat’	‘cats’
‘dog’	‘dogs’
‘mouse’	‘mice’
‘louse’	‘lice’
etc.	etc.

Ferguson & Farwell (1971) argue that categories emerge as words are learned. Although I am extremely sympathetic to this point of view, as I mentioned above, the acquisition procedure introduced in this section is unfortunately not meant to be isomorphic with the stages of acquisition observed by linguists working with actual acquisition data. First, we must show that there exists at least one procedure that shows the morphological lexicon as characterized by TCWC to be learnable. Obviously, if it weren't learnable at all, TCWC would face a serious problem. The question of how well the actual learning mechanism proposed matches what we know about language acquisition is an important one, but it would be unrealistic in a single dissertation to propose a theory, formalize it and make sure that its acquisition matches all of those facts. This kind of abstract learning theory is standardly used in linguistics—see Tesar & Smolensky (1998), Albright & Hayes (2002) for Optimality Theory—at least as a first step. I will thus limit myself to showing that there is a systematic way to arrive at the CWCs I use in the dissertation.

II The Connection Step

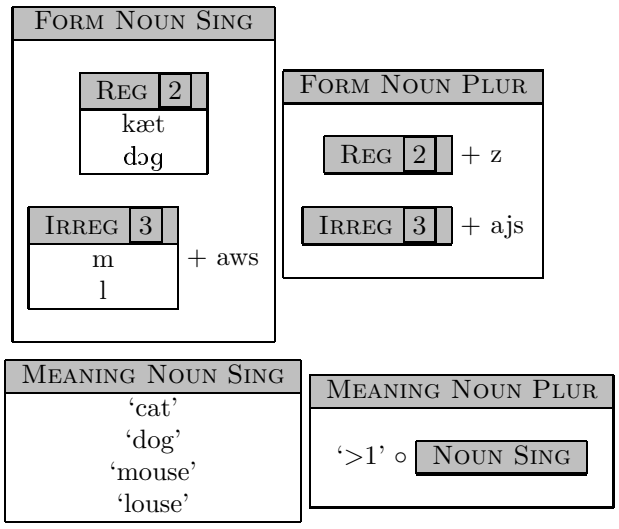
The second step is called the Connection Step and for it to take place, two conditions must be met. 1) The form of a word (w1) in CWC A is included in the form of another word (w2) in CWC B or there is a string of phonemes s1 strictly included in w1 and a non empty string s2 strictly included in w2 such that the substitution of s1 by s2 in w1 gives w2. 2) There is a single categorial relationship between the meanings associated with w1 and w2. When these two conditions are met, all pairs of words that share the same formal difference are embedded in LexiBlocks within their respective CWC, factoring out the shared string and meaning.

For example, in (101), either the Singular forms are included in the Plural ones (that have an extra /z/), or replacement of the string /aw/ by /aj/ yields the Plural forms and there is a single categorial relationship between the Singular and Plural meanings. Therefore transforming (101) into (102) is licensed.

The requirement that *the form of a word (w1) in CWC A be included in the form of another word (w2) in CWC B* accounts for simple cases of suffixation, as we have just seen, but also identity relationships, circumfixation and obviously prefixation. The alternative requirement that there be *a string of phonemes s1 strictly included in w1 and a non empty string s2 strictly included in w2 such that the substitution of s1 by s2 in w1 gives w2* accounts for all the other cases. As we have just seen, vowel change can be abstractly described this way. The same goes for infixation: replacement of the empty string in *sulaat* by *-um-* yields the infixed *sumulaat*. More complex cases of segment

alternating affixes as well: replacement of the string /ɪk/ in /ilɛktrɪk/ by /ɪsiti/ yields /ilɛktrɪsiti/. I could go on, but crucially, this requirement limits the form relationships that are considered morphological. For example, it prevents an affix on one edge to trigger a segment replacement on the other edge. Thus, imagine a fictional language similar to English, where suffixing the PLURAL /z/ triggers a replacement of the first segment by /k/: SINGULAR /dɔg/, PLURAL /kɔgz/. Obviously, none is included in the other, but also, there is not a string strictly included in /dɔg/ (empty or otherwise), such that its replacement by another string would yield /kɔgz/.

(102)



As for the "single categorial relationship" requirement, it is intended to guide the learner in establishing connections. For example, while there is a single categorial relationship between *dog* and *dogs* (SINGULAR/PLURAL), or between *dog* and *dog breath* (a compounding relation), there are two such relationships between *dog* and *dog breaths* (SINGULAR/PLURAL and compounding). The requirement thus helps the learner limit the number of connections s/he looks for.

III The Sharing Step

This leads us to the third step called the SHARING STEP. In this step, LexiBlocks are used within the already constructed CWCs to group words that are similar in form and meaning together. There are three phases to this step: 1) in the FORM constructions, edge segments shared by a significant number of words or LexiBlocks in a LexiBlock established in the previous step are factored out, while in the MEANING constructions, words with similar meanings are grouped together; 2) edge segments shared by a significant number of words or LexiBlocks in a LexiBlock that are grouped

together in the corresponding MEANING construction are factored out; 3) edge segments shared by a significant number of words or LexiBlocks in the remaining LexiBlocks are factored out.

I will leave open how significant the number of items must be for each phase to apply, because we will not deal with crucial cases in this dissertation anyway. The first phase of the step may be illustrated with a language that has theme vowels, for example Armenian. In Armenian, the CONNECTION STEP yields the following CWCs for the INFINITIVE and the 3SING SUBJUNCTIVE PRESENT:[8]

(103)

FORM SUBJ PRES 3SING
xɔsi
bɑrgi
udɛ
xmɛ
bɔrɑ
xntɑ

FORM INFINITIVE
SUBJ PRES 3SING + l

The first phase of the SHARING STEP recognizes that the words in the SUBJUNCTIVE CWC all end in either of three vowels: /i/, /ɛ/ or /ɑ/. Though I realize that there are surely some ambiguous cases out there, in this case, we definitely have a significant number of words that share their final vowel.

(104)

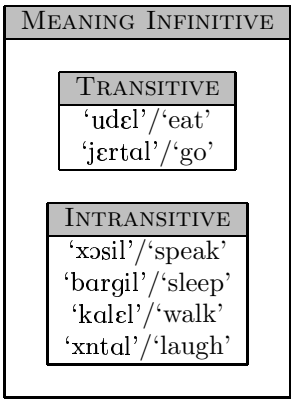
FORM SUBJ PRES 3SING
xɔs / bɑrg + i
ud / kɑl + ɛ
bɔr / jɛrt + a

FORM INFINITIVE
SUBJ PRES 3SING + l

During the same phase however, the meanings of these Armenian verbs will not be grouped in exactly the same manner. Although lexical semantics is not the focus of this dissertation, it is nevertheless safe to assume that the transitive/intransitive distinction is one that is relevant for Armenian:

[8]The facts are simplified for our purposes. See Chapter 4 for more detail.

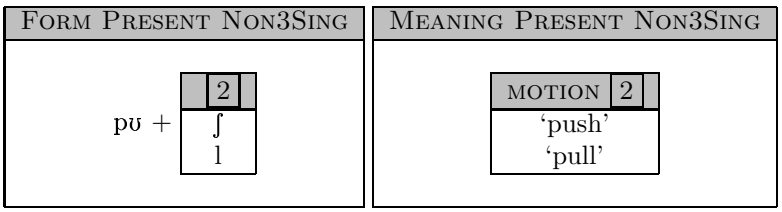
(105)



Of course, there may very well be a tendency for words in a same semantic class to also share morphemes such as theme vowels. Such a situation could be a relic of a previous stage where the relationship between form and meaning was more transparent, and that has been altered by an independent event such as sound change. But it could also speak of a tendency in languages to make the Form and Meaning constructions match as much as possible, something easily expressible in TCWC.

The second phase yields what is called phonesthemes in the linguistic literature. For example, the words *push* and *pull* behave the same way morphologically (they both have regular Past and PastParticiple forms), thus they end up in the same LexiBlock by the Connection Step. It is probably not the case that there is a significant number of words in the regular verb LexiBlock that begin with the string /pʊ-/. However, the two verbs are semantically very similar, they are both motion verbs. The second phase of the Sharing Step allows us to represent them as follows in the CWCs:

(106)



Perhaps grammar itself does not make much use of the initial string /pʊ/. However, the theory predicts its availability for use in domains such as poetry and marketing. Bergen (2004) shows how phonesthemes help speakers recognize words beyond mere form or meaning resemblance. Phonesthemes have no special status in this theory. They are groupings made according to the principles that guide the acquisition of the lexicon, just like morphemes, except that the grammar doesn't

make direct use of them, though we'll see later how useful similar groupings can be in accounting for paradigm gaps and some morphological generalizations.

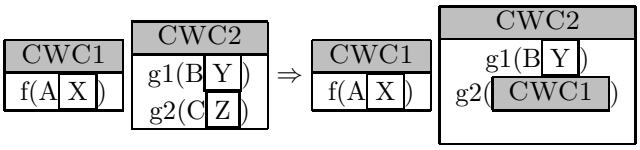
Finally, in the third phase, the remaining words may be grouped by rhyme or by other recurring patterns in the remaining LexiBlocks.

To summarize, the SHARING STEP groups words by phonological and semantic similarity. It gives priority to similarities that correlate with morphological groupings, then to similarities that correlate with semantic groupings, and finally to the remaining purely phonological similarities.

IV The Elsewhere Step

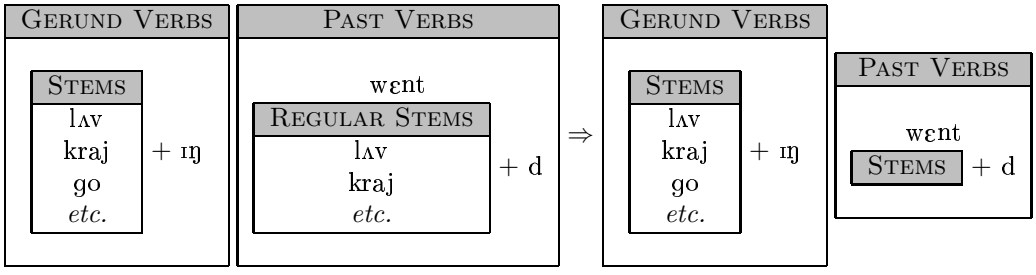
The fourth step is the ELSEWHERE STEP. This step is used in cases of suppletion and general vs. specific rules. Its formal statement is as follows:

(107) Let A X, B Y and C Z be FORM LexiBlocks.
Let |B| < |C|
Let A = B ∪ C



For example, in English, most verbal stems can take the suffix *-ing* to form the gerund, but in the past tense, there is a large regular class that takes the suffix *-ed*, and several irregular classes that either change their root vowel, stay the same, etc. At first, learners may try to keep all the classes separate, but it soon becomes obvious that one class of past tense verbs is much larger than the others. Taking only the example of the suppletive past tense of *go* below, they then modify the gerund/past tense pair of LexiBlocks as follows:

(108)



Hence, because it would be very tedious to repeat a second stem list with just those stems that use the regular past tense suffix, it is more economical to simply tag all the verbal stems, and only

list the smaller irregular classes above. Because the irregulars will thus be ranked higher on the Expanded Lexicon (see previous chapter), they will be chosen before their regular equivalents.

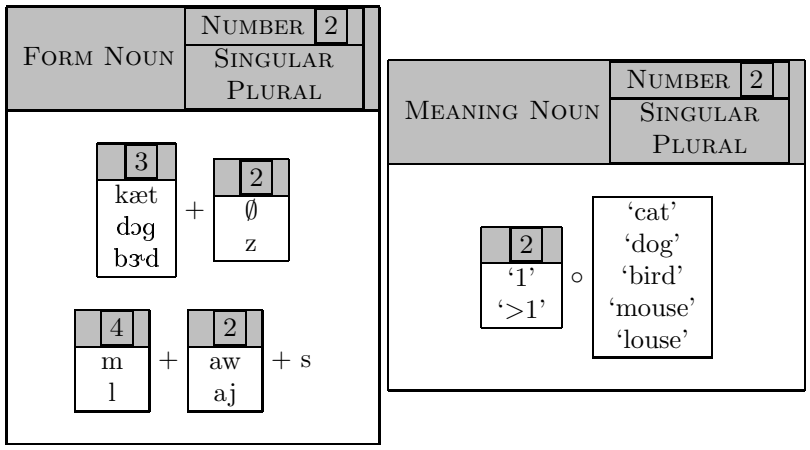
Ordering specific cases before the general case is often called Pāṇini's principle, and it is used by theories as different as Distributed Morphology and Paradigm Function Morphology.

V The Integration Step

The fifth and final step is called the Integration Step and involves merging several Word Constructions in order to achieve more economy. Three conditions must be met. 1) The resulting CWC must be well-formed. 2) There should be a non empty string s1 that occurs in the same prosodic position of all the relevant words in both CWCs. 3) The "stem structure", that is, those LexiBlocks that encapsulate the phonemic strings that are associated with the lexical meanings, must be the same.

For example, in (102), the strings in Reg and Irreg respectively occur in the same prosodic positions in both the singular and plural CWCs and the stem structure is identical. And since (109) is well-formed, it is then licensed:

(109)



The CWCs in (108) could not be integrated once the Elsewhere Step has applied, because they do not have the same stem structure. For convenience, I summarize below the five steps of the acquisition procedure described so far.

(110) **CWC Acquisition steps**

a. The Word Step: Words that share exactly the same morphosyntactic categories are stored in common Form and Meaning CWCs.

b. The Connection Step: 1) The form of a word (w1) in CWC A is included in the form of another word (w2) in CWC B or there is a string of phonemes s1 strictly included in w1 and a non empty string s2 strictly included in w2 such that the substitution of s1 by s2 in w1 gives w2. 2) There is a single categorial relationship between the meanings associated with w1 and w2. When these two conditions are met, all pairs of words that share the same difference are embedded in LexiBlocks, with the relevant strings and parts of meanings factored out.

c. The Sharing Step: Three successive phases: 1) in the Form constructions, edge segments shared by a significant number of words or LexiBlocks in a LexiBlock established in the previous step are factored out, while in the Meaning constructions, words with similar meanings are grouped together; 2) edge segments shared by a significant number of words or LexiBlocks in a LexiBlock that are grouped together in the corresponding Meaning construction are factored out; 3) edge segments shared by a significant number of words or LexiBlocks in the remaining LexiBlocks are factored out.

d. The Elsewhere Step: When a class of words or stems in a CWC A must be split into several classes in another CWC B, forcing one to re-list them, the exceptional (i.e., less numerous) classes are listed first in B, and then the entire class of words or stems from A is repeated by tagging.

e. The Integration Step: Two CWCs are merged using LexiBlocks in the label rectangle if 1) the words in two CWCs all share a non empty string that occurs in the same prosodic position; 2) the resulting Word Construction is well-formed; 3) their stem structure is the same.

3.3.2 Lexical Insertion Conditions

So far we have accounted for how patterns are extracted from learned words. This implies that TCWC offers two ways of accounting for the relationships between words. Either the words forming the patterns are learned and the pattern is extracted, or newly learned words are inserted into existing CWCs, generating related words, with the pattern that had been previously extracted.

(111) Two ways to account for word relationships in TCWC

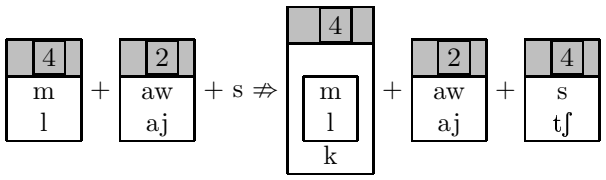
Learn *bird*, *birds*	→	Extract CWC-1
Learn *dog*	→	Insert in existing CWC-1 to generate *dogs*

If we didn't have the second possibility (of inserting words in existing CWCs to derive and inflect them), we would basically be claiming that speakers must learn every word in their language before applying the five steps discussed above. What we need to assume then, is that the acquisition steps take place after a certain number of words are learned, or as words are learned. Once the acquisition

steps have taken place and the speaker is equipped with CWCs, we need to account for how new words may be inserted inside a CWC in which they fit. In order to do so, I propose three Lexical Insertion Conditions, the first of which I state below:

(112) Generalization Preservation: Do not insert a word in a LexiBlock if this would result in removing the concatenation of a LexiBlock with a phoneme.

This condition allows us to preserve the concatenation of a LexiBlock with a phoneme string. For example, inserting the word *couch* in the LexiBlock that hosts *mouse* and *louse* would violate Generalization Preservation. Indeed, in order to insert *couch* in the same LexiBlocks as *mouse*, we would need to modify the structure as follows, losing the generalization that all the words in this class ends in /-s/:

(113) 

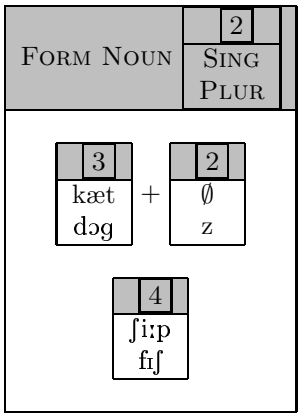
However, if there were a language called pseudo-English where the plural of *couch* was /kajtʃ/, then this form would be learned and would be stored during the Word Step with the other plurals, and then by the other steps, the LexiBlock above would have a different shape.

Our second Lexical Insertion Condition is Counter-Evidence Respect:

(114) Counterevidence Respect: If there is evidence in the CWCs of a language that a given word should not be inserted in a given LexiBlock, then it isn't.

It is predicted that some words will ambiguously fit two or more CWCs. For example, the word *deer* could be inserted either with the regular nouns or with the nouns that have identical singulars and plurals.

(115)



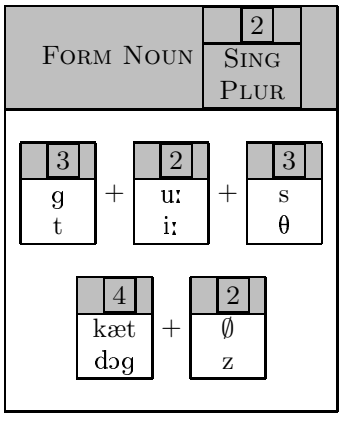
However, when speakers learn the *plural deer*, then they have no choice but to insert it with *sheep* and *fish*, because of Counterevidence Respect.[9]

The third and final Lexical Insertion Condition is Localized Generalization:

(116) Localized Generalization: When inserting a word in a CWC, it is preferred to insert phonological information in as few LexiBlocks as possible.

The CWC in (117) stores the forms *goose*, *geese*, *tooth* and *teeth*. Crucial to understanding it is to pay attention to the boxed numbers 2 and 3. The two LexiBlocks tagged with 3 associate /g/ with /s/, and /t/ with /θ/. Because the middle LexiBlock is numbered 2, any of the two vowels may appear between the consonant pairs. (The number-attributing LexiBlock is also numbered 2, thus giving a value to the inflected words that use the corresponding vowels).

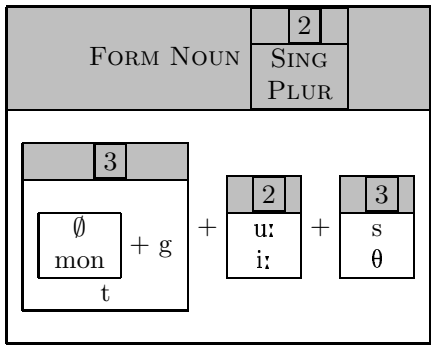
(117)



[9] As we will see in §3.12, the principles by which words get attracted to one class instead of another probably have to do with factors such as frequency, class size (productivity) and a LexiBlock's specificity.

In the case of a discontinuous LexiBlock such as (117), LOCALIZED GENERALIZATION states that new phonological information should not be inserted in two LexiBlocks if it can be inserted in only one. For instance, the word *proof* /pru:f/ could not be inserted with *goose* and *tooth* in (117) (and yield a plural **preef* /pri:f/) because strings would have to be inserted in two LexiBlocks: the string /pr/ would have to be inserted in the leftmost LexiBlock (with /g/ and /t/) and the string /f/ would have to be inserted in the rightmost LexiBlock (with /s/ and /θ/). By inserting *proof* with *cat* and *dog*, we do not violate LEXICAL INSERTION. However, the word *mongoose* could be inserted anywhere in (117), and allow a plural *mongeese* as attested in the Oxford English Dictionary:

(118)



Another word that is attested with such a plural is *moose/meese*. What LOCALIZED GENERALIZATION does in this case, is that it gives an extra choice to words that are more similar to *goose* and *tooth*. Because *mongoose* and *moose* rhyme with *goose*, LEXICAL GENERALIZATION cannot prevent them from using the vowel changing strategy. Of course, speakers eventually learn that *moose* and *mongoose* have a regular plural, and the regular class, with its higher numbers, has an advantage, but the possibility of using vowel change is not ruled out. Theories that do not make room for similarity of form to influence the choice between strategies cannot explain why speakers are tempted to form the plurals *meese* and *mongeese*, but not **preef* or **beet*.

(119) **Lexical Insertion Conditions**

a. GENERALIZATION PRESERVATION: New words can only be inserted in LexiBlocks in which they fit, without modifying factored out segments.

b. COUNTEREVIDENCE RESPECT: If positive counterevidence exists that a word should not be inserted in a given LexiBlock, then it isn't.

c. LOCALIZED GENERALIZATION: When inserting a word in a CWC, it is preferred to insert phonological information in as few LexiBlocks as possible.

3.4 North American French dialects

I have chosen for this chapter to examine morphological changes that have shaped the French dialects of North America, based on my recent fieldwork[10] in Louisiana on Cajun and Creole varieties and on the literature on Canadian and Mississippi Valley varieties. Also, these changes are not well known in the general linguistic literature and since the traditional analogical terminology is so well-known, I thought it would at least make it more interesting to examine "new" data.

I will examine several morphological changes that have occurred in the verbal system of the Acadian French variety spoken in Pubnico, Nova Scotia.[11] These changes all exist in other varieties of Acadian French, but I have chosen Pubnico because its verbal morphology is described by Gesner (1985), who has the advantage of listing entire paradigms.[12] Two of these changes have similar but not identical counterparts in Québec French, and I will also examine these when relevant.

Mississippi Valley French—or Missouri French—is a little known variety that was once spoken in the area around St. Louis—area known by the French-Canadians as *Le Pays des Illinois*—where it even had newspapers until the 19th century. The best described oral variety is the one from Old Mines: see Dorrance (1935), Carrière (1937, 1939, 1941), and Thogmartin (1970). In Missouri French, I will examine a change in some Feminine Adjectives. Finally, I will examine a change that has affected some verbs of Cajun French.

3.5 The analysis summarized

Earlier, I introduced five steps by which I claimed Connected Word Constructions (CWCs) are learned and gradually take the shape they have. These five steps are repeated in (120) for convenience. I intend to show that errors on each one of these five steps, as well as errors in Lexical Insertion in CWCs, yield six categories of morphological change that overlap with the traditional notions of folk etymology, contamination, leveling and proportional analogy, but offer a *more restrictive* theory of morphological change that is borne out by the ways in which the French dialects examined have changed (and the ways in which they have not changed).

[10]I acknowledge a Stanford Graduate Opportunity Grant for this purpose.

[11]I will henceforth refer to this variety of Acadian French simply as *Pubnico*.

[12]Another important fact about Pubnico, is that it is one of the rare communities in Nova Scotia where native French speakers still formed the majority in the 2001 census and it is the oldest Acadian settlement that is still inhabited by the descendents of the first settlers (1653). Indeed, most Acadian communities were dispersed between 1755 and 1763. While many eventually ended up in Louisiana, only about 20% of the then 13,000 Acadians later settled what is now Northern New Brunswick, various Quebec counties, Northeastern Maine, and elsewhere in the Gulf of St Lawrence. Over 10,000 Acadians were deported, though some eventually found their way back and were allowed to remain. (Numbers from Cormier 1999).

(120) **CWC Acquisition steps**

a. The Word Step: Words that share exactly the same morphosyntactic categories are stored in common Form and Meaning CWCs.

b. The Connection Step: 1) The form of a word (w1) in CWC A is included in the form of another word (w2) in CWC B or there is a string of phonemes s1 strictly included in w1 and a non empty string s2 strictly included in w2 such that the substitution of s1 by s2 in w1 gives w2. 2) There is a single categorial relationship between the meanings associated with w1 and w2. When these two conditions are met, all pairs of words that share the same difference are embedded in LexiBlocks, with the relevant strings and parts of meanings factored out.

c. The Sharing Step: Three successive phases: 1) in the Form constructions, edge segments shared by a significant number of words or LexiBlocks in a LexiBlock established in the previous step are factored out, while in the Meaning constructions, words with similar meanings are grouped together; 2) edge segments shared by a significant number of words or LexiBlocks in a LexiBlock that are grouped together in the corresponding Meaning construction are factored out; 3) edge segments shared by a significant number of words or LexiBlocks in the remaining LexiBlocks are factored out.

d. The Elsewhere Step: When a class of words or stems in a CWC A must be split into several classes in another CWC B, forcing one to re-list them, the exceptional (i.e., less numerous) classes are listed first in B, and then the entire class of words or stems from A is repeated by tagging.

e. The Integration Step: Two CWCs are merged using LexiBlocks in the label rectangle if 1) the words in two CWCs all share a non empty string that occurs in the same prosodic position; 2) the resulting Word Construction is well-formed; 3) their stem structure is the same.

For each of the five steps and Lexical Insertion Conditions, I will follow the same procedure: 1) based on the nature of the step, I will ask what a potential error in this step might be and what consequences it would have on a morphological system; 2) I will illustrate how specific changes in North American French dialects can be attributed to errors in the said step.

3.6 Word Step changes

The Word Step consists in listing in separate LexiBlocks all forms that bear the same morphosyntactic categories. A potential error in this step might be either to list together forms that should be listed separately (because they have different morphosyntactic categories) or to list separately forms that should be listed together.

The latter case doesn't seem to cause a problem. Suppose that 2Sing Present *ate* in English is classified in two CWCs, one for Terminative aspect, one for NonTerminative aspect:[13]

[13]Here, the fact that *ice cream* is a mass noun allows these two readings by the use or non-use of the determiner. The point is to imagine that a speaker considers these two readings to also depend on the aspect attributed to *ate*.

(121) a. I ate the ice cream. (i.e. I ate all of it)

b. I ate ice cream. (i.e. some was left)

Certain languages, like Mayan,[14] mark this distinction overtly with a separate morpheme on the verb. If an English learner were to classify the two *ate* forms above in two separate LexiBlocks, they would later get merged by the INTEGRATION STEP, because they would be identical, so it wouldn't have any impact on the (form-generating) morphology of this speaker.[15]

If however a speaker were to start putting in one LexiBlock forms that bear a different morphosyntactic category in the target language, then this might lead to a significant change. This is what must have happened in Pubnico where the SUBJUNCTIVE and the CONDITIONAL are falling out of use.

First, I will explain how these tenses differ respectively from the INDICATIVE and the FUTURE in Standard French for 1STGROUP and 2NDGROUP verbs (the two biggest classes of verbs). In Standard French, 1STGROUP verbs differ in the SUBJUNCTIVE PRESENT from the INDICATIVE PRESENT only in the 1PLUR and 2PLUR, by the presence of a morpheme /-j-/:

(122) **Present Indicative and Subjunctive of French *manger* 'eat'**

Present	Indicative	Subjunctive
1Sing	mãʒ	mãʒ
2Sing	mãʒ	mãʒ
3Sing	mãʒ	mãʒ
1Plur	mãʒ + õ	mãʒ + j + õ
2Plur	mãʒ + e	mãʒ + j + e
3Plur	mãʒ	mãʒ

For verbs of the 2NDGROUP (over 300 verbs), the SUBJUNCTIVE PRESENT is more distinct in that these verbs use an alternate stem bearing an additional suffixed /-s/ that is also used in the plural persons of the indicative present.

(123) **Present Indicative and Subjunctive of French *finir* 'finish'**

Present	Indicative	Subjunctive
1Sing	fini	fini + s
2Sing	fini	fini + s
3Sing	fini	fini + s
1Plur	fini + s + õ	fini + s + j + õ
2Plur	fini + s + e	fini + s + j + e
3Plur	fini + s	fini + s

[14] I thank Judith Tonhauser for sharing her knowledge of Mayan with me.

[15] I do not wish to imply that categories such as TERMINATIVE are universal (nor that they aren't). All I am saying is that if for one reason or another English learners at some point assume a TERMINATIVE/NONTERMINATIVE distinction, this will have no serious consequence on their CWCs.

As for the SUBJUNCTIVE IMPERFECT, most speakers of French do not use it in everyday speech, and it is considered a highly literary tense. Instead, most French speakers use forms identical to the SUBJUNCTIVE PRESENT. Some Acadian dialects have preserved this tense, but this is not the case in Pubnico, according to Gesner (1985:13), who only found one instance of it in his corpus.

For the SUBJUNCTIVE PRESENT, Gesner (1985:14) gives a table, where it is shown that 60% of the time,[16] Pubnico speakers use PLURAL SUBJUNCTIVE forms identical to the INDICATIVE PRESENT. He attributes the presence of the standard forms (and also of some "mixed" forms in the 3PLUR) to the influence of Standard French in school. While there is reason to doubt this claim,[17] it appears to be at least true that the SUBJUNCTIVE PRESENT is indeed being overtaken by the INDICATIVE PRESENT in Pubnico, much like the SUBJUNCTIVE IMPERFECT has given way to the PRESENT in Pubnico as in many French-speaking areas.

(124) **Main Pubnico Present Indicative and Subjunctive pattern**

Present	Indicative	Subjunctive
1Sing	fini	fini + ∅/s
2Sing	fini	fini + ∅/s
3Sing	fini	fini + ∅/s
1Plur	fini + s + ɔ̃	fini + s + ɔ̃
2Plur	fini + s + e	fini + s + e
3Plur	fini + s	fini + s

The first observation I want to make is that the SUBJUNCTIVE, by virtue of occurring in subordinate clauses, is less frequent than the INDICATIVE. As Clark (1985:700-701) notes, the SUBJUNCTIVE is acquired later than the INDICATIVE in French. Secondly, as the paradigms in (122) show, among the verbs of the 1STGROUP, only the 1PLUR and 2PLUR have a different SUBJUNCTIVE form from the INDICATIVE PRESENT. The 1STGROUP contains thousands of verbs. The 2NDGROUP contains about 300 verbs, while the so-called 3RDGROUP contains about 350 verbs. The 2NDGROUP, as we saw, has more distinct SUBJUNCTIVE forms. Traditionally, the remaining verb classes are lumped together under the label 3RDGROUP. This is not a group or class in the same sense as the other two however, since it actually consists of about 60 irregular classes that behave differently. Nevertheless, about 10% of 3RDGROUP verbs behave like the 1STGROUP as far as a distinctive SUBJUNCTIVE

[16] I computed the statistics myself (n=229).

[17] At least one speaker uses the SINGULAR SUBJUNCTIVE PRESENT forms of Standard French and has even extended the consonant-final pattern to some vowel final 1STGROUP verbs. Also, in the introduction to the study, it is explicitly said that Roselle d'Entremont, the main interviewer on the project, recognizes this pattern (apparently from her wider experience with the community).

is concerned, while the rest behave more like the 2ndGroup. It is true that 3rdGroup verbs represent the majority of produced verbs by children acquiring French[18] (which is what allows them to preserve their irregular conjugation). However, because the 3rdGroup is actually a collection of 60-odd conjugations that must be learned independently, the 1stGroup is in fact the consistent *pattern* that children are exposed to more often. It is not surprising then that children first tend to regularize 3rdGroup verbs on the model of 1stGroup verbs (Clark 1985:702-703). Given that the 1stGroup only marks an Indicative/Subjunctive distinction in the 1Plural and 2Plural, it is not a stretch to conclude that at least some learners might not notice at first that there is an Indicative/Subjunctive distinction in French.

In TCWC, this would mean that during the Word Step, some speakers would not have a separate LexiBlock to store the Subjunctive forms. If this situation is not corrected before the other steps apply, the learners are left with a grammar without a distinctive Subjunctive. If the learners notice the distinctive Subjunctive forms later, it is not too late to build a new CWC, however, especially if this exposure occurs in school, the CWC might be marked with a special category such as Formal or Literary.

In the case of Pubnico, the existence of some non-standard yet distinctive Subjunctive forms suggests that the less frequent Subjunctive forms are not strictly attributable to the influence of Standard French in school.[19] Nevertheless, the explanation offered above still holds: once some learners have concluded that there is no Indicative/Subjunctive difference, they will generate forms which will feed the CWCs of other learners and eventually, some speakers will at least relax the syntactic requirements in a way that both the Indicative and Subjunctive forms are allowed to occur in subordinate clauses, thus accounting for the variation observed.

In the case of the Pubnico Conditional, Gesner (1985:19) notes that 79.6% of the time, the forms are identical to the Future forms. Another interesting fact is that, unlike the Subjunctive, there are very few Conditional forms in Pubnico that are not either identical to the Pubnico or Standard French Future or Imperfect form, or identical to the Standard French Conditional. For example, look at the inflections of the verb *dire* 'say':

[18] See the table in Clark (1985:704), data from Guillaume (1927).
[19] We will come back to this in §3.10

(125)

	Pubnico Future	Standard Future	Pubnico Conditional	Standard Conditional
1Sing	diz + r + e/a	di + r + e	diz + r + a/di + r + e/a	di + r + ɛ
2Sing	diz + r + a	di + r + a	diz + r + a/di + r + a	di + r + ɛ
3Sing	diz + r + a	di + r + a	diz + r + a/di + r + a	di + r + ɛ
1Plur	diz + r + ɔ̃	di + r + ɔ̃	diz + r + ɔ̃/di + r + (j) + ɔ̃	di + r + j + ɔ̃
2Plur	diz + r + e	di+ r + e	diz + r + e/di + r + (j) + e	di + r + j + e
3Plur	diz + r + ɔ̃	di + r + ɔ̃	diz + r + ɔ̃/di + r + (j) + ɔ̃	di + r + ɛ

In (125), it is striking that the non-standard stem /diz-/ is never used with the standard CONDITIONAL person suffixes.[20][21] (The form diz-rjõ never occurs). Therefore, given that 1) homophony between the CONDITIONAL and FUTURE forms are 33% more common than homophony between SUBJUNCTIVE and INDICATIVE forms; 2) there does not appear to be any indication of a traditionally inherited CONDITIONAL that builds on the innovative local stems; I conclude that the CONDITIONAL in Pubnico has merged with the FUTURE and that the distinctive CONDITIONAL forms observed belong to the learned speech.

The problem now is that unlike for the SUBJUNCTIVE, we do not have a form-based explanation for the merger of the FUTURE and CONDITIONAL, since these two tenses are never identical in Standard French.[22] However, although the CONDITIONAL is traditionally described as a mood rather than a tense, the FUTURE and the CONDITIONAL have similar semantics, both being irrealis: they are realized in a time that has not come yet, and in the case of the CONDITIONAL that may or may not come. As we will see in the chapter on Armenian, they both exclusively use the same auxiliary in this language. In both English and French, they are sometimes used in subtly different constructions:

(126) Si j'avais faim, je mangerais.
'If I were hungry, I would eat.'

Si j'ai faim, je mangerai
'If I am hungry, I will eat.'

[20] A notable exception in Gesner's paradigms is /asir/, the local form of standard /aswar/ 'sit', which in the CONDITIONAL shows up with the non standard stem and the standard CONDITIONAL suffixes, e.g. 2PLUR /asi-r-j-e/. However, unlike the non standard stems of *dire*, *finir*, etc., /asir/ is a very wide-spread form in all of the Francophone world, thus it is entirely conceivable that /asi-r-j-e/ is a much older change, or that it penetrated Pubnico via television, teachers who came from outside the community, etc.

[21] Also, the 3PLUR CONDITIONAL form /di-r-j-ɔ̃/ instead of /dirɛ/. This is a separate phenomenon (on which more later), by which Pubnico speakers almost always have homophonous 1PLUR and 3PLUR forms.

[22] I obviously do not wish to imply that (Modern) Standard French is the immediate ancestor of Pubnico. However, Pubnico is a variety of French, and therefore must have originated from a variety whose forms were at some point fairly similar to the less innovative forms of Standard French. Using Standard French as a point of comparison is thus not irrelevant.

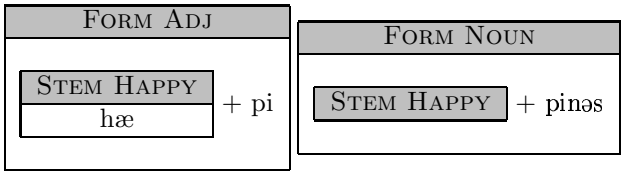
It therefore does not seem implausible that the FUTURE forms' semantics gradually shifted and took over the uses of the CONDITIONAL. Since these two tenses are very infrequent,[23] the same scenario as proposed for the SUBJUNCTIVE is also valid. At some point in the history of Pubnico, speakers categorized futures as if they were also used as conditionals, which prevented them from constructing a separate LexiBlock for futures and conditionals, and therefore started generating futures in contexts that previously required the conditional. This did not necessarily happen at once; it could be that over a few generations of learners, the conditional got pushed back to more and more restricted contexts until it disappeared (or nearly so?) and it was later reintroduced via Standard French.[24]

3.7 Connection Step changes: folk etymology

The second step consists in establishing connections between CWCs by sharing information using LexiBlocks. As in the previous step, there are two types of errors that could occur, only one of which may have diachronic consequences. First, speakers may overlook a difference or a similarity between two words.

For example, if a speaker does not notice that *happy* and *happiness* share not only their first two segments, but their first four segments, then s/he would posit the following CWCs:

(127)



The only consequence this would have for this speaker would be that from *happy*, s/he could not generalize the "suffix" *-(pi)ness* to as many other adjectives.[25] Of course, in the case of this suffix, the speaker will hear so many *-ness* nouns that it would be a particularly odd situation if s/he systematically failed to notice that the relevant string to be factored out is *-ness*; the mistake should easily be fixed after having learned other words. Therefore, failing to see every similarity between *happy* and *happiness* does not entail any significant diachronic morphological consequences: the morphological strategy would be limited to adjectives in /-pi/, but the learning of many other

[23] In Gesner's corpus, they each represent less than 1% of the attested forms.

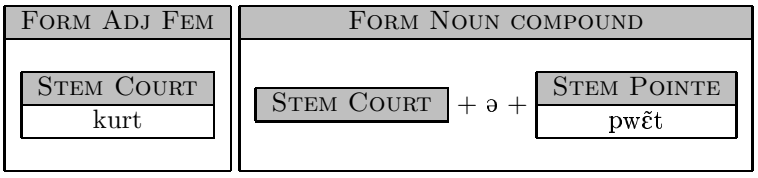
[24] It is well-known that Nova Scotia Acadians are relatively isolated from the main French-speaking areas of Canada located essentially in Quebec and adjacent areas of Ontario and New Brunswick.

[25] This is due to the Lexical Insertion Condition called GENERALIZATION PRESERVATION.

pairs would make it improbable that a learner would systematically fail to isolate the suffix *-ness* correctly.[26] In the case of a less productive suffix though, the speaker would simply be using it in an even less productive fashion, because of the additional limitations s/he is imposing on it.

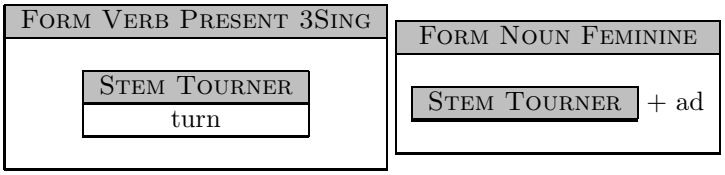
On the other hand, if the speaker fails to notice a difference between two words, this can lead to a situation that is known as folk etymology. For example, in Standard French, there is a word *courte-pointe* meaning 'quilt', whose form used to be *coute-pointe*. Here, speakers have noticed that *pointe* is the FEMININE of the word *point* (meaning 'stitch') and have assumed that the first part is the FEMININE ADJECTIVE *courte* meaning 'short', thus yielding the CWCs below. They have then failed to see a difference between the strings /kut-/ and /kurt/.

(128)



Folk etymology is not restricted to compounds. For example, I have heard more than once the word *tournade* instead of standard *tornade* 'tornado'. Here, these speakers have assumed that the stem is the same as in the verb *tourner*, meaning to 'turn' or 'spin'.

(129)



Given that changes due to the CONNECTION STEP are necessarily perception errors, it is not surprising that folk etymology typically affects derivational morphology (which is typically less productive). Indeed, for folk etymology to affect the inflectional system of a language, the same perceptual errors would have to occur over and over again to several lexical items, a situation which should normally be rarer.

Paul Kiparsky (personal communication) points out the case of the PAST and PASTPARTICIPLE suffixes *-ed*, which are etymologically unrelated. This is potentially a good example of an inflectional folk etymology if it is the case that the shapes have been altered because of their semantic relatedness, rather than by coincidence or by the application of regular sound change.

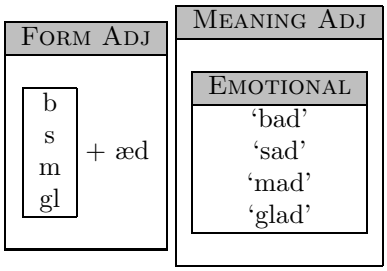
[26] Although in the particular case of *happy*, it could be argued for example that this word might then get a different stress pattern, because of a different word structure.

I also want to stress that four-part analogy would miss the point in a case like *tournade*. Sure, we could write up a proportion *accoler*:*accolade*/*tourner*:X. But this innovation is not an independent one; it is a reinterpretation of *tornade*.

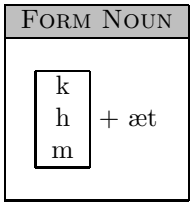
3.8 Sharing Step changes: contamination

Errors in the Sharing Step may lead to contamination. This third step allows speakers to store word forms in a more economical way, by sharing part of affixes or stems. It is akin to the Connection Step, except that it is active within CWCs, instead of across them. Sometimes shared segments occur between words that also belong to the same semantic category, as in (130), or sometimes the words may only share a rhyme without sharing any semantics, as in (131).

(130)



(131)



As in the previous step, if a speaker fails to see that some segments could be shared between two stems or two affixes, the only consequence is that s/he will have stored forms less economically and, in the case of stems as above, perhaps will be a poorer rhymer or will be less able to make word associations, etc.

If, on the other hand, a speaker fails to see a difference between two stems, then this leads to contamination. A standard example is Proto-Romance **gravis* 'heavy' which allegedly became **grevis* under the influence of **levis* 'light'.

(132)

FORM ADJ
gr l + evis

This grouping is favored by Phase 2 of the SHARING STEP, where learners look for similarity of forms based on similarity of meaning. In this case, two adjectives relating to the semantic category of Weight have been grouped together. I found a very similar case in my fieldwork on Cajun and Creole French in Louisiana, where the FEMININE ADJECTIVE for 'light' was sometimes pronounced [leʒɛrd] instead of [leʒɛr] (cf. [lurd] 'heavy-FEM'), and the MASCULINE was sometimes pronounced [leʒɛr] instead of [leʒe], (cf. [lur] 'heavy-MASC').[27]

(133) **Contamination of French *léger* 'light' by *lourd* 'heavy'**

FORM ADJ MASC		FORM ADJ FEM	
WEIGHTS lu leʒɛ	+ r	ADJ MASC WEIGHTS	+ d

(134) **Standard French *léger* 'light' and *lourd* 'heavy'**

FORM ADJ MASC	FORM ADJ FEM
lur	lurd
leʒe	leʒɛr

There are at least two other similar pairs in Québec, Missouri and Louisiana. The first is /myr-t/ 'ripe-FEM' and /puri-t/ 'rotten-FEM', whose standard FEMININE forms lack the /-t/ and are homophonous to the masculine forms /myr/ and /puri/. However, this pair is different in that one of the two FEMININES needs to first have acquired a /t/ by something like four part analogy (which, as we will see shortly, in this theory is accounted for by lexical insertion in the wrong LexiBlock). Another pair is attested in Quebec: /kryt/ 'raw-FEM', which can be analyzed as having acquired its FEMININE /-t/ by contamination from Standard /kɥi-t/ 'cooked-FEM'.

Thogmartin (1979) cites a FEMININE /ʒɔlit/ for /ʒɔli/ 'pretty' in Missouri, for which I have also a single attestation in my Louisiana fieldwork. This could be a similar case, because in Québec,

[27] I should point out that most speakers had at most one of these innovative forms, so the structure below is not meant to be representative of Cajun. It is only meant to represent the innovations. Also, some speakers had a FEMININE form [leʒɛrt], also found in Missouri (Thogmartin 1979) and in Québec, which might be best analyzed otherwise, though the FEMININE [lurt] is also attested in French dialects.

we have /lɛt/ instead of /lɛd/ for 'ugly-FEM', though Thogmartin doesn't cite the FEMININE form of this adjective in Missouri. Unfortunately, we cannot say anything about Louisiana, because the adjective /lɛ(d/t)/ is so scarcely attested and /vilɛ̃-vilɛn/ is the most common adjective used.[28]

Finally, two other words cited by Thogmartin (1979): the FEMININES /mɛjørt/ 'better-FEM' and /nwart/ 'black-FEM', both of which showed up in my Louisiana fieldwork. While I know of no form /blɑ̃ʃt/ for the FEMININE of 'white', the FEMININE /vɛrt/ of /vɛr/ 'green' is standard in French and a FEMININE /pirt/ of /pir/ for 'worse' is also attested in France.

(135) **Contamination in North American French Adjective pairs**

lur	lur-d	'heavy'
leʒɛr	*leʒɛr*-**d**	'light'
myr	*myr*-**t**	'ripe'
puri	*puri*-**t**	'rotten'
kɥi	kɥi-t	'cooked'
kry	*kry*-**t**	'raw'
lɛ	lɛ-d/**t**	'ugly'
ʒɔli	*ʒɔli*-**t**	'pretty'
pir	pir-**t***	'worse'
mɛjør	*mɛjør*-**t**	'better'
vɛr	vɛr-t	'green'
nwar	*nwar*-**t**	'black'

*Only attested in France.

In the previous section, I stated that the errors in establishing LexiBlock connections are due to phonemic misperceptions. While this is certainly a factor, the examples of contamination we know indicate that semantic similarity also plays a role. This is not surprising in TCWC, given Phase 2 of the SHARING STEP, where learners seek similarities of form where there is already similarity of meaning. Although form and meaning are stored in separate constructions, it should be clear that when corresponding FORM and MEANING constructions are more similar, they are easier to process (because their expansions are then more similar). Thus speakers assume that words similar in their semantics should be similar in form and vice versa. Hence, when learning a new word, they first try to store it with semantically similar words, which triggers errors of perception, and conversely when they hear a word whose meaning is unclear, they may assume it has something to do with another word that looks like it in form.

Once again, four-part analogy would miss the point. The innovations examined in this section could be accounted for by proportional analogy, but such an analysis could not capture the semantic

[28] In my dialect, *vilain(e)* normally means 'bad' or 'evil', though the 'ugly' meaning persists in the title *Le vilain petit canard* and the expression *Il/Elle n'est pas vilain(e)* 'S/He is not ugly'.

similarity that correlates with the errors. Why aren't there any innovations */fut/ 'crazy-FEM', or */bõt/ 'good FEMININE'? Granted, these are possible innovations, but the North American French facts suggest that they usually come in semantic pairs. Four-part analogy has no answer.

3.9 Elsewhere Step changes

The ELSEWHERE STEP is the one that allows us to account for facts of suppletion and specific/general strategy ordering. By placing the special cases above the general cases, it is possible to achieve economy of representation and to generate the suppletive or special forms higher up on the EXPANDED LEXICON list, which allows them to be used by syntax instead of their doublets generated lower down by the general strategy.

For example, in Standard French, there is one 2NDGROUP verb that behaves differently in the SINGULAR persons of the INDICATIVE PRESENT. In this inflection, the verb *haïr* 'hate', pronounced in two syllables [a.ˈir], takes on the form /ɛ/, while the stem /ai/ is used for every other inflection of the verb.[29] This suppletion of /ai-/ by /ɛ/ is represented as follows:

(136)

FORM VERB IND PRESENT SING	1 HAÏR STEM	1

ɛ

STEM 2NDGROUP
ai
fini
bati
etc.

If a speaker were to pull out another 2NDGROUP verb, let's say /fini/, and place it above the main LexiBlock with /ɛ/, then this speaker would generate /fini/ twice on the EXPANDED LEXICON list, which does not result in a detectable difference in the output. If however, a speaker fails to notice the special stem used by /air/, then this speaker will only generate /ai/ and never /ɛ/. This seems to be what has happened to virtually every dialect of North American French I am aware of, where /hai/[30] is the normal form for the INDICATIVE PRESENT SING.

Another example comes from the verb /ale/ 'go'. In Standard French, this verb has /vɛ/ as 1SING INDICATIVE PRESENT form, but /va/ for the other SINGULAR persons. This verb then is

[29] Except the 2SING IMPERATIVE, where it is also /ɛ/.
[30] With preservation of an initial etymologically Germanic /h/.

quite exceptional, since not only does it use a different stem in the INFINITIVE and the INDICATIVE PRESENT, but it is also the only verb, apart from 'be' and 'have', to have a distinct 1SING form in this inflection from the other SINGULAR persons.

(137)

FORM VERB INDICATIVE PRESENT	1	2
	1SING SING	GO HAVE BE

2
vɛ
e
sɥi

VERB STEMS
va
a
ɛ
mãʒ
etc.

The form /vɛ/ has been replaced by /va/ for many French speakers in Louisiana, Missouri, Quebec and Pubnico. This again can be accounted for by an error in the ELSEWHERE STEP, whereby speakers failed to single out the suppletive form /vɛ/, or failed to refer to it instead of the lower ranked /va/, which may then be used for any SINGULAR person.

3.10 Integration Step changes

In order to illustrate how errors in this step can lead to morphological change, I will once again resort to the verbal system of Pubnico Acadian French. In this variety, we will be concerned with three changes. First, the suffix /-õ/, which, in Standard French, marks the 1PLUR of most tenses and the 3PLUR of the FUTURE, spreads to mark the 3PLUR of all tenses. The second change is that those classes of verbs that oppose a SHORT STEM to a LONG STEM have generalized the use of the LONG STEM to almost every inflection, except the SINGULAR persons of the INDICATIVE PRESENT. Finally, the 1SING FUTURE suffix /-e/ has been leveled out and replaced by the suffix /-a/, already used by the 2SING and 3SING FUTURE.

As mentioned earlier, the CONDITIONAL and the SUBJUNCTIVE are falling out of use in Pubnico and so, for ease of reading, I will concentrate on the three simple tenses that are the most stable,

that is, the INDICATIVE PRESENT and IMPERFECT, as well as the FUTURE tense. In order to make the demonstration even clearer, I will also only discuss the two largest verb classes, the 1STGROUP and 2NDGROUP verbs. Only about 350 verbs do not belong to one of these two classes, however, for the phenomena examined, they behave either like the 1STGROUP or 2NDGROUP, so ignoring them will allow a simpler demonstration without harming its validity.

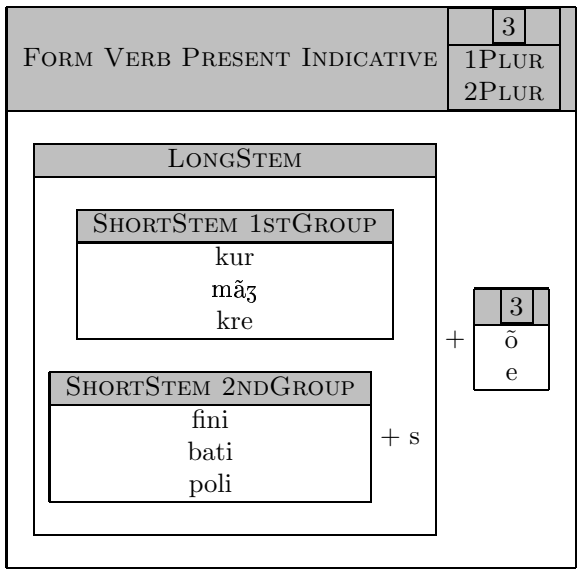
In the INDICATIVE PRESENT of Standard French, both the 1STGROUP and 2NDGROUP use the bare stem for the singular inflections. The PLURAL inflections use suffixes, and the 2NDGROUP also uses a different stem, which I call the LONGSTEM, and which ends in /-s/.

(138) **Standard French Present**

Present	Indicative Gloss	First Group 'run'	Second Group 'finish'
	Sing	kur	fini
	1Plur	kur-õ	fini-s-õ
	2Plur	kur-e	fini-s-e
	3Plur	kur	fini-s

These two paradigms can be integrated by distinguishing between SHORTSTEM and LONGSTEM and defining these as equivalent to respectively SING and 3PLUR.

(139)



ShortStem ≡ Present Indicative Sing
LongStem ≡ Present Indicative 3Plur

The IMPERFECT differs from the PRESENT in that the 2NDGROUP verbs use their LONGSTEM

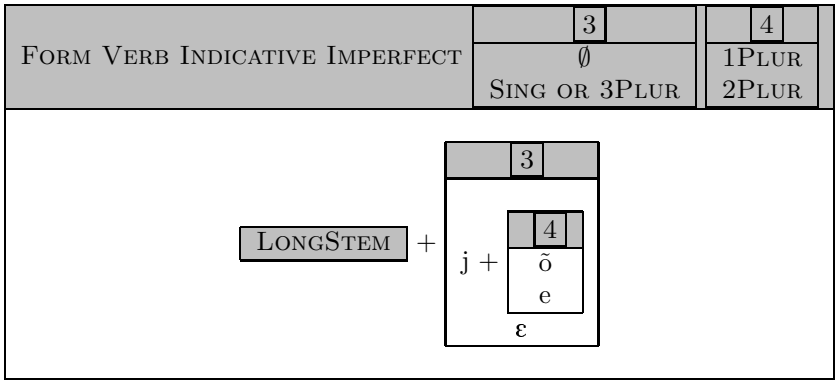
throughout, and in that all groups of verbs use the IMPERFECT suffix /-ɛ/ which has an allomorph /-j/ before 1PLUR and 2PLUR suffixes.

(140) **Standard French Imperfect**

Imperfect	Indicative Gloss	First Group 'run'	Second Group 'finish'
	Sing	kur-ɛ	fini-s-ɛ
	1Plur	kur-j-õ	fini-s-j-õ
	2Plur	kur-j-e	fini-s-j-e
	3Plur	kur-ɛ	fini-s-ɛ

The CWC that describes the IMPERFECT is thus as follows:

(141)



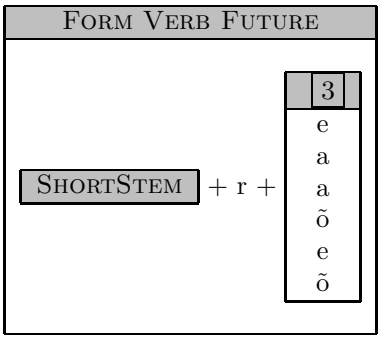
The FUTURE tense of Standard French uses the SHORTSTEM followed by a suffix /-r/ and person-number suffixes that are mostly distinct from the ones used by other tenses.

(142) **Standard French Future**

Future Gloss	First Group 'run'	Second Group 'finish'
1Sing	kur-r-e	fini-r-e
2Sing	kur-r-a	fini-r-a
3Sing	kur-r-a	fini-r-a
1Plur	kur-r-õ	fini-r-õ
2Plur	kur-r-e	fini-r-e
3Plur	kur-r-õ	fini-r-õ

The FUTURE CWC of Standard French is thus described below, and it is the simplest so far.

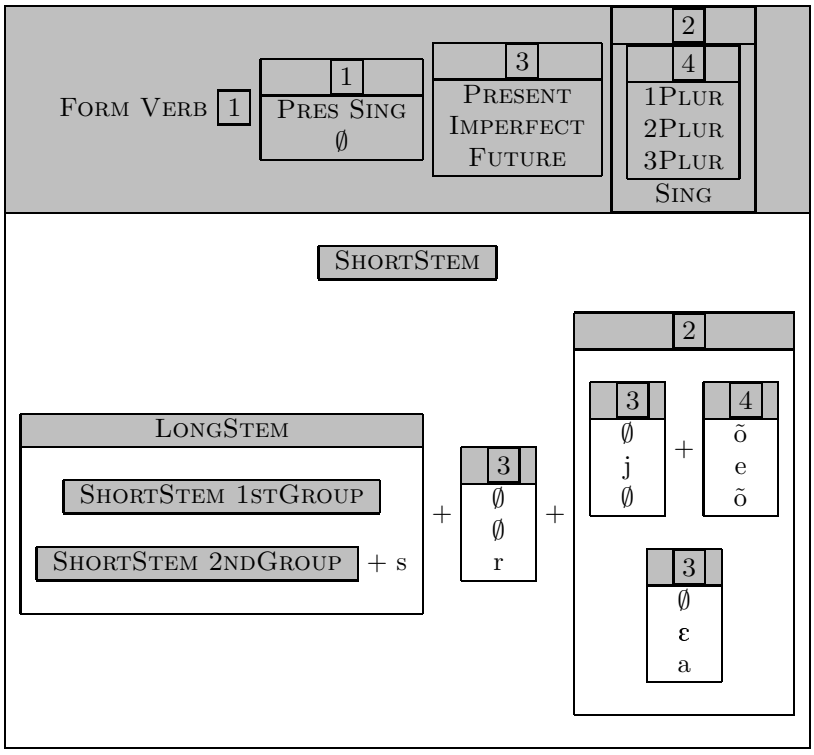
(143)



The Integration Step allows us to integrate (139) and (141), but (143) may not be integrated with either, because it has a different stem structure. If we loosened the requirements for the Integration Step to apply and allowed the integration of the three CWCs seen so far into a single bigger CWC, this could be done, but the resulting structure would be hard to read. This is because the stem structure and the suffix structure are different in all three: 1) while the Imperfect uses the LongStem, the Future uses the ShortStem and the Present uses a combination of both; 2) the person-number suffixes of the Future are quite different from the ones of the two other tenses. For these two reasons, it would be necessary to introduce a lot of extra LexiBlocks in the resulting CWC, which would make it more difficult to read. Nevertheless, even if it is hard for us linguists to read a CWC rich in LexiBlocks, TCWC assumes that learners apply the last step when they can and struggle to achieve some economy of representation.

The changes of Pubnico discussed earlier can then be attributed to speakers simplifying CWC structure with three innovations: 1) by systematically using the LongStem (except for the Sing of the Present Indicative, easily handled by suppletion); 2) by generalizing the use of the Plural set of suffixes of the Future to the other tenses; 3) by eliminating the idiosyncratic 1Sing Future suffix. These three changes each have a reasonable independent motivation. 1) The stem simplification may result from the fact that the ShortStem and LongStem are identical at least for the 1stGroup of verbs; 2) the Plural set of suffixes is nearly identical in all tenses, only the Future set is different in the 3Plur and this 3Plur /-õ/ suffix is actually also used in the Indicative Present of three highly frequent verbs: have, be and go; 3) since only the Future distinguishes between the 1Sing and the other Sing persons, the elimination of the idiosyncratic Future 1Sing allows the Future CWC to be integrated with the others more economically (without an extra LexiBlock). What is striking is that it seems like all these independent changes have conspired to allow the integration of the Pubnico CWC for the three most stable tenses:

(144)



In order to explain the extension of the use of the 2ndGroup's LongStem, let us first recall some observations. The pattern of the 1stGroup is the consistent pattern that learners are most exposed to. In the 1stGroup, the ShortStem and LongStem are identical. Learners may then treat 2ndGroup LongStem as if they should be identical to the ShortStem, and hence generate plurals with the LongStem instead of the ShortStem. Since the Future is rare, and since the 2ndGroup is also not as common as the 1stGroup, it is not unexpected that the 2ndGroup would be modeled on the 1stGroup.

Spreading of 3Plur /-õ/ has a similar explanation. Since three highly frequent verbs have this suffix, it is expected that in the first stages of acquisition, learners might model other verbs on them. Similarly, loss of the 1Sing Future suffix can be attributed to the fact that nearly all verbs in nearly all tenses do not have a separate category for the 1Sing.

Without an Integration Step in the learning procedure, the co-occurrence of these three changes would be coincidental. In traditional terms, the suffix changes examined in this section would be described as some form of leveling, that is, the elimination of "unimportant" paradigm

alternations. They could also be described with four-part analogy, but this would be a very unconstrained version of this mechanism, whereby some tenses would serve as models for others:

(145)

Present 2Sing	Present 1Sing
kur	kur
Future 2Sing	**Future 1Sing**
kurra	_

Such a version of four-part analogy would be too powerful, and would allow for improbable changes, such as the following:

(146)

2Plur Future	mangerez	1Sing Future	mangerai
2Plur Present	mangez	1Sing Present	???mangeai???

If we could use proportional analogy by making two categories vary, then not only would it give it too much power, but such a general notion of four-part analogy would fail to explain why it is the 2Sing and 3Sing that served as model for the 1Sing and not the other way around. In TCWC, it at least makes sense that the 1Sing, being more exceptional, in that it could be considered a case of suppletion, would regularize by letting the general case take over. In the case of the generalization of the 3Plur suffix, certainly it makes more sense to group all the plural persons together, instead of grouping the 3Plur with the Sing persons.

The change affecting the stem is a little trickier. Four-part analogy would account for it nicely as well (see 147); however, it is not clear from either theory alone why it would be that the LongStem used in the Imperfect spread to the Future and not the other way around. The key seems to be in the relative infrequency of the Future. Gesner notes that although this tense is very stable and speakers are much more confident about it than the Conditional, it is one of the most poorly attested in his corpus. This low frequency might help explain why the LongStem of the Imperfect (which, for the 1stGroup verbs is the same as the ShortStem) spread to the Future. This is then an additional reason to look into an incorporation of frequency as a factor influencing CWCs in future research.

(147)

3Sing Imperfect	3Sing Future
mɑ̃ ʒɛ	mɑ̃ ʒra
finisɛ	**finisra**

In conclusion, three peculiarities of Pubnico French morphology can be derived from the integration of CWCs that are not integrated in Standard French. Had Pubnico speakers not integrated

CWCs that are integrated in Standard French, this would only have led to a less economical grammar, but to no noticeable different morphological behavior.

3.11 Lexical Insertion changes

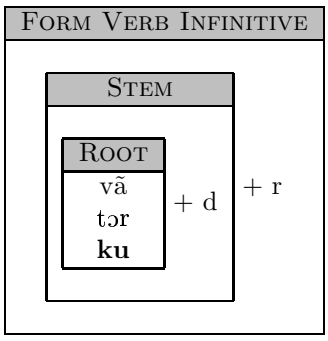
Finally, once the CWCs are built up, learning new words involves inserting them in the existing CWCs in order to store them and be able to inflect them and derive new words from them. Insertion of newly learned words yields changes that are traditionally handled by four-part analogy. For example, the Pubnico inflection of the verb /kudr/ 'sew' may be modeled on the inflection of the verb /vãdr/ 'sell'.[31]

(148)

Infinitive	Pres 1Plur
vãdr	vãdõ
kudr	**kudõ** (instead of standard /kuzõ/)

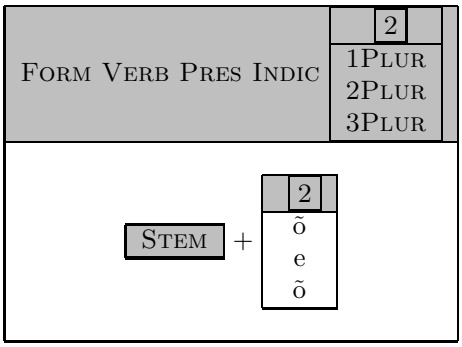
In this section, I intend to show that TCWC, with its Lexical Insertion Conditions, yields once again a more constrained and accurate model of morphological change than four-part analogy. For example, in the case in (148), a four-part analogy account would have to posit the same proportional change for all person-numbers of the verb /kudr/. In TCWC, since the person-numbers are already integrated in a single CWC, it is expected that inserting a stem with other stems will yield the same change for all inflections that use this stem. For example, below, insertion of the Stem /kud-/ with the others automatically yields the three plural persons with the innovative stem.

(149)



[31] I give the innovative form in bold. The standard form is /kuzõ/.

(150)



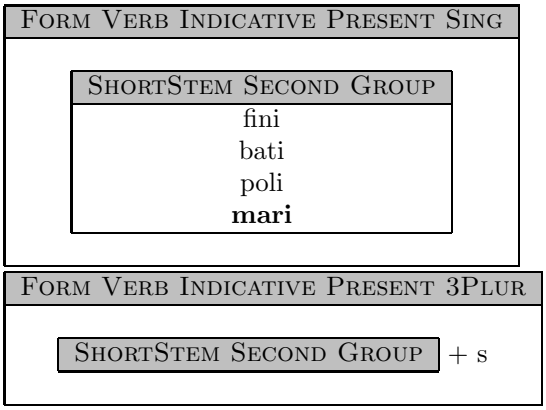
Here, four-part analogy would have to apply between stems, which is undesirable from the perspective of a Word-and-Paradigm type of framework assumed by most linguists who use four-part analogy. TCWC handles the change in a way that preserves the word as the basic unit of morphology, while still accounting for the systematic spread of the innovative stem of /kudr/ to all plural persons.

Such mis-insertions are frequent in Pubnico. However, it is not necessary that the innovative stem spread to all uses. For example the verb /marie/ 'marry' has developed a new stem /maris-/, on the model of /finir/ 'finish' (see paradigm in (123)), but has not completely become a 2NDGROUP verb: its INFINITIVE is still /marie/. If it so happens that the learner learns the form /marie/ after having derived the new stems, then it simply becomes a new verb class. TCWC accounts for the cases where an innovative stem completely takes over the uses of a former stem, but it is not necessary that this always happens. And so, although analogy is usually taken to be a mechanism that simplifies structure, it may actually sometime complicate the system, because it occurs in the course of real-time language acquisition.

(151) **Standard French paradigms**

Infinitive	Present Sing	Present 3rdPlural	Gloss
finir	fini	finis	'finish'
marie	mari	mari	'marry'
mãʒe	mãʒ	mãʒ	'eat'

(152) **Misinsertion of *mari***



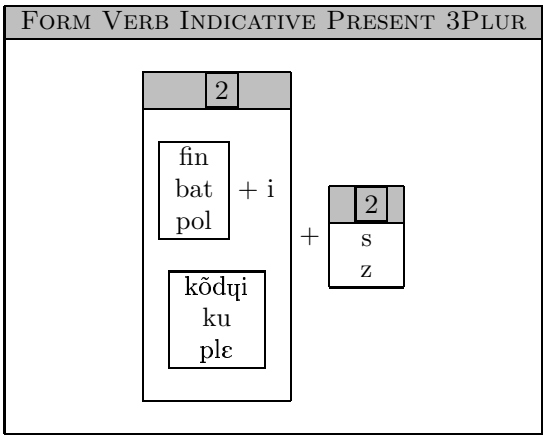
Hence : Sing /mari/ → 3Plur /maris/

(153) **New paradigms after *marier* is learned**

Infinitive	Present Sing	Present 3rdPlural	Gloss
finir	fini	finis	'finish'
marie	mari	maris	'marry'
mãʒe	mãʒ	mãʒ	'eat'

Also, TCWC is once again more constrained than four-part analogy. While the latter would allow for verb stems ending in other vowels to be modeled on /finir/, TCWC would not, because of Generalization Preservation. There is a nice illustration of this in Quebec French. In this dialect, a stem /maris-/ is also attested for /marie/, but the verb /ʒue/ 'play' has an innovative stem /ʒuz-/. The innovative LongStem is formed not by adding an /-s-/ as for 2ndGroup verbs, but a /-z-/, which is not restricted to verbs having a verb stem that ends in /-i/. This is exactly what is to be expected under Generalization Preservation, because inserting the root-stem /ʒu-/ with the stems of the 2ndGroup (which all end in /-i-/) would violate this Lexical Insertion Condition. Likewise, the verb /kree/ does not have an innovative stem */kres-/ or */krez-/, because there are no other roots ending with the vowel /e/ elsewhere in the lexicon of French.

(154)



Finally, in Cajun French, the verb /ekrir/ 'write' has switched classes, going from a verb that forms its LongStem by suffixing /-v-/ to one that suffixes /-z-/. Given the larger 2ndGroup class of verbs, one might expect that /ekrir/, if it were to get rid of its idiosyncratic /-v-/ would use an /-s-/, just like /finir/ and 300-odd other 2ndGroup verbs. A stem /ekris-/ however is blocked by Counterevidence Respect because of the special Feminine PastParticiple /ekrit/, with a suffix /-t/ that this verb shares with a smaller subclass of verbs that use /-z-/, but not with the 2ndGroup.[32]

3.12 Accounting for frequency effects

Studies coming from a cognitive/psycholinguistic perspective have given frequency and related effects a more central role than it had played so far in linguistics—for a basic introduction and referenced, see Fromkin et al. (2003:404-405,436-437). For example, it has been discovered that recently heard words, words that have been primed by a semantic or phonological context, frequent words and word forms that behave in an irregular or suppletive fashion are accessed more rapidly by speakers. That recent words and frequent words are accessed more rapidly is perhaps not surprising from a cognitive perspective, and could be ignored by rule-based frameworks under the provision that such phenomena are indeed real, but are of no concern to the core description of a grammar. The fact that irregular and suppletive morphological forms are also accessed more rapidly on average than the output of productive regular morphological pattern however, has been taken as an argument showing that the line between linguistic and (other) cognitive processes is not so clear-cut. For example, it

[32]Cajun has lost the Feminine suffix /-t/. It is hence necessary to assume that the change affecting /ekrir/ happened before the loss of /-t/. This is not controversial at all, since /ekriz-/ is attested in Pope (1934) for previous stages of French.

has convinced Pinker (1999) and Blevins (2003), two researchers from very different theoretical perspectives, that irregular and suppletive morphology should not be accounted for in the same way as regular productive morphology. While both disagree on how to account for productive regular morphology, they both consider that irregular and suppletive morphological forms are simply listed in the speaker's lexicon and that no rule or constraint accounts for them. Instead, they argue, the forms are listed, and an analogical/connectionist model can handle this type of pattern very well. Of course, connectionist/analogical models can handle both irregular and regular morphology in a unified fashion, but the point that is made by Pinker and Blevins independently is that because they behave differently, regular and irregular morphology should be handled differently (by different mechanisms).

These midway solutions between a traditional rule/constraint-based model that handles all of morphology and a full connectionist/analogical model of morphology sensitive to frequency and other non core linguistic factors do not comply at all with the spirit of TCWC that aims at unifying morphology.

In TCWC, I can show that the theory allows for both a unified account, like full analogical models and traditional rule or constraint-based theories, and that, yet, it can accommodate the observed differences between irregular and suppletive morphology on the one hand, and regular productive morphology on the other, just like the intermediate proposals of Pinker and Blevins. TCWC, I then argue, has the best of all worlds.

The key resides in the fact that TCWC integrates the lexicon within the grammar. Because the lexicon is not listed separately, when we assign a frequency value to a lexical item, it is not difficult to percolate this value to the morphological strategies in which the lexical item is placed. A straightforward way to do this would be that each time a word is used/heard, it is ranked on top of the immediate LexiBlock in which it is located. In this way, when the speaker searches for this word shortly after having used/heard it, s/he can access it faster in the Expanded Lexicon, which is a list of the forms and meanings of the languages lexicon. A frequent word will also usually be ranked fairly high on the Expanded Lexicon list, because there is a good chance it will have been used recently—by virtue of being a frequent word. The Elsewhere Step already ensures that irregular and suppletive forms will be ranked higher in the LexiBlock, so it should come as no surprise that they would be accessed faster by speakers.

Other interesting frequency and related effects exist. For example, Ramscar (2002) shows that words can chose the morphological strategy in which they participate, based on semantic or formal

similarity to other lexical items.[33] Let us not either forget the basic fact that the most productive strategies tend to be the ones that contain a larger class of lexical items.[34] Again, an account of these facts would seem to fit naturally within TCWC. I tentatively propose the following four preference statements on inserting new words in the CWCs of the language:

(155) **Tentative preference statements on lexical insertion**

a. Insert the form of a word by preference in a LexiBlock with similar sounding forms.

b. Insert the meaning of a word by preference in a LexiBlock with similar meaning.

c. Insert the form and meaning of a word by preference in the largest compatible LexiBlock.

d. Insert the form and meaning of a word by preference in the compatible LexiBlock that was used more recently.

This account is obviously not complete. These preferences can clash in various ways: the most recently used LexiBlock is not necessarily the largest, and the latter does not necessarily contain the most similar word forms or meanings. Unfortunately, I must leave to future research to find a way to resolve these conflicts, but at least I have shown that TCWC can in principle account for such phenomena.

To summarize, traditional rule-based models, such as Distributed Morphology, have the advantage of offering a unified account of morphology, but give up on even trying to account for frequency and related effects by not recognizing them as valid linguistic evidence. At the other end of the spectrum, a connectionist analogical model like that of Ramscar (2002) recognizes frequency effects as valid evidence, but questions the distinction between productive regular strategies and irregular or suppletive ones as too simplistic. Ramscar shows evidence that "unproductive" strategies can be quite productive under the correct circumstances.[35] In spite of this, the fact remains that some strategies are more productive than others *in general*, and that it is an advantage for a theory to be able to account for this fact. As for the intermediate Dual-Route model of Pinker (1999) and Blevins' take on Paradigm Function Morphology, their rise is their fall: by sharing the tasks between

[33] Ultimately, these observations should be recognized as reflexes of Saussure's *relations associatives*—see Chapter 5 for more details.

[34] This is not a tautology. A strategy can be productive in the sense that it is the one currently used by speakers to create new forms, but not (yet) contain the largest subgroup of the lexicon. For example, Brown (2003) reports that Louisiana French speakers tend to borrow English verb stems without the suffixes of the larger 1stGroup of verbs. Morin (to appear) reports a similar tendency in the French *langue des cités*. I have noticed the same phenomena in my Louisiana fieldwork, as well as in Montreal French rap songs.

[35] For example, Ramscar shows that the i/a alternation in *drink/drank* can productively be used by speakers with novel words referring to an activity related to drinking.

an analogical/connectionist model and a rule-based theory, they give up any hope of providing a unified account of morphology. In this case then, TCWC succeeds where other theories have failed:[36] by incorporating the lexicon inside the grammar, we can provide a unified account of morphology that allows in a natural way for frequency factors in the lexicon to influence the choice between morphological strategies and accounts for frequency and related effects.

(156)

Model	Frequency effects	Unified treatment	Accounts for different behavior
Distributed Morphology	Peripheral	Yes	No
Paradigm Function M.	Analogy	No	Yes
Dual-Route Model	Analogy	No	Yes
Ramscar	Within the model	Yes	?No
TCWC	Within the model	Yes	Yes

3.13 Summary

In this chapter, I have accomplished three things. First, I have proposed a five-step learning procedure that shows how Connected Word Constructions, as I use them in this dissertation, are learnable. The procedure isn't intended to correspond exactly to the acquisition data available in the literature; it is a first step merely intended to show that the theory is learnable under certain assumptions, some of which might admittedly turn out to be problematic upon further investigation. Along with the five steps, I have proposed three Lexical Insertion Conditions that constrain the ways in which newly learned words may be inserted in the CWCs built by the acquisition procedure. The acquisition procedure and the three insertion conditions by themselves constrain the possible CWCs of the grammar by integrating form and meaning compatibility with locality conditions.

Second, I have shown how errors in the five acquisition steps correlate with traditionally recognized types of analogies: category merger, as folk etymology, contamination, loss of suppletion and leveling, and how the three Lexical Insertion Conditions constrain what is traditionally called proportional analogy. This accomplishment is probably the most significant one, as it unifies various diachronic analogical phenomena, such as folk etymology, contamination and leveling with a single acquisition procedure, it establishes a link between phonesthemes and contamination, and makes much more accurate predictions than traditional four-part analogy.

[36] It would not be too hard for a model like Ramscar's to gain the same advantages by recognizing some general frequency function over the lexicon. TCWC would then be closer to Ramscar's model, though there still remains some hard constraints, like the Lexical Insertion Conditions that have no counterpart in Ramscar's model.

Finally, I have also made some proposals regarding the future developments of TCWC in integrating frequency factors. While much work remains to be done in this area, I have at least shown that TCWC can in principle be sensitive to frequency and related effects, while still providing a unified account of morphology, thanks to its integration of the lexicon inside the grammar. This confirms TCWC as a promising framework.

Chapter 4

Western Armenian Verbs

The first goal of this chapter is to provide an exhaustive account of a morphological system with the Theory of Connected Word Constructions (TCWC). Standard Western Armenian[1] (henceforth Armenian) has a rich verbal morphology that serves this purpose well. In addition to this complexity, another advantage of Armenian verbal morphology, is that although it is well-described by traditional grammars (Kogian 1949, Gulian 1965, Bardakjian & Thomson 1977), a formal account is lacking in the literature, so this chapter brings a novel contribution to formal linguistics in general.[2]

The second goal is to illustrate how some fundamental morphological phenomena work in TCWC. Armenian will allow me to illustrate how the theory can account for facts of allomorphy, morphologically conditioned vowel changes, suppletion, syllable-based morphological phenomena, as well as a highly marked case of double morphology. In the latter case, it will turn out that a principle that I have already introduced, the SHARING STEP in the acquisition of Connected Word Constructions (CWCs), is sufficient to explain why this particular type of double morphological marking is rare cross-linguistically.

[1]Standard Western Armenian is the standard language used mainly by Armenians of the Near East (Lebanon, Jerusalem, Syria, Turkey, Egypt...), Europe and North America. Standard Eastern Armenian is the standard among Armenians of the Republic of Armenia and the former USSR (Nagorno-Karabakh, Russia, Georgia...), Iran and India. In the last 15 years, many Eastern speakers have immigrated to the West, most notably in the area of West Hollywood/Glendale in California.

[2]A much more rudimentary version of this chapter was published as Baronian (2002).

4.1 Armenian

Armenian is mainly a suffixing language. While Classical Armenian[3] is considered largely inflectional, it is sometimes said that Modern Armenian (both Eastern and Western) has taken an agglutinative direction, apparently under the influence of Turkish. For example, Trask (1996) cites examples from the nominal number and case system. In this case, highly fusional suffixes have been replaced by rather clear-cut morphemes.

(157) **Armenian case-number changes** adapted from Trask (1996:310-311)

Inflection of "cer" 'old man'

	Classical			Modern Eastern	
Case	Sing	Plur	Case	Sing	Plur
Nom	cer	cerk'	Nom, Acc	cer	cer-er
Acc, Loc	cer	cer-s	Gen, Dat	cer-i	cer-er-i
Gen, Dat, Abl	cer-o-y	cer-o-c	Abl	cer-ic	cer-er-ic
Instr	cer-ov	cer-ov-k'	Instr	cer-ov	cer-er-ov
			Loc	cer-um	cer-er-um

Whether or not this is truly due to an influence of Turkish, the verbal system does not provide us with such an overt resemblance with Turkish. Modern Armenian verbs have undergone changes, but if these were influenced by Turkish, it must be in a more subtle manner. Modern Armenian has preserved the complex aorist system with its suppletive roots and stems (though this category is probably best characterized, semantically, as a simple past now), as well as passive and causative morphology, though their use is now more limited, not to mention the basic Indo-European verb structure: Root + Theme Vowel + Suffixes.

Western Armenian has innovated by introducing a new INDICATIVE prefix /g(u)-/, and has replaced the Classical future with an isolating structure using the auxiliary /bidi/. The Classical subjunctive has been replaced by the Classical INDICATIVE.

Before diving into my theory-specific account of Armenian, I will go through the conjugation of three sample verbs in a traditional fashion, in order to familiarize the reader with the structure of the language and its relative complexity.

The Armenian stem consists of the root, followed by a theme vowel. For example, the stem of the verb 'sleep' consists of the root /barg-/ with the theme vowel /-i-/. The verb 'eat' has the root /ud-/ and the theme vowel /-ɛ-/, while the verb for 'cry' consists of the root /l-/, followed by /-ɑ-/. With the addition of the suffix /-l/, we get the INFINITIVE, also known as the verbal noun.

[3]Also known as *Grabar* 'the written word'.

(158) **Infinitive**

	Stem (Root+ThV)		
Gloss	Root	Theme vowel	Infinitive
'sleep'	barg-	-i-	-l
'eat'	ud-	-ɛ-	-l
'cry'	l-	-ɑ-	-l

The simplest conjugation is the SUBJUNCTIVE PRESENT. In this tense, the verb stem is used with different person-number suffixes. The 3SING does not have any suffixes (and thus uses the bare stem).

(159) **Subjunctive Present**

	Sing	Plur	Sing	Plur	Sing	Plur
1	barg-i-m	barg-i-nk	l-ɑ-m	l-ɑ-nk	ud-ɛ-m	ud-ɛ-nk
2	barg-i-s	barg-i-k	l-ɑ-s	l-ɑ-k	ud-ɛ-s	ud-ɛ-k
3	barg-i-	barg-i-n	l-ɑ-	l-ɑ-n	ud-ɛ	ud-ɛ-n

The SUBJUNCTIVE IMPERFECT has different SINGULAR suffixes. Here, it is the 1SING that uses no overt person-number suffix and the 2SING and 3SING both use /-r/. To mark this tense further, an extra suffix /-i/ is inserted before the person-number suffixes (except for the 3SING). The verbs with a theme vowel /-i-/ have the extra peculiarity of changing this theme vowel to -ɛ- in this tense.

(160) **Subjunctive Imperfect**

	Sing	Plur	Sing	Plur	Sing	Plur
1	barg-ɛ-i	barg-ɛ-i-nk	l-ɑ-i	l-ɑ-i-nk	ud-ɛ-i	ud-ɛ-i-nk
2	barg-ɛ-i-r	barg-ɛ-i-k	l-ɑ-i-r	l-ɑ-i-k	ud-ɛ-r	ud-ɛ-i-k
3	barg-ɛ- -r	barg-ɛ-i-n	l-ɑ- -r	l-ɑ-i-n	ud-ɛ- -r	ud-ɛ-i-n

The INDICATIVE PRESENT and IMPERFECT differ from their SUBJUNCTIVE counterparts only in the presence of a prefix /g-/. Thus, /g-barg-i-m/, /g-barg-i-s/, etc. for the INDICATIVE PRESENT of this verb. In this particular case, however, a schwa is inserted by the phonology[4] to break up the initial cluster. Monosyllabic stems, which all have the /-ɑ-/ theme vowel,[5] use the allomorph /gu-/ instead of /g-/.[6]

[4]See Vaux (1998) for convincing arguments in favor of the phonological character of schwa epenthesis in Armenian.
[5]Though only three of the verbs with an /-ɑ-/ theme vowel have monosyllabic stems.
[6]There are two further idiosyncrasies. The verb 'be' uses a different root and does not use the prefix /g(u)-/. There are a handful of verbs that do not use the prefix /g(u)-/, among which 'have', 'know' and 'can', and are then identical to their SUBJUNCTIVE forms.

(161) **Indicative Present**

	Sing	Plur	Sing	Plur	Sing	Plur
1	g(ə)-barg-i-m	g(ə)-barg-i-nk	gu-l-a-m	gu-l-a-nk	g-ud-ɛ-m	g-ud-ɛ-nk
2	g(ə)-barg-i-s	g(ə)-barg-i-k	gu-l-a-s	gu-l-a-k	g-ud-ɛ-s	g-ud-ɛ-k
3	g(ə)-barg-i-	g(ə)-barg-i-n	gu-l-a-	gu-l-a-n	g-ud-ɛ-	g-ud-ɛ-n

The AORIST is the tense with the most morphological phenomena involved. They are summarized below in (162) and can be checked against the examples in (163).

(162) **Facts about the Aorist**

a. Concerning the root:

i) Many verbs such as 'eat' below use a suppletive root or stem in the aorist.[7]

b. Concerning theme vowels:

i) Suppletive roots select the theme vowel a.

ii) As in the IMPERFECT, if the theme vowel is /-i-/ in the present, it changes to -ɛ-.

c. Concerning AORIST suffixes:

i) Unless the root/stem is suppletive, a suffix /-ts/ follows the theme vowel.

ii) When used, the suffix /-ts/ is followed by /-a-/ if the verb belongs to the I-CLASS or by /-i-/ if the verb belongs to any other class.

iii) 3SING verbs that are not suppletive or of the I-CLASS do not use any suffix after /ts-/.

d. Concerning person-number suffixes:

i) 3SING suppletive and I-CLASS verbs use the suffix /-v/.

ii) Other classes use no suffix to mark 3SING.

iii) Otherwise, the same person-number suffixes as in the imperfect are used.

(163)

					Suppletive	
	Sing	Plur	Sing	Plur	Sing	Plur
1	barg-ɛ-ts-a-	barg-ɛ-ts-a-nk	l-a-ts-i-	l-a-ts-i-nk	gɛr-a	gɛr-a-nk
2	barg-ɛ-ts-a-r	barg-ɛ-ts-a-k	l-a-ts-i-r	l-a-ts-i-k	gɛr-a-r	gɛr-a-k
3	barg-ɛ-ts-a-v	barg-ɛ-ts-a-n	l-a-ts-	l-a-ts-i-n	gɛr-a-v	gɛr-a-n

This completes my brief overview of the Armenian verbal system. There are also causative and passive suffixes, a negative prefix, as well as many tenses realized periphrastically with an auxiliary/participle construction, but I consider we have enough of the core data to begin the analysis of Armenian in TCWC.

[7] In the case of 'eat', the root /ud-/ is suppleted by /gɛr/.

4.2 Infinitive, Subjunctive and theme vowels

As seen in the previous section, Armenian has three main classes of verbs distinguished by their respective theme vowel, which may be either /ɛ/, /i/ or /ɑ/.[8] As stated earlier, this theme vowel appears between the root of the verb and the suffixes (when suffixes are required). For example, the infinitive forms of the A-Class of verbs are constructed with the root (the A-Class roots), followed by the vowel /ɑ/, followed by the infinitive suffix /l/:

(164)

Form Verb a-Class Inf Pos	
Root a-Class: unɛn, bɔr, xnt, k, *etc.*	+ ɑl

FORM	GLOSS
unɛnɑl	'have'
bɔrɑl	'scream'
xntɑl	'laugh'
kɑl	'come'
etc.	*etc.*

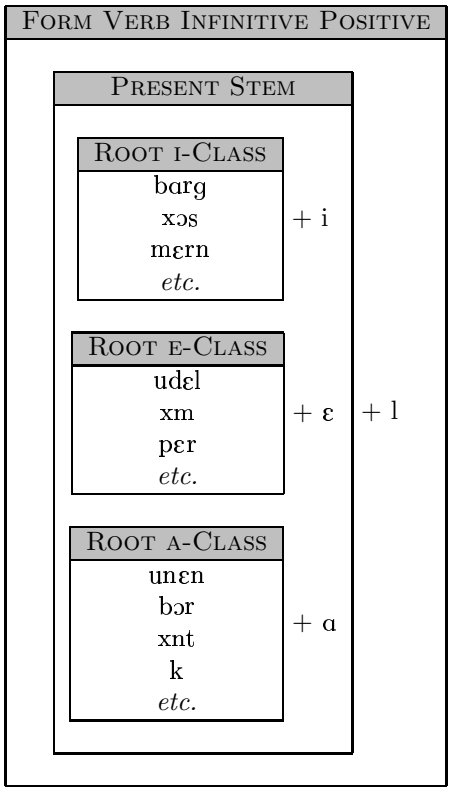
The CWC in (164) stores the verb forms /unɛnɑl/ 'have', /bɔrɑl/ 'scream', /xntɑl/ 'laugh', /kɑl/ 'come' and others. The two other classes are naturally integrated in this CWC since they share the same infinitive suffix /-l/:[9] [10]

[8] There is also a small class of literary verbs with the theme vowel /u/ that were "borrowed" from Classical Armenian and that I will not discuss.

[9] In the Word Constructions after (165), I will not include a line *etc.*; it should nevertheless be understood that I assume that the other verbs that are members of the particular classes are present.

[10] The I-Class verbs are often intransitive.

(165)



FORM	GLOSS
bɑrgil	sleep
xɔsil	speak
mɛrnil	die
etc.	*etc.*
udɛl	eat
xmɛl	drink
pɛrɛl	bring
etc.	*etc.*
unɛnɑl	have
bɔrɑl	scream
xntɑl	laugh
kɑl	come
etc.	*etc.*

Following a traditional conception of Indo-European word structure, I have termed the grouping of the root and theme vowel Present Stem.[11] As we will see in §4.4, the acquisition steps introduced in Chapter 2 naturally bring us to the *Root* and *Present Stem* groupings, which, as we will see throughout this chapter, are used in various morphological strategies of Armenian.

The first illustration of this is the Subjunctive Present 3Sing Positive forms. In this inflection, the bare stem (root + theme vowel) is used. There are two completely equivalent ways of representing this fact in TCWC. First, building on (165), we can simply define Present Stem as being equivalent to Form Subjunctive Present 3Sing Positive. This may seem arbitrary to some and in fact it is. What we are saying is simply that Present Stem is a more convenient expression to use than Form Subjunctive Present 3Sing Positive.

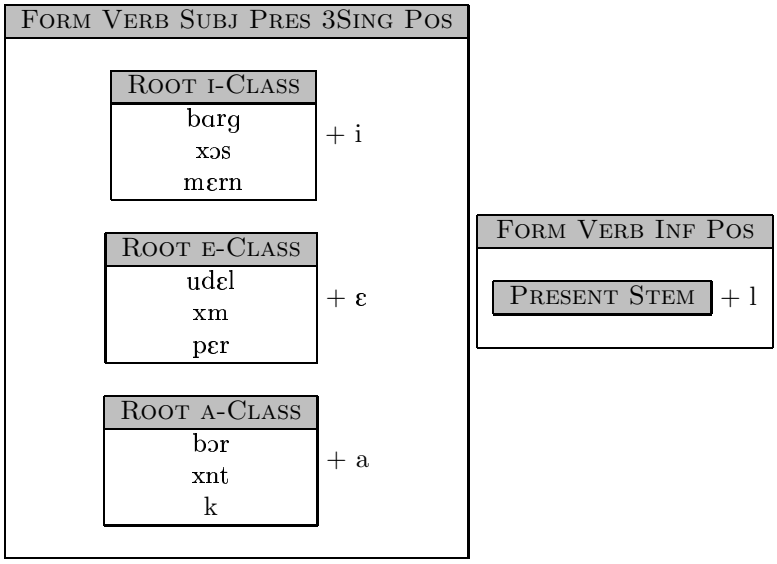
(166) Present Stem ≡ Form Verb Subjunctive Present 3Sing Positive

A second and equivalent way would be to start by describing the Subjunctive Present 3Sing Positive forms and refer to them in the Infinitive construction. The Present Stem equivalence

[11] Halle & Vaux (1998) call *stem* what I call *root*, and *base* what I call *stem*. The difference seems to be simply one of terminology.

introduced above may still be used for convenience, instead of repeating the lengthy categories.

(167)

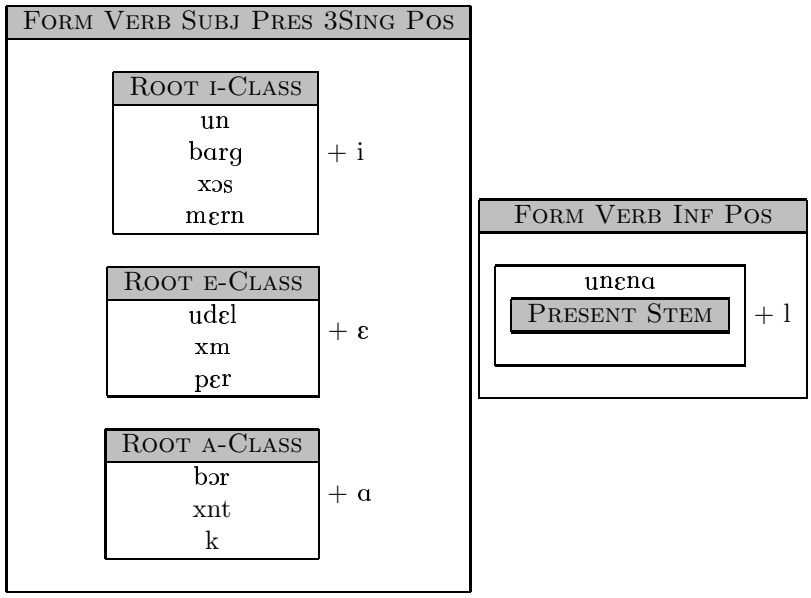


Present Stem ≡ Form Verb Subjunctive Present 3Sing Positive

Thus the full lexical description of the verbs needs only to be stated once. It does not matter if the roots are listed in the Infinitive construction or in the Subjunctive construction. Conceptually, the stems are *shared information* between the two constructions. Such information creates a network of connected constructions.

There is a further complication among the two CWCs we have seen so far. The verb /unɛnɑl/ 'have' uses a different stem in the Subjunctive: /uni/. This is a case of suppletion. In order to account for it, we rewrite (167) as (168):

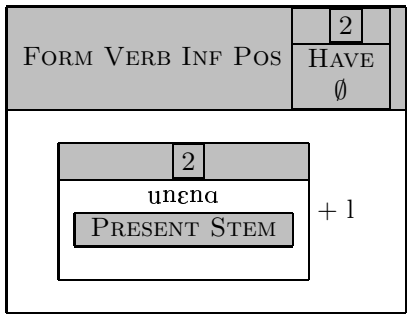
(168)

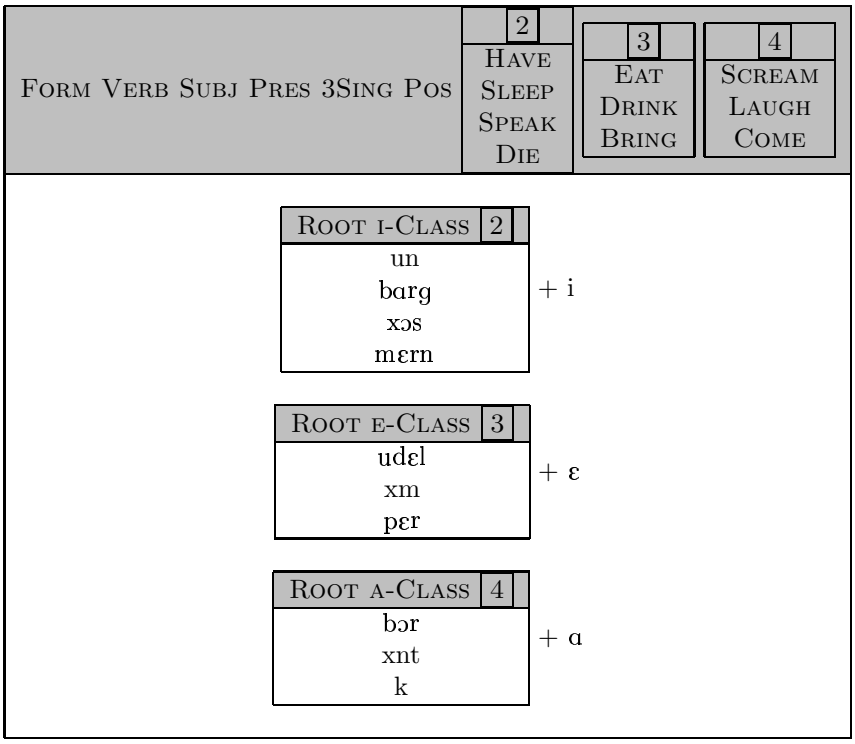


Present Stem ≡ Form Verb Subjunctive Present 3Sing Positive

Before explaining this in greater detail, remember from the abbreviation conventions in Chapter 2, that the label rectangle of the relevant LexiBlocks should have a tagged category associated with lexical items:

(169)





In (168), rewritten as (169), the Subjunctive form /uni/ corresponds to a form /unil/ in the Infinitive CWC. However, a suppletive form /unɛɳɑl/ is listed on top of the other forms in the Infinitive CWC. Given the expansion algorithm assumed for CWCs given in Chapter 2, /unɛɳɑl/ will show up higher than /unil/ on the generated lexical list and will thus be used instead of /unil/.

(170) Subj list = <uni, bɑrgi, xɔsi, mɛrni, udɛ, xmɛ, pɛrɛ, bɔrɑ, xntɑ, kɑ>
Inf list = <unɛɳɑl, unil, bɑrgil, xɔsil, mɛrnil, udɛl, xmɛl, pɛrɛl, bɔrɑl, xntɑl, kɑl>

The two CWCs in (169) generate respectively the lists in (170). As can be seen, both /unɛɳɑl/ and /unil/ are generated on the infinitive list, however, /unɛɳɑl/ is higher up on the list and thus gets chosen instead. This way of handling suppletion, and specific/general cases more generally, is found in many theories of morphology since Pāṇini. The first to recognize this in the generative tradition and to formulate it most explicitly was Kiparsky (1973).[12] Pāṇini's principle is also used in Paradigm Function Morphology and Distributed Morphology. I will illustrate in more detail how TCWC handles suppletion in the section on the Armenian Aorist.

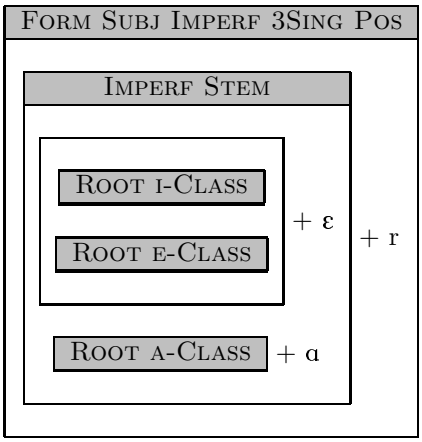
[12]Kiparsky (1973) credits Anderson (1971/1969) and Koutsoudas et al. (1974/1971) for using the principle that in general, specific rules are ordered before general ones.

4.3 Vowel change in the Subjunctive Imperfect

I have described so far the CWCs that account for the INFINITIVE POSITIVE and the SUBJUNCTIVE PRESENT 3SING POSITIVE, two rather simple forms of the Armenian verb. When describing a language in this model, it is useful for pedagogical reasons to start with the barest words, the ones that carry the fewest affixes, then moving along gradually to morphologically more complex forms. For this reason, I will start by showing how the 3SING inflects in each simple tense, before moving along to the other person-numbers. However, keep in mind that the learning procedure is independent of the order in which the words are learned.

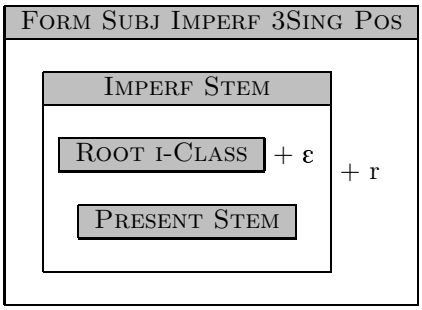
In the SUBJUNCTIVE IMPERFECT now, the I-CLASS roots and the E-CLASS roots are grouped together, sharing a same theme vowel /ɛ/. The three groups of roots then share a final suffix /r/. It would be possible to describe this as follows:

(171)



This representation can be compressed by using the same Pāṇinian mechanism that we used to deal with suppletion. The I-CLASS roots can be singled out (by the ELSEWHERE STEP) and concatenated with their special theme vowel, and then we simply use as the elsewhere case the same PRESENT STEMS that have been defined earlier:

(172)



Thus (172) generates the list in (173). Although the I-CLASS verbs are generated twice, once with the theme vowel /ɛ/, once with /i/, the forms with /ɛ/ are ranked higher on the list and are thus the ones that speakers use.

(173) Subj Imperfect list = <unɛr, bɑrgɛr, xɔsɛr, mɛrnɛr, unir, bɑrgir, xɔsir, mɛrnir, udɛr, xmɛr, pɛrɛr, bɔrɑr, xntɑr, kɑr>

We have just treated a phoneme alternation (that of the theme vowel) in a morphological context (a morphophonological alternation) at the same time as a purely morphological one (suffixation of /-r/). Earlier, I did the opposite in following Vaux (1998) in relaying schwa epenthesis to the phonology of the language. I realize that at this point my choices may seem arbitrary to some. I invite the impatient reader to skip ahead to Chapter 6, devoted to the relation between phonology and morphology assumed in TCWC.

4.4 Acquisition demonstration

At this point, it will be useful to show how the particular shapes I have given to the three constraints follow from the acquisition procedure introduced in Chapter 2. In order to do so, we will work with a reduced version of Armenian, called mini-Armenian, which consists of six verbs, each of which has three forms, the INFINITIVE and the 3SING SUBJUNCTIVE PRESENT and IMPERFECT:

(174) **Mini-Armenian Verbal Lexicon**

Gloss	'scream'	'speak'	'eat'	'have'	'sleep'	'drink'
Infinitive	/bɔrɑl/	/xɔsil/	/udɛl/	/unɛnɑl/	/bɑrgil/	/xmɛl/
Subj Pres 3Sing	/bɔrɑ/	/xɔsi/	/udɛ/	/uni/	/bɑrgi/	/xmɛ/
Subj Imperf 3Sing	/bɔrɑr/	/xɔsir/	/udɛr/	/unɛr/	/bɑrgɛr/	/xmɛr/

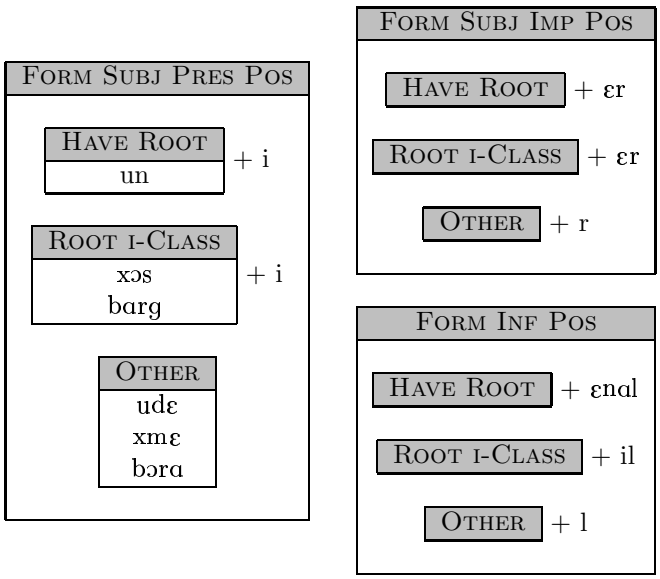
According to the WORD STEP of the acquisition procedure, the forms are first grouped by category, as follows:

(175)

FORM SUBJ PRES POS	FORM SUBJ IMP POS	FORM INF POS
bɔrɑ	bɔrɑr	bɔrɑl
uni	unɛr	unɛnɑl
xɔsi	xɔsɛr	xɔsil
bɑrgi	bɑrgɛr	bɑrgil
udɛ	udɛr	udɛl
xmɛ	xmɛr	xmɛl

According to the CONNECTION STEP now, words are grouped according to their behavior across CWCs. Thus, the A-CLASS and the E-CLASS will be grouped together, while the I-CLASS and 'have' will each stand on its own.

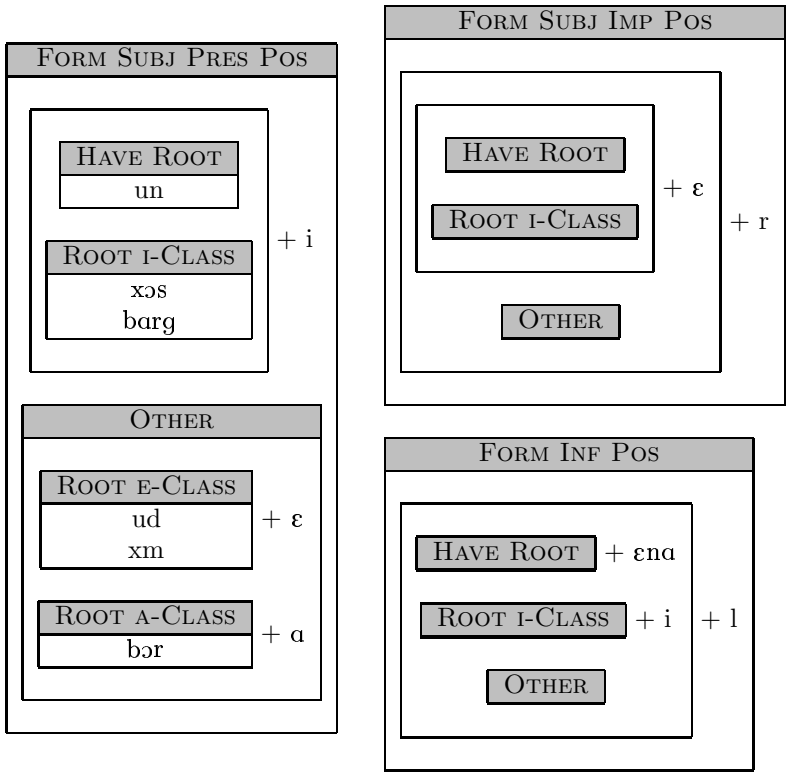
(176)



It is now the SHARING STEP which comes into play and groups similar affixes and subparts of stems or roots together in a Saussurean associative way:[13]

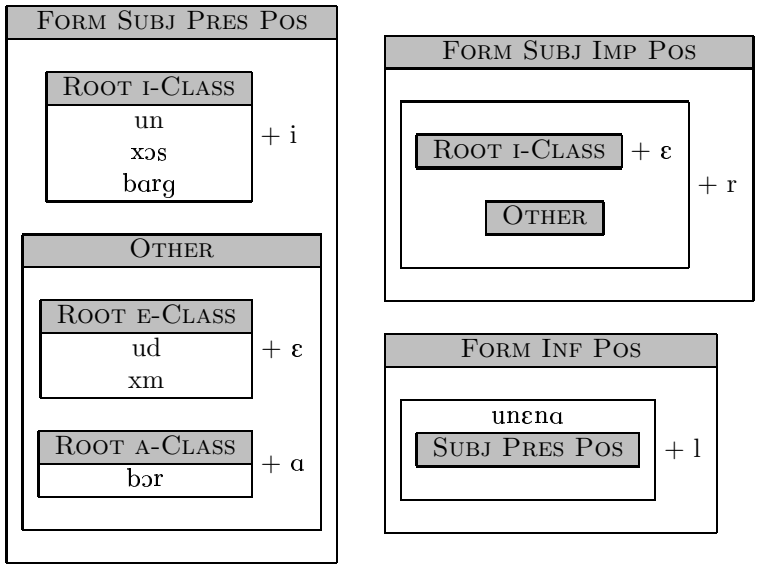
[13]Technically speaking, we are not warranted here to create the A-CLASS ROOT LexiBlock, however, it should be clear to the reader that with the other A-CLASS verbs of regular Armenian, this LexiBlock will be created.

(177)



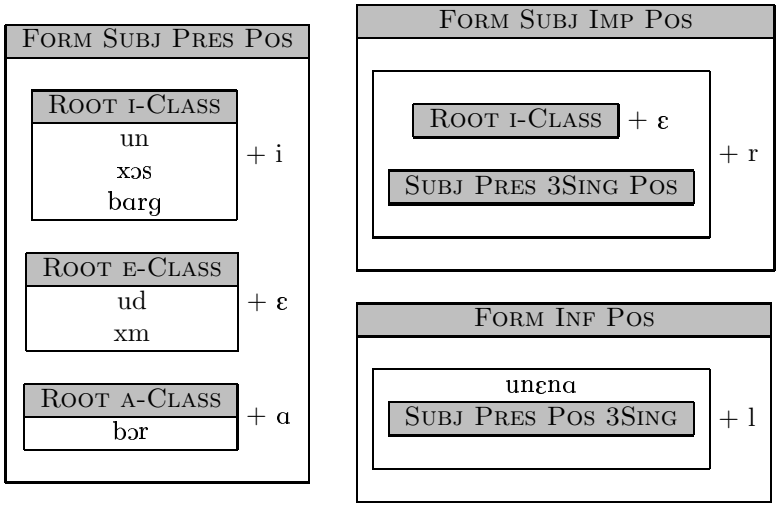
Next, by the Elsewhere Step, we make use of the Pāṇinian principle to gain some more economy of representation. There are two places where we apply this principle. First, in (178), the two Subjunctive CWCs are simplified by "moving up" /unɛnɑl/ in the Infinitive CWC.

(178)



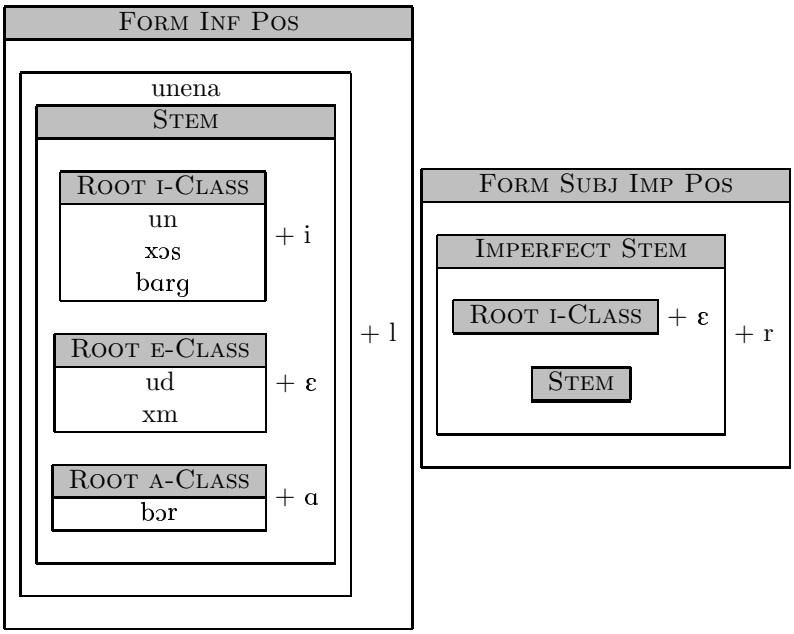
Then, in a second application of the ELSEWHERE STEP,[14] which I have represented separately in (179) for convenience, the INFINITIVE and SUBJUNCTIVE PRESENT CWCs are simplified by applying the Pāṇinian principle to the SUBJUNCTIVE IMPERFECT CWC.

(179)



Finally, the PRESENT and INFINITIVE are merged by the INTEGRATION STEP, but not the imperfect, since it does not share the same stem structure.

(180)



STEM ≡ SUBJ PRES 3SING

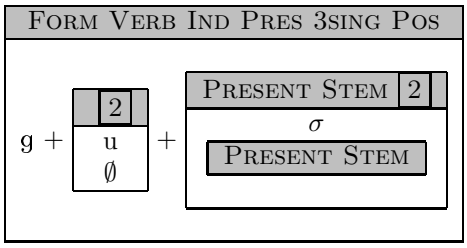
[14]The order does not matter.

4.5 Syllable-based allomorphy in the Indicative

4.5.1 The Indicative Present

In the 3SING INDICATIVE PRESENT, monosyllabic stems select the prefix /gu-/,[15] while the others select the allomorph /g-/ of the same prefix.[16] The LexiBlock with the two allomorphs is co-indexed with the LexiBlock containing the stems, which ensures that the correct pairing will be made during the expansion of the CWC. I use the same mechanism as previously, which should be familiar by now, of singling out the exceptional (monosyllabic) forms. The monosyllabic forms are then generated twice, but since the ones concatenated with the prefix /gu-/ are ranked higher on the list, they are the ones used by speakers.[17]

(181)



This structure makes a prediction as to the types of errors that speakers may make, either in the learning process or simply while mentally computing the forms: a monosyllabic form may "slip" in the general case, because it matches its description, but a disyllabic form may not take the prefix /gu-/ since it does not match the formal description σ.[18] This should not cause any controversy, as it is well known that overall irregular/specific cases tend to generalize, rather than the opposite.

In this inflection also, there are some idiosyncratic suppletive forms. The CWC in (181) is thus rewritten as (182). First the verb 'be' is described. Then, a subset of the stems (or FORM SUBJUNCTIVE PRESENT 3SING POSITIVE) are just straightforwardly used without any modification. Finally, the stems using /g(u)-/ are listed.

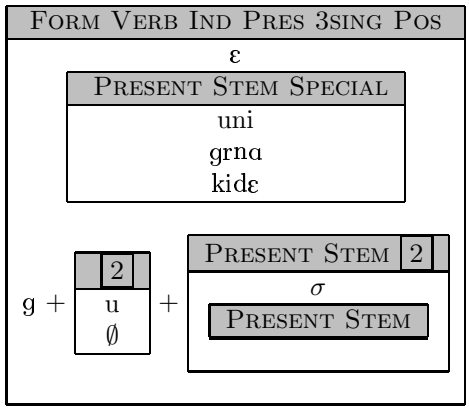
[15] Synchronically, this does not appear to be linked to any prosodic or phonological condition of the type *stems must be disyllabic*, because simple monosyllabic stems are used in the SUBJUNCTIVE, FUTURE and CONDITIONAL tenses. Further, the epenthetic vowel is schwa, not [u] in Armenian.

[16] A schwa is often inserted by phonology after /g-/ when the stem is consonant-initial. This follows from more general principles of phonological epenthesis in Armenian. For the phonological character of schwa epenthesis in Armenian, see Vaux (1998).

[17] Since PRESENT STEM is a set name, the fact that the larger PRESENT STEM LexiBlock includes a second PRESENT STEM LexiBlock is not a problem.

[18] This does not predict that speakers never make wrong generalizations. It predicts that once speakers make this generalization, the errors produced will go one way.

(182)



A clarification is in order concerning the prefix /g(u)-/. While some traditional grammars do refer to it as a "prefix" (e.g. Gulian 1965:49-50), the standard orthography spells it as a separate word. There are at least two indications that the string is considered a prefix by speakers: 1) it only occurs in the context of the indicative present and imperfect, having no independent existence; 2) it is impossible to insert anything between the string /g(u)-/ and the rest of the verb:

(183)

g-ɛrta-m	*g anbajman ɛrta-m	anbajman g-ɛrta-m
IND-go-1Sing	IND necessarily go-1Sing	necessarily IND-go-1Sing
'I am going'	'I am necessarily going'	'I am necessarily going'
bidi ɛrta-m	bidi anbajman ɛrta-m	anbajman bidi ɛrta-m
FUT go-1Sing	FUT necessarily go-1Sing	necessarily FUT go-1Sing
'I will going'	'I will necessarily go'	'I will necessarily go'

In (183), the adverb /anbajman/ 'necessarily' cannot be inserted between /g-/ and the rest of the verb, however, it may be inserted between a "true" independent word like the auxiliary or tense marker /bidi/ and the verb. Therefore, because of lack of evidence for the wordhood status of /g(u)-/, I assume that the orthography is archaic (reflects a previous stage of the language) and that for modern speakers of Armenian, /g(u)-/ is a full-fledged prefix.[19]

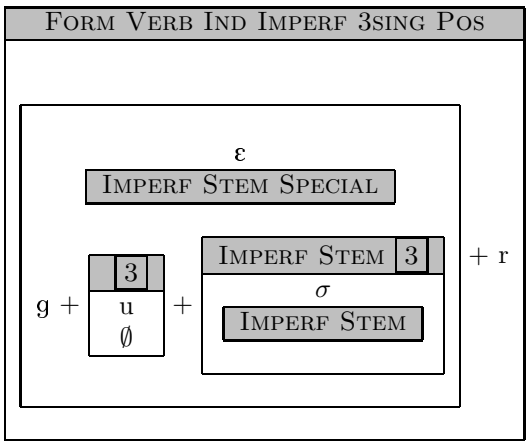
4.5.2 The Indicative Imperfect

The Indicative Imperfect is not much different from the Indicative Present. The same verbs have idiosyncrasies and the same prefix is used for the regular verbs. The only two differences are that the Imperfect Stem is used (which, if you recall (172) is only different for the i-Class

[19] I checked the intuitions with three native speakers of Armenian born in Egypt and now living in Canada.

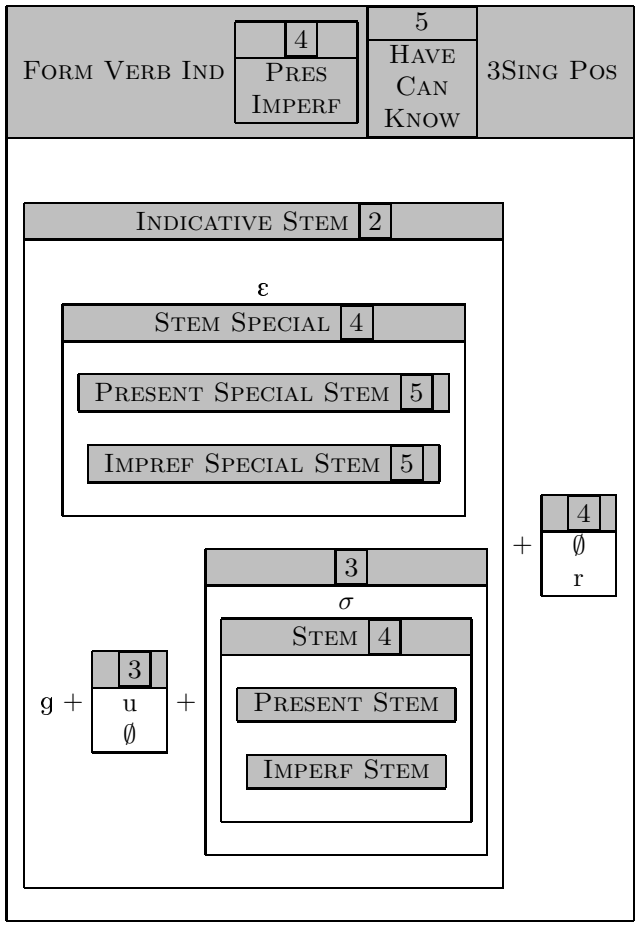
verbs that use a different theme vowel than in the PRESENT) and the /-r/ suffix marking 3SING IMPERFECT.

(184)



The CWCs for the INDICATIVE PRESENT and IMPERFECT are nearly identical in (182) and (184), so we can merge them (by the INTEGRATION STEP):

(185)

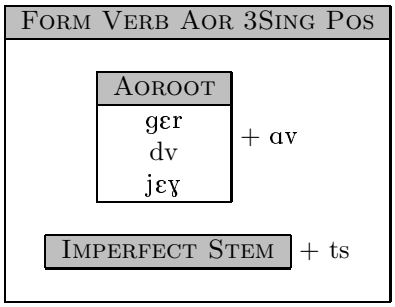


In (185), the LexiBlocks with the index 4 indicate which stem and suffix to use for, respectively, the Indicative Present and Imperfect. The LexiBlock labeled 5 indicates which exceptional verbs do not have an Indicative different from the Subjunctive.

4.6 The Aorist and double morphology

In the Aorist inflection, there are three main peculiarities. First, there is a large class of verbs that use a suppletive root. As we have seen earlier, this is straightforwardly accounted for in TCWC by listing them above the other forms. Remember that the suppleted forms use a different Aorist suffix than the other forms.

(186)

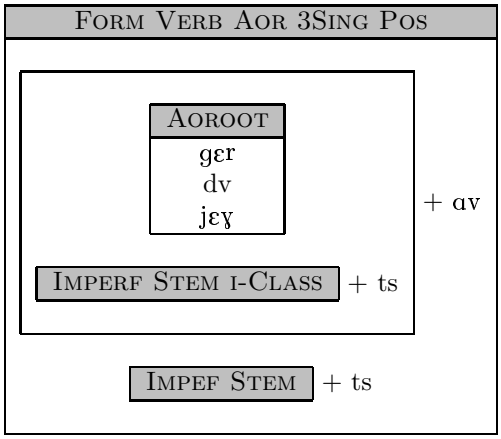


The second peculiarity concerns the i-Class verbs once again. In the Aorist they use the same stem as in the Imperfect (with the theme vowel /ɛ/ instead of /i/). It is important to note that in TCWC, the alternation between /i/ and /ɛ/ is not only an alternation between the surface forms, but an alternation between the underlying representations. A classic argument against treating this type of morphologically conditioned phonemic alternation in morphology is that it is more economical to state it once as a phonological alternation, rather than specifying which vowel to use each time you add a suffix. For example, Lexical Phonology and Morphology would use a phonological rule as in (187):

(187) i → ɛ / _[Aorist] or _[Imperfect]

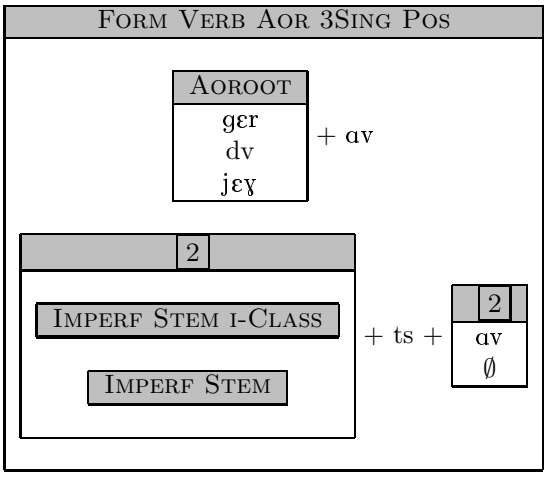
However, in TCWC, since the i-Class roots concatenated with different theme vowels can be given different stem names, it is not more complicated to refer to these different stems in different CWCs, and the "change" from /i/ to /ɛ/ needs only to be stated once. Thus, we refer to the Imperfect Stem from (180) in (188):

(188)



Finally, the I-CLASS verbs also have the peculiarity of using both AORIST suffixes used by the other classes. There are two ways of representing this double marking, both of which are unsatisfactory; either we group the I-CLASS with the suppletive roots, as in (188), or we group them with the other stems, as in (189):

(189)



Here the CWCs in (188) and (189) are problematic because the acquisition procedure is incapable of storing the forms in a maximally economical way. No matter how one tries to group the verb classes, either the suffix /-ts/ or the suffix /-ɑv/ must be repeated in the CWC. In (188), in order to gain more economy, we should try to group the two /-ts/ suffixes, but this is impossible, because the one suffixed to the I-CLASS is already in a LexiBlock. In (189), we avoid repeating the /-ts/ suffix, but then this forces us to repeat the /-ɑv/ suffix. In order to represent both suffixes economically, we would have to imagine some sort of intersecting LexiBlocks, something that has not been defined in this formalism. Therefore, I conclude that TCWC does not provide a maximally economical way to represent the Armenian AORIST suffixes.

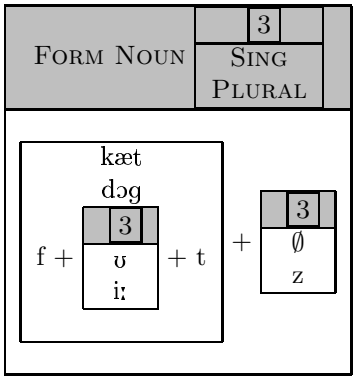
This however can be viewed as good news, because not only is double-marking the exception rather than the rule in morphological systems, but this specific kind of double-marking seems to be especially rare. Take for example some non-standard plurals of Western Armenian nouns (data from Vaux 2003:115):

(190)

	Singular	Plural	Gloss
a.	mɑd	mɑd-vi (rare), mɑd-və-nɛr	finger
b.	dəɣɑ	dəɣɑ-kʰ, dəɣɑkʰ-nɛr	boy
c.	mɑɾtʰ	mɑɾtʰ-ig, mɑɾtʰig-nɛr	man

In (190), what we see is variation between an older plural marked by what is now perceived as an idiosyncratic suffix and another plural that adds on a more productive suffix /-nɛr/ to the older plural form. This is similar to an English dialect that would have variation between *children* and *childrens* or *feet* and *feets*, as the plurals of *child* and *foot*. This type of double marking is not problematic for TCWC. The representation is not as elegant as for simple-marking morphology, but it does not yield the same problem of repeating a suffix:

(191)



In the Armenian AORIST, the two suffixes used to mark this category are added onto an entire class of verbs, the I-CLASS verbs, not just a few random verbs. Furthermore, the two suffixes continue to exist *independently* on separate classes of verbs.

Given that the acquisition procedure fails to maximally reduce the representation for the AORIST, it is to be expected that this kind of double-marking would be rarer in the world's languages.

4.7 Other person-numbers

So far I have shown how TCWC handles the infinitive form of the Armenian verb, as well as the 3SING inflections of the SUBJUNCTIVE PRESENT and IMPERFECT, the INDICATIVE PRESENT and IMPERFECT, and of the AORIST. If we ignore for now the smaller classes of irregular verbs, the paradigms so far look like the table below. I have hyphenated morphemes and boldfaced the class irregularities of the I-CLASS for ease of reading.

(192) **Sample Paradigms of Regular Infinitive and 3Sing Forms**

VERB CLASS	i					ɛ					a				
GLOSS	'sleep'					'drink'					'scream'				
INFINITIVE		bɑrg-	i-	l			xm-	ɛ-	l			bɔr-	ɑ-	l	
SUBJ PRES		bɑrg-	i				xm-	ɛ				bɔr-	ɑ		
SUBJ IMPERF		bɑrg-	**ɛ**	-r			xm-	ɛ-	r			bɔr-	ɑ-	r	
INDIC PRES	g-	bɑrg-	i			g-	xm-	ɛ			g-	bɔr-	ɑ		
INDIC IMPERF	g-	bɑrg-	**ɛ-**	r		g-	xm-	ɛ-	r		g-	bɔr-	ɑ-	r	
AORIST		bɑrg-	**ɛ-**	**ts-**	**ɑv**		xm-	ɛ-	ts			bɔr-	a-	ts	

The other person-number suffixes of Armenian come in a couple of patterns. For the regular verbs of the E-Class and A-Class, they are always the same. In the table below, I give a sample paradigm for /xmɛl/ 'drink'. Although the Indicative tenses are not illustrated, they differ from the Subjunctive only in the presence of the prefix /g-/.

(193)

Tense	Pres	Imp	Aor
1Sing	xm-ɛ-m	xm-ɛ-i	xm-ɛ-ts-i
2Sing	xm-ɛ-s	xm-ɛ-i-r	xm-ɛ-ts-i-r
3Sing	xm-ɛ	xm-ɛ- -r	xm-ɛ-ts
1Plur	xm-ɛ-nk	xm-ɛ-i-nk	xm-ɛ-ts-i-nk
2Plur	xm-ɛ-k	xm-ɛ-i-k	xm-ɛ-ts-i-k
3Plur	xm-ɛ-n	xm-ɛ-i-n	xm-ɛ-ts-i-n

As mentioned above, the regular A-Class verbs are conjugated in the same manner (except, of course, the theme vowel there is /ɑ/). The regular I-Class verbs follow the same pattern for the Indicative and Subjunctive Present. In the Indicative and Subjunctive Imperfect, the I-Class verbs change their theme vowel to /ɛ/, but take the exact same person-number suffixes as the two other regular classes. However, in the Aorist, given the double-marking discussed above, the paradigm for the I-Class verbs is different:

(194)

Tense	Pres	Imp	Aor
1Sing	bɑrg-i-m	bɑrg-ɛ-i	bɑrg-ɛ-ts-ɑ
2Sing	bɑrg-i-s	bɑrg-ɛ-i-r	bɑrg-ɛ-ts-ɑ-r
3Sing	bɑrg-i	bɑrg-ɛ- -r	bɑrg-ɛ-ts-ɑ-v
1Plur	bɑrg-i-nk	bɑrg-ɛ-i-nk	bɑrg-ɛ-ts-ɑ-nk
2Plur	bɑrg-i-k	bɑrg-ɛ-i-k	bɑrg-ɛ-ts-ɑ-k
3Plur	bɑrg-i-n	bɑrg-ɛ-i-n	bɑrg-ɛ-ts-ɑ-n

As is to be expected, the Aorist conjugation of the I-Class verbs is identical to the conjugation of the suppletive verbs (modulo the presence of the suffix /-ts/). However, there are actually two classes of suppletive verbs that conjugate slightly differently in the Aorist. The first class,

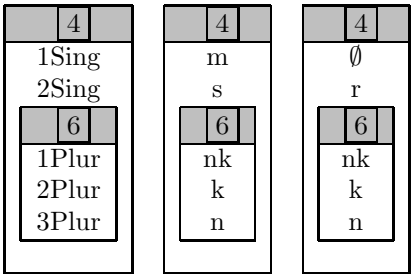
represented below by /udɛl/ (suppleted by /gɛr-/) 'eat', conjugates exactly like the I-CLASS verbs. The second class, represented by /dɑl/ (suppleted by /dv-/) 'give', uses /i/ instead of /ɑ/ in every person except the 3SING.

(195)

	SUPPLETIVE AORISTS	
	1st Class	2nd Class
1Sing	gɛr-ɑ	dv-i
2Sing	gɛr-ɑ-r	dv-i-r
3Sing	gɛr-ɑ-v	dv-ɑ-v
1Plur	gɛr-ɑ-nk	dv-i-nk
2Plur	gɛr-ɑ-k	dv-i-k
3Plur	gɛr-ɑ-n	dv-i-n

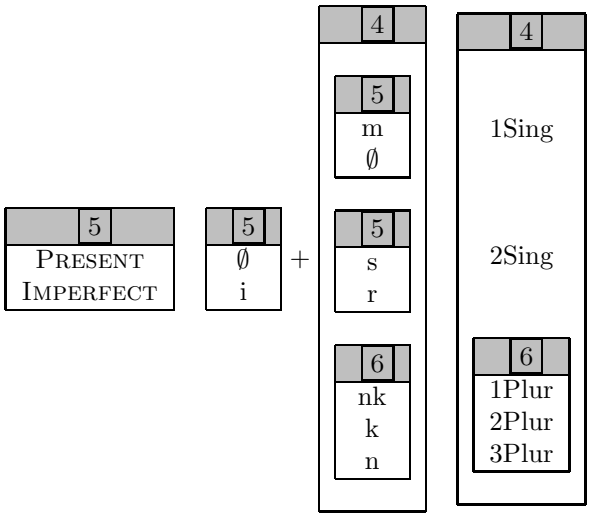
In all the paradigms elicited so far, the plural person suffixes are always the same: 1PLUR /-nk/, 2PLUR /-k/ and 3PLUR /-n/. Further, the 1SING/2SING pair is always realized by either /-m/ and /-s/, or the null suffix and /-r/ (respectively). The 3SING is the most idiosyncratic person-number. It will therefore be useful to have two structures for the person-number suffixes other than the 3SING. These two structures should vary according to the differences in the 1SING and 2SING.

(196)



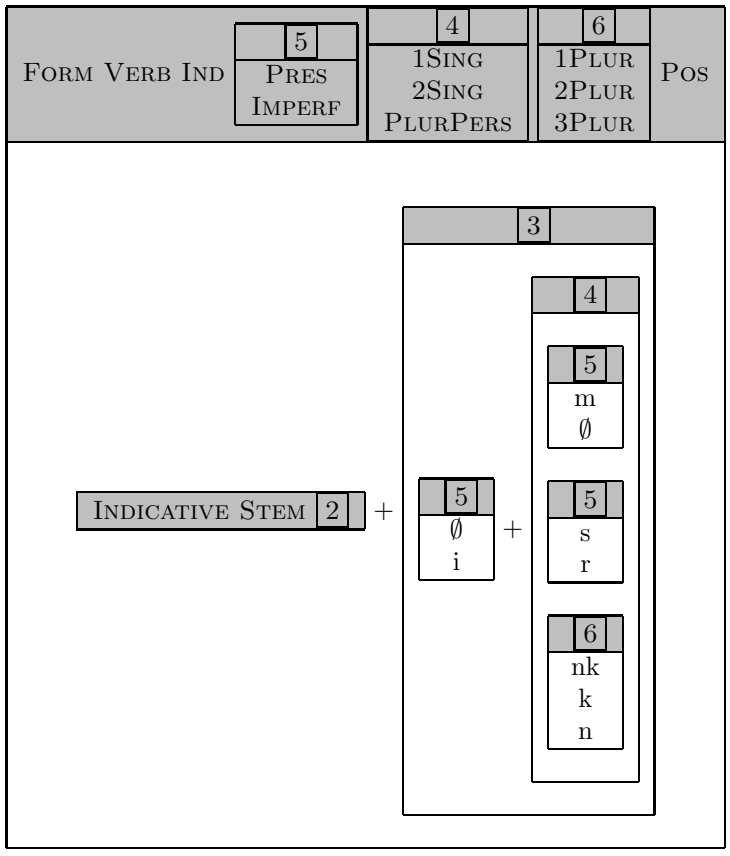
In the case of the indicative, as we just saw, the first set of suffixes is associated with the PRESENT, while the second is associated with the imperfect. Remember that the IMPERFECT suffixes were also preceded by the suffix /-i/. Hence, by using the appropriate index numbers to link the relevant suffixes to their associated category, the INTEGRATION STEP merges the two suffix sets as follows:

(197)



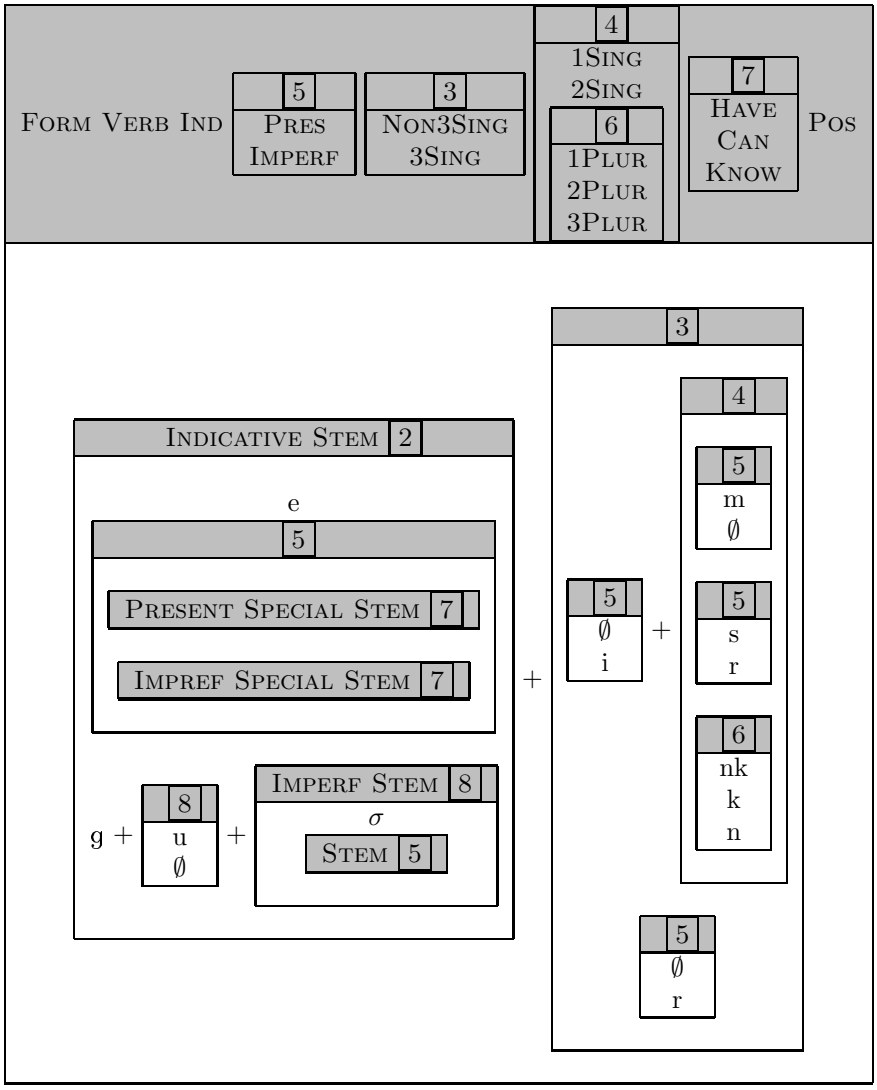
Since the stem for these person-numbers is exactly the same as the 3Sing stem, we then simply concatenate the suffix structure above with the stem structure from (185):

(198)



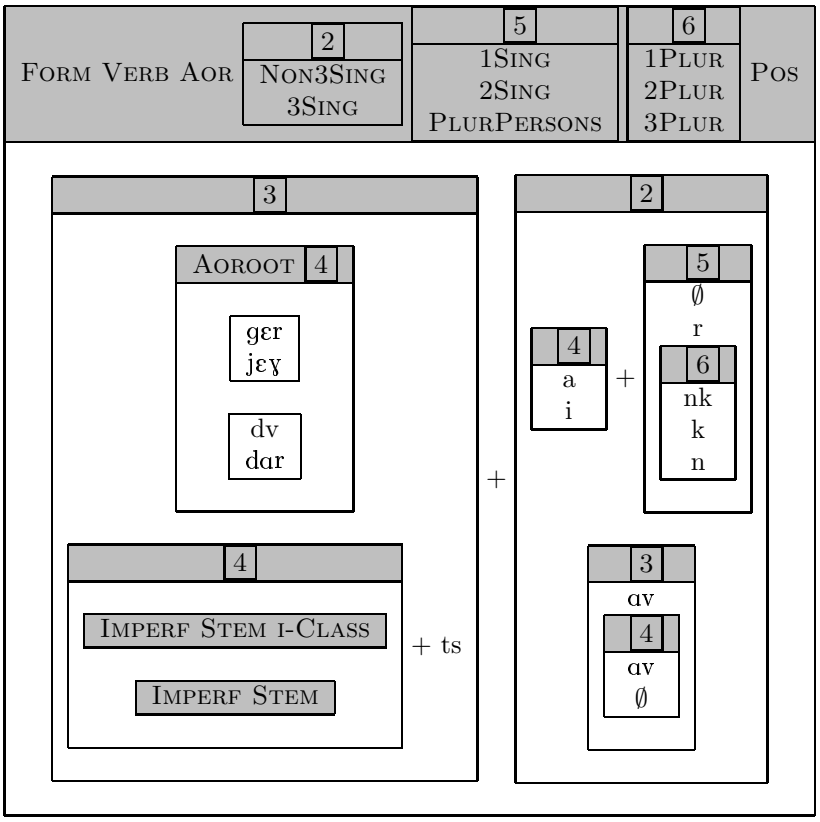
This CWC and the one in (185) can be merged by the INTEGRATION STEP:

(199)



In the case of the AORIST, all the verb classes select the same NON3SING suffix set. However, the complication that requires a repetition of the AORIST suffix and 3SING suffix remains:

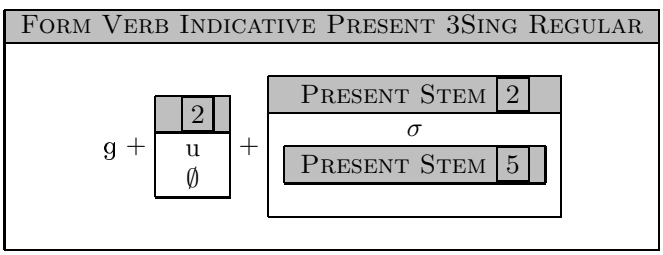
(200)



Although the acquisition procedure of TCWC requires that huge CWCs like the last two be constructed, for descriptive (and pedagogical) purposes, this is definitely not the best option. The larger the construction, the harder it is to keep the indices straight and to "see" structure.

The advantage of TCWC is that when describing a language, one can use a CWC that is as specific or as general as needed to illustrate some property. For example, if one wanted to illustrate how the lexicon is divided into verb classes using different theme vowels, then the constraint in (165) serves this purpose well. If one wanted to illustrate the syllable-sensitivity of the Indicative prefix, then one could single out the following structure, representing the regular Indicative Present 3Sing, ignoring the irregular classes and the Non3Sing persons:

(201)

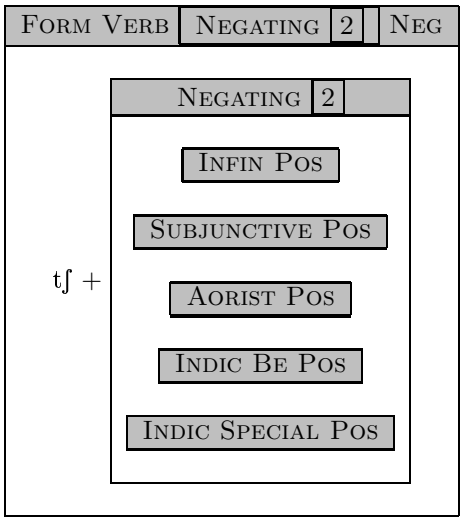


This flexibility comes naturally to many frameworks. For example, a linguist working with an OT grammar, never lists all the assumed constraints and candidates in a tableau. Typically, only the constraints that are relevant to the phenomenon examined and the candidates that have the best chance of beating the winner are given. Things should be no different in TCWC: one can zoom in to that part of the Morphological Lexicon that is relevant.

4.8 Negation and phrasal blocking

Some Armenian tenses are negated with the prefix /tʃ/. These are accounted for below.

(202)



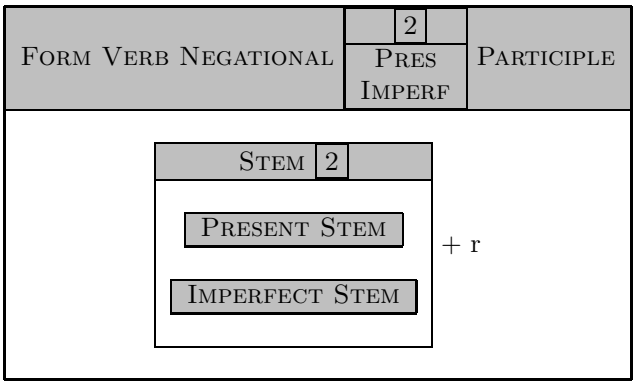
In (202), we prefix /tʃ-/ to a subset of the Positive forms generated so far. The outermost LexiBlock takes all the same categories as those forms, but replaces the category Pos with Neg. The forms that take this suffix are termed "Negating". These are the Infinitive, as well as all the person-numbers of the Aorist, the Subjunctive Present and Imperfect, as well as the irregular (Special) Indicative forms (those that don't take the prefix /g(u)-/).

The Regular Indicative forms do not have a Negative form. Instead, they are negated phrasally. Since TCWC is compatible with a lexicalist model of syntax, morphology happens "before" syntax, in the sense that syntax is assumed to only have access to full words, not to any of their subparts. Hence, it is impossible for a phrase to block a word from being generated. However, the opposite—a word blocking a phrase—is entirely possible, if we only assume that speakers "look up" their Morphological Lexicon before forming phrases. Kiparsky (2004), though working in a different framework and with different languages, comes to the same conclusion.

4.9 Phrasal negation and other participles

As mentioned in the previous section, most INDICATIVE forms negate phrasally. The NEGATIVE PRESENT or NEGATIVE IMPERFECT form of the verb 'be' is used before a special participle, which I will call the "NEGATIONAL" PARTICIPLE, and which also varies according to the PRESENT or NEGATIVE tense of the verb:

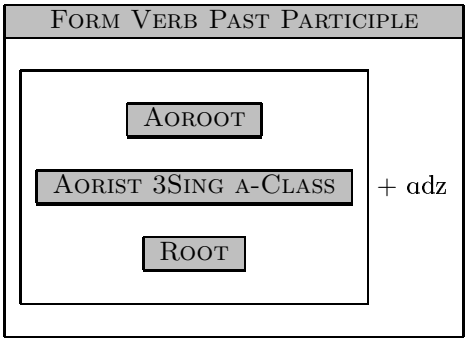
(203)



The participle is thus built on the PRESENT STEM or the IMPERFECT STEM, depending on the tense of the preceding auxiliary and of the whole phrasal construction. Since the goal of the dissertation is not to account for syntactic facts, I will not speculate on which way is best to formalize the combination of the auxiliary with the participle.[20]

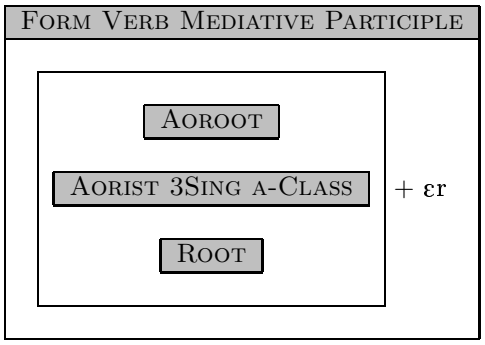
The other participles used in various phrasal constructions are also straightforwardly accounted for below. The only aspects of Armenian verbal morphology that have not been treated in this chapter are the passive, the causative, the imperative, as well as deverbal nouns and adjectives. None of these however require different principles than the ones we have seen so far, so I will not go into them here.

(204)

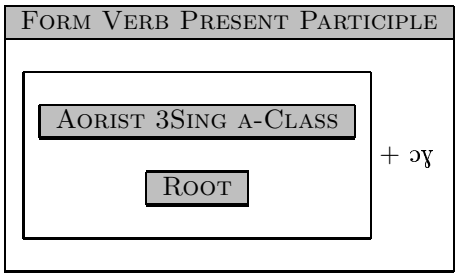


[20] For a suggestion on how to handle this, see Baronian (2004/2005).

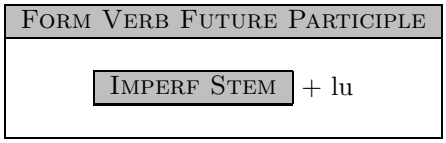
(205)



(206)



(207)



4.10 Conclusion

We have successfully accounted for the Western Armenian verbal morphology system with relatively few CWCs. The same principles introduced for building CWCs in the previous chapter that have allowed us to deal with facts of French morphological change, have also proven to be sufficient to account for the synchronic morphology of the more complex Armenian language. Further, the system fails to provide a maximally economic way to account for a case of double morphology, which reflects the scarcity of this type of double morphology cross-linguistically.

Chapter 5

Paradigm Gaps of Defective Verbs

5.1 Introduction

5.1.1 The problem

The distinction between syntagmatic and paradigmatic relations is often quoted in linguistics. According to Saussure (1995[1916]), syntagmatic relations hold between elements within a phrase (e.g., a verb and its complement), or within a word (e.g., a stem and an affix). Saussure however opposes syntagmatic relations not to paradigmatic ones, but to associative ones. Associative relations encompass all form and meaning relations between words. For example, the word *teaching* may evoke *learning* and *education* on the semantic dimension, but *teacher*, *teaches* and *taught* on the formal side, as well as *praising*, *freeing* or *sing*, with which it shares an ending, whether this ending is a suffix or not. Paradigmatic relations are then a special case of associative relations, according to Saussure (1995:175[1916:253]). While Saussure does not define paradigmatic relations precisely, they are always finite, according to him.

Because the distinction between inflection and derivation is not a crucial one in TCWC, the distinction between paradigmatic and associative relations should not be crucial either. How then are we to even talk about paradigms and paradigm gaps in TCWC?

First, let us examine what paradigms and paradigm gaps are in theories for which paradigms are crucial. If we recognize the distinction between lexemes (objects stored in the lexicon) and word-forms (the realization of a lexeme in a given syntactic context), an inflectional relation is one between the word-forms of a lexeme, or the lexeme and its word-forms, while a derivational one is

one that exists between lexemes. A paradigm then, is the set of word-forms of a lexeme.

(208)

Lexeme		*give*	*have*	*help*	*stride*
Paradigm	Infinitive	/gɪv/	/hæv/	/hɛlp/	/strajd/
	Non3Sing Present	/gɪv/	/hæv/	/hɛlp/	/strajd/
	3Sing Present	/gɪvz/	/hæz/	/hɛlpz/	/strajdz/
	Gerund	/gɪvɪŋ/	/hævɪŋ/	/hɛlpɪŋ/	/strajdɪŋ/
	Past	/geɪv/	/hæd/	/hɛlpd/	/strajdd/ or /stroʊd/
	PastParticiple	/gɪvn/	/hæd/	/hɛlpd/	/strajdd/, /stroʊd/, /strɪdn/ or GAP

In these theories, a paradigm gap is an observable phenomenon: it is a missing word-form in a paradigm for a given exponent. For example, some speakers have no PastParticiple for the verb *stride*, as shown in the table above. A word—or, properly speaking, a lexeme—is said to be defective when it lacks one or more forms from its paradigms. The gaps are not merely forms of rare words that happen to be unattested, they are forms that speakers are unable to generate, finding unacceptable all the forms that linguists propose as possible candidates. For example, the French verb *frire* 'to fry' has no Plural Present forms, nor does it have any of the Imperfect forms. Speakers do not find any of the possibilities acceptable (*ils *frient, *frisent, *frissent*) and typically say that they would paraphrase the intended meaning by constructions such as *Ils sont en train de frire* 'they are in the process of frying'.

Since the distinction between inflection and derivation is not clear-cut in TCWC, paradigm and paradigm gaps cannot be defined as above. This can be viewed as a good thing, because "derivational gaps" are a very well known phenomenon.[1] We will discuss the derivation/inflection distinction in more detail in Chapter 7, but for now let us simply note that the rarity of inflectional paradigm gaps when compared to derivational gaps is to be expected given that inflectional patterns are recognized as more productive than derivational ones. Therefore, the phenomenon we are studying in this chapter, properly speaking, is the impossibility to generate one word from another, though, for space reasons, we limit our domain of study to the more surprising productive inflectional relations.

Defective words pose a problem for most theories of morphology, because so many of them rely on an "elsewhere" or "default" case to account for the more productive morphological strategies. If there were always a default, then speakers would always have a way of generating the various forms of a noun or verb.

In TCWC, there are "defaults" or "elsewhere cases" only when the Elsewhere Step applies. Otherwise, speakers may always exercise a free choice between the strategies in which the words

[1]For several English examples, see Aronoff (1976)

"fit", according to the Lexical Insertion Conditions proposed in Chapter 3. It is assumed that the greater the number of words participating in a strategy (or the less restrictive the strategy), the more attractive it is, which gives the impression of the existence of a "default" in many cases. However, in TCWC, it is only when the ELSEWHERE STEP has applied that we really have the true equivalent of a default. In this respect, it is significant that I know of no cases of paradigm gaps in Armenian, where the ELSEWHERE STEP applies pretty much across the board, according to the definitions we established in Chapter 3—see Chapter 4.

In this chapter, I will limit myself to the examination of the inflectional paradigms of French, Spanish and Russian verbs, three systems for which the paradigm gaps are well-known and already discussed in the literature. In the domain of paradigm gaps, it is essential to start with linguistic systems that are already carefully documented, because it is often not clear if a form is just unattested because of too small a corpus, or if we are truly dealing with a gap.

5.1.2 Previous accounts

The earliest generative attempt to account for paradigm gaps was made by Morris Halle. Halle (1973) proposed a straightforward account for the defective verbs of Russian that lack a 1SINGULAR PRESENT form: these verbs are marked [-lexical insertion]. Albright (2003) advances three arguments against this approach. First, the approach is unrestricted as to which forms could be lexically marked, but in Russian for example, it is consistently the 1SINGULAR PRESENTS that are lacking. Second, one might wonder how speakers know which words cannot be used in the 1SING. Third, the gradient nature of the uncertainty associated with gaps is not accounted for.

I believe the first two arguments given by Albright are extremely convincing, and will return to the third after discussing the Spanish gaps. We could rephrase Albright's first argument in terms of lexical variation: the diacritic approach states that words are arbitrarily marked for lexical insertion, hence, in principle, different dialects could mark any word or any form of any word as [-lexical insertion]. As we will see in French, Spanish and Russian, in a given language, the gaps occur in a restricted set of inflections and for a conjugational class often associated with some morphophonological properties. The diacritic approach misses those generalizations entirely.

The second argument is the most serious one and can be restated in terms of negative evidence. Given that we know that speakers are able to produce or inflect words that they have never heard before, the default setting cannot be [-lexical insertion]. If that were the default setting, then this would amount to saying that every word is learned, a conclusion that would go against any generative

model of morphology. But then, if words start out with the value [+ lexical insertion], what evidence could allow learners to know that a given word is to be marked [-lexical insertion]? Only a specific instruction to that effect of the sort "do not use this form" could justify such a change. As we know, this is the kind of negative evidence that most theories of language acquisition do not recognize as valid. Even if they did, the possibility that this is how these words get ruled out is highly unlikely. If speakers had been told that a given inflection does not exist, they would say "there is no form for this verb in this inflection" (and some might add "but I use this or that form"). Anybody who has studied paradigm gaps knows that their reaction is rather different: speakers are typically puzzled and unsure, and claim to not know the correct form. They are genuinely unsatisfied with the ones they come up with or the ones that are proposed to them.

The second argument can thus be summarized as follows. On one side, it is impossible that speakers have a gap "because they never heard a form before", because under most circumstances, they are able to generate forms they have never heard before. On the other side, the gap cannot be due to having been told that there is no form for a given verb in a given inflection, because then speakers would not display the reaction they do when faced with gaps.

Another account often proposed by speakers or linguists who haven't given the question much thought is "homonymy avoidance". For example, as Albright reports and as I've experienced, some Spanish speakers may tell you that the reason there is no 1SING PRESENT for *abolir* 'abolish' is because **abuelo* already means 'grandfather'. This however does not explain why the option **abolo* is also rejected, nor does it explain why *abolir* is lacking some other forms that have no homonym or why other verbs like *aguerrir* also have gaps that wouldn't cause homonymy at all. Similarly, some French speakers claim they reject *ils *frisent* for 'they fry' because it would be homonymous with 'they curl'.

A third type of approach is represented by Dell (1970) who argues for French *frire*, that speakers do not know which conjugation to choose. Morin (1987) argues against this position since speakers obviously do make some choice for the inflections that don't show a gap. In TCWC, this cannot be a valid argument either, because when speakers have a choice between morphological strategies, the result is variation, not gaps. Again, when speakers are confronted with a gap, their overt reaction is that *no form is good*, not that they do not know which form to pick.

Plénat (1981) adopts a more articulated view for the French gaps where speakers are only unable to pick a consonant to link the stem *fri-* to the various suffixes. Thus only those inflections that use the bare stem show up. This type of approach is closer to the one I adopt, though, again, in TCWC,

it is crucial to note that we cannot say that speakers "hesitate" between strategies, since giving the speaker a choice between strategies is at the heart of the theory. Rather I will argue that no lexical insertion condition allows speakers to choose any strategy.

Finally, in Optimality Theory, researchers have made the use of a constraint MPARSE (Prince & Smolensky 1993) that assigns a violation mark to the NULL PARSE, and is the only one to do so. As Rice (2005:17) points out, this is a stipulative approach, and we could add to Rice's, the same criticisms that Albright (2003) had for Halle (1973). Another OT approach, proposed by Orgun & Sprouse (1999) and also used by Hansson (1999), adds a CONTROL component to OT, where non violable constraints are stored: if an output of the set of regular constraints CON violates a constraint in CONTROL, then there simply is no output. This approach is again somewhat stipulative, but crucially, we will show that in TCWC, we can do with the tools already in hand, namely the Lexical Insertion Conditions from Chapter 3, whereas the control approach needs to justify the addition of an entire component.[2]

5.1.3 Types of defective verbs

There are different types of defective verbs and TCWC does not claim to account for all of them. From the languages I examined (French, Spanish and Russian), I have classified the gaps observed according to three types:

(209) **Defective Verb Types**

a. Argumentless verbs: Verbs that require an expletive subject.

b. Paradigmless verbs: Verbs that have no finite forms.

c. Paradigm gap verbs: Verbs that lack only certain finite forms.

Although TCWC has nothing special to say about the first type, an independent syntactic and semantic explanation is quite straightforward. A verb like 'rain' or 'snow' in French, as in English, selects for an expletive argument: a syntactic argument that does not have a semantic correspondent. In a famous poem, Émile Nelligan willingly violates this syntactico-semantic constraint in order to convey how much the snow has "snowed":

[2]Rice (2005) proposes another OT approach to paradigm gaps, but unfortunately, this paper came to my attention too late for me to discuss it fairly here.

(210) Ah ! comme la neige a neigé !
Ma vitre est un jardin de givre.
Ah ! comme la neige a neigé !
Qu'est-ce que le spasme de vivre
À la douleur que j'ai, que j'ai !

Émile Nelligan, *Soir d'hiver*

Once the syntactic constraint requiring an expletive is violated this way, it is not a problem for speakers to attribute the words *je neige* 'I snow', *je neigeais* 'I snowed' or *je neigerai* 'I will snow' to an imaginary snowflake falling from the sky. Since in French there is only 3SING *il* that may be used as an expletive subject, these verbs in French are said to have only 3SING forms. TCWC however rightly predicts that speakers know what the other forms of these verbs would be, and thus may use them in poetry or in humorous situations.[3]

The second type is more problematic. For example, the French verb *douer* meaning 'endow' only has non finite forms. TCWC provides no explanation for this. The verb *douer*, just like *nouer* 'tie (a knot)', is an otherwise perfectly normal FIRST GROUP verb, and by virtue of belonging to this most productive class, its forms should be easily predictable. The (nonexistent) forms in fact, don't seem to pose any problem for the speakers I have consulted, the speakers just claim, like the prescriptive grammars, that the forms are just not in use. There are similar cases in Spanish (*adir* 'accept'; *usucapir* 'acquire by prescription'). A similar case in English would be *beware*, which may be used in the INFINITIVE (*It's important to beware of this danger*), the IMPERATIVE (*Beware of dogs*) and the SUBJUNCTIVE (*He asked that I beware of strangers*), but not in the PRESENT or PAST (**I beware of the danger*). *Beware* is a little different, however, as it has no GERUND **bewaring* either, though the GERUND is usually considered a non finite form—see Fodor (1972) for more details.

It seems to me that these cases are different, and may be more closely related to derivational gaps. Like adjectives and nouns are not values of a common feature under most analyses, neither are finite and non-finite the values of the feature verb. I recognize that at this point, definite answers to the question of why some verbs do not show *any* finite forms would be purely speculative.[4]

The third type is the one that will concern us in this chapter. In this type, some person-number(s) of some inflected tense(s) is or are missing for a given verb or for a given class of verbs. Sometimes it is an entire tense that is missing, but never all inflected tenses, or else we would be dealing with cases from the second type. My explanation of this type of gap is that the Lexical Insertion Conditions

[3] A quick search on google.com for *je neige* showed primarily poetic hits.
[4] One area of investigation might be considering semantic conditions on lexical insertion.

introduced in Chapter 3 prevent the verb from being inserted in any of the morphological strategies available in the CWCs.

After an introductory English example, I will show how CWCs can account for two French paradigm gaps, followed by a comparison with Morin's account of these facts. I will then illustrate how I account for two types of Spanish gaps followed by a comparison with Albright's account of these facts.[5] I will then proceed to account for the numerous 1SING gaps in the Russian verbal system.

5.2 English *stride*

Before getting into the heart of this chapter, I will show how the Lexical Insertion Conditions can help us account for a gap in English PASTPARTICIPLES. Some speakers of English have a gap for the PASTPARTICIPLE form of *stride*. I will show why theories using default principles across the board are doomed and cannot hope to explain paradigm gaps.

The *stride* gap is well-known, but for the benefit of all readers, some of whom may not have a gap for the verb *stride*, I attempted to obtain some form of objective confirmation. In particular, several speakers may hesitate for the right PASTPARTICIPLE form of the very similar and also rare verb *strive*, but I will argue that this hesitation is not a gap at all. Using the search engine google.com, I calculated the number of hits for the words or sequences of words in the example below. The goal was to compare the similar and rare verbs *stride* and *strive*, along with a non rare rhyming verb (*drive*) and another control verb (*give*). I wrote down the total hits given by the search engine for three possible choices in forming the past of these verbs (even the ones most native speakers would consider ungrammatical), as well as the number of hits for these forms when immediately preceded by *has* or *have*. I then calculated the ratio of the total *has/have...* over the possible PAST forms. The assumption is that this ratio should reflect roughly the proportion of attested PASTPARTICIPLE uses over simple PAST uses.[6] As the numbers eloquently attest, according to this simple test, the PASTPARTICIPLE of *stride* is more than ten times less frequent that of any of the other verbs, confirming that there is a paradigm gap for the PASTPARTICIPLE of *stride*.

[5] In spite of my objections, I would like to give credit to both Albright and Morin for taking paradigm gaps more seriously than many linguists and for giving them the central role they should rightly occupy in any theory of morphology. The account presented here shares a lot in spirit with each of these two accounts.

[6] Actually the real ratios are probably higher given that adverbs can be inserted between *has/have* and the PASTPARTICIPLE as in *I have always given birthday presents to my friends*. In principle though, this distortion should be the same for all verbs.

(211)

		Ratio
has strided 12		
have strided 98	strided 18,300	.006
has strode 112		
have strode 280	strode 421,000	.001
has stridden 148		
have stridden 98	stridden 4930	.05
Total 748	Total 444,230	.002
has strived 20,100		
have strived 26,300	strived 125,000	.37
has strove 208		
have strove 461	strove 244,000	.003
has striven 17,500		
have striven 29,300	striven 87,700	.53
Total 93,869	Total 456,700	.206

		Ratio
has drived 67		
have drived 109	drived 13,200	.013
has drove 536		
have drove 4770	drove 4,390,000	.001
has driven 231,000		
have driven 246,000	driven 11,700,000	.041
Total 482,482	Total 16,103,200	.03
has gived 75		
have gived 148	gived 9340	.024
has gave 3850		
have gave 12,600	gave 16,000,000	.001
has given 2,780,000		
have given 2,590,000	given 73,700,000	.073
Total 5,386,673	Total 89,709,340	.06

Google hits compiled February 17 2004.

However, *strive* has a ratio 4 to 7 times higher than *drive* or *give*. This I believe is due to the fact that *driven* and *given* have a much more frequent adjectival use than *striven* (both in absolute numbers and proportionately). A good indication of this is that if we remove the hits for *stridden*, *striven*, *driven* and *given* in the denominators of the four total ratios calculated above, we get much more comparable ratios for the last three verbs: *stride:0.002*, *strive:0.25*, *drive:0.11*, *give:0.34*. I believe I have provided enough evidence to the effect that several speakers are lacking a PastParticiple for the verb *stride*, in a way that is significantly different from any hesitation they may have as to the correct form of the similar-sounding and also rare verb *strive*.

5.2.1 Distributed Morphology

Distributed Morphology (DM) uses morpheme-based rules that are ordered from the most specific to the most general. This version of the Pāṇinian principle ensures that the more specific rules "bleed" the more general. For example, take the way in which DM accounts for English past participles

(212) **The English Past Participle in Distributed Morphology**
(adapted from Halle & Marantz 1993)

[+participle, +past] ↔ /-n/ / X+_
where X = see, go, beat, ...

[+past] ↔ /-d/

In (212), a certain number of roots (X=see, go, beat...) first select the suffix /-n/ in the context of being [+past, +participle]. The second rule states that any leftover [+past] form takes the suffix /-d/. There are of course other rules specific to the past of *see*, *go*, etc. But the idea is that any verb that is left, for example *love*, will form both its [+past] and its [+past participle] with the most general rule, unless one of these forms has already been generated with another higher ranked (and more specific) rule.

Thus by learning the verb *stride*, whether or not English speakers have learned the simple past *strode*, they should form its PastParticiple by suffixing /-d/. Unfortunately, this is not always what happens. While several speakers use the inherited form *stridden* or the innovative *strided*, many speakers, as we saw in the previous section, do not know how to form the past participle of *stride*: they have a gap.

Another problem with the DM approach is that it predicts that speakers will never use the higher ranked rules to inflect new verbs. When a speaker learns a new verb, its root cannot logically already be specified in the grammar, hence it cannot be pre-specified to select the suffix /-n/, and every speaker should use the default suffix /-d/. This of course is a wrong prediction, as we know from the rich literature on the topic (Albright & Hayes 2002, Bybee 1985, 2001, Ramscar 2002)—see the discussion in Chapter 3.

5.2.2 Paradigm Function Morphology

Another popular theory that heavily relies on the Pāṇinian principle is Paradigm Function Morphology (PFM). Its analysis of the English PastParticiple, in spite of the deep theoretical issues that separates PFM from DM, is surprisingly similar to the previous account.

In PFM, a cluster of properties are realized with the stem *seen*. This highly specific realization rule is ordered higher than the more general one that states that dental stems take on the suffix /-d/. The difference with DM is essentially that *seen* is not seen as the concatenation of the root *see-* with the suffix /-n/, but rather as an entity in itself.

In any case, the result is the same: when learning the verb *stride*, speakers should be able to generate the dental[7] stem *strided*. But, as we know, some speakers have a gap.

(213) **The English Past Participle in Paradigm Function Morphology**
(adapted from Blevins 2003)
R([see, verb, dent, part]) = seen
R([dent]) = Xd

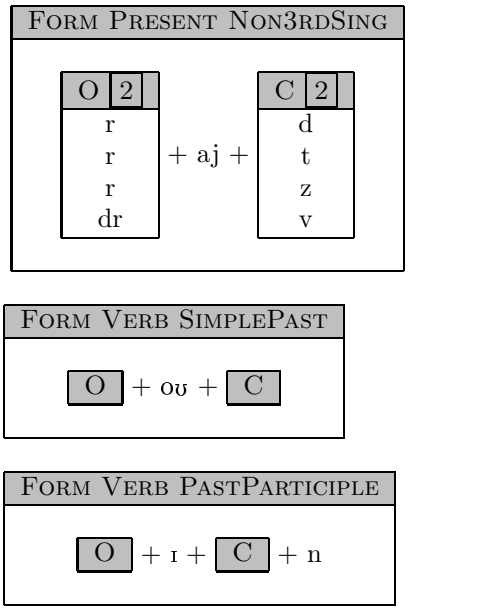
PFM also runs into the problem of predicting that the higher ranked rules are always "unproductive". Blevins (2003) goes into great length to explain that the higher ranked rules operate differently, therefore (implicitly?) recognizing some form of a dual-route model such as Pinker's (1999). For a more detailed discussion of the issues involved, see Chapter 3.

5.2.3 Connected Word Constructions

In TCWC, the words *ride*, *rise*, *write* and *drive* would be stored together, by the Connection Step, because they always behave in the same way.[8] [9]

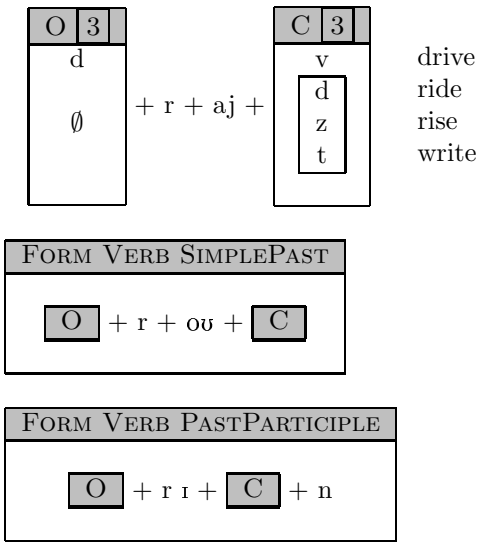
[7] Blevins (2003) uses this term for the cognate Germanic stems that end in a dental (coronal) stop.
[8] I will ignore *thrive*, *shine*, *smite* and *dive* for ease of reading.
[9] I have labeled the repeated LexiBlocks "O" and "C" as mnemonic for onset and coda of the relevant verbs. This label is purely arbitrary.

(214)

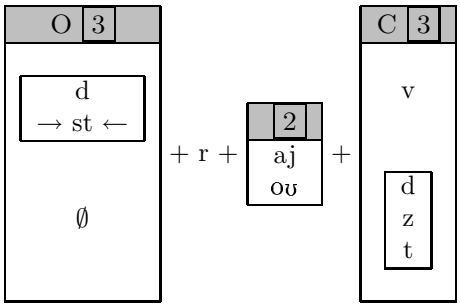


By the Sharing Step, some economy may be reached in the representation:

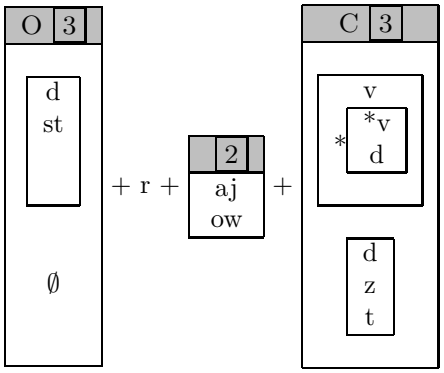
(215)



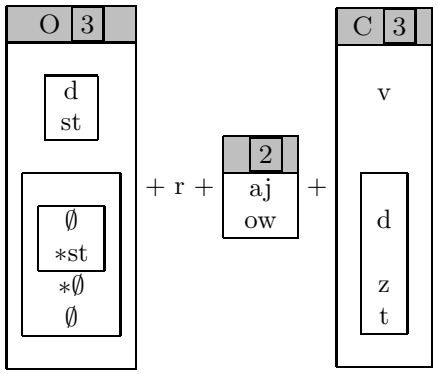
Before dealing with *stride*, let's take a look at the similar verb *strive* and show why it does not have any gaps. The verb *strive*, if learned later than the verbs listed so far, can neatly be inserted next to *drive*, without violating any of the Lexical Insertion Conditions:

(216) 

As for *stride*, things are quite different. No matter where we try to insert it, we need to insert phonemic material in at least two places, thus violating Localized Generalization each time:[10]

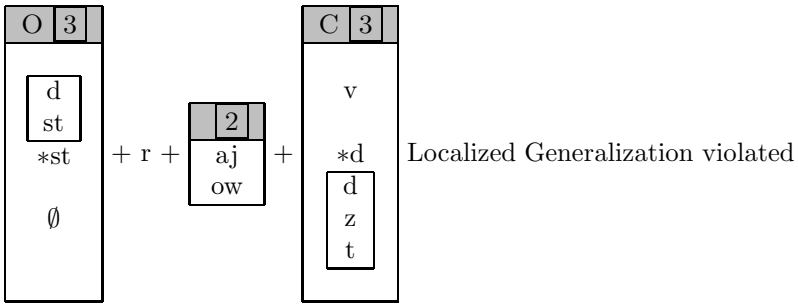
(217) 

Localized Generalization violated



Localized Generalization violated

[10] Localized Generalization is violated because we could insert *stride* in a single LexiBlock between the irregular and regular verbs.



As long as the simple past form *strode* is also learned, the verb cannot be inserted along with the regular verbs either, because it would then violate COUNTER-EVIDENCE RESPECT. TCWC then predicts that only speakers who have learned both the forms *stride* and *strode* may have a gap in the PASTPARTICIPLE (unless they have also learned the inherited form *stridden*). Speakers who have a simple past form *strided* are predicted not to have a PASTPARTICIPLE gap. This, I believe, is a correct prediction, because, in (211) the ratio of *(has strided + have strided)/strided* is six times higher than the corresponding ratio for *strode*. Thus, it seems plausible that speakers who use a simple past *strided* can use this form as a PASTPARTICIPLE, while speakers who use a simple past *strode*, unless they have learned a PASTPARTICIPLE, can have a gap (which explains the lower ratio).[11]

Now that I have shown how TCWC accounts for this simple gap and why DM and PFM both fail in this respect, let us turn to a language with a bit more morphology and more interesting gaps.

5.3 French

5.3.1 The structure of the French verbal system

Because the LexiBlock structure of the lexicon that I am assuming is crucial in explaining the gaps, it will be useful to first demonstrate that the acquisition procedure introduced in Chapter 3 provides us with just the right structure. I will provide this demonstration only for French. In the Spanish and Russian examples, I will simply start with the lexical verb structure and the reader is free to verify that I am not using a structure that is any different from what the acquisition procedure would have produced.

The relevant part of the lexicon that we need to examine concerns mainly verbs whose theme

[11] Obviously, the *stridden* ratio is the highest, since people who have the inherited *stridden* should not have a gap at all. The numbers for *stridden* may be misleading however, because the *stridden* hits without *has* or *have* are probably adjectives, rather than simple past uses, which are either rare or impossible.

vowel is /-i-/. I will thus assume the following Mini-French lexicon, consisting of sample verbs from the classes that are relevant to our purposes.

(218) **Mini-French Verbs**

Verb	Form	Gloss
plaire	plɛr	please
taire	tɛr	silence
conduire	kɔ̃dɥir	drive
déduire	dedɥir	deduct
construire	kɔ̃strɥir	build
instruire	ɛ̃strɥir	inform
prédire	predir	predict
dédire	dedir	retract
inscrire	ɛ̃skrir	inscribe
proscrire	proskrir	proscribe
finir	finir	finish
bâtir	batir	build

According to the WORD STEP of the acquisition procedure, the learned verbs are first grouped together by inflection. Thus the INFINITIVES above are stored in one LexiBlock, and the PRESENT SINGULAR, the 3PLUR, the PASTPARTICIPLE, etc. are each stored in separate LexiBlocks.

(219)

FORM VERB PRES SING	FORM VERB PRES 3PLUR	FORM VERB PAST PART FEM
plɛ	plɛz	ply
tɛ	tɛz	ty
kɔ̃dɥi	kɔ̃dɥiz	kɔ̃dɥit
dedɥi	dedɥiz	dedɥit
kɔ̃strɥi	kɔ̃strɥiz	kɔ̃strɥit
ɛ̃strɥi	ɛ̃strɥiz	ɛ̃strɥit
predi	prediz	predit
dedi	dediz	dedit
ɛ̃skri	ɛ̃skriv	ɛ̃skrit
proskri	proskriv	proskrit
fini	finis	fini
bati	batis	bati

Then, according to the CONNECTION STEP, verbs that behave in the same way across CWCs are grouped together. I will represent this in three phases. First, I represent the grouping of the PRESENT 3PLUR and FEMININE PASTPARTICIPLE in relation to the PRESENT SING. Then, because the groupings are intersecting, when the 3PLUR and the FEMININE PASTPARTICIPLE are compared, the first groupings will split further.[12] Finally, I name the LexiBlocks and factor out the relevant segments.

[12]The sequence of comparison is irrelevant, since all pairs must be compared eventually.

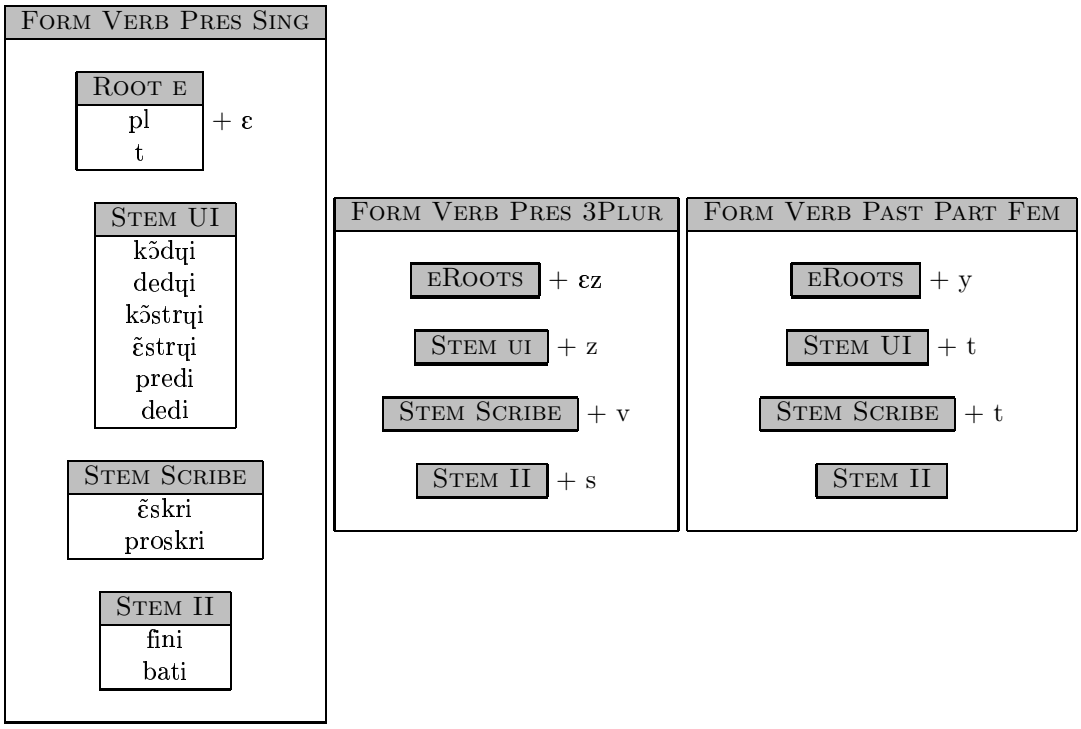
(220)

Form Verb Pres Sing	Form Verb Pres 3Plur	Form Verb Past Part Fem
plɛ	plɛz	ply
tɛ	tɛz	ty
kɔ̃dɥi	kɔ̃dɥiz	kɔ̃dɥit
dedɥi	dedɥiz	dedɥit
kɔ̃stɾɥi	kɔ̃stɾɥiz	kɔ̃stɾɥit
ɛ̃stɾɥi	ɛ̃stɾɥiz	ɛ̃stɾɥit
predi	prediz	predit
dedi	dediz	dedit
ɛ̃skɾi	ɛ̃skɾiv	ɛ̃skɾit
proskɾi	proskɾiv	proskɾit
fini	finis	fini
bati	batis	bati

(221)

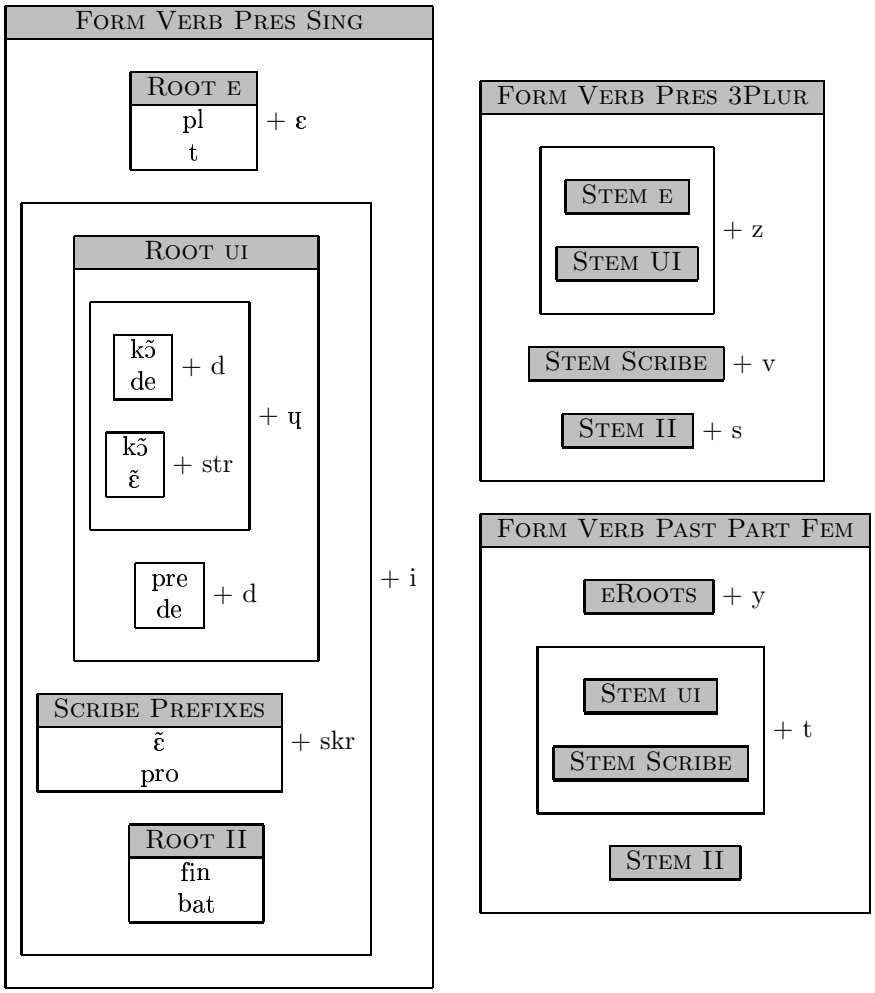
Form Verb Pres Sing	Form Verb Pres 3Plur	Form Verb Past Part Fem
plɛ	plɛz	ply
tɛ	tɛz	ty
kɔ̃dɥi	kɔ̃dɥiz	kɔ̃dɥit
dedɥi	dedɥiz	dedɥit
kɔ̃stɾɥi	kɔ̃stɾɥiz	kɔ̃stɾɥit
ɛ̃stɾɥi	ɛ̃stɾɥiz	ɛ̃stɾɥit
predi	prediz	predit
dedi	dediz	dedit
ɛ̃skɾi	ɛ̃skɾiv	ɛ̃skɾit
proskɾi	proskɾiv	proskɾit
fini	finis	fini
bati	batis	bati

(222)



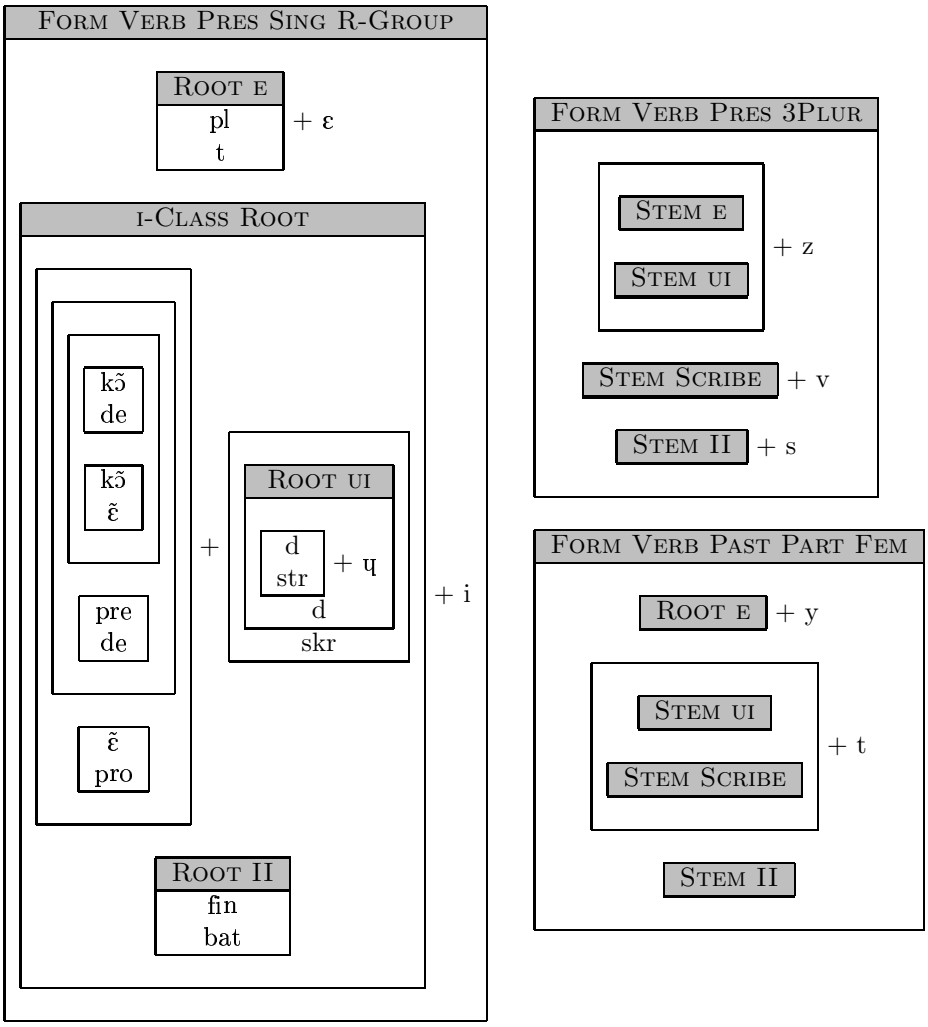
Next, by the Sharing Step, segments that are most frequently shared within LexiBlocks are factored out. Since the contents of some of the remaining LexiBlocks show up several times, they are recognized as prefixes and factored out as well.

(223)



Stem ≡ Form Verb Pres Sing

(224)



Stem ≡ Form Verb Pres Sing

The Elsewhere Step and Integration Step do not apply in this part of the lexicon, because the formal conditions for them to apply are not met. There are no suppletive forms and the stems are grouped in various intersecting ways such that the Elsewhere Step would yield no economy. Also, the stem structure is differently organized in all three CWCs, so it is impossible for the Integration Step to apply. Hence, by following the acquisition steps described in the chapter on analogy, we have obtained the lexical verb structure that is necessary to account for the paradigm gaps of French.

5.3.2 The verb *clore*

According to the Bescherelle (1992) verb conjugation tables, the French verb *clore* 'bring closure' has no 1PLUR or 2PLUR PRESENT or IMPERATIVE forms, nor any IMPERFECT or SIMPLE PAST forms. According to Morin (1987), some speakers are also lacking a 3PLUR PRESENT form. Morin also notes that no speaker has the 1PLUR and 2PLUR PRESENT, while lacking a 3PLUR PRESENT.

Bescherelle (1992) also gives information concerning some prefixed verbs. The verb *éclore* 'hatch' is said to be used only in the 3SING, but this is most likely a semantic restriction due to the fact that animates cannot hatch. The verbs *forclore* 'exclude' and *déclore* 'open up' are said to only have INFINITIVE and PASTPARTICIPLE forms, making them paradigm-less verbs, as defined in §5.1.3, thus not being relevant to the present analysis. Finally, *enclore* 'fence (e.g. a property)' is said not to have gaps for the 1PLUR and 2PLUR PRESENT and IMPERATIVE. This fact came to my attention rather late, and is not discussed elsewhere, as far as I know. If indeed some speakers have a gap for 1PLUR and 2PLUR PRESENT and IMPERATIVE *clore*, but not for *enclore*, then this yields a problem for the analysis presented in this section.

Morin accounts for these facts with a set of implicational statements about verb forms. For example, he proposes that the stem used by the 1PLUR is also used by the 3PLUR by default, but not vice versa. Thus if a speaker learns a 1PLUR form, s/he may use the same stem to form the 3PLUR, but learning the 3PLUR form does not entail that the same stem may be used in the 1PLUR. In turn, the 3PLUR stem may be used for the SING persons, but not vice versa.

(225) **A sample of Morin's suppletion implicational statements**

By default:

1PLUR STEM is used in the 3PLUR

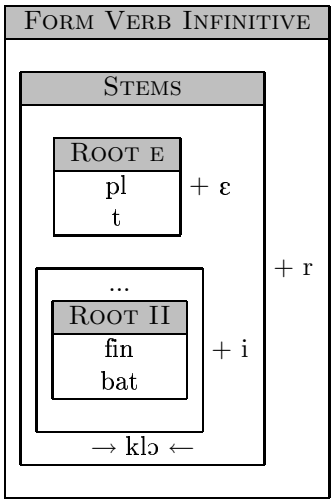
3PLUR STEM is used in the SING

This implicational system allows speakers to learn different stems. For example, the verb *finir* 'finish' uses a stem /finis-/ in the 1PLUR, and because speakers never learn a stem specific to the 3PLUR, they are able to use this same stem in that person. However, learning that the SING persons use the stem /fini-/, overrides the second implicational statement above.

Thus, by simply assuming that speakers never learn a 1PLUR form, Morin accounts for the fact that some speakers have a gap in the 1PLUR, but not in the 3PLUR, but that the opposite situation is not observed. I will return to Morin's analysis, after showing how TCWC accounts for these facts, and those of the verb *frire*.

First, the storage of the French INFINITIVE *clore* is easily done. Indeed, there exists exactly one LexiBlock (shown below) in which it is possible to store *clore*. The INFINITIVES differ from the PRESENT SINGULAR verbs shown in (224) only by the suffixation of /-r/. Since I equated the PRESENT SINGULAR with R-GROUP STEM, we can then simply suffix /-r/ onto it. Ending in /-r/ is the only requirement an INFINITIVE needs to satisfy to be a member of the LexiBlock. However, its theme vowel /ɔ/ being unique,[13] *clore* is not stored with the verb stems of any of the subclasses (which end in other vowels), or else it would violate GENERALIZATION PRESERVATION.

(226)



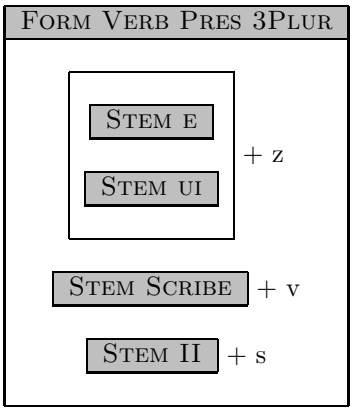
Given that /klɔ-/ is now considered an R-GROUP STEM and given that the R-GROUP STEM is in fact a PRESENT SINGULAR, this latter inflection is easily predictable for *clore* (it is /klɔ/).[14]

This however is not the case for the 3PLURAL forms. In this person, the subclasses of R-GROUP STEM behave differently. Unlike for the INFINITIVE, the larger LexiBlock with all the stems is not referred to, instead each individual subclass is described separately.

[13] Modulo the prefixed verbs *forclore*, *enclore*, *déclore* and *éclore* discussed above, which are also defective in some way. The only potential problem lies with *enclore*, but at least as far as the author is concerned, it shows the same gaps as *clore*.

[14] For *clore*, I assume an underlying form /klɔr/, where the /ɔ/ is automatically tensed by phonology when word-final. Thus the surface PRESENT SINGULAR form is actually [klo].

(227)



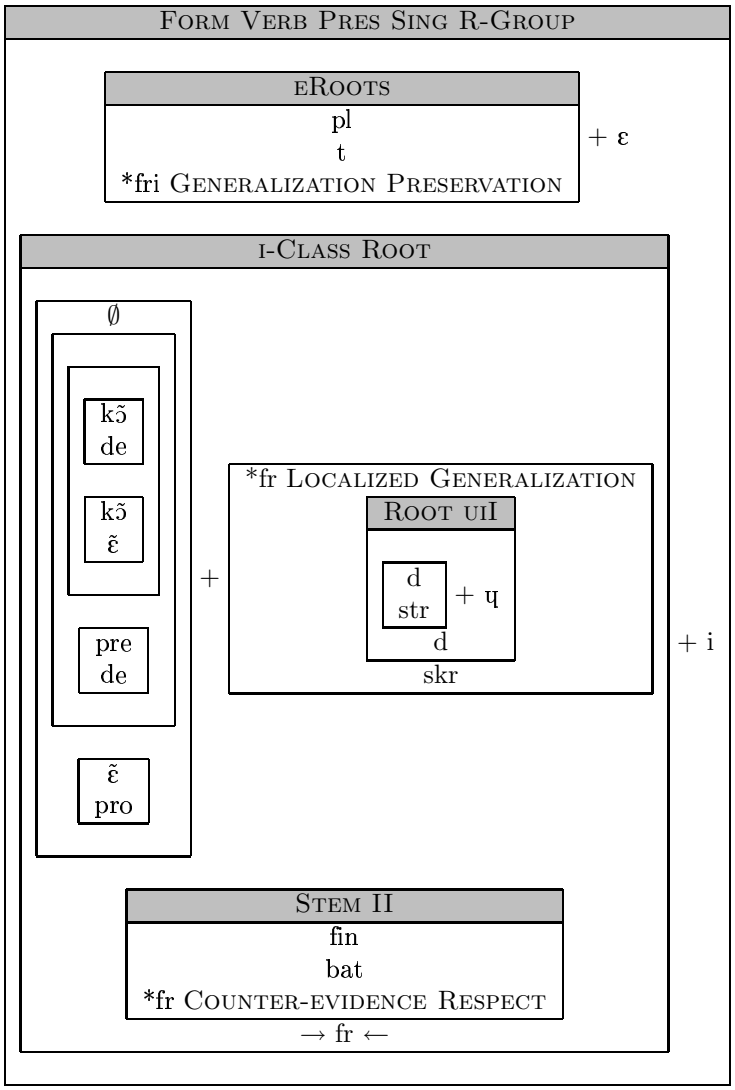
Hence, since a 3Plur form for *clore* is never learned and since there is no general way of forming a 3Plur in French, the 3Plural form of *clore* cannot be generated. This explains a gap for many French speakers (see Morin 1987, 1995). While some speakers may have encountered the inherited historical form *closent* /kloz/, its low frequency makes the 3Plural form unavailable to many speakers. The CWC patterns outlined above predict that if the 3Plural form is not learned, speakers will have a gap in the paradigm of *clore* precisely in this person, which is why this verb is traditionally called "defective". However, the Indicative Present Singular and the Infinitive are mutually predictable.

As mentioned above, the verb *clore* also has gaps in the 1Plur and 2Plur Present, as well as all the person-numbers of the Imperfect and of the Simple Past. As expected, these are also inflections where the stem behaves differently for every class. While the verbs listed in the CWCs so far all use the same stem structure in these inflections as the 3Plural, there are several other subclasses for which this is not the case, for example *tenir*, which has the stem /tən-/ in the 1Plur and 2Plur, but the stem /tjɛ̃-/ in the 3Plur. Therefore, there is no general way of forming the 1Plur or 2Plur Present in French, nor the Imperfect or the Simple Past. Hence, *clore* has gaps for these inflections, because it is part of no specific class.

5.3.3 The verb *frire*

The French verb *frire* also exhibits gaps similar to those of *clore*. However, unlike *clore* that has a unique theme vowel /ɔ/, it is an i-Class verb like the many verbs in (224). If *frire* fit in any of the subclasses of the verbs whose theme vowel is /-i-/, then its 3Plural should be predictable. I argue that *frire* only fits in the larger LexiBlock that includes all the subclasses whose theme vowel is /-i-/.

(228)



In the case of *frire*, each of the three Lexical Insertion Conditions plays a role. First, it cannot be stored with /tɛr/ or /kɔ̃dɥir/, because of Generalization Preservation. Indeed, these subclasses have a phoneme factored out (respectively /ɛ/ and /ɥ/), so that storing *frire* with them would ruin this generalization.

Secondly, it seems *frire* would fit very well with the 300-odd 2ndGroup verbs in the Stem II LexiBlock. However, speakers receive positive counterevidence that *frire* cannot be stored with 2ndGroup verbs: its Feminine PastParticiple is *frite* (like Stem UI and Stem Scribe and instead of **frie*). Thus, by Counterevidence Respect, *frire* cannot be a 2ndGroup verb either.

Finally, the remaining options are ruled out by LOCALIZED GENERALIZATION. Indeed all the other subclasses would require breaking up *frire* into two phonemic strings (even if one of them is the null string). Hence, *frire* remains floating between the LexiBlocks that have /-i-/ as a theme vowel in (224), forming a class of its own. Since a 3PLURAL form for *frire* is never learned, the gap is maintained.

The prediction is that—other things being equal—there shouldn't be a dialect of French where *frire* does not have an exceptional FEMININE PASTPARTICIPLE *frite*, but maintains a gap. Support for this claim comes from fieldwork I conducted in Louisiana during the Summer of 2003 with approximately 50 Cajun and Creole speakers of Acadiana French[15]. The results of my survey show that not a single one of them used Standard *frite*,[16] and indeed it seems the gap does not exist in Louisiana.[17] I thus predict the typological pattern in (229) within French dialects for *frire*.

(229)

	masc. [fri]~fem. [frit]	[fri] only
gap	Standard French	*
no gap	Middle French	Acadiana French

5.3.4 A comparison with Morin's account

Morin (1987, 1995) proposes an account in terms of stem suppletion of the French gaps. According to Morin, speakers extract implicational statements from the words they learn. For example, in French, the 3PLURAL stem serves as 3SING stem unless speakers learn another stem for the 3SING. This directionality makes it such that in the case of *frire*, since speakers only learn a 3SING stem, they never use it in the 3PLURAL.

I have three objections to this view of things. First, Morin predicts that speakers will *never* guess a 3PLUR from a 3SING. Though hard to prove in real discourse situation (since we don't know what words speakers have heard before), it is definitely not true in experimental situations where speakers are asked to conjugate non-sense words. Typically, if a speaker is presented with a 1SING non-verb ending in /-i/, s/he will simply add an /-s/ to form the 3PLUR: /kotri/ → /kotris/. This

[15]By Acadiana French, I mean here what is usually termed Cajun French. In this particular case, I follow Klingler (2003) in avoiding the latter term, since the variety is spoken both by Cajuns and Creoles. Louisiana French would not be a good term either, since it can be confused with the more Standard variety once spoken in New Orleans. Acadiana French refers to the Louisiana parishes officially designated as Acadiana where this variety is still spoken by Cajuns and Creoles, so it has less of an ethnic connotation.

[16]Speakers were asked to translate the phrase *fried potatoes*, which invariably showed up as *patates fries* [fri] and never **patates frites* [frit].

[17]This is an extrapolation: speakers were not consulted on 3PLUR PRESENT forms, but on the IMPERFECT SINGULAR, which uses the same stem as the 3PLURAL in both Standard and Acadiana French and which are also defective in Standard French. Louisianans used either *frisait* or *friait*, which are both attested in Middle French—see Pope (1934).

productive strategy is based on the SECOND GROUP, which includes over 300 verbs. Second, though Morin is careful to state that his model is one of adult language, his model, if it were transferred to child language would predict error patterns that do not conform to what is actually observed. For example, some Acadiana French speakers have extended the Standard SING stem /tjɛ̃-/ of *tenir* to all persons, while Morin's statements imply that it is the SING persons that are predictable from the PLURAL persons. (Thus Morin needs a separate model for language acquisition and change, while TCWC is an integrated model). [18] Third, it is not clear what, apart from paradigm gaps, helps speakers establish the directionality of the implicational statements. The model is thus potentially circular and opens the door to admitting negative evidence in accounting for the gaps (unless the directionality can be established independently).

Like Morin's (1987, 1995), my account is one that uses explicit symbolic representation. However, the model I use has more ambitious aims of not being solely a model of adult grammar, but also of morphological acquisition. It also makes the implicational statements of Morin emerge (in some contexts only) from independently needed Lexical Insertion Conditions.

One interesting fact about *clore* that Morin gets right is that, because his implicational statements are set up such that the 2PLUR stem may be used for the 1PLUR, and in turn the 1PLUR stem may be used for the 3PLUR, Morin predicts (rightly, as far as I know) that there shouldn't be a French dialect with a gap in the 1PLUR or 2PLUR, but with no gap in the 3PLUR. However, if we were to take this prediction seriously, then we should wonder why is it that no dialect has a gap in the 2PLUR, but not in the 1PLUR or 3PLUR...

In TCWC, we must say that these observations are probably due to the fact that the 1PLUR and 2PLUR are less common than the 3PLUR, and hence, are less likely to have retained the historical forms if the 3PLUR hasn't done so itself.

Finally, Morin reports one last gap, for which neither of us has a satisfactory answer. Morin reports that some speakers have no PRESENT SINGULAR forms for the verb *colorier* 'color'. His explanation relies on the assumption that these speakers postulate a phonemic representation /kɔlɔrje/ instead of /kɔlɔrie/. The form /kɔlɔrje/ is closer to the surface realization and there is indeed a phonemic contrast in French between /i/ and /j/: /pɛj/ 'pay' vs, /pɛi/ 'country'.

Thus when speakers try to form the PRESENT SINGULAR by removing the final /e/, they are left with a phonologically ill-formed */kɔlɔrj/. TCWC predicts exactly the same thing:

[18] In any case, analogical change is not necessarily based in language acquisition—see Hopper & Traugott (1993:204-209, 2003:43-50). Hence, Morin's model might not work for the Acadiana change either.

(230)

VERB INFINITIVE 1STGROUP	
STEM	
parl tõb mãʒ kɔlɔrj	+ e

In the example above, the STEM is equal to the VERB PRESENT INDICATIVE 1STGROUP SING for all 1STGROUP verbs. Thus *colorier* is analyzed as being underlyingly /kɔlɔrje/, the STEM or PRES SING should be */kɔlɔrj/, a sequence unpronounceable as such.

The problem—if there is one—is the reliance on a theory of phonology that rules out certain underlying forms. This runs counter to a standard assumption in current phonology called RICHNESS OF THE BASE (Prince & Smolensky 1993), which states than there is absolutely no restriction on underlying forms. The insight, which can be traced at least back to Stampe (1973), is that phonology's job is to make pronounceable any input by various substitutions, insertions or deletions. Thus, if [kɔlɔrj] is unpronounceable in French, an output [kɔlɔri], [kɔlɔr] or [kɔlɔrje] should be available.

If RICHNESS OF THE BASE is a correct assumption for phonology, I am forced to admit that TCWC, like Morin's account, fails to account for the *colorier* facts.[19]

5.4 Spanish

5.4.1 Type 1: *abolir*

Albright (2003) notes that there are two types of paradigm gaps in Spanish. The first type consists of verbs like *abolir* and *aguerrir* that have gaps for all the forms for which all other verbs with a root in /o/ or /e/ do not preserve these vowels by diphthongizing them to /we/ or /je/ respectively. However, in most cases where /o/ or /e/ raise to /u/ or /i/ respectively,[20] defective verbs like *abolir* and *aguerrir* do not raise their root vowel. In line with the analysis of the French gap for *frire*, this provides speakers ample evidence that *abolir* and *aguerrir* are not to be inserted with the major

[19]This does not imply that TCWC adopts OT phonology. TCWC is relatively neutral as to how "pure" phonology is to be dealt with. However, since I concluded that the *colorier* facts involve phonology, it is my responsibility to point out that there is a problem with this conclusion at least in the most widely used phonological framework at the moment. I would like to also acknowledge the possibility that I am wrong and the *colorier* facts should really be treated in morphology exclusively.

[20]PRETERIT 3PLUR, PASTPARTICIPLE, IMPERFECT SUBJUNCTIVE, but not PRESENT SUBJUNCTIVE 1PLUR AND 2PLUR.

classes of roots with the vowels /o/ and /e/, repectively:

(231)

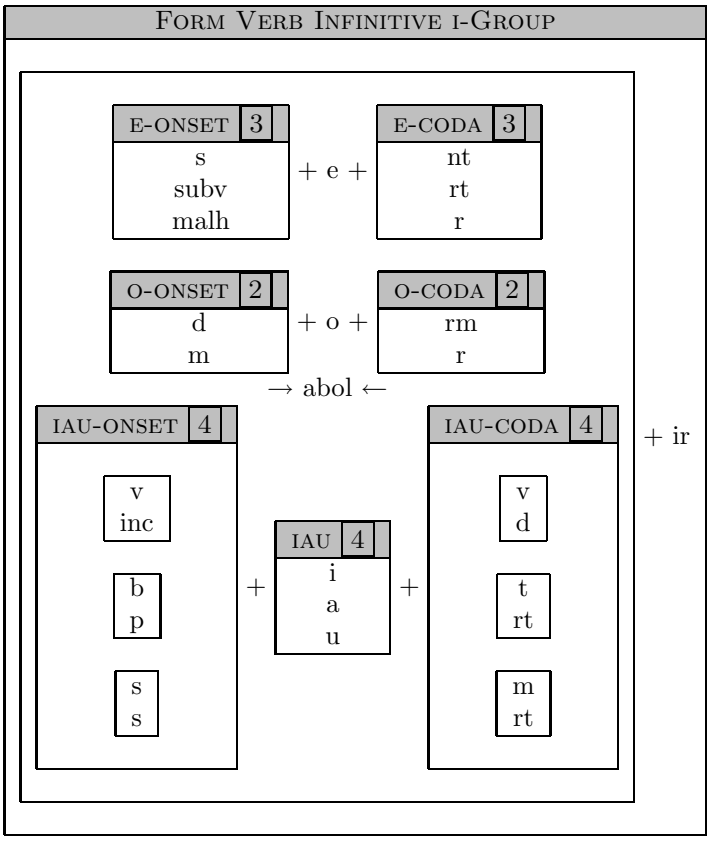
INFINITIVE	PASTPARTICIPLE	1SING
dormir	durmjendo	dwermo
abolir	aboljendo (counter-evidence)	GAP
sentir	sintjendo	sjento
agerrir	agerrjendo (counter-evidence)	GAP

In (232), the INFINITIVE forms of some classes of Spanish verbs are listed. In (233), the 1SING PRESENT forms of these classes are listed. Since *abolir* has an o-root, the INFINITIVE could be stored with *dormir* and the like and its 1STSING PRESENT should thus be **abuelo*. However, speakers receive evidence not to store *abolir* with the other o-rooted verbs: this evidence is that the PRESENTPARTICIPLE form is *aboliendo* and not **abuliendo* as it should be, in order to be stored with *dormir*. Thus, by COUNTEREVIDENCE RESPECT, *abolir* is not stored with *dormir*.

Given that *abolir* has an o-root, by GENERALIZATION PRESERVATION, it cannot be stored with verbs like *sentir* either. And finally, by LOCALIZED GENERALIZATION, *abolir* cannot be included in the class that contains *vivir*, etc., since it would require inserting three strings in the three different LexiBlocks (*ab-*, *-o-* and *-l-*), when it is possible to insert it elsewhere in a single LexiBlock.

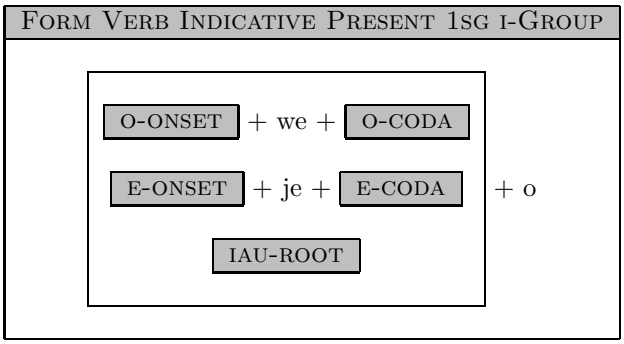
Since the INFINITIVE does end in /-ir/ though, it is stored in the larger LexiBlock as a class of its own and the inflections that are common to all classes in this LexiBlock such as the 1STPLURAL PRESENT pose no problem: *abolimos*, *dormimos*, *vivimos*. The same reasoning applies to verbs like *aguerrir*, which could fit the e-rooted class of *sentir*, but because of their learned PRESENTPARTICIPLE *aguerriendo* (and not **aguirriendo*), they are left floating between classes and only those inflections that apply to all /-ir/ verbs exist.

(232)

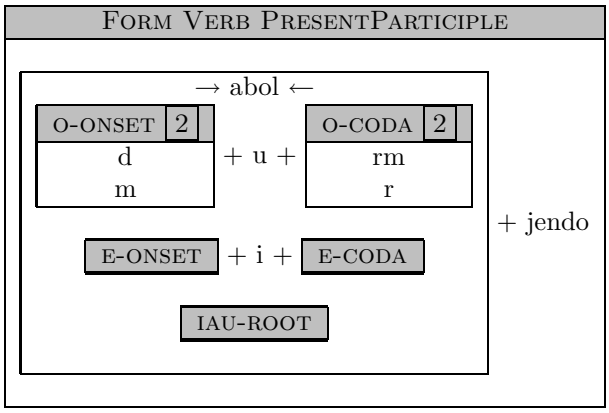


Def.: IAU-ROOT ≡ IAU-ONSET + IAU + IAU-CODA

(233)



(234)



There are seven verbs like *aguerrir*, with an /e/ root, for a total of nine defective Spanish verbs that fall in this class: *agredir*, *arrecir se*, *aterir se*, *denegrir*, *empedernir*, *transgredir* and *trasgredir*.

5.4.2 Type 2 *balbucir*

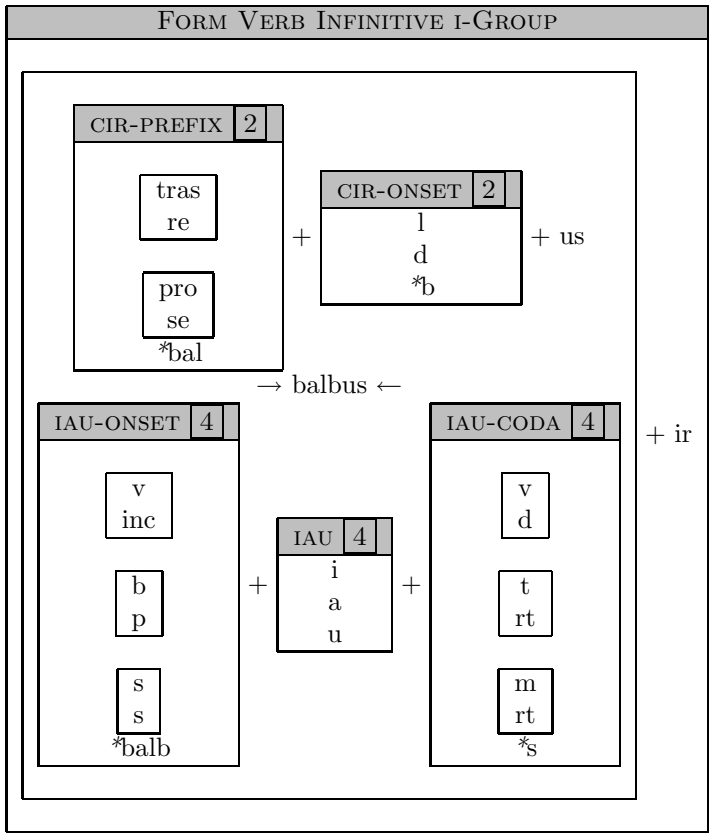
The other type of gap in Spanish concerns verbs whose INFINITIVE ends in /-usir/, like *balbucir*. The gap here occurs in those inflections where some verbs in *-cir* have velar insertion, e.g. for *lucir*, the 1SING PRESENT is *luzco*, but *balbucir* has no form for this person-number. The PRESENT SUBJUNCTIVE and POLITE IMPERATIVE forms also have velar insertion for *lucir*-type verbs and show gaps for *balbucir*-type verbs.

(235)

INFINITIVE		lucir	balbucir
PRESENT	1SING	luzco	GAP
	2SING	luces	balbuces
	3SING	luce	balbuce
	1PLUR	lucimos	balbucimos
	2PLUR	lucis	balbucis
	3PLUR	lucen	balbucen

The two facts that make *balbucir* float between the IAU-ROOT and CIR verbs in (236) are the following. 1) Though there are u-roots among the IAU-ROOT verbs, there are none ending in /-sir/; thus by LOCALIZED GENERALIZATION, *balbucir* cannot be stored there. 2) The only two classes belonging to the CIR-ROOT verbs are those in *-lucir* or *-ducir*, two subclasses using prefixes; thus again by LOCALIZED GENERALIZATION, *balbucir* cannot be stored there either, because it can be stored in a single LexiBlock elsewhere.

(236)



5.4.3 A comparison with Albright's account

Albright (2003) does not really provide an account of Spanish paradigm gaps. He notices a correlation between familiarity with a verb and confidence on its inflectional forms. This leads him to suggest that familiarity and confidence are factors that should be encoded directly in the grammar in a quantitative model such as the one proposed in Albright & Hayes (2002).

My main concern with this suggestion is that Albright equates the combination of unfamiliarity and uncertainty about a verb form (the opposite of confidence) with defectivity. We have seen evidence in this chapter that unfamiliarity and uncertainty do not directly translate into gaps. Remember that *strive* showed a comparable distribution to *drive* and *give*, although it is a rarer verb. Admittedly, it is not as rare as *stride*, but remember that in French, it would be hard to argue that speakers are unfamiliar with the verb *frire*. Further, it is not clear why uncertainty would not yield more gaps and why gaps occur under the conditions they do: morphophonological alternation, combined with specificity of phonological form, like *clore* and *balbucir*, and existence of idiosyncratic

forms elsewhere in the paradigm, like the *frire* and *abolir* type gaps. Albright does notice the first two factors, but it is not clear why and how uncertainty and unfamiliarity are limited to acting when these two factors come in. Speakers are always confronted with a series of options to inflect a new word, but in the end they make choices.

There are both similarities and differences between this account and the one presented in Albright (2003). The similarities include rejection of principles like homonymy avoidance and the acceptance of the fact that what causes the gap can be a combination of an exceptional morphophonological pattern combined with a generalization about the phonological class.

The main difference is that I do not suggest incorporating directly in the grammar knowledge about familiarity with and uncertainty about lexical items *to explain* gaps (though this knowledge may independently be included, if other evidence requires it). Rather, I account for the gaps with principles independently needed to insert new words in the existing CWCs. This follows from the integrated character of morphology and the lexicon in this theory.

Another difference is that paradigm behavior, through e.g. exceptional participles, plays a role in explaining a gap (by COUNTEREVIDENCE RESPECT). This follows from the Word-and-Paradigm origins of TCWC.

5.5 Russian

Of the languages examined in this chapter, Russian is the one with most paradigm gaps. Halle (1973) mentions that there are about 100 verbs in Russian that lack a 1SING PRESENT form. Using a reverse Russian dictionary by Zalizniak (1977),[21] I was able to identify 59. They are listed in (237). The three Lexical Insertion Conditions were sufficient to explain them all.

Some observations about Russian defective verbs are in order. First, all of the defective Russian verbs use either the theme vowel /-e-/ or /-i-/. Second, all the Russian defective verbs are either prefixed or fall into a subclass formed mainly or exclusively of prefixed verbs. Third, when a Russian verb is defective, all of its prefixed forms are also defective. Finally, the 1SING PRESENT uses a different stem structure in that many verbs ending in /t/, /d/, /s/ or /z/ replace these consonants respectively with /tʃ/, /dʒ/, /ʃ/ and /ʒ/.

[21] I thank Lev Blumenfeld for pointing out this reference to me.

(237) **Russian verbs with a 1Sing Present gap**

bdetʲ	obezloʃaditʲ	kudesitʲ
galdétʲ	pobeditʲ	nakudesitʲ
zagaldetʲ	ubeditʲ	tʃudesitʲ
pogaldetʲ	predubeditʲ	natʃudesitʲ
dudetʲ	pereubeditʲ	oblesitʲ
podydetʲ	razybeditʲ	obezlesitʲ
prodydetʲ	ugorazditʲ	lisitʲ
	sbrenditʲ	lʲɪlesositʲ
	sbonditʲ	prolʲɪlesositʲ
oburʒuazitʲ	erunditʲ	ljapsitʲ
ugobzitʲ	naerunditʲ	sljapsitʲ
ljamzitʲ	tʃyditʲ	parusitʲ
sljamzitʲ	natʃuditʲ	obrusitʲ
derzitʲ	potʃuditʲ	rʲɪsitʲ
naderzitʲ	ottʃuditʲ	zarʲɪsitʲ
merzitʲ	lixoraditʲ	prorʲɪsitʲ
buzitʲ		
nabuzitʲ		pretitʲ
	ʃelestetʲ	ʃerstitʲ
	zaʃelestetʲ	pereʃerstitʲ
tmitʲ	poʃelestetʲ	tʃtitʲ
zatmitʲ	proʃelestetʲ	potʃtitʲ

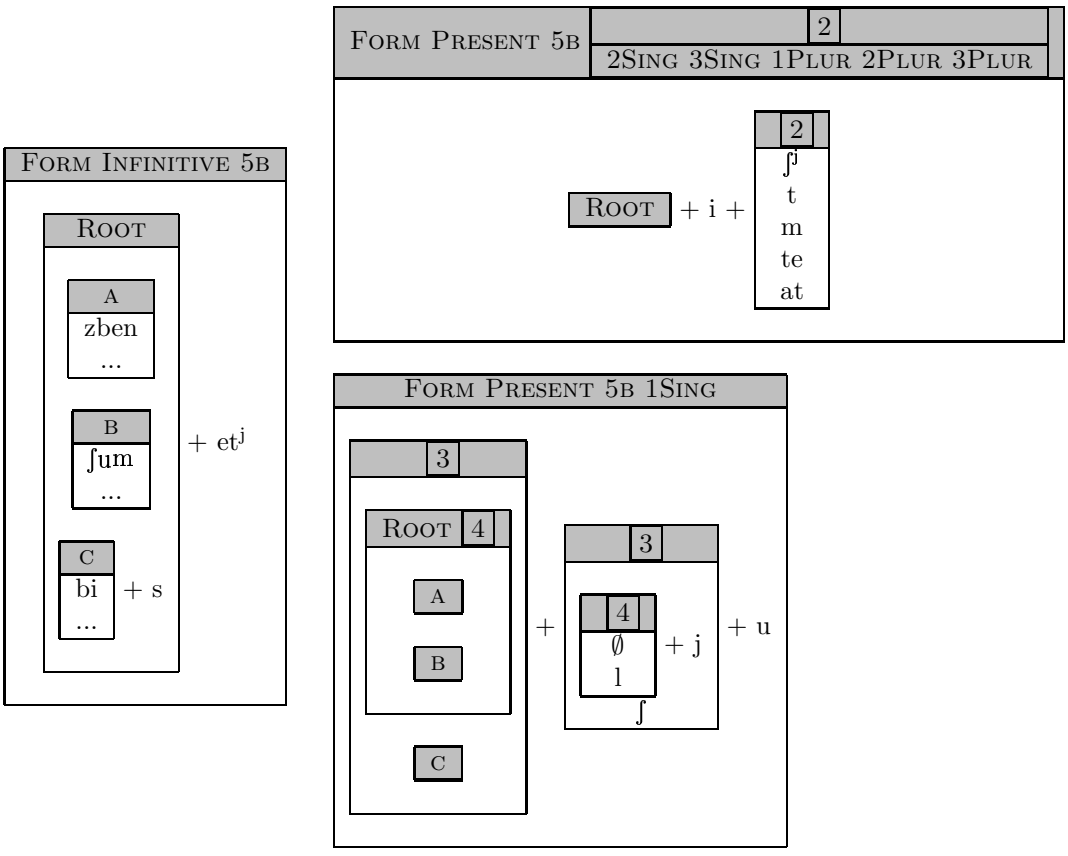
Hence, LOCALIZED GENERALIZATION should be sufficient to account for the Russian defective verbs. The defective verbs using the theme vowel /-e-/ are less numerous, so I will discuss them first. All the defective Russian verbs with the theme vowel /-e-/ belong to Zalizniak's inflectional class 5b. Below I give the PRESENT paradigms of three sample verbs from this class without paradigm gaps.

(238) **Russian Present; Zalizniak Class 5b; e-Stems**

Infinitive	zben-e-t^{j}	ʃum-e-t^{j}	bis-e-t^{j}
1Sing	zben-i-u	ʃum-lju	bi-ʃu
2Sing	zben-i-ʃ^{j}	ʃum-i-ʃ^{j}	bis-i-ʃ^{j}
3Sing	zben-i-t	ʃum-i-t	bis-i-t
1Plur	zben-i-m	ʃum-i-m	bis-i-m
2Plur	zben-i-te	ʃum-i-te	bis-i-te
3Plur	zben-i-at	ʃum-i-at	bis-i-at

As we can see, the verbs in the person-numbers other than the 1Sing use the root, followed by the theme vowel /-i-/, followed by the same set of suffixes. In the 1Sing however, the pattern is different: the root is concatenated with slightly different suffixes for each subclass, and undergoes palatalization in the case of /biset^{j}/. This situation should be familiar to us by now: the stem structure being different, the Integration Step will fail to group the 1Sing with the other person-numbers.

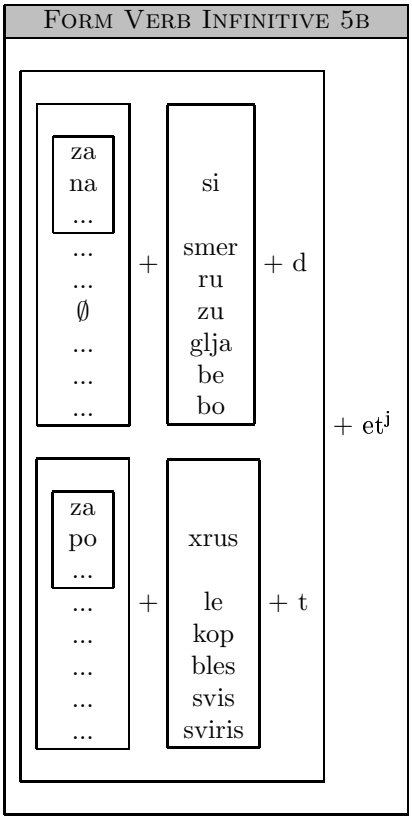
(239)



In (240), the relevant part of the lexicon with the non defective verbs of class 5b is illustrated. In the case of e-stem verbs, only certain verbs whose roots end in either /-d/ or /-t/ are defective, so I

have only listed the non defective in /-d/ and /-t/ in (240). I have enumerated all such lexical roots, but for simplicity, I have left blank the slots where the idiosyncratically selected prefixes should be specified.

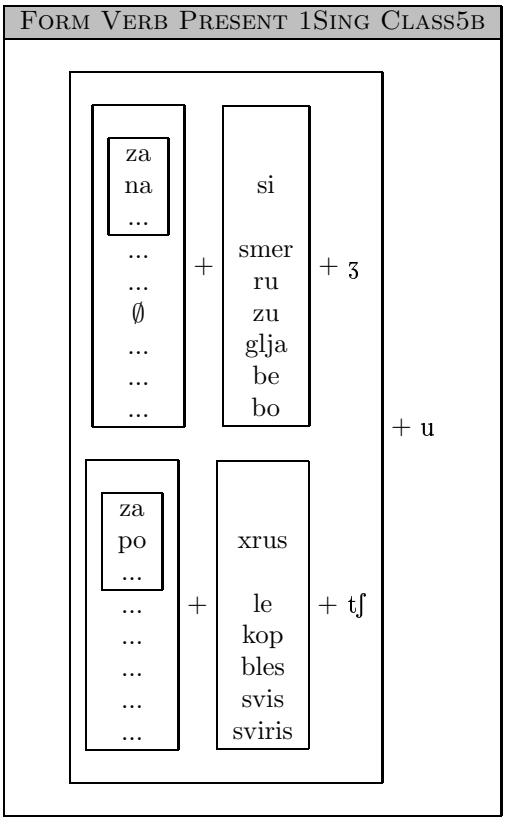
(240)



All of the relevant non defective verbs except one (/zudet^{j}/) select for a set of prefixes, which I have only partially specified only for the first roots of each group.

Therefore, when learning a new verb, by Localized Generalization, the speakers will prefer to insert it between subclasses. Hence, since the formation of the Present 1Sing refers to each subclass independently, the newly learned verbs, stuck in between classes are not selected.

(241)

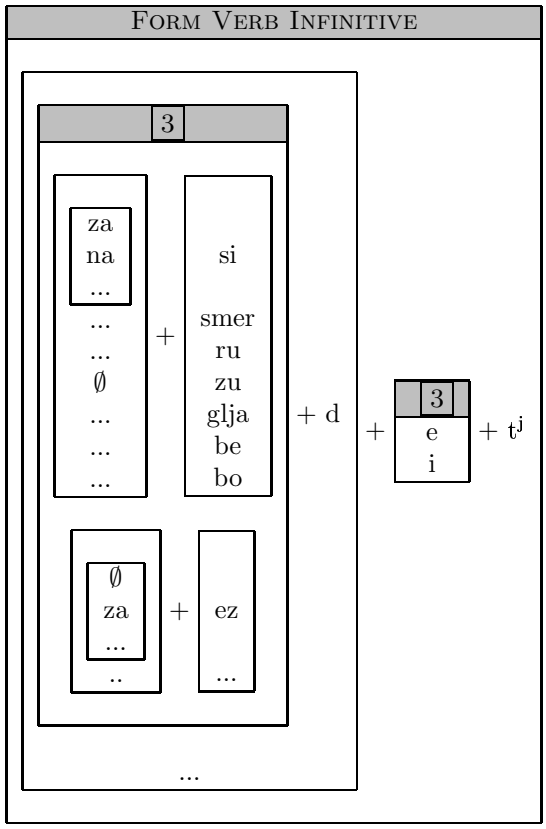


In (241) then, the individual classes are referred to separately, the root-final consonants are replaced, and the suffix /-u/ is used.[22]

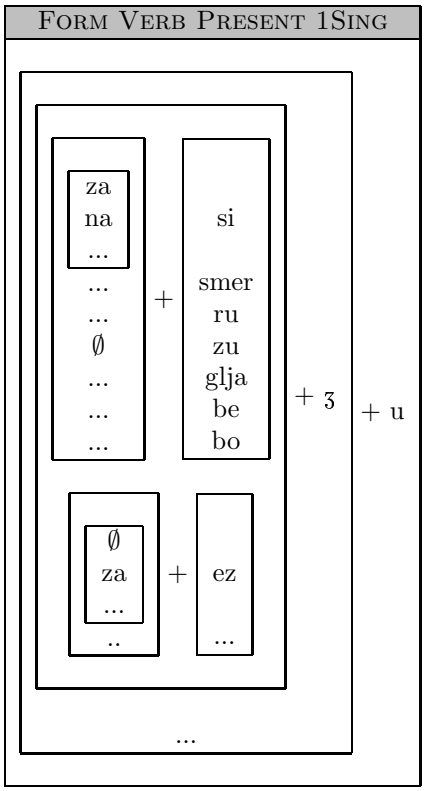
Verbs that use the theme vowel /-i-/ behave essentially in the same way. There are however, more of them. The defective ones all belong to Zalizniak's classes 4a and 4b. Apart from the theme vowel, this is not a different class from 5b. Hence, we can place the theme vowel in a LexiBlock, and thus avoid repeating the alternation between /d/ and /ʒ/, /t/ and /tʃ/, etc., and rewrite (240)-(241) as (242)-(243)

[22]The roots ending in /-st/ should actually be replaced by /-ʃtʃ/, but I leave this matter aside for simplicity and I assume it can be handled by the language's synchronic phonology, though if that turned out not to be the case, it wouldn't be a stretch to partition the /-t/ class further in (240) and (241).

(242)


Form Verb Infinitive
3
za
na
...
...
...
∅
...
...
...
+
si
smer
ru
zu
glja
be
bo
+ d
+
3
e
i
+ tʲ
∅
za
...
..
+
ez
...
...

(243)



This account of the Russian defective verbs must also explain why it is that precisely these verbs have no inherited forms for the 1Sing Present. Lev Blumenfeld (personal communication) explains that they are all either denominal verbs (thus of more recent existence) or borrowings from Old Church Slavonic. Hence, it is not surprising that Russian speakers have no inherited 1Sing Present forms for these verbs. In the case of the denominal, a 1Sing form never existed, while in the case of the Slavonic borrowings, they were probably borrowed via their Infinitive forms, and given the modern Russian inflections, which is problematic for the 1Sing Present.

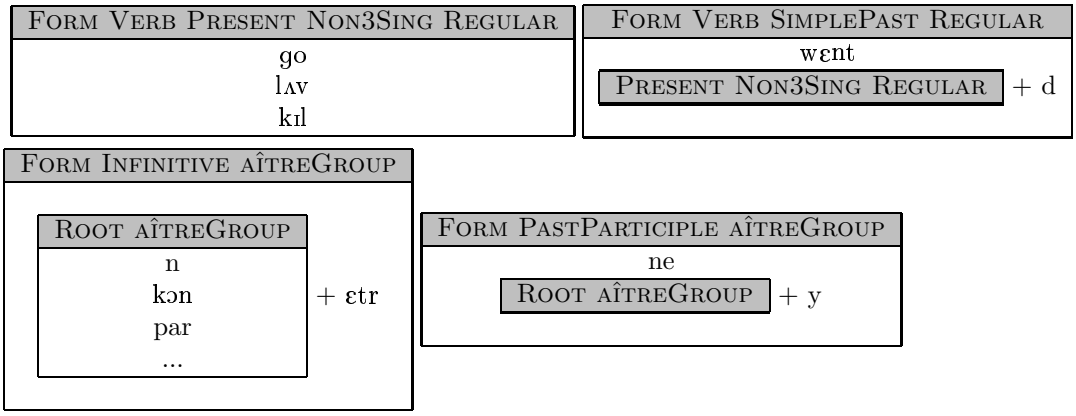
5.6 A problematic type

In conclusion, let us examine a type of paradigm gap that TCWC aims to account for, but for which it fails. I should state that I know of no other theory that can account for this type of paradigm gap. I know of one example in English and one in French. The English example is *forgo*, for which many English speakers have a gap in the Past (**forwent*, **forgoed*), but, as far as I know, not for the PastParticiple, which is invariably *forgone*. The French example is *renaître* 'reborn', which,

according to the Bescherelle (1992) verb conjugation tables, shows a gap in its PastParticiple form (**rené*, **renu*).

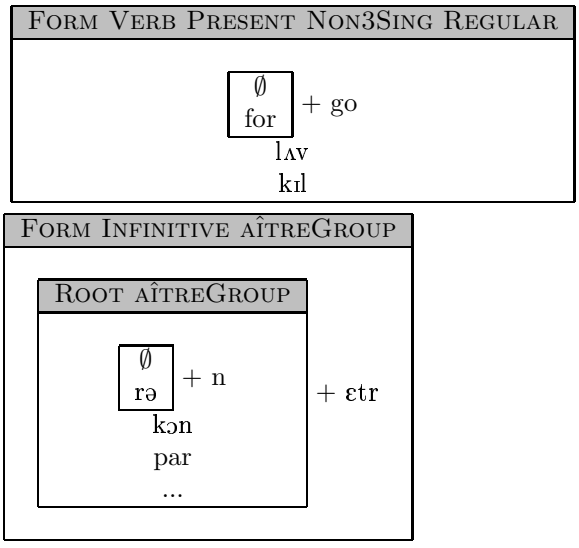
There are two things common to both these cases: 1) the gaps occur on prefixed verbs, but not on their unprefixed counterparts; 2) the gaps occur when the unprefixed verbs use a suppletive form. The common situation of both these verbs is well illustrated in TCWC, as shown in (244), the problem is that the Lexical Insertion Conditions as stated so far are not sufficient to account for the gap.

(244)



The Lexical Insertion Conditions do not prevent a prefix such as English /for-/ or French /rə-/ from being grafted onto a root:

(245)



This only requires creating and inserting phonemic material in one LexiBlock, thus not violating Localized Generalization. The good news is that this rightly predicts most of the inflections

of *renaître* and *forgo*. The bad news is that it wrongly generates the non suppletive forms */rəny/ and */forgoʊd/.

5.7 Conclusion

In this chapter, I have been able to show that the very same Lexical Insertion Conditions independently motivated to account for facts of analogical change and acquisition are enough to explain paradigm gaps in English, French, Spanish and Russian. True, some types of paradigm gaps remain problematic, but they remain so for every theory proposed. As was made clear while discussing English *stride*, Distributed Morphology and Paradigm Function Morphology have no handle on any type of paradigm gaps, because of their systematic reliance on default or elsewhere principles. By contrast, the Elsewhere Step of TCWC only applies under restricted conditions, so that we do not have defaults systematically. We have also compared TCWC in two specific case studies with the accounts of Morin (1987, 1995) and Albright (2003). While close in spirit, Morin's account is potentially circular in motivating its implicational statements with paradigm gaps themselves, and does not have the same breadth as TCWC. TCWC is also sympathetic to Albright's suggestion of making paradigm gaps fall under notions of unfamiliarity and uncertainty that would ultimately be directly encoded in the grammar, though some clear-cut cases remain that such an eventual account could not handle.

It is true that TCWC cannot (yet) account for paradigmless verbs (French *douer*, English *beware*), gaps yielded by an apparently unacceptable underlying form (French *colorier*), or gaps of the type exhibited by English *forgo* or French *renaître* that seem to involve the Elsewhere Step. However, one should keep in mind that any progress in accounting for paradigm gaps is significant, especially since the two leading theories, Distributed Morphology and Paradigm Function Morphology, have no handle on defective verbs at all.

Chapter 6

Phonology and Morphology

In this chapter, I tackle the difficult task of defining the border between phonology and morphology for TCWC. This is a difficult task, because TCWC claims to be strictly a model of morphology, and it is compatible with more than one model of phonology. Therefore, the reader should expect some variation concerning the exact place where the line falls between phonology and morphology, depending on the model of phonology one decides to use along with TCWC. TCWC *can* account for a lot of morphonological facts, but if a given model of phonology has better reasons for treating a given phenomenon within phonology, then TCWC does not have to account for these facts. We could imagine, for example, that a given phenomenon is first analyzed by speakers as morphological, but in later stages of acquisition, it becomes more economical for the speaker to treat the phenomenon within his/her phonology. Finally, it could be that some phenomena are in a transition stage for a generation of speakers, and that both a phonological and a morphological analysis of a phenomenon are necessary to account for the facts at hand. Linguists who are strictly interested in the most economical account possible may find this reasoning suspicious, but who is to say that speakers do not reinforce certain constraints in more than one way? This chapter will illustrate all of these situations.

6.1 Morphonology as morphology

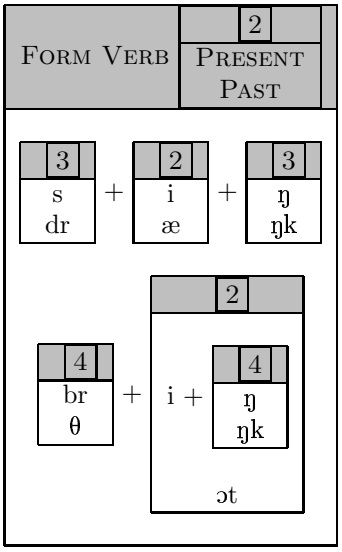
In the struggle to define the place of morphonology in TCWC, there are clear cases.[1] For example, in English, the vowel alternations between *sing/sang* and *bring/brought* are, as far as TCWC is

[1] See the collection of papers and debates in Singh (1996) for a wide range of views on the question.

concerned, clearly in the realm of morphology.

For *sing/sang* to be considered a phonological alternation in TCWC would require a very abstract analysis, accompanied by a very limited phonological rule, or a series of formal tricks on the underlying representation to allow a general phonological constraint to become active. The most straightforward way in TCWC to account for these alternations is as follows:

(246) **English past vowel alternation in TCWC**

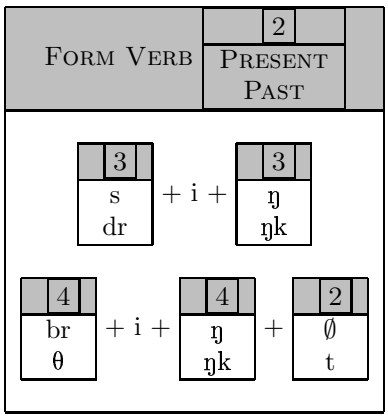


One alternative in a rule-based phonological framework would be to propose a single underlying representation for both *sing/sang*, along with a lexically restricted phonological rule. This rule would have to be lexically restricted to a subset of the lexicon, and not any morphological level, because there are verbs such as *bring* whose past is not **brang*, but *brought*. This would yield the analysis in (247).

In TCWC, this would not be a great analysis, no matter what the accompanying phonological framework is, because it would complicate the phonology by the addition of three limited rules, while not simplifying the morphology. Further, it would complicate the acquisition procedure of the CWCs. Remember that TCWC assumes that learners first store fully inflected words and then collapse CWCs and LexiBlocks, using the steps of Chapter 3. Hence, after changing the underlying form of the past tense of *drink*, an extra step of restructuring the LexiBlocks in the relevant CWCs would be required to arrive at the alternative analysis above. As stated in the chapter on analogy and acquisition, the monotonic nature of the acquisition procedure is a simplification, and reanalyses surely do take place in real-time language acquisition, but in this case it seems unjustified, because

it does not help simplify anything. Why complicate something that has worked so well so far?

(247) **Alternative English past vowel alternation in TCWC**



i → æ/ Past of *sing* and *drink*
i → ɔ / Past of *bring* and *think*
(C)C → ∅ / _t, Past of *bring* and *think*

6.2 Morphonology as phonology

I will now introduce a case of morphonology that would clearly have to be treated by the phonology under TCWC. In Quebec French, we find an alternation between tense and lax high vowels:

(248)

[vɪt]	'quick'	[vitɛs]	'speed'
[lʏk]	'Luc'	[lykɑs]	'Lucas'
[trʊp]	'troop'	[trupo]	'herd'

This alternation however is not at all limited to these morphological contexts; it concerns the whole lexicon of the language. The lax high vowels are not phonemic and no word of Quebec French can have a tense high vowel in a checked syllable.[2] Thus by any phonological framework of which I am aware, the tense and lax counterparts of the high vowels would be considered part/instantiations of the same phoneme. So far in TCWC, we have only used phonemes in our representations, and there is no reason to start doing otherwise. If we did, we would effectively have to multiply the number of CWCs of each language by the number of allophones this language has, thus immensely complicating the morphological system, eliminating the need for phonology and obscuring the physiological and acoustic motivations that are so much more transparent in phonology than in morphology.

However, this analysis assumes that the allophonic status of the lax vowels is detected early on in acquisition, or else a restructuring of the CWCs will be necessary once the speaker realizes [i] and [ɪ] are part of the same phoneme. Under certain phonological assumptions though, they will

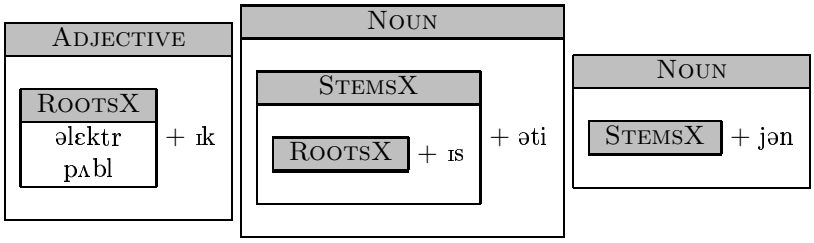
[2]Unless the said vowel is also long, in which case it is usually a borrowing from English, such as *cheap*, *cool*, etc.

be part of the same phoneme by default. Such was the insight of Natural Phonology (Donegan & Stampe 1979) carried over in the versions of OT where the markedness constraints are ranked higher than faithfulness constraints in the initial stages of acquisition (Gnanadesikan 1995).[3] It is not our purpose here to discuss whether this is a well-founded position in OT, but to the extent that TCWC is an appealing model of morphology, the simplicity that the initially highly ranked markedness constraints brings to TCWC morphology is an argument for this position within OT, and its higher compatibility with TCWC.

6.3 Morphonology as probably morphology

In TCWC, velar softening in pairs like *electric/electricity* is probably better treated within morphology. The phones [k/s] involved here both have phoneme status, the alternation is not general (automatic) in the language: we do not find it elsewhere than before a set of suffixes. Clearly then, our acquisition procedure will yield the following CWCs:[4]

(249)



A predictable argument against a morphological treatment of velar softening is that the alternation will have to be repeated in every morphological context that uses it. But this argument does not hold in TCWC as illustrated above, since it is possible to tag the alternate stem of *electricity* and use it in *electrician*.

In cases of this type, I can think of three categories of arguments for a compatible theory of phonology to convince me that, for example, English velar softening is better treated in phonology than in morphology. Since the acquisition procedure yields the CWC above, we need convincing arguments that a simpler phonological analysis exists that can justify the restructuring of the CWC. We also need to show that the alternation behaves like truly phonological ones; in other words,

[3] For a different opinion, see Hale & Reiss (1995).

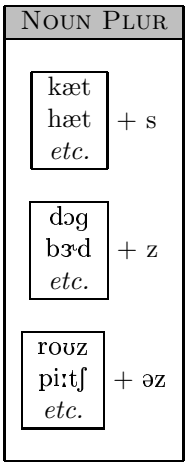
[4] I am obviously concentrating on the relevant part of the lexicon. There is no need to state all the *-ity* and *-ian* taking verbs, though it should be clear by now that those facts could be integrated here. If we did not limit the scope of CWCs in our examples, we would end up repeating the entire lexicon on every page. Likewise, in OT, if one were really to be exhaustive, one would state the hundreds of constraints in a language for every little phenomenon.

since CWCs relate underlying forms, is there external evidence—psycholinguistic or otherwise—that speakers have a /k/ in *electricity* that comes out as an [s], just like they have a /t/ in *matter* that comes out as a flap?[5] We also need proof that the domain of the phenomenon (apparently the stem) is readily and unmistakably identifiable by speakers.

6.4 Morphonology as probably phonology

The case of the English plural suffix alternation in words such as *dogs/cats* is probably best treated by a phonological framework, although it is not a problem to account for it in TCWC. In this case, the phones [z]/[s] involved do have phoneme status, but the alternation is not limited to some morphological contexts;[6] the surface sequences [tz] and [ɡs] are just impossible in English. Nevertheless, in the first stages of acquisition, learners have no way of knowing what the correct underlying form for the plural suffix is in a given word, and because both /s/ and /z/ are phonemes, it is simpler to always posit them as underlying. Thus the initial CWC below:

(250)



Crucially in this case however, a phonological account is independently motivated by the absence of the word final sequences [tz], [ds], [ss], etc. elsewhere in the lexicon. Hence, restructuring the CWC above by introducing a single plural suffix /z/ for these three cases is justified, because it simplifies the morphology while allowing for a more elegant phonological analysis.

We can't exclude the possibility however that some speakers may never restructure their plural CWC, and have two constraints in their grammar (one in the phonology, one in the morphology)

[5] At least Stampe (1987:290-293) comes to the conclusion that there are no such arguments for velar softening.

[6] The same goes for the epenthesis of schwa in *roses* and other words ending in a coronal fricative. See Zwicky (1975) for a thorough review of the possible analyses of these and related facts.

that reinforce this state of affairs. In fact, in the long run, for a suffix as common as the plural suffix, it may very well be that it is more convenient to have a readily accessible allomorph rather then deriving the correct phonetic form each time. Pushed to its extreme, this logic would argue that the same thing is possible for the tense/lax alternation in Quebec French, and indeed, while it clearly is not the case in the synchronic system of the language, this is how I believe phonological alternations grammaticalize and eventually become morphological.

6.5 Ambiguous cases

Most cases of allomorphy studied in the previous chapters have been dependent on arbitrary classes, such as the French conjugational classes. But we know that allomorphy may also be dependent on the phonological make-up of the stem. A simple example is the French cognate to the English repetitive prefix *re-* /ri-/.

The French prefix surfaces as [re-] before vowels and as [rə-] before consonants. (Schwa and [e] are not the same phoneme in French). Sometimes, [r-] surfaces before low vowels (nasal or non nasal), or front midvowels. The only other exceptions are semantically opaque, for example some verbs with bound roots show [re-] before a consonant.

(251) **General distribution of the French repetitive prefix**

[r-] before low vowels

[re-] before other vowels

[rə-] before consonants

(252) **Exceptions to the general distribution of the French repetitive prefix**

Consonant-initial verbs with [re-] (semantically opaque uses of the prefix)

récapituler réceptionner récidiver réciter réclamer récliner récolter récompenser réconcilier réconforter récrier récriminer récupérer récurer récuser rédiger rédimer réduire référencer référer réfléchir réformer[7] réfrigérer réfugier réfuter régir régaler régénérer régresser régulariser régurgiter réifier réjouir rélargir rémunérer rénover répandre réparer répartir[8] répercuter répertorier répéter répliquer répondre réprimander réprimer réprouver républicaniser répudier répugner réputer réquisitionner réserver résider résigner résilier résiner résinifier résister résonner résorber résoudre respecter respirer resplendir resquiller ressusciter ressuyer restaurer rester restituer restreindre résulter résumer rétablir rétamer rétorquer rétracter rétrécir rétreindre rétribuer révéler réverbérer révérer réviser révolter révoquer révulser

[7]This verb means 'to reform'. There also exists *reformer* meaning more transparently 'to form again'.

[8]This verb means 'distribute'. There also exists *repartir*, meaning more transparently 'to go/start again'.

Low vowel-initial verbs with [re-]

réabonner réabsorber r(é)accoutumer réactiver réadapter réadmettre réaffirmer r(é)affûter réagir r(é)ajuster réaléser réamorcer réanimer réapparaître r(é)apprendre r(é)approvisioner réargenter réarmer réarranger réassigner r(é)assortir réassurer réhabiliter réhabituer réembaucher r(é)employer r(é)engager réensemencer réentendre

Mid front vowel initial verbs with [r-]

r(é)écrire r(é)essayer

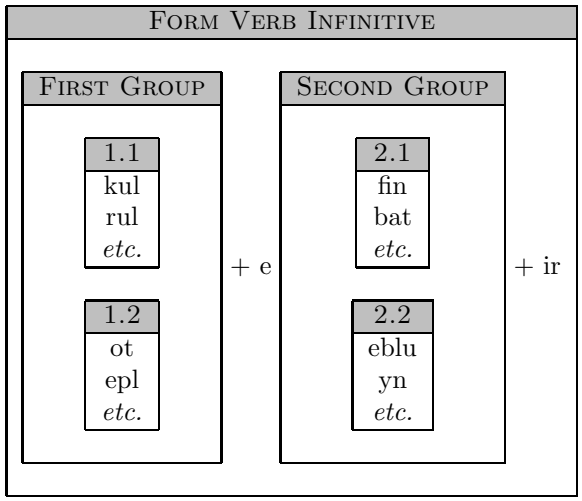
Other

rouvrir

Under a morphological analysis of this case, the dependency on the phonological make-up of words illustrated above raises important questions for TCWC. Is it necessary to presort the verbs in the CWC in which they are stored according to what allomorph they select? If so, do these classes crosscut with previous classes, generating a multiplication of classes? If not, how do we account for the phoneme-based generalizations?

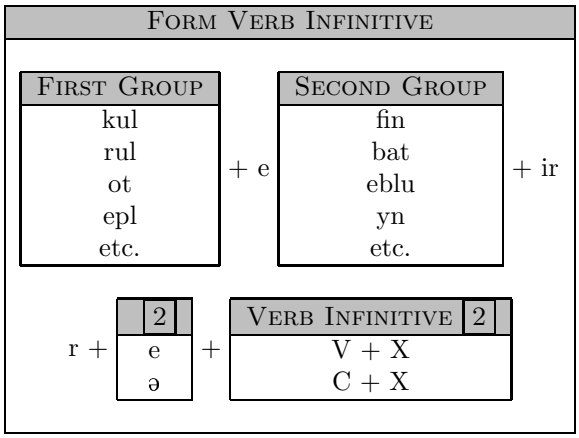
Of course, presorting the verbs according to which repetitive allomorph they select would multiply the classes uselessly. Take for example the two main classes of verbs, the first and second groups. They both contain vowel and consonant initial verbs. Thus, presorting them for the right allomorphs /re-/ and /rə-/ would require creating four verb classes (1.1, 1.2, 2.1 and 2.2).

(253)



Partitioning each one of the dozens of verb classes in this way would definitely not be an economical solution. Instead, we will make use of the different tags used by LexiBlocks to refer to the set of objects they contain and the specific ordered list that contains them.

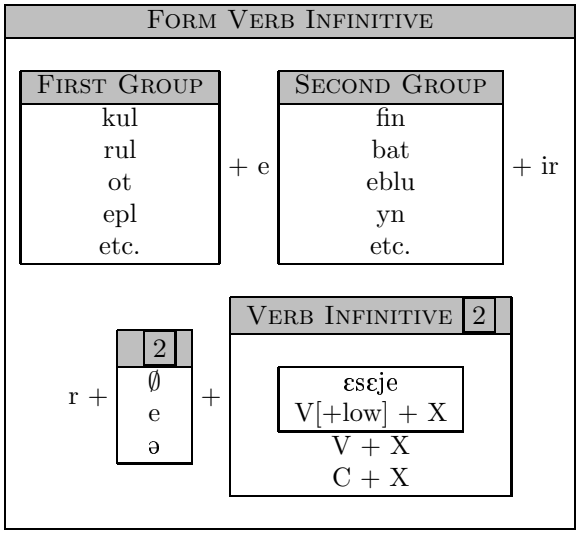
(254)



At the bottom of (254), the Infinitive verbs are referred to recursively, and divided up differently than in the conjugational classes stated above. While the set Infinitive refers to all the verbs stored above, the LexiBlock allows us to reorder them by the category of their first segment (C or V). Then the concatenation of this reordered set of infinitives yields more infinitives, which are in turn inserted in the embedded Infinitive LexiBlock, on the C+X line, because they will start with the consonant /r/. Thus the C+X line of this LexiBlock contains a countable infinity of verbs.

As mentioned earlier though, some vowel initial verbs take a special allomorph /r-/. Thus we need to modify slightly the CWC in (254) to accommodate them:

(255)



This case is not as clear-cut as the previous ones though, because some frameworks may use a rule whereby a vowel is postulated for the prefix and realized differently in front of vowels and consonants, while others may choose to simply have the stems select for different allomorphs.

(256) ə → e / #*r*_#V

or

re + $[VX]_{Stem}$
rə + $[CX]_{Stem}$

An argument for the phonological analysis would be that schwa is never found before another vowel in French within a word. However, unlike the case of the English plural, it is not clear to me that this is not simply a distributional gap in the lexicon of French that is not motivated synchronically by any phonological constraint.[9] Further, schwa is so often deleted in French that one would have to justify why deletion is not the preferred repair in this case as well. Perhaps we are dealing again with a doubly motivated case: there is still a constraint against schwa+V in French, and an old alternation between [ə/e] helped what was once two surface variants of a single prefix *re-* grammaticalize as two separate allomorphs. Or perhaps some phonological framework could really account for these facts more elegantly than TCWC can.

I am aware of the fact that several phonologists will view the double analysis option as a conspiracy problem (à la Kisseberth 1970), since we would have a phonotactic distribution and an allomorphy case apparently working towards a common goal (avoiding schwa+V). But really, if there is such a synchronic phonological constraint, what is wrong with having two corners of a single grammar each coming with a different reflex of the more general constraint? There are other complex systems that behave this way. Gravity is a single constraint that prevents most animals from flying. Should all the characteristics that allow birds to fly (hollow bones, feathers, wings) stem from a single genetic mutation? Survival is a single motivation that drives humans and other animals to adopt a million different behaviors. Are they all the result of a single process? Recognizing a unifying constraint is one thing, but it is entirely possible for the different components of a system to each come up with their own answer.

Vowel harmony cases are likewise ambiguous. While they are almost always tied to the morphological structure (spreading from the stem vowel to the suffix vowels), they are phonetically well motivated (they could be viewed as the phonological reflex of coarticulation), and they can even be exceptionless, thus motivating a simpler phonological analysis. My only concluding thoughts on the matter are that in TCWC, we have to look at these cases one by one, and not prejudge of their general nature.

[9] As a native speaker, I feel that such sequences don't sound foreign or "unpronounceable".

6.6 Morphoprosody as part phonology, part morphology

It is not difficult to name prosodic phenomena that are best treated within phonology. French stress assignment for example is completely regular (it falls on the last syllable of the word without a schwa) and it would complicate the CWCs immensely to encode this information within them. Likewise, a phonological (rather than lexical) analysis of Armenian syllabification, as elegantly demonstrated by Vaux (1998, 2003), allows one to capture several generalizations and should thus be favored

The argument runs as follows. Armenian has nominal possessive suffixes (1SING /-s/, 2SING /-t/). When suffixed to a noun ending in a consonant, a schwa is inserted between that consonant and the possessive suffix. When the noun ends in a vowel, no schwa appears on the surface.

Arguing for the phonological character of schwa epenthesis in Armenian, Vaux (1998) shows that it occurs both within and across stems, in known and unknown words, according to predictable rules. Vaux however points out an interesting distinction between Standard Western Armenian (SWA) and Standard Eastern Armenian (SEA). In SWA, the word-final sequences /ns/ and /ms/ always surface as [nəs] and [məs], while in SEA, this only happens when there is a morpheme boundary between /m/ or /n/ and /s/.

Thus we find a contrast in SEA between words like /tɔms/=[toms] 'ticket' and /mɔm-s/=[mɔməs] 'my candle'. While it is tempting to simply give up and admit that SEA has developed an allomorph /-əs/ for the 1SING POSSESSIVE suffix, Vaux persists and argues that syllabification occurs in cycles, first syllabifying the stem in the lexicon, then syllabifying the affixes to this stem:

(257) a. toms → $[\text{toms}]_\sigma$ (lexical cycle of syllabification)

b. mom
   → $[\text{mɔm}]_\sigma$ (lexical cycle of syllabification)
   → $[\text{mɔm}]_\sigma$ + s (suffixation)
   → $[\text{mɔm}]_\sigma$əs (epenthesis)
   → $[\text{mɔm}]_\sigma[\text{əs}]_\sigma$ (second syllabification)
   → $[\text{mɔ}]_\sigma[\text{məs}]_\sigma$ (adjustment rule)

Vaux (1998:29) argues further that traditional analyses by Armenologists, who assume a 1SING POSSESSIVE allomorph /-əs/, besides not accounting for the spontaneous use of epenthesis by native speakers must postulate an extra deletion rule in certain contexts. For example the word /ɑrɛv-n/[10]

[10]Vaux assumes that the underlying form is actually /areu-n/, for reasons I will not get into here.

'the sun', with the definite suffix /-n/ surfaces as [ɑrɛvən] in isolation, but as [ɑrɛvn] when the next word starts with a vowel.

(258)

[ɑrɛv]	[ɑrɛvən]	[ɑrɛvnɛ]
'sun'	'the sun'	'it is the sun'

However, at least in my dialect, this is not true of the possessive suffixes. The schwa present there is never deleted:

(259)

[ɑrɛvəs]	[ɑrɛvəsɛ]
'my sun'	'it is my sun'

In TCWC, it is therefore better to adopt an intermediate position where some schwas are underlyingly specified (and in this case are part of an allomorph) and some schwas are epenthesized by phonology. Vaux (personal communication) recognizes this possibility and cannot cite a speaker who pronounces [tɔmʂ] 'ticket', [ɑrɛvəs] 'my sun' and [ɑrɛvsɛ] 'it is my sun'. This is however what one would expect if the schwas in both [ɑrɛvən] and [ɑrɛvəs] are epenthesized. Why not recognize that the schwa in [ɑrɛvən] is indeed epenthesized, but the POSSESSIVE suffix has two allomorphs (at least for the speakers studied so far)?

(260)

FORM NOUN POSSESSED 1SING				
NOUN INDEFINITE 2		+	2	
X + V X + C			∅ ə	+ s

Vaux's non recognition of the phonemic status of schwa leads to another problem. For example, because the word /ngar/ 'picture' surfaces as [nəgar], but the word for 'friend' surfaces as [ənger], Vaux posits an empty vowel slot in the beginning of the underlying representation of this second word: /Vnger/, vowel slot later filled by a schwa. This is an unnecessary abstraction: it is not simpler to have an empty vowel slot than a full schwa, and in fact, it adds a burden on the learner who must make an extra hypothesis about the underlying form.

In conclusion then, TCWC has no problem in accepting Vaux's demonstration that Armenian schwa epenthesis can be phonological, and is happy to let phonology do the work in most cases. However, for the opaque cases, such as the possessive, where a consonant cluster sequence is pronounceable by speakers without a schwa, it is simpler to admit that the language has developed a

separate allomorph, while continuing to reinforce the phonological constraint, and that the schwa is now an underlying phoneme. On the surface, we can find schwas originating from three sources: sometimes it is part of an allomorph, as in the case of 1SING POSSESSIVE /-(ə)s/; sometimes it is epenthesized, as the schwa in DEFINITE [ɑrɛvən]; and sometimes it is just one of the phonemes of a lexical root, as in /ənger/ 'friend'.

6.7 Morphoprosody as morphology

In this section I will argue that in TCWC it is much more natural to treat Margi's famous tone alternations within morphology, rather than within phonology. Further, I will show how a melody-based morphological analysis of Margi's verbal tonology can help us gain some insight on this language that phonological analyses have failed to capture. This analysis is inspired by the melodies of Leben (1978), and the proposals of Donegan & Stampe (1983) on how to treat rhythm.

Based on Hoffmann's (1963) grammar of Margi, Williams (1976) recognized the necessity of roots unspecified for tone in their underlying representation, which contrast in their morphological behavior with roots bearing tonal specification. Assuming a default low tone, the underlying contrast between unspecified and specified roots allows one to neatly describe the tonal patterns of Margi verbs. For example, when a toneless root is put together with a high-tone suffix, the suffix's tone spreads to the root, while underlying root tones are preserved. Toneless roots pronounced in isolation, however, get a default low tone.

(261) a. /ɗəl bá/ → [ɗə́lbá] b. /ɗəl/ → [ɗə̂l]
c. /ndàl bá/ → [ndàlbá] d. /ndàl/ → [ndàl]

Pulleyblank (1986) incorporated this insight in the architecture of Lexical Phonology and Morphology. This latter analysis requires a more complex set of formal tools, including morphological levels and extratonal segments. Pulleyblank (1997) recasts the Margi facts in a simpler Optimality Theory (OT) analysis. However, a disadvantage of this latter analysis is that it wrongly predicts the absence of falling tones and high-low patterns in Margi roots. Pulleyblank is aware of this prediction (p. 94), but does not seem to be aware of the falling tones in the nominal system. In Pulleyblank's account, the same ranking that accounts for rising tones, accounts for the "absence" of falling tones. The following is adapted from Pulleyblank (1997:97):

(262)

LH /vəl/	AlignR [H]	Faith [H]	Faith [L]	No Contour	AlighL [H]
→ LH vəl				*	*
H vəl			* !		
L vəl		* !			

(263)

HL CaC	AlignR [H]	Faith [H]	Faith [L]	No Contour	AlighL [H]
HL CaC	* !			*	
L CaC		* !			
→ H CaC			*		

The alternative analysis that follows rests on the hypothesis that there exists a verbal melody in Margi illustrated with the function below. The interpretation of this function is that the tone of a bare infinitive is low until a specified high tone (H) gives the "cue" to raise the tonal melody for the rest of the verb. Thus, it is L(ow) from "moment zero" until the specified H, where it raises to H(igh) for as long as is required.

(264) **The Margi infinitive melody**

f(x) = {L,]0, H[; H, [H, + ∞ },

where x ranges from the first syllable of bare infinitives to the last.

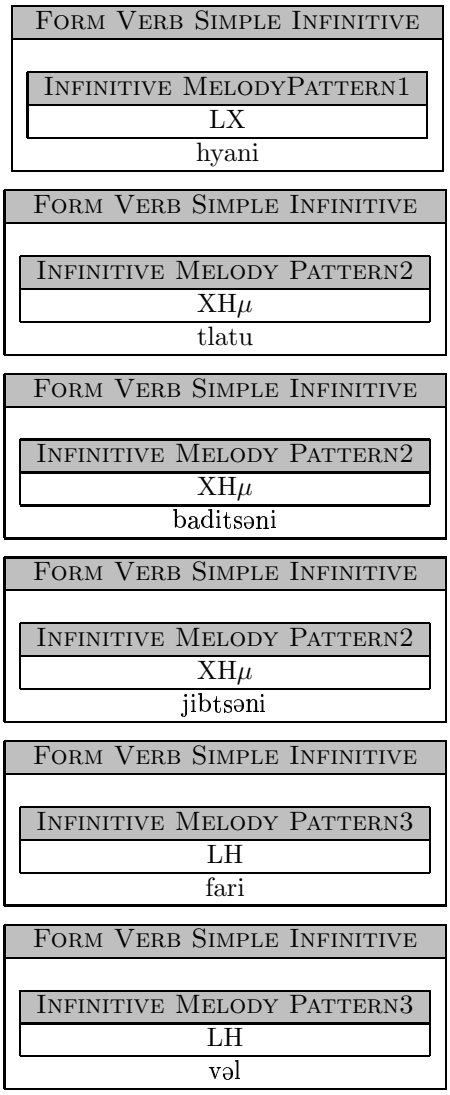
LLL...HHH...

In order to understand how words are mapped onto this melody, let us first look at different surface patterns of Margi infinitives:

(265) hyànì, tlátú, bàdìtsénί, jìbtsénί, fàrί, vəl

Given the architecture of TCWC and the formulation of the infinitive melody, we do not have to specify the entire part of the melody associated with the verb in the lexical entry. All we need for the preceding examples is the following information:

(266)

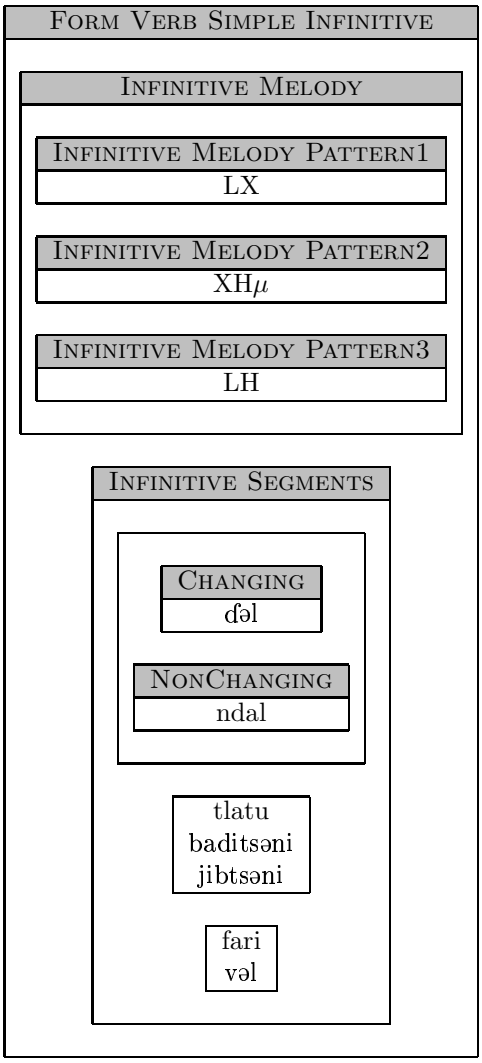


A low-tone verb needs only to specify that its first tone is L, and since no H cue is ever given, the tone will always remain low. Assigning a H tone to the syllable with the penultimate mora, can have different effects: on a disyllabic verb, it produces a H tone verb (because the penultimate is then also the first syllable); on a three or four-syllable verb, it creates an LH pattern. Finally, specifying LH is possible on both monosyllabic and disyllabic verbs, corresponding respectively to raising and LH tone patterns. The melody therefore states that by default a bare-infinitive's tone is low, until it raises.

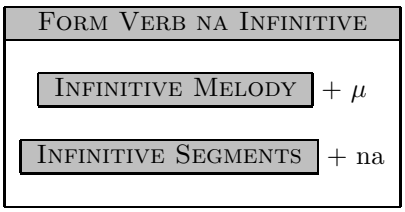
Our first CWCs are the ones for Simple Infinitive and Na Infinitive. The tonal melody being the same, the last tone on the Simple Infinitive will be copied onto the /-na/ syllable.

Possibly, we need to restrict via a more specific type the class of infinitives that can use this /-na/ suffix, however, since it is not clear from Hoffmann how productive it is, I leave the question open.

(267)



(268)



Thus, by adding a mora to the tonal melody, at the same time as we add the segments /-na/, the stems will simply keep the same tonal pattern and the last tone will be copied to the syllable

containing the final mora.

According to Hoffman (1963), in Margi, as in related languages, words often have distinct phrase final and non-final forms. As far as verbs are concerned, it seems that in most cases, these forms are identical, but that some but not all of the verbs which show up with low tone(s) finally change this to high non-finally. This is accounted for below, by forming a class of Margi verbs called the Changing class:

(269)

Form Verb Infinitive Changing NonFinal
Infinitive Melody Pattern4
HX
Segments Infinitive Changing Final

For example, final /ɗə̂l/ becomes /ɗə́l/ non-finally, because it is categorized as Changing, but NonChanging /ndàl/ remains the same, just like /tá/, /və̌l/ or /ptsàbá/. We need the Changing and NonChanging categories, because Changing verbs also fit the description of the lower part of the CWC. The CWC in (269) applies to both Simple and Na-Infinitives.

Some suffixes behave differently than /-na/. For example, the suffix /-ba/ imposes its own H tone at the end of the melodic pattern used by the Simple Infinitive. The analyses using defaults would have the unspecified roots obtain their tone by spreading from the tone of the suffix, but what happens here is that the CWC is built on the NonFinal form, hence we get /ɗə́lbá/, from the changing root /ɗəl-/.

(270)

Form Verb ba Infinitive	
Infinitive Melody NonFinal	+ H
Infinitive Simple Segments	+ ba

The derivational CWCs I have proposed have allowed me to reduce Williams' and Pulleyblank's underlying contrast of /L, H, ∅/ to /L, H/, i.e. what is visible on the surface, thus avoiding underspecification and the ternary use of a binary feature criticized by Lightner (1963) and Stanley (1967). Let us now turn to inflectional morphology and see what advantages can be gained in this area. In order to maintain the claim that some stems are underlyingly toneless, Pulleyblank has to

posit some extratonal affixes, e.g. the PRESENT prefix /a-/ and the object pronoun suffix /-nyi/, which are marked with the diacritic [+ex], so that their tone won't spread to toneless stems.

There is no need for an ad hoc extratonal diacritic with CWCs. For the present tense, where the prefix /a-/ takes the opposite value of the first tone, this can be directly incorporated in a CWC, using a reordering of LexiBlocks based on form:

(271)

FORM VERB PRESENT	
-T +	INFINITIVE MELODY: T + X
a +	NONFINAL INFINITIVE SEGMENTS

So for example, changing verbs like /ɗəl/, which has a final form /ɗə́l/, has a present form /àɗə́l/. Non changing verbs /sá/ and /wì/ have the present forms /àsá/ and /áwì/.

Pulleyblank (1997) requires unspecified tones because his analysis is morpheme-based and he has no other way to make tones interact correctly after putting morphemes together. Below, unspecified /fa-na/ get its default low tone, from the ranking of Faith[H] above Faith[L]—from Pulleyblank (1997:92):

(272)

fa-na	Interp	Faith[H]	Faith[L]
fa-na	*!*		
→ L fa-na			*
H fa-na		* !	

Pulleyblank (1986:130) recognizes that "Lightner's (1963) and Stanley's (1967) criticisms of underspecification are to be disregarded not because there is a way to avoid the pitfalls they describe, but because the extra power of a ternary feature system is required." The fact that TCWC doesn't require unspecified tones should be a good thing then, even according to Pulleyblank. This result is resonant with results obtained in more recent work in OT where whether features are specified or not does not matter, cf. Archangeli (1997), Bakovic (2000).

Pulleyblank (1997:95) also makes a wrong prediction: no words in Margi with a falling tone and no HL patterns within morphemes. The analysis presented here makes this prediction only for verbs and indeed Hoffmann gives examples of falling tones and HL pattern roots in Margi, like

[ə́mnà] 'yesterday', [lák'u] 'weak' [pâm] 'sterling pound'. It is true that Pulleyblank's formulation of Align constraints makes them apply only to verb roots. However, if that is truly the way of dealing with the difference between verbs and other types of words, his criticism of melody-based accounts (p. 95) does not hold anymore, because his formulation of Align constraints are no less a "stipulation" than the melodies of melody-based accounts. Indeed, he states that *for Margi, it would be stipulated that the inventory of tonal melodies is {L; H; LH}, with the additional possibility of leaving morphemes toneless.* I hope to have convinced the readers that a tonal melody approach needs not be so stipulative. The one proposed here, has a single INFINITIVE melody and *no* toneless morpheme possibility.

A further problem with Pulleyblank's OT analysis is that it is too powerful: because AlignR[H] is highly ranked and AlignL[H] is lowly ranked, he gets the desired result for verbs, but this implies that languages with rising tones but no falling ones are as frequent as languages with falling tones but no rising ones. This, according to Leben (personal communication), is definitely not the case; falling tones are much more frequent than rising tones in the world's languages. The present analysis does not have this problem, since the reason why rising tones are so frequent in Margi under this account is because of a lexicalized melody, but the phonology of the language need not forbid falling tones, it only needs to allow for rising ones.

A very attractive feature of CWCs is that they derive tonal properties about roots/stems, without presupposing such constructs. For example, because it is the INFINITIVES which map on the melody, we can describe the pattern of verbal roots. Also, this analysis rightly predicts that there couldn't be a low-tone suffix on Margi verbs deriving other INFINITIVES, because this would violate the melody.

The lexical character of the analysis predicts that some verbs might end up lexicalized in a more productive pattern or with idiosyncrasies. Hoffman notes that the verb [lṵ̀] 'to go' has a raising tone non finally (is toneless under Williams' and Pulleyblank's analysis), but that the corresponding /-ba/ derivative is [làbá] and not *[lábá] as the underspecification theory of these authors would predict. Likewise for the verb [pṵ̀] 'to put (many)', which also raises non finally, but for which the non final /-na/ derivative is [pànà] and does not raise to *[pánà]. Given the lexical character of the TCWC analysis, it is to be expected that some words might slip into a more productive pattern at some point and historically get lexicalized with an idiosyncrasy such as this.

6.8 Conclusion

To summarize, the Theory of Connected Word Constructions (TCWC) uses a powerful tool, the LexiBlock, and thus *can* account for several morphonological and morphoprosodical facts. The acquisition procedure from Chapter 3 is designed in such a way that it captures all the morphonological generalizations it can. Later though, if it turns out that the model of phonology used along with TCWC is better suited to account for the facts at hand, then we have to admit a reanalysis of the facts leading to a restructuring of the Connected Word Constructions (CWCs). In some cases, it is possible that at least some speakers entertain a double analysis of the facts, an intermediate stage between the phonological status of an alternation and its grammaticalization as a morphological one. The sequence of conditions in (273) summarizes the reasoning followed in order to determine whether an alternation should be treated in TCWC morphology or in the accompanying phonological framework. The latter two, (c) and (d), properly speaking, are tests that allow the linguists to determine what the speakers do, rather than conditions allowing the speaker choose an analysis.

(273) **Deciding on the status of morphonological alternations in TCWC**

a. Are the phones involved part/instantiations of different phonemes?

No: The alternation is phonological.

Yes: Move on to condition (b).

b. Are there words in the lexicon that violate the phonological conditions under which the morphological conditions hold?

No: The alternation is most likely phonological, though a double analysis should not be excluded.

Yes: move on to the conditions in (c).

c. i Would the resulting phonological analysis be more elegant than the morphological one?

ii. Would the resulting phonological analysis simplify the morphological account?

Yes to both: The correct analysis is phonological.

No to both: The correct analysis is morphological.

Yes to one, no to one: The answer will depend on the relative simplicity and elegance of the two accounts. See also the conditions in (d).

d. i) Is the morphological domain under which the alternation holds identifiable by speakers?

No: The alternation is probably morphological.

Yes: The alternation is probably phonological.

ii) Is there external—psycholinguistic or otherwise—evidence that this alternation is treated by speakers on par with other clearly phonological ones?

Yes: The alternation is phonological.

No: The alternation is morphological.

In closing, let's try to apply these conditions to the Armenian THEME VOWEL alternation between [i]/[ɛ] we saw in Chapter 4. There, it was assumed that the THEME VOWEL of I-CLASS verbs, which changes to [ɛ] in the IMPERFECT or AORIST tense, displays an allomorphic relationship by doing so, rather than a neutralizing phonological one. At least in the IMPERFECT, a phonological analysis is tempting, because i alternates with ɛ in front of the IMPERFECT suffix /-i-/

First, both vowels are phonemes of Armenian, so we must move on to condition (b). Condition (b) is not entirely satisfying either: while I know of no Armenian word with two [i] in a row, the [-ɛts] sequences of the AORIST context do not replace an impossible [its] sequence, nor does the 3SING IMPERFECT's ɛr context (e.g. Armenian /bɑnir/ 'cheese'). thus, we move on to condition (c). Here, the alternative phonological analysis would definitely not be very elegant: it would either have to posit an underlying /i/ that gets systematically deleted in the AORIST and IMPERFECT 3SING, or it would have to apply in a disjunction of contexts that don't share much. In addition, very little simplicity would be gained in the morphology: as we saw, we only need to state the alternation once in the IMPERFECT construction. Hence, the trade-off favors a morphological analysis.

Chapter 7

Generalizations

In this chapter, I examine several generalizations that have been made about morphology and discuss the ease with which TCWC can account for them. In some cases, it will turn out that principles independent from the theory need to be adopted. In some cases, the generalizations simply fall out of the architecture of TCWC, including its acquisition procedure and Lexical Insertion Conditions. In other cases, the generalizations make sense in TCWC, if we take into account considerations of economy (e.g. rarer linguistic facts require more mental processing of LexiBlock, or more complex CWCs with more embedded LexiBlocks are rarer in the world's languages).

7.1 On stems

The concept of Stem is widely used in linguistics and in this theory it seems to amount to no more than a LexiBlock that contains those parts of words that are associated with lexical meanings and that is embedded in the Form LexiBlock. The question one might ask then is: is there any generalization that is lost with such a diluted concept of Stem?

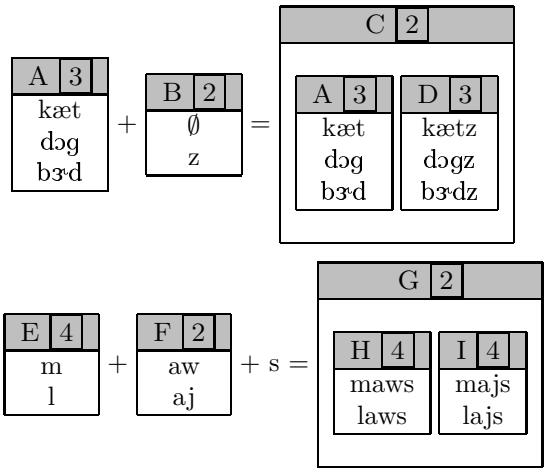
7.1.1 Kiparsky's observation on English compounds

A frequently quoted example is Kiparsky's *mice-infested* vs. **rats-infested*. Kiparsky (1982a) notes that while *mouse-infested* and *mice-infested* are both possible compound words, *rat-infested* is possible, while **rats-infested* is ruled out. The conclusion Kiparsky comes to is that while *mouse*, *mice* and *rat* are all stems, *rats* is a word (and, crucially, not a stem). While there surely are attested

compounds where the first member is a plural noun with /-s/,[1] Kiparsky's generalization, whether it needs to be watered down to a tendency or made subject to subtler rules, is certainly still a valid observation.

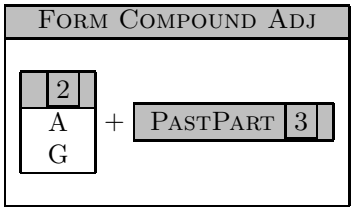
As illustrated in Chapter 2, we have the following equivalences between the phonological forms of nouns (the labels are arbitrarily distributed here for our purposes):

(274)



A CWC on N+PastParticiple compounds then could in principle refer to any of the LexiBlocks on any side of the equations. To capture Kiparsky's observation however, the CWC must be:

(275)



The question one is tempted to ask now is: why doesn't the compound refer to C instead of A? The answer is that CWCs prefer referring to morphologically simpler forms when they can. Referring to C would require some extra processing because it is obtained by decompressing the information stored in the CWC. The Singular form of regular nouns is thus directly available to speakers in the CWC. In the case of the irregular nouns, both the Singular and Plural are equally

[1]For example, the search engine google.com shows approximately 18 times more hits for "weapons hunt" than "weapon hunt". Kiparsky (personal communication) points out that these compounds select a noun phrase, rather than a noun, as illustrated by the possible *dirty bomb hunt* or *nuclear weapons hunt*. By contrast, one cannot say **black rat infested*.

morphologically complex (they require just as much processing).[2] We could propose the following principle to account for these facts:

(276) **LexiBlock reference principle:** CWCs prefer to refer to LexiBlocks that are obtained directly in other CWCs without needing to expand a CWC. This is only a preference that accounts for a tendency, not an absolute principle.

However, the reasoning makes sense irrespective of whether we adopt this principle or not, so assuming a general theory of cognition may suffice.

7.1.2 French V+N compounds

If the previous analysis is right, then one is tempted to ask: could there be cases where only certain classes are selected (e.g. only regular nouns or only certain irregular classes of nouns)? Such a case can be found in French V+N compounds. French V+N compounds are formed with the verb stem followed by a noun. The productivity of this morphological strategy is well-known, as attested by the following two stanzas, where author-singer Boris Vian coins no less than nine such well-formed compounds (in bold).

(277) Ah ! Gudule, excuse-toi ou je reprends tout ça
Mon frigidaire, mon armoire à cuillers,
Mon évier en fer et mon poêle à mazout
Mon **cire-godasses**, mon **repasse-limaces**
Mon tabouret à glace et mon **chasse-filous**
La tourniquette à faire la vinaigrette
Le **ratatine-ordure** et le **coupe-friture**

Et si la belle se montre encore rebelle
On la fiche dehors pour confier son sort
Au frigidaire, à l'**efface-poussières**, à la cuisinière
Au lit qu'est toujours fait, au **chauffe-savates**, au canon à patates
À l'**éventre-tomates**, à l'**écorche-poulets**

Boris Vian, *La complainte du progrès "Les arts ménagers"*, 1956

In this song, Vian couldn't have used the 1PLURAL, 2PLURAL or the INFINITIVE forms of the verbs, which would have produced unacceptable compounds to a native French speaker:

(278) coupe-friture, *coupons-friture, *coupez-friture, *couper-friture

[2]Between *mouse-infested* and *mice-infested*, some speakers I consulted do prefer one of the two forms over the other, suggesting that they connect to either H or I, rather than G.

Furthermore, all the compounds coined by Vian belong to the class of French verbs called the 1stGroup (*premier groupe*) in the literature. (This is the most productive and regular class, forming its Infinitive in *-er* [e]). In fact, and I believe no one has ever noticed this before, V+N compounds cannot be formed with most classes of French verbs.

(279) a. *bâtit-maison 'house builder' *clôt-débat 'debate ender' *boit-alcool 'alcohol drinker'

b. casse-barraque 'shack breaker' ferme-boîte 'box closer' mange-patate 'potato eater'

c. *cassons-barraque *fermer-boîte *mangeait-patate

In (279a), we can see that French verbs forming their infinitives in *-ir*, *-ore*, or *-oire* cannot be used to form new compounds. By contrast, the semantically similar compounds that I coined in (279b) are perfectly acceptable. These are not simply restricted morphological classes. They include the second biggest class of French verbs, the 2ndGroup, with over 300 verbs! Also, again, compounds formed with other forms of the verb are not acceptable (279c). Apart from the 1stGroup, the only other classes of verbs I found with which V+N compounds may be formed are the ones whose infinitive ends in *-ouvrir/-ffrir*[3], *-battre*, *tordre*[4] or *valoir*[5] as the following lexicalized compounds attest:

(280) a. ouvre-bouteille 'bottle opener', ouvre-boîte 'can opener', ouvre-gant 'glove opener', couvre-chef 'head cover'='hat', souffre-douleur 'pain suffer'='someone who's always being picked on'

b. abat-jour 'lamp shade', rabat-joie 'party pooper'

c. tord-boyau 'pipe twister'='bad liquor', tord-nez 'nose squeezer'

d. vaurien 'good-for-nothing'

The 1stGroup verbs and the classes stated above all have in common that they do not use a theme vowel in the Present Indicative Singular (which is identical to the form used in the compound). The class containing *tenir*, *venir*, etc., which also does not use a theme vowel in the

[3]These two subclasses of verbs actually form a single conjugation.

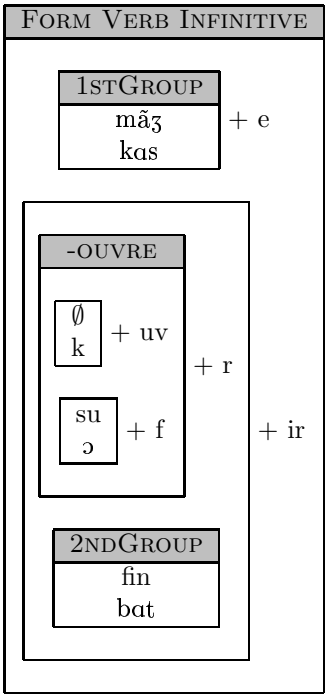
[4]I thank Yves Charles Morin for pointing this class out to me.

[5]It could be argued that *vaurien* (from *vaut-rien*) is no longer analyzed as a compound by at least some French speakers, given the feminine *vaurienne*.

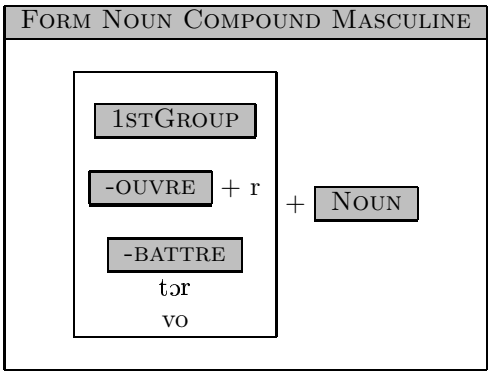
Present Indicative Singular, does not form compounds, but it uses an alternate stem to form this inflection.

We could then consider the French V+N compounds as a case where only certain subclasses are selected in a CWC for use in another CWC. Such a CWC for V+N compounds is given in (282) and a sample of verb classes is given in (281)

(281)



(282)



It is very tempting to try and unify the verb classes that can be compounded with a noun in French. If we consider the Present Indicative Singular of French to be equivalent to the Stem (as we did in the paragraph on paradigm gaps), then the two main "compoundable" classes have in

common that their STEM has no THEME VOWEL and is invariant. For this generalization to hold for *valoir*, we would have to admit that the THEME VOWEL is /-wa-/ and that the [o] of the SING is phonologically derived from /al/ synchronically. As for *tordre*, since the PRESENT INDICATIVE SING is /tɔr/, we would have to admit that the /d/ is not part of the stem in order to maintain our claim that this inflection is equivalent to the stem. Certainly, there are other French verbs ending in *-dre* (*mordre*, *prendre*, *vendre*...), and this /d/ is isolated in the CWC acquisition procedure, so maybe speakers consider it a "theme consonant". However, while *prendre* has an alternating stem, *vendre* and *mordre*[6] do not compound, nor do they have an alternating stem.[7]

Any framework that divides up the lexicon into classes should in principle be able to account for these facts. However, I believe it is not a coincidence that the French compounding facts had not been noticed before. The most striking feature that distinguishes TCWC from other frameworks is that the lexicon is fully integrated within the grammar. Thus it is easier to (literally) see the divisions of the lexicon and their different behaviors.

7.2 The diachronic stability of morphophonology

In this section, I discuss an observation made by Ford & Singh (1983) concerning morphophonological alternations in diachrony. The authors note that they know of no case where a morphophonological alternation spread from one morphological context to another. They cite in support of their thesis, the stability of German umlaut and Spanish diphthongization, always associated with the same morphological contexts through the various stages of these languages:

(283) "It is important to point out that morphophonological alternations are associated with a specific group of morphological operations. For example, in German, umlaut characterizes certain morphological relations such as the diminutive of nouns, the comparative of adjectives and the conjunctive of verbs, but never characterizes the formation of adverbs, the imperfect of verbs, verbalization in -ieren or morphological operations with -bar, -los, -schaft or -ung. In Spanish, diphthongization characterizes the formation of adjectives from nouns, the link between the first and second person plural and the other persons of the verb, but never in adverb formation nor in the diminutive formation in nouns. This same pattern is verified in all other cases mentioned in (1).

[6]Here is one example of a compound with *mordre* found on google.com: *Elle a fini par s'endormir dans son transat après avoir bien machouillé son mord-dent.* Here it seems *mord-dent* means 'a baby's sucker'. Perhaps future investigations will reveal compounds with *vendre* as well.

[7]Nor do verbs in *-oir* such as *voir*, if we do not consider /a/ to be a valid THEME VOWEL of French verbs.

> ... a morphophonological alternation is not dynamic in the sense that it can wander from one morphological context to another. For example, we'd like to claim that the mark of the imperfect or the morpheme -los could not trigger umlaut in German and that diminutive -ito/-ita could not trigger a monophtongization in the Spanish noun. We think, until proof to the contrary is offered, that historical changes of this form are impossible and that a morphophonological process cannot be mobile in this sense."
>
> Ford & Singh (1983), pp. 66-68.

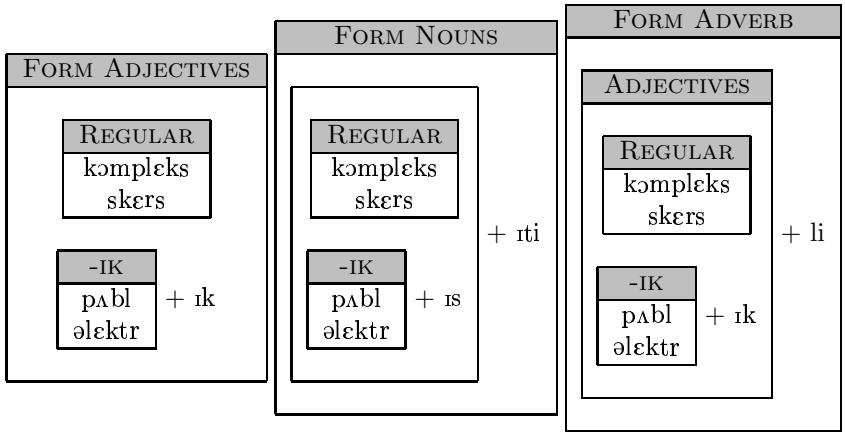
We can thus state their observation as follows:

(284) **Ford & Singh's observation**

A phoneme alternation associated with a set of morphological contexts at a given stage in the history of a language may not be associated with new morphological contexts at a later stage of the language.

This generalization is accounted for in TCWC by the tagging of parts of words together across the lexicon. Take for example velar softening in English, whereby the final /-k/ of certain stems alternates with an /-s/ before certain suffixes: *electric*:*electricity*/*electrician*. If a speaker were to innovate by creating a new adverb *electric-ly*, the stem-final consonant would be a /-k/, not an /-s/, according to Ford & Singh's generalization:

(285)



In (285), because the adverbial CWC refers to adjectives, the stem in the nominal *-ity* construction is not carried over directly. Suppose we have a situation where a speaker has built the CWCs in (285), but has not yet learned the words *electric*, *electricity* or *electric-ly*. Suppose now that this speaker learns the word *electricity* first. In (285), if *electricity* was inserted in the LexiBlock labeled Regular, then its corresponding adverb would be *electri[s]ly*, but its corresponding adjective

would be *electri[s]*, not *electric*. Thus, misinsertion would yield a historical change in the stem, not the generalization of a morphophonological alternation to a new morphological context. But what if *electric* is leaned after a speaker has (wrongly) generated *electri[s]ly* from *electricity*? We would then have the triplet *electric/ electricity/ electri[s]ly*. Wouldn't that be the spreading a morphophonoloical alternation to a new context?[8]

In Chapter 3, we saw precisely such a case. Pubnico Acadian French has generalized an alternation ∅/s from the PRESENT PLURAL to all persons of the FUTURE. This is illustrated in (286). The alternation exists in Standard French for the 2NDGROUP of verbs, between the PRESENT SINGULAR and PLURAL, the /-s-/ being restricted to the PLURAL persons. In Pubnico, the /-s-/ is generalized to the FUTURE forms, a morphological context that did not require this /-s-/ in Standard French.

(286)

Verb	Standard French		
Group	Pres Sing	Pres 1Plur	Future 1Plur
First	mãʒ-∅	mãʒ-∅-õ	mãʒ-r-õ
Second	fini-∅	fini-s-õ	fini-r-õ
Verb	**Pubnico French**		
Group	Pres 1Sing	Pres 1Plur	Future 1Plur
First	mãʒ-∅	mãʒ-∅-õ	mãʒ-∅-r-õ
Second	fini-∅	fini-s-õ	→ fini-s-r-õ ←

Therefore I think we need to weaken Ford & Singh's generalization to a tendency, or limit it by certain conditions. It seems to be rare, but not impossible, for a morphophonological alternation to spread from one morphological context to another. TCWC is in line with this weaker version. The explanation I provided in Chapter 3 was that the change had to have happened in a given sequence: 1) speaker learns 1PLUR /finisõ/ and inserts it with the 1STGROUP, which temporarily generates a SING /finis/ and a FUTURE /finisrõ/; 2) speaker learns SING /fini/, which replaces the form s/he had generated; 3) speaker never learns the "correct" FUTURE form (futures are very rare in the corpus); 4) the dialect has thus spread the ∅/s alternation to the FUTURE tense. Hence, in TCWC, spreading of a morphophonological alternation can be caused by a partial (uncompleted) change in the system due to the relatively low frequencies of just the right forms.

Not any such migration of a morphophonological alternation is possible in TCWC. For example, I cannot find a migratory path for the /s/ of /finisõ/ to reach the stem of *manger*. In Chapter 2, we saw an example of this /s/ analogized in another 1STGROUP verb, *marier*, but it was the fact that this verb's stem ends in /i/ like that of /finir/ that allowed this change to take place. We saw

[8]This is strictly a hypothetical case; whether speakers would rather generate *electrically* or *electricly* is not what is at stake. See the next paragraph for a real case.

further that an /s/ could not migrate to roots ending in /u/ like *jouer*, or to roots ending in /e/ like *créer*, by GENERALIZATION PRESERVATION, though the former may gain a /z/ on the model of *coudre*. It is a great advantage of TCWC to allow these generalizations to simply fall out from the structure of the theory. This is due to the foundations of TCWC in analogical pattern spreading that is not a common feature of current theories.

7.3 The two "Laws" of the Root

Aronoff (1976) claims that there are two "laws" that apply to roots, for which there are no exceptions.[9] As we will see shortly, there actually are some exceptions, and once again we should thus speak of tendencies rather than "laws". Aronoff's observations can be formulated as follows:

(287) **Aronoff's first root observation**

If a word with a given root selects for a given allomorph, another word in the same root may not select for a different allomorph.

Aronoff's second root observation

If a word with a given root has a phonemic alternation between two given morphological contexts, another word in the same root must also show that same alternation in the same contexts.

Let us illustrate the first observation. The English word *de-scrip-tion* is taken to be formed of the prefix *de-*, the bound root *-scrip* and the allomorph *-tion* of the suffix *-(a)tion*. According to Aronoff's first observation then, the form of the noun *pre-scrip-tion* is rightly predicted, as opposed to **pre-scrip-ation*.

The second observation may be illustrated with the same root. The root alternates between a form *-scribe* and a form *-scrip*. Thus we have the verb *de-scribe* (with a diphthong and a /b/), but the noun *de-scrip-tion* (with a short vowel and a /p/). It is then predicted that the noun corresponding to *pre-scribe* is *pre-scrip-tion*, not *pre-scribe-tion*.

Counter-examples to the two observations can be found in French. For the first observation, the verbs /sɛrv-ir/ 'serve' and /de-sɛrv-ir/ 'service' indeed always select for the same allomorphs: 1PLUR /sɛrv-õ/, 2PLUR /sɛrv-e/, 3PLUR /sɛrv-/; 1PLUR /de-sɛrv-õ/, 2PLUR /de-sɛrv-e/, 3PLUR /de-sɛrv-/. However, the historically related /a-sɛrv-ir/ 'enslave' behaves like 2NDGROUP verbs by selecting a suffix /-s-/ before the plural person suffixes: 1PLUR /a-sɛrv-i-s-õ/, 2PLUR /a-sɛrv-i-s-e/,

[9] It is not clear whether Aronoff intends these laws to apply only to English or not.

3PLUR /a-sɛrv-i-s-/. Another example is the verb /mo-dir/ 'damn'.[10] While all the other verbs with the root /dir-/ 'say' (e.g. *prédire* 'predict', *médire* 'say bad things about someone') form their PRESENT PLURALS in the same way (/-di-z-õ/, /-di-z-e/, /-di-z-/), /mo-dir/ has PLURAL forms on the pattern of SECOND GROUP verbs: 1PLUR /mo-di-s-õ/, 2PLUR /mo-di-s-e/, 3PLUR /mo-di-s-/; thus it selects the allomorph /-s-/ instead of /-z-/.

A violation of Aronoff's second observation can be found in the PRESENT SINGULAR forms of /a-sɛrv-ir/: PRESENT SINGULAR /a-sɛrv-i/ does not truncate its root, while /sɛrv-ir/ and /de-sɛrv-ir/ become /sɛr/ and /de-sɛr/.

In all the prefixable verbal roots of French however, these are the only exceptions I have found. Thus Aronoff's observations reflect a true tendency. The two observations are straightforwardly accounted for in TCWC because the acquisition procedure stores all words with a given root together, as seen throughout this dissertation. Therefore, other CWCs can only refer to the whole set at once and cannot split them in an incoherent fashion. One or two words may with time slip into a more productive class of verbs, as /a-sɛrv-ir/ and /mo-dir/ have done, but in general they will stick together.

In closing, it is always possible to have an idiosyncratic suppletive form for only one of the verbs with a given root. For example, the 2PLUR PRESENT of /dir/ is /dit/ in Standard French, while the 2PLUR PRESENT of all the other verbs in /-dir/ have a /-dize/ form. This is of course compatible with TCWC and weakens the interpretation of Aronoff's laws.

7.4 The Adjacency Condition

Siegel (1977) proposes the Adjacency Condition,[11] whereby an affix can be sensitive to an embedded morpheme, but only if that morpheme is one cycle away. Thus the prefix *un-* is sensitive to the presence of the prefix *dis-*, but only if *dis-* was the last affix attached:

(288) un[[distinguish]ed] *un[dis[honest]]
un[[dismay]ed] *un[discrete]

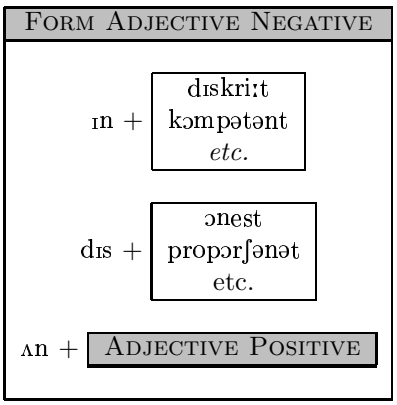
The explanation is that *un-* may not be attached to *discrete* or *dishonest* because the prefix *dis-* is only one cycle away. There is however an entirely different explanation available for these facts.

[10] The allomorph /mo/ is one of the forms of /mal/ 'evil or bad or ache' and is present is such words as /mɔvɛ/ 'bad', /mogree/ 'complain' and /mo/, the PLURAL of the noun /mal/, in the sense of 'ache'. The al/o alternation is found elsewhere in French morphology, as exhibited by the SINGULAR/PLURAL pair for 'horse': *cheval/chevaux*.

[11] Essentially, Williams' (1981) Atom Condition had the same purpose as the Adjacency Condition, with the additional ability to handle the systematic behavior of Latinate bound roots (Aronoff's "laws"), by referring to head features. Since I have already shown how TCWC handles these cases, I will not go into the Atom Condition.

First, the *dis-* in *dishonest* negates an independently existent adjective. It could then be that *un-* only selects for positive adjectives. Second, it may be that **undiscrete* is ungrammatical only by virtue of the existence of *indiscrete* (by blocking). These two explanations come for free with TCWC and there is no need for the Adjacency Condition.

(289)



In (289), the prefixes *in-* and *dis-* are attached to lexically specified sets of forms, while *un-* is attached to any positive (not already negated) adjective, thus excluding *dishonest*. Since *in-* and *dis-* are also stored above the *un-* LexiBlock, this prevents forms such as **uncompetent* and **undiscrete* from being generated.

If the Adjacency Condition were right, however, it should be rare for words such as *discrete* and *distinct*, which normally form their negative with the less productive prefix *in-*, to slip into the more productive *un-* negating pattern. However, a search on google.com on November 28 2005 showed 3800 hits for *undiscrete* and over a thousand for *undistinct*. This does not come as a surprise in TCWC, since the *dis-* in *discrete* and *distinct* is not equated with a negating semantics. We also saw similar changes in Chapter 3 (French *vas* for *vais*, or *haïs* for *hais*). By contrast, the form **unirregular* only has three distinct hits on the same search engine, which can be explained in TCWC by the fact that this innovation requires a much more significant change, the generalization of the domain of application of the prefix *un-* from positive adjectives to any adjective. Likewise, Fromkin et al. (2003:89) claim that **unsad*, **unbrave* and **unobvious* are ungrammatical, yet we find several hundred hits for each of them on google.com (107 000, in the ase of *unobvious*.

7.5 Inflection and derivation

In TCWC, there is no theoretical distinction made between inflection and derivation. Linguists (myself included) do however find this distinction useful, because it correlates with a set of characteristics, some of which are listed below:[12]

(290)

Inflection tends...	Derivation tends...
to be productive	to be category-changing
to be semantically transparent	to be ordered before inflection
to apply to most of the lexicon	to have more morphophonological alternations

Experience shows that there are exceptions to all of these statements: paradigm gaps are counter-examples to the productivity of inflection; English comparatives, if we consider them a case of derivation, form a counter-example to the category-changing nature of derivation; the plural of mass nouns is a counter-example to the semantic transparency of inflection; Breton shows derivative affixes outside inflectional ones; English feminine gender applies to only a few inanimate objects (*ship*, *car*, etc.).

In spite of these counter-examples, we can consider inflection and derivation as poles around which the above properties cluster. Changing the category of a word, sometimes leads to more semantic opacity in itself. For example, when going from a verb to a noun, speakers must figure out whether the new noun refers to the action, the person performing the action or the instrument usually used to perform the action. Semantically and phonologically transparent relations should naturally be more productive: in a semantically opaque relation, more choices need to be made, thus relying on a greater degree of social conventions, on a shared context between the speaker and the hearer or on common pragmatic assumptions; phonemically complex relations require more form manipulation. In turn, productive relations, with time, end up accounting for the biggest subset of the lexicon.

7.5.1 Affix ordering

One often-cited tendency in affix ordering is that derivational affixes tend to be closer to the root than inflectional ones. Joseph Greenberg stated the following universal:

(291) Universal 28. If both derivation and inflection follow the root, or they both precede the root, the derivation is always between the root and the inflection. (Greenberg 1966: 93).

[12] See Rice (1999:295-296), Stump (1990:98) or Bybee (1985) for similar lists.

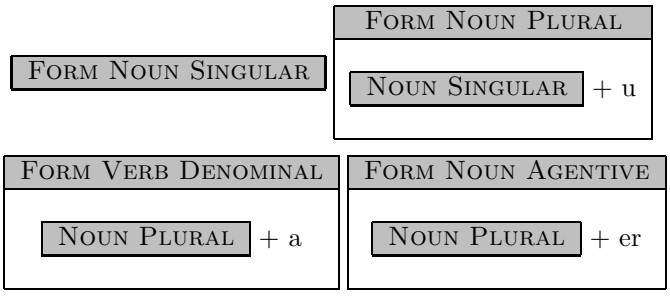
Stump (1990) presents a some counter-examples from the Celtic language Breton. For example, denominal verbs, agentive nouns, and denominal adjectives are formed on the suffixed plural forms:

(292)

korn	kernioù	kerniaouek	
'horn'	'horns'	'having many horns'	
Noun Singular	Noun Plural	Adjective	
aval	avaloù	avalaoùa	avaloùer
'apple'	'apples'	'to look for apples'	'one who looks for apples'
Noun Singular	Noun Plural	Noun Agentive	Verb

These examples however are not really any different from the case on English compound seen at the beginning of this chapter, where it is possible, though rarer, to have compounds formed with plural nouns, rather than singular ones: *weapons hunt*. It involves less processing for CWCs to refer to fully inflected words, than to stems, but it's not impossible. Not wanting to be distracted by the morphophonological issues, we can sketch the Breton cases as follows:

(293)



Similarly, Fulmer (1991) shows that in Afar, the agentive derivative affix is realized outside the inflectional gender affix, though she notes that only the deverbal nouns get inflected for gender in Afar:[13]

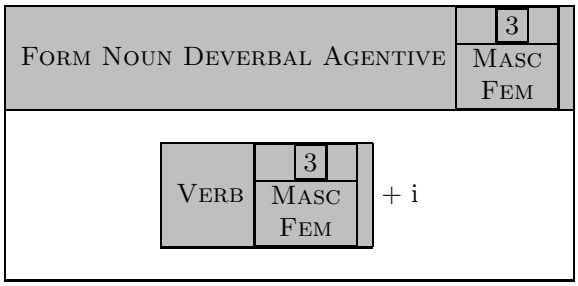
(294)

y-árd-i	t-árd-i
Masc-'run'-Agentive	Fem-'run'-Agentive
y-aaxíg-i	t-aaxíg-i
Masc-'kill'-Agentive	Fem-'kill'-Agentive

Again abstracting away from morphophonological facts for simplicity, this is not a very hard situation to account for in TCWC:

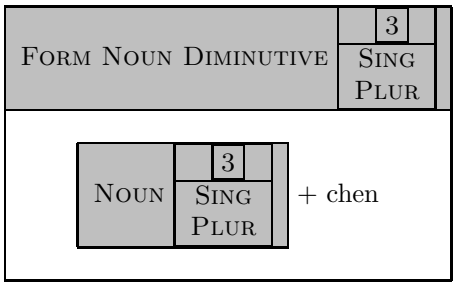
[13]Technically speaking, Fulmer's examples do not violate Greenberg's Universal 28, which states that an inflectional affix may not occur between the root and a derivational one. Fulmer however does present arguments for consider the Agentive affix as added "after" the gender ones.

(295)



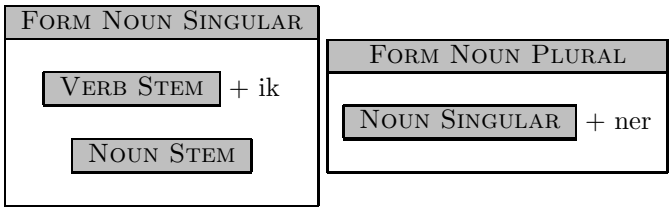
Stump's and Fulmer's cases are not really what people are looking for, when they talk about inflection never or rarely preceding derivation. In Stump's case, the semantics of the derived words refer to the Plural of the Noun, but does not represent the Plural form of the Verb. In Fulmer's examples, though it is the case that the Masculine and Feminine categories properly inflect both nouns and verbs for gender, we are told that other nouns don't inflect for gender... The standard example of the German diminutive suffixes lying outside number suffixes is also not problematic in TCWC, because diminutive formation is not category-changing, and can then be simply added to a previously defined Noun construction:

(296)



To understand what would be a real counter-example to Greenberg's Universal 28, take an imaginary language where a suffix *-ner* marks plural nouns, and a suffix *-ik* is a nominalizing suffix applied to verbs. If *-ner* is ordered after *-ik*, we get the following system:

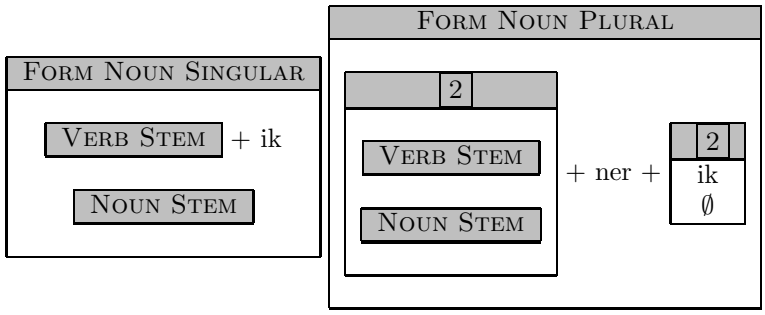
(297)



Though I haven't defined rigorously what complexity is in TCWC, even by a loose metric based on the number of embedded LexiBlocks, it is clear that the structure below, where the Plural suffix

is ordered before the nominalizing one would be more complex, as it has four LexiBlocks instead of one:

(298)



This imaginary example would be a true counter-example to Greenberg's Universal 28, because the Plural suffix would really be inflecting the nouns (unlike in Stump's examples) and it would be a suffix applicable to all nouns of the language (unlike Fulmer's example). Whether this type of system is impossible or just very rare, the explanation in TCWC for this observation is one of economy. LexiBlocks, as I have mentioned before, are a powerful tool and they do allow us to represent many morphological structures, but the representations come at different levels of complexity and it is expected that more complex structures will be rarer in the world's languages. In the case at hand, a CWC that would allow for a non category-changing affix to be truly ordered closer to the root than a category-changing one, would yield a more complex structure, because what we would be doing in fact would be to specify the morphemes between which the derivational affix is ordered, while not having morphemes as primitives in TCWC.

Furthermore, given what we know of grammaticalization facts (that affixes originate as independent words), such a scenario would be extremely difficult to construct. In the imaginary case stated above, a Plural suffix /-ner/ would have to first be shared by both nouns and verbs. While this does happen,[14] we should be quick to point out that plurality of the verb is semantically very different from the plurality of the noun. In English for example, plurality of the verb means that the external argument is plural, not that the action happens many times. Therefore, it is not so intuitive to formulate a single CWC that would pluralize both nouns and verbs in form and meaning. Secondly, as we saw in the case of compounding, derivation typically builds on stems, because stems are typically accessible with less processing. Therefore, even if nouns and verbs can on occasion share a suffix, the derivation from noun to verb is more likely to build on the (simpler) suffixless form. Hence, there are a number of internal and external reasons in TCWC why inflectional affixes

[14]For example, in Classical Armenian, the suffix /-k^{h}/ marked both plural persons of verbs and plural nouns.

would rarely—if ever—appear before category-changing affixes.

7.6 The Peripherality Constraint

Carstairs-McCarthy (1987) formulates several generalizations on morphological systems, of which the Peripherality Constraint is one of the most interesting. I summarize it as follows:

(299) **Carstairs-McCarthy's Peripherality Constraint**

An affix may not have a special allomorph that depends on an outer affix.
Caveats: The allomorph however may depend on an outer category.
An allomorph may depend on an outer affix if there is systematic homonymy between the forms that violate the Peripherality Constraint and other forms in the paradigm of this verb.

For example, in Armenian, we have seen that the allomorphs of the outer SINGULAR person suffixes depend on the inner tense suffixes of the verb. In (300), for example, the 1SING, 2SING and 3SING person sets {/-m/, /-s/, -∅} and {-∅, /-r/, /-r/} depend on whether the verb is PRESENT or IMPERFECT:

(300)

	Present	Imperfect
1Sing	uz-e-m	uz-e-i-
2Sing	uz-e-s	uz-e-i-r
3Sing	uz-e-	uz-e- -r

In Armenian, the realization of the IMPERFECT suffix, which is either ∅ or /-i-/, depends on the particular person-number inflection (it is null only in the 3SING). This 3SING IMPERFECT is however homonymous with the NEGATINGPARTICIPLE, which allows Armenian not to violate the Peripherality Constraint by the second caveat.

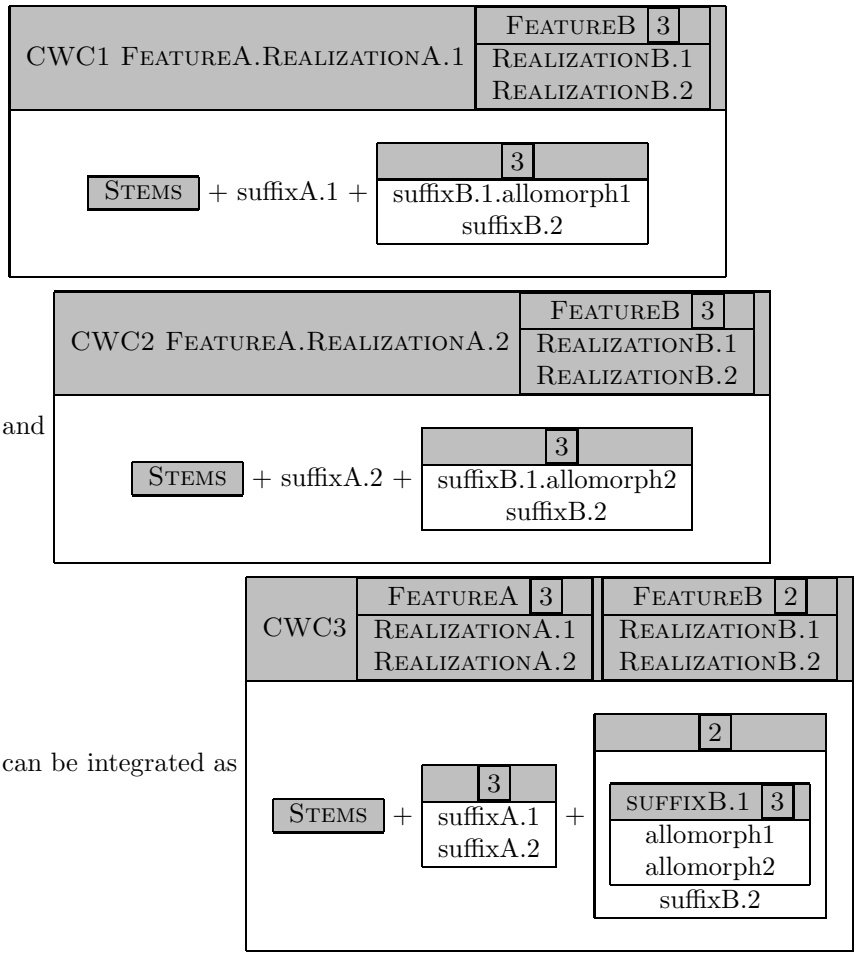
We have seen another such case in French: the IMPERFECT suffix, which is usually /-ɛ/, has an allomorph /j/ before the 1PLUR and 2PLUR suffixes.[15] Here again, these forms are syncretic with others in the paradigms: the French IMPERFECT 1PLUR and 2PLUR are homonymous with their SUBJUNCTIVE PRESENT correspondents. At least for French and Armenian though, it is indeed more common not to have an allomorph dependent on an outer affix.

In this case, TCWC unfortunately does not help us gain much insight on why this state of affairs should be. Take the case where an outer allomorph depends on an inner allomorph. Formulation of such a case, not said to be problematic by the Peripherality Constraint, is indeed unproblematic in

[15] I thank Yves Charles Morin for pointing out the relevance of this case to me.

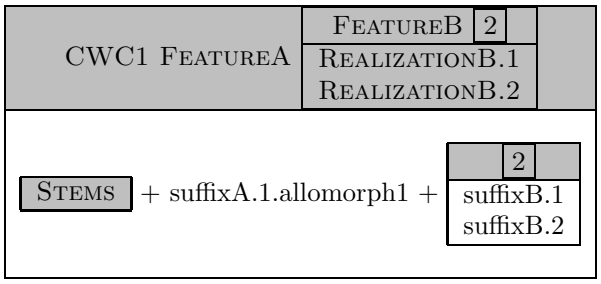
TCWC. As illustrated below, two different CWCs, with suffix1 and suffix2, are simply followed by two different sets of allomorphs, and can be integrated into a single CWC.

(301)

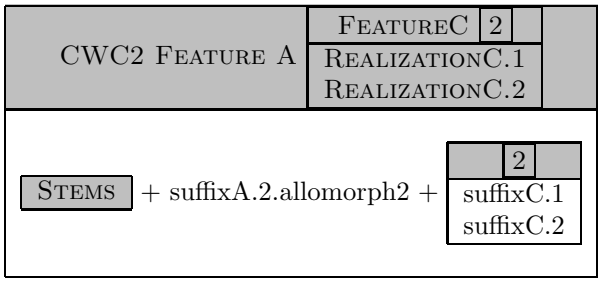


An allomorph of an inner suffix depending on the realization of the feature associated with an outer suffix does not yield more complexity either (and is not problematic for the Peripherality Constraint by the second caveat), the allomorphy is simply displaced:

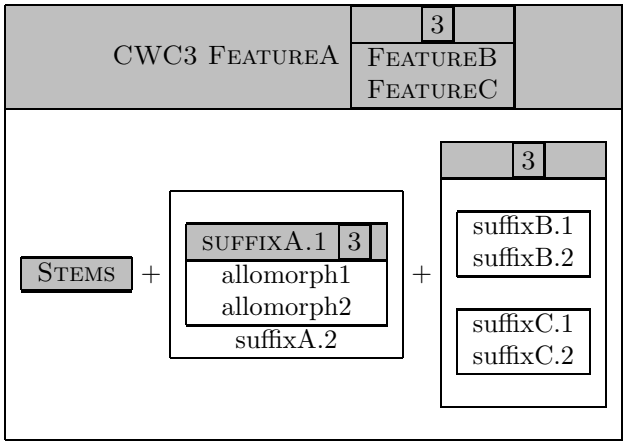
(302)



and

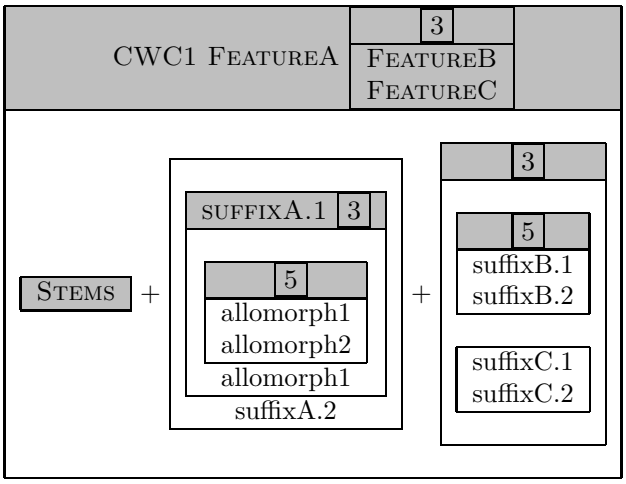


can be integrated as



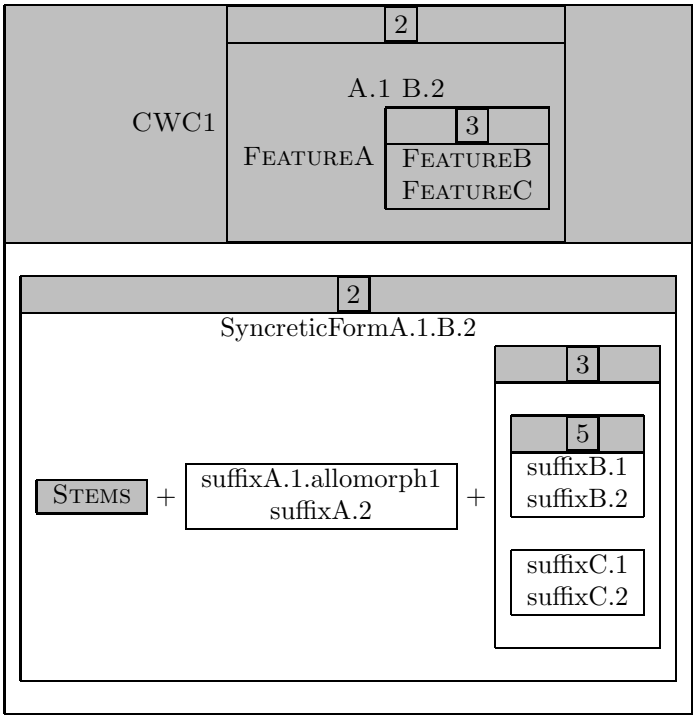
If, however, the allomorph of an inner suffix depends on the particular realization of a certain feature, this still does not yield more complexity, though these cases are ruled out by the Peripherality Constraint:

(303)

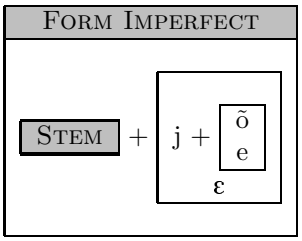


However, in the above situation, if the special allomorph is syncretic with another inflection of the language, this does help reduce the complexity by the Elsewhere Step:

(304)

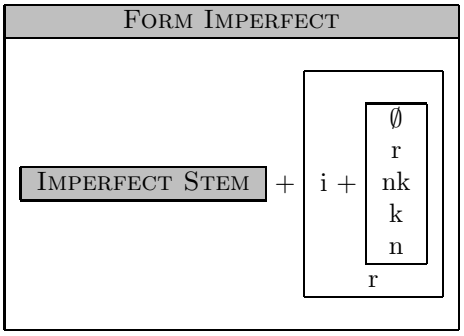


Turning back to our French and Armenian Imperfects:



French
1Plur
2Plur
3Plur, Sing

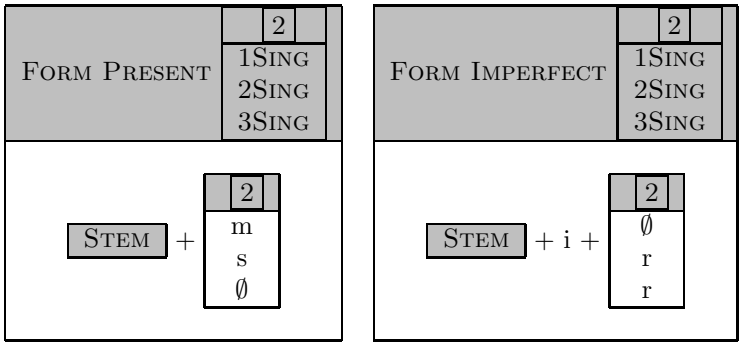
(305)



Armenian
1Sing
2Sing
1Plur
2Plur
3Plur
3Sing

In the two cases, there is a stranded person that is not integrated in the same LexiBlock as the rest of the paradigm. Looking back on the facts in (300), if the allomorphy were strictly sensitive to an inner category, the LexiBlocks would be organized in a more orderly fashion and would thus be easier to parse:

(306) **Pseudo-Armenian CWCs**



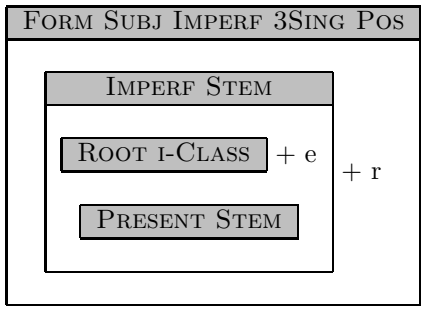
However, this only accounts for the fact that it is simpler to have allomorphy "looking" one way (inwards or outwards), rather than looking both ways. Carstairs-McCarthy's generalization has proven to be very robust over the years. Apparent counter-examples have been demonstrated to involve morphemes that are part of the stem, such as theme vowels or noun classifiers.[16] Unlike tense

[16]See for example Adger et al. (2003), and Carstairs-McCarthy's convincing reply in the same volume. See also Kiparsky (1996:22).

morphemes that are concatenated between the STEM LexiBlock and the person-number suffixes, theme vowels are part of the stem. Theme vowels in TCWC are isolated in a different way than affixes: the acquisition procedure, for independent reasons, groups the various parts of the stems inside a LexiBlock, so that these cases don't yield the kind of complexity that we saw in (305).

For example, recall the theme vowel of the i-Stem verbs, which alternates between an /-i-/ in the Present and an /-e-/ in the Imperfect. Although the theme vowel's shape (/i/ or /e/) depends on a specific category that is realized on an outer layer (the imperfect tense), because the theme vowel is part of the stem, we can represent the phenomenon with the well-known Pāṇinian mechanism:

(307)



Therefore Carstairs-McCarthy's impressive Peripherality Constraint resonates well with TCWC, but explanations in terms of complexity of embedding and representation are not sufficient to account for it here. Hence, we must, at least for the time being, import the constraint directly into the framework.

7.7 Conclusion

There are two strategies that linguists use to account for cross-linguistic tendencies or generalizations. One way is to build a framework that is naturally suited in its formalism and its principles to account for the empirical facts. Such was the strategy behind feature-geometry representations in phonology, and such is the strategy behind the formulation of universal constraints in Optimality Theory. In this chapter, I have shown how TCWC is naturally suited to account for several morphological generalizations or tendencies reported by other linguists. Some generalizations fall out from the architecture of the theory. Examples violating Ford & Singh's observations concerning the diachronic stability of morphophonology require a set of very special circumstances, while the acquisition procedure from Chapter 3 provide the necessary CWCs for Aronoff's two laws of the root to hold. Siegel's Atom Condition is also explained by the suppletion mechanism available in TCWC

and a reanalysis of some of the facts. Other generalizations come from economy considerations. Kiparsky's observation about English compounds can be explained by ease of processing: referring to a stem, rather than a word, does not require to expand the compressed lexicon. Default ordering of category-changing derivational affixes before inflectional ones correlates with less complex representations. If we take TCWC seriously as a theory of mental representation of the lexicon and generating the possible words of a language, then it makes sense that more complex representations, associated with more complex mental parsing, correlate with a rarer or impossible situation.

On the other hand, no theory derives all of the observed tendencies and generalizations from the top down. Most generalizations are first made by observation of empirical facts, rather than by logical derivation from the theory and later empirical confirmation. Sometimes the true explanation for a generalization lies outside the domain studied. For example, morphophonological alternations often occur in a position adjacent to the affixes that trigger them. Since we know that the historical source of morphophonological alternations is from purely phonological ones, and that the principles of phonetics and phonology already favor locality, it does not seem necessary for a theory of morphology to also encode these principles. Sometimes however, the motivation may be obscure in the current state of research and a theory must adopt an external principle to limit its expressive power. For example, Carstairs-McCarthy (1987) formulates the Paradigm Economy Principle, which tries to limit the number of possible *macroparadigms* for a given part of speech. Matching up Carstairs-McCarthy's concept of macroparadigm with something identifiable in TCWC is an extremely difficult task, not to mention looking for potential counter-examples in the world's languages, and at this point in the development of TCWC, I have no idea how the theory could account for it on independent grounds. Another interesting claim is the Systematic Homonymy Claim, which tries to circumscribe the ways in which language may display syncretism (the identity of two forms) in a paradigm. To the extent that these generalizations hold true, TCWC needs to adopt them as external principles, much like any theory of morphology I am aware of.

There is of course nothing wrong with having both internal and external explanations for a generalization. Affix ordering is a case in point. Semantic and grammaticalization principles by which independent words become affixes make it such that it is harder to construct a scenario where category-changing affixes end up outside non category-changing ones. Even when this happens, TCWC does not favor such CWCs, because they require extra complexity. Therefore, I hope to have convinced the reader with this chapter that several generalizations about morphology naturally fall out from the way TCWC is constructed, and that TCWC is not incompatible with those that don't.

Chapter 8

Conclusion

In this dissertation, I have proposed the Theory of Connected Word Constructions (TCWC), a theory of morphology consisting of a set of FORM constructions and a set of MEANING constructions. The Connected Word Constructions (CWCs) form a compressed lexicon. The two sets of constructions are expanded by an algorithm, producing two lists, the elements of which are then unified into fully inflected words. However, I have limited the object of study to the FORM constructions, because this dissertation is concerned with the formal side of morphology and its relationship with phonology. We can capture the uniqueness of TCWC by saying that *the lexicon is done in the morphology.* In Distributed Morphology (Halle & Marantz 1993), the lexicon is a separate list, and morphology is done with syntax, whereas in Paradigm Function Morphology (Stump 2001), the morphology is done over the lexicon (separately from syntax, as in TCWC).

After formalizing the theory in Chapter 2, I proposed an acquisition procedure that correlates with types of historical morphological changes. Chapter 4 dealt with the verbal morphology of Armenian, while Chapter 5 showed that TCWC provides an analysis of the neglected phenomenon of paradigm gaps in English, French, Spanish and Russian. Chapter 6 established criteria for distinguishing between morphophonological alternations that should be treated in phonology and those alternations that should be treated in morphology. Finally, in Chapter 7, we saw how TCWC accounts for several morphological generalizations by providing more complex representations for rare or impossible situations.

8.1 Advantages of the theory

The main advantage of TCWC is the amount of facts and generalizations it can explain with a very small set of strong falsifiable assumptions. The independently motivated acquisition procedure runs through the grammar explaining several facts and generalizations: new word formation, analogical change, paradigm gaps, Aronoff's laws of the root and the morphological structure of the languages examined in this dissertation.

The separation of the CWCs into FORM and MEANING constructions allows TCWC to account for mismatches between the FORM and MEANING of words, while organizing each side economically. When the organization of submorphemic phonemes match up with semantic organization, we get phonesthemes. It is also possible for a FORM construction to generate a form that does not have a MEANING correspondent to unify with. For example, the pluralia tantum such as English *scissors* comes from the unification of a PLURAL FORM with an idiosyncratic PLURAL MEANING;[1] a SINGULAR FORM *scissor* is generated, but it has no MEANING with which to unify. The FORM *scissor* though is available for compounding or other word formation strategies, as in *scissor hands* or the verb *to scissor*.

The set theory formalization of the main tool used by TCWC, the LexiBlock, grounds TCWC in familiar mathematical notions with well-known properties. On the other hand, the feature-structure implementation allows one to easily compare and test the compatibility of TCWC with other formally-oriented theories.

The five step acquisition procedure proposed for the formal CWCs allows speakers to build them in an economical fashion, and the Lexical Insertion Conditions that go along with them prevent the CWCs from overgenerating. Further, I demonstrated how errors on each of the five steps correlate with five types of morphological change: category merger, folk etymology, contamination, loss of suppletion and some types of leveling. The Lexical Insertion Conditions on the other hand allow TCWC to make much more accurate predictions than traditional four-part analogy.

TCWC also provides a satisfying analysis for the complex morphological system of Western Armenian verbal morphology, which includes both suffixation and prefixation, vowel alternations, suppletion, etc. We saw that the acquisition procedure also held for that language and that TCWC provided an accurate analysis of a very special case of double morphology: 1) The suppletive roots

[1] More precisely, the meaning is semantically singular, but is associated with the category PLURAL, so that it can unify with the PLURAL FORM.

select for a special aorist suffix that is also selected by vowel alternating verbs; 2) The vowel alternating verbs however also select the regular aorist suffix, thus doubly marking this tense; 3) This is a special case of double morphology because each suffix is used by independently by a class of verbs; 4) This yields extra complexity, because the vowel alternating verbs are then at the intersection of two classes, which in TCWC translates into having to repeat information in separate LexiBlocks; 5) Thus, the rarity of such cases of double morphology translates into extra complexity in TCWC.

As mentioned above, one of the biggest accomplishments of TCWC is that the very same Lexical Insertion Conditions that help limit historical four-part analogy and prevent over-generation also serve to explain paradigm gaps in English, French, Spanish and Russian. The TCWC account is superior to those of Morin (1987, 1995) and Albright (2003) in different ways. Morin's analysis of French paradigm gaps generates the right facts, but is admittedly not designed to predict the inflection of new words. Albright's account's main problem is that it equates the combination of unfamiliarity and uncertainty about a word's inflection with gaps. This is far from being an obvious equation, since, for example, French *frire* 'fry' is not at all an unfamiliar verb, but it shows gaps, and there are lots of unfamiliar verbs that don't show gaps. Speakers may also be uncertain about what the correct or standard inflection of a certain form may be, but that is very different from having a gap. A gap is the impossibility to generate a form, while uncertainty and unfamiliarity yield variation in TCWC, corresponding by to the hesitation between two choices provided by the grammar.

Several generalizations made by other linguists also fall out from the principles of TCWC, or from principles of economy within TCWC. For example, the fact that compounds prefer to use bare stems rather than fully inflected forms, as observed by Kiparsky (1982a), comes from the fact that it requires less parsing for a speaker to tag the stem, usually fully described in a CWC, while inflected word forms are more often obtained by the expansion of the compressed lexicon. As this pick-and-choose account of compounding predicts, it is possible to have a compound construction referring only to some classes of verbs: in French, the Verb+Noun compounds can only be formed from 1stGroup verbs or from a few other small classes. Most verb classes, including the large 2ndGroup, cannot compound. The diachronic stability of morphophonological alternations stated by Ford & Singh (1983) is also predicted by TCWC. In this case, it is possible for a morphophonological alternation holding between contexts A and B to become associated with a new morphological context C over time, but only under a very special set of circumstances, whereby a form in context A is wrongly generated from context B, spreading the alternation to context C. Then, the correct

form for context A is learned later, but not the correct form for context C. Aronoff's (1976) two Laws of the Root, whereby all words sharing a root select the same allomorphs and undergo the same morphophonological alternations, are shown to have a few exceptions, but the general tendency is nevertheless recognized and explained in TCWC by the storing of all the common roots of prefixed words together, something done simply by following once again the acquisition steps from Chapter 3. Siegel's (1977) Adjacency Condition, which states that affixes may only be sensitive to material added immediately before them, is handled by a combination of blocking, available in TCWC by the ELSEWHERE STEP and the ordered lists generated by the expansion algorithm, and a reanalysis of the facts concerning the prefix *un-*, which, I claim, may only be prefixed to unnegated adjectives. As in the Armenian double morphology case, TCWC accounts for the more common ordering of derivational affixes before inflectional ones by independently assigning more complexity to cases violating this principle. As for Carstairs-McCarthy's (1987) challenging generalizations, TCWC offers no insight, and must resort to adopting his principles straight into its premises, but crucially, TCWC is not incompatible with those principles.

8.2 Falsifiablility

TCWC defends the strongest most falsifiable positions possible until proven wrong. In the introduction chapter, we started out by assuming no distinction between types of phenomena, and that TCWC should be responsible for as many morphological facts as possible. We also proposed a very simple and unified tool for morphological analysis accompanied by a five-step acquisition procedure and three Lexical Insertion Conditions. We adopted the word as our basic unit, because it is the only one admitted by all theories, as opposed to morphemes, stems or lexemes.

In the end, our assumptions held relatively well. First, very little distinctions between types of phenomena have been made in the dissertation. True, we have often referred to inflection and derivation, but it was in a traditional sense, and it was often simply to delimit the object of study. In fact, TCWC was even able to account for a generalization on the ordering of derivation with respect to inflection, without even having those categories. Second, as the preceding section enumerates, TCWC has been able to explain quite a large and diverse set of morphological facts. Third, the LexiBlock tool and the word unit we assumed from the start have not been seriously challenged by the facts examined. We had however to admit the generalizations of Carstairs-McCarthy (1987) as independent principles, as well as tentatively sketch lexical insertion preferences, to account for productivity and class attraction.

One way to falsify TCWC would be to provide enough empirical data showing that each one of the four theoretical choices is completely ill-founded. Another way would be to justify a complication of one of these dimensions that would lead to a contradiction elsewhere. For example, nothing in the formal tool allows it to "count" morphemes, and in fact, this inability of the theory was what allowed us to account for Greenberg's Universal 28, on the order of derivation and inflection. If it could be proven that in some language, an affix is systematically inserted three morphemes from the right of the word, and that the morphemes counted do not always represent the same features, then this would cause serious problems for TCWC. Since morphemes are derived from words in TCWC, there is no way to represent them formally as a countable entity. Another example would be to show that the acquisition steps I propose in Chapter 3 do not correspond at all to the real-time acquisition of language, and that there logically could not exist an acquisition algorithm leading to the CWCs I propose that would correspond to those real-time acquisition facts.

8.3 Remaining issues and future research

As with any ambitious linguistic theory, there is always a set of outstanding problems. In the case of TCWC, first, the acquisition procedure is not yet formalized in the most rigorous manner and it has not yet been proven that it corresponds to the actual stages of the acquisition of morphology observed by linguists working in language acquisition. The first step was to show that there exists a coherent logic that allowed me to posit the CWCs I did; that they were not just conveniently set up to account for everything that they have. Also, the entire relationship between morphology and syntax/semantics has been ignored in this dissertation.

There are also some types of paradigm gaps that TCWC has not succeeded in explaining, namely gaps involving a suppletive form, or gaps involving an underlying form with unattested phonotactics in the language.

In the upcoming research, it would be important to implement the theory computationally, and not let it remain one manipulated by linguists only. It would be quite an achievement to see the CWCs and LexiBlocks encode and generate a lexicon on a computer. Once implemented, it should be also easier to formalize a true acquisition algorithm. In a related vein, formalizing the lexical preference statements from Chapter 3 (preference to insert words the similar-sounding classes, or in a related semantic class; insert words in larger LexiBlocs, or in LexiBlocs used more recently) would also be useful.

On a strictly linguistic side though, I feel the neglected phenomena of phonesthemes and paradigm

gaps are the ones most exciting to pursue. These are two original sets of problems, and TCWC is probably the only theory of morphology that is so centrally concerned with them.

Because, TCWC was built from a clean slate of assumptions, I had to concentrate on defining it carefully and providing explanations for phenomena central to morphology, so I also plan to better delimit the border between morphology and phonology, as well as account for more morphological generalization in the future.

The main strength of TCWC is its simplicity. After three to four decades of generative morphology, with LexiBlocks as a new tool, TCWC offers a fresh look on old problems, and takes on the challenging task of treating "morphology by itself", as Aronoff best expressed it, but with minimal assumptions. This simplicity, combined with a novel formalism, give it a definite originality among the current frameworks. However, I hope the reader won't get the wrong impression and equate its radically different look with a theory that would be dismissive of previous and contemporary frameworks. I cannot stress enough how most of the elements used in TCWC have been inspired by work in other frameworks. The integration of the lexicon within the morphology, as opposed to the morphology being done within the lexicon, the very principle at the foundation of TCWC, is directly inspired by Kiparsky's (1982b) Lexical Morphology, along with the ambitious aim to account for morphology hand in hand with morphophonemics. The recognition of the word as the basic unit of morphology, as well as the reductionist nature of TCWC, deriving from a feeling that several frameworks unnecessarily import complications that were justified in other theories stem from Ford et al.'s (1997) and Singh & Starosta's (2003) Seamless Morphology, and the Neogrammarian school, as represented by Paul (1888). Like Halle & Marantz' (1993) Distributed Morphology, Mel'cuk's (1993-2001) Meaning-Text Theory, or Baker's (1988) Incorporation Theory, TCWC aims to account for as many intricate patterns as possible. It is possible that TCWC will need more inspiration from these theories when it tries to tackle morphosyntactic and morphosemantic facts. With realizational frameworks such as Zwicky (1992), Matthews (1991), Anderson (1992), or Stump (2001), TCWC also shares a form of lexicalism and a moderate approach to morphonology, where ambiguous cases may exist that represent a transitional state of grammaticalization from phonology to morphology. In fact, the title of this dissertation, attributable to David Stampe, can be interpreted in two ways, inclusive or not of (part of) phonology, and is a reference to this unclear boundary between phonology and morphology that has haunted these two disciplines.

References

[Adger et al., 2003] Adger, D., Bejar, S., and Harbour, D. (2003). Directionality of allomorphy: A reply to Carstairs-McCarthy. *Transactions of the Philological Society*, 101(2).

[Albright, 2003] Albright, A. (2003). A quantitative study of Spanish paradigm gaps. In Garding, G. and Tsujimura, M., editors, *WCCFL 22 Proceedings*, pages 1–14, Somerville MA. Cascadilla Press.

[Albright and Hayes, 2002] Albright, A. and Hayes, B. (2002). Modeling English past tense intuitions with minimal generalization. In *Proceedings of the Sixth Meeting of the ACL Special Interest Group in Computational Phonology*.

[Anderson, 1971] Anderson, S. R. (1971). *West Scandinavian vowel systems and the ordering of phomological rules*. Ph.D. dissertation, Massachusetts Institute of Technology.

[Anderson, 1975] Anderson, S. R. (1975). On the interaction of phonological rules of various types. *Journal of Linguistics*, 11:39–62.

[Anderson, 1992] Anderson, S. R. (1992). *A-Morphous Morphology*. Cambridge University Press, Cambridge England.

[Anderson and Kiparsky, 1973] Anderson, S. R. and Kiparsky, P. V., editors (1973). *A Festschrift for Morris Halle*. Holt, Rinehart & Winston, New York.

[Archangeli, 1997] Archangeli, D. (1997). *Optimality Theory: An Introduction to Linguistics in the 1990s*, pages 1–32. In [Archangeli and Langendoen, 1997].

[Archangeli and Langendoen, 1997] Archangeli, D. and Langendoen, D. T., editors (1997). *Optimality Theory: An Overview*. Explaining Linguistics. Blackwell Publishers, Oxford.

[Aronoff, 1976] Aronoff, M. (1976). *Word formation in generative grammar*. Linguistic Inquiry Monograph 1. MIT Press, Cambridge MA.

[Aronoff, 1994] Aronoff, M. (1994). *Morphology by Itself*. MIT Press, Cambridge MA.

[Backofen et al., 1990] Backofen, R., Euler, L., and Görz, G. (1990). Towards the integration of functions, relations and types in an AI programming language. In *Proceedings of GWAI-90*, Berlin. Springer.

[Backofen et al., 1991] Backofen, R., Euler, L., and Görz, G. (1991). Distributed disjunctions for LIFE. In Boley, H. and Richter, M. M., editors, *Proceedings of the International Workshop on Processing Declarative Knowledge 91, Kaiserslautern Germany*, number 567 in Lecture Notes in Artificial Intelligence, pages 161–170, Berlin. Springer-Verlag. Subseries of Lecture Notes in Computer Science.

[Baker, 1988] Baker, M. (1988). *Incorporation: a theory of grammatical function changing*. University of Chicago Press, Chicago.

[Bakovic, 2000] Bakovic, E. (2000). *Harmony, Dominance and Control*. Ph.D. dissertation, Rutgers University.

[Bardakjian and Thomson, 1977] Bardakjian, K. B. and Thomson, R. W. (1977). *A Textbook of Modern Western Armenian*. Delmar, New York.

[Baronian, 2002] Baronian, L. V. (2002). No morphemes in my pockets, lexemes up my sleeves or stems under my hat: Western armenian verbal morphology. In *Papers from the 37th Meeting of the Chicago Linguistic Society*, pages 53–66, Chicago. Chicago Linguistic Society.

[Baronian, 2005] Baronian, L. V. (2004/2005). Armenian negation with word constructions. *Annual of Armenian Linguistics*, 24-25:1–11.

[Benua, 1997] Benua, L. (1997). *Transderivational identity: phonological relations between words*. Ph.D. dissertation, University of Massachusetts, Amherst.

[Bergen, 2004] Bergen, B. K. (2004). The psychological reality of phonaesthemes. *Language*, 80(2):290–311.

[Bescherelle, 1992] Bescherelle (1992). *L'art de conjuguer; Dictionnaire de 12 000 verbes*. Éditions Hurtubise HMH Ltée, La Salle QC.

[Bird, 1992] Bird, S. (1992). Finite-state phonology in HPSG. In *Proceedings of the Fifteenth International Conference on Computational Linguistics (COLING-92)*, pages 74–80.

[Blevins, 2003] Blevins, J. P. (2003). Stems and paradigms. *Language*, 79(2):737–767.

[Brown, 2003] Brown, B. (2003). Code-convergent borrowing in louisiana french. *Journal of Sociolinguistics*, 7(1).

[Bybee, 1985] Bybee, J. (1985). *Morphology: A study of the relation between form and meaning*. John Benjamins, Amsterdam.

[Bybee, 2001] Bybee, J. (2001). *Phonology and Language Use*. Cambridge University Press, Cambridge England.

[Carrière, 1937] Carrière, J. M. (1937). *Tales from the French Folk-lore of Missouri*. Northwestern University, Evanston.

[Carrière, 1939] Carrière, J. M. (1939). The creole dialect of missouri. *American Speech*, 14:109–119.

[Carrière, 1941] Carrière, J. M. (1941). The phonology of missouri french: A historical study. *French Review*, 16:410–415, 510–515.

[Carstairs-McCarthy, 1987] Carstairs-McCarthy, A. D. (1987). *Allomorphy in inflexion*. Croom Helm, London & Wolfeboro NH.

[Chafe, 1997] Chafe, W. (1997). How a historical linguist and a native speaker understand a complex morphology. In Schmid, M. S., Austin, J. R., and Stein, D., editors, *Historical Linguistics 1997: Selected Papers from the 13th International Conference on Historical Linguistic, Düsseldorf, 1-17 August 1997*, pages 101–116, Amsterdam. John Benjamins.

[Chomsky, 1957] Chomsky, N. A. (1957). *Syntactic Structures*. Mouton, The Hague.

[Chomsky and Halle, 1968] Chomsky, N. A. and Halle, M. (1968). *The Sound Pattern of English*. Harper & Row, New York.

[Clark, 1985] Clark, E. V. (1985). Acquisition of romance, with special reference to french. In Slobin, D. I., editor, *The crosslinguistic study of language acquisition*, volume 1, pages 687–782. Lawrence Erlbaum Associates, Hillsdale NJ.

[Cormier, 1999] Cormier, Y. (1999). *Dictionnaire du français acadien*. Éditions Fides, Montréal.

[Cruse, 1986] Cruse, D. A. (1986). *Lexical Semantics*. Cambridge University Press, Cambridge England.

[de Saussure, 1995] de Saussure, F. (1995). *Cours de linguistique générale*. Payot, Paris. Original edition is 1916.

[Dell, 1970] Dell, F. (1970). *Les règles phonologiques tardives et la phonologie dérivationnelle du français*. Ph.D. dissertation, Massachussetts Institute of Technology.

[Donegan and Stampe, 1979] Donegan, P. J. and Stampe, D. K. (1979). The study of natural phonology. In Dinnsen, D. A., editor, *Current Approaches to Phonological Theory*, pages 126–173. Indiana University Press, Bloomington & Londres.

[Donegan and Stampe, 1983] Donegan, P. J. and Stampe, D. K. (1983). Rhythm and the holistic organization of language structure. In *Papers from the Parasessions*, pages 337–353, Chicago. Chicago Linguistic Society, Chicago Linguistic Society.

[Dorrance, 1935] Dorrance, W. A. (1935). The survival of French in the old district of Sainte Geneviève. Master's thesis, The University of Missouri, Columbia.

[Dressler et al., 1987] Dressler, W. U., Luschützky, H. C., Pfeiffer, O. E., and Rennison, J. R., editors (1987). *Phonologica 1984*. Cambridge University Press, Cambridge England.

[Dörre and Eisele, 1989] Dörre, J. and Eisele, A. (1989). Determining consistency of feature terms with distributed disjunctions. In Metzing, D., editor, *Proceedings of GWAI-89, 13th German Workshop on Artificial Language*, number 216 in Informatik-Fachberichte, pages 270–279, Berlin. Springer-Verlag.

[Ferguson and Farwell, 1975] Ferguson, C. A. and Farwell, C. B. (1975). Words and sounds in early language acquisition. *Language*, 51:419–439.

[Fodor, 1972] Fodor, J. D. (1972). Beware. *Linguistic Inquiry*, 3(4):528–534. Squib.

[Ford and Singh, 1983] Ford, A. and Singh, R. (1983). On the status of morphophonology. In *Papers from the Parasessions*, pages 63–78, Chicago. Chicago Linguistic Society, Chicago Linguistic Society.

[Ford et al., 1997] Ford, A., Singh, R., and Martohardjono, G. (1997). *Pace Pāṇini: Towards a Word-Based Theory of Morphology*. Peter Lang, New York.

[Fromkin et al., 2003] Fromkin, V., Rodman, R., and Hyams, N. (2003). *An introduction to language*. Thomson Wadsworth, Boston, seventh edition.

[Fulmer, 1991] Fulmer, S. (1991). A case of inflection before derivation. In *WCCFL Proceedings 10*, pages 151–162.

[Gazdar et al., 1985] Gazdar, G., Klein, E., Pullum, G., and Sag, I. (1985). *Generalized Phrase Structure Grammar*. Harvard University Press, Cambridge MA.

[Gesner, 1985] Gesner, B. E. (1985). *Description de la morphologie verbale du parler acadien de Pubnico (Nouvelle-Écosse) et comparaison avec le français standard*. Centre international de recherche sur le bilinguisme, Québec.

[Gnanadesikan, 1995] Gnanadesikan, A. E. (1995). Markedness and faithfulness in child phonology. Rutgers Optimality Archive.

[Greenberg, 1966] Greenberg, J. H. (1966). Some universals of grammar with particular reference to the order of meaningful elements. In Greenberg, J. H., editor, *Universals of language*, pages 73–113. MIT Press, Cambridge MA, second edition.

[Gulian, 1965] Gulian, K. H. (1965). *Elementary Modern Armenian grammar*. Frederick Ungar Publishing Co, New York.

[Hale and Reiss, 1995] Hale, M. and Reiss, C. (1995). On the initial ranking of ot faithfulness constraints in universal grammar. Poster presentation at Stanford Child Language Research Forum. Also available in Concordia University Working Papers in Language and Linguistics and Rutgers Optimality Archive.

[Halle, 1973] Halle, M. (1973). Prolegomena to a theory of word formation. *Linguistic Inquiry*, 4(1):3–16.

[Halle and Marantz, 1993] Halle, M. and Marantz, A. (1993). Distributed morphology and the pieces of inflection. In Hale, K. and Keyser, S. J., editors, *The View from Building 20*, pages 111–176. MIT Press, Cambridge MA.

[Halle and Vaux, 1998] Halle, M. and Vaux, B. (1998). Theoretical aspects of indo-european nominal morphology: The nominal declensions of latin and armenian. In Jay, H. J., Melchert, C., and Olivier, L., editors, *Mir Curad: Studies in Honor of Clavert Watkins*, pages 223–240. Innsbrucker Beitraege zur Sprachwissenschaft, Innsbruck.

[Hansson, 1999] Hansson, G. O. (1999). When in doubt...: Intraparadigmatic dependencies and gaps in icelandic. In Tamanji, P., Hirotani, M., and Hall, N., editors, *Proceedings of the 29th meeting of the North Eastern Linguistic Society*, pages 105–119, Amherst. GLSA Publications, University of Massachusetts.

[Hoffmann, 1963] Hoffmann, C. (1963). *A Grammar of the Margi Language*. Oxford University Press, London.

[Hopper and Traugott, 1993] Hopper, P. J. and Traugott, E. C. (1993). *Grammaticalization*. Cambridge University Press, Cambridge England, first edition.

[Hopper and Traugott, 2003] Hopper, P. J. and Traugott, E. C. (2003). *Grammaticalization*. Cambridge University Press, Cambridge England, second edition.

[Jusczyk, 1997] Jusczyk, P. W. (1997). *The Discovery of Spoken Language*. MIT Press, Cambridge.

[Kaplan and Bresnan, 1982] Kaplan, R. and Bresnan, J. (1982). Lexical-functional grammar: A formal system for grammatical representation. In Bresnan, J., editor, *The Mental Representation of Grammatical Relations*, pages 173–281. The MIT Press, Cambridge MA.

[Kiparsky, 1973] Kiparsky, P. V. (1973). "Elsewhere" in phonology. In [Anderson and Kiparsky, 1973], pages 93–106.

[Kiparsky, 982a] Kiparsky, P. V. (1982a). *Explanation in Phonology*. Foris Publications, Dodrecht.

[Kiparsky, 982b] Kiparsky, P. V. (1982b). Lexical phonology and morphology. In Yang, I.-S., editor, *Linguistics in the Morning Calm*, pages 3–91. Hansin, Seoul.

[Kiparsky, 1995] Kiparsky, P. V. (1995). The phonological basis of sound change. In Goldsmith, J., editor, *The Handbook of Phonological Theory*. Blackwell, Oxford.

[Kiparsky, 1996] Kiparsky, P. V. (1996). Allomorphy or morphophonology. In [Singh, 1996]. Edited with the collaboration of Richard Desrochers.

[Kiparsky, 2000] Kiparsky, P. V. (2000). Opacity and cyclicity. *The Linguistic Review*, 17:351–367.

[Kiparsky, 2004] Kiparsky, P. V. (2004). Blocking and periphrasis in inflectional paradigms. In *Yearbook of Morphology 2004*, pages 113–135. Springer, Dordrecht.

[Kiparsky, 2005] Kiparsky, P. V. (2005). Grammaticalization as optimization. Available on author's website.

[Kisseberth, 1970] Kisseberth, C. W. (1970). On the functional unity of phonological rules. *Linguistic Inquiry*, 1(3):291–306.

[Klingler, 2003] Klingler, T. A. (2003). Language labels and language use among cajuns and creoles in louisiana. *Penn Working Papers in Linguistics*, 9.

[Kogian, 1949] Kogian, F. S. L. (1949). *Armenian Grammar (West Dialect)*. Mechitharist Press, Vienna.

[Koutsoudas et al., 1971] Koutsoudas, A., Sanders, G., and Noll, C. (1971). The application of phonological rules. *Language*, 50(1):1–28.

[Krieger et al., 1993] Krieger, H.-U., Nerbonne, J., and Pirker, H. (1993). Feature-based allomorphy. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 140–147, Columbus OH.

[Kurylowicz, 1949] Kurylowicz, J. (1949). La nature des procès dits 'analogiques'. *Acta Linguistica*, 5:121–138.

[Labov, 1994] Labov, W. (1994). *Principles of linguistic change*. Blackwell, Oxford England & Cambridge MA.

[Leben, 1978] Leben, W. R. (1978). The representation of tone. In Fromkin, V. A., editor, *Tone: a linguistic survey*, pages 177–219. Academic Press, New York.

[Levin, 1985] Levin, B. C., editor (1985). *Lexical Semantics in Review*. Lexicon Project Working Papers. Center for Cognitive Science MIT, Cambridge MA.

[Lightner, 1963] Lightner, T. M. (1963). A note on the formulation of phonological rules. In *Quarterly Progress Report of the Research Laboratory of Electronics*, number 68, pages 187–189. MIT.

[Marantz, 1998] Marantz, A. (1998). No escape from syntax: Don't try morphological analysis in the privacy of your own lexicon. In Dimitriadis, A., editor, *Proceedings of the 1998 Penn Linguistics Colloqium*. Available from Penn Working Papers in Linguistics.

[Matthews, 1991] Matthews, P. (1991). *Morphology*. Cambridge University Press, Cambridge England & New York NY.

[McCarthy, 1999] McCarthy, J. (1999). Sympathy and phonological opacity. *Phonology*, 16:331–399.

[McCarthy and Prince, 1993] McCarthy, J. and Prince, A. (1993). Prosodic morphology i: Constraint interaction and satisfaction. Technical Report 3, Rutgers University Center for Cognitive Science.

[Mel'cuk, 2001] Mel'cuk, I. A. (1993, 1994, 1996, 1997, 2001). *Cours de morphologie générale, tomes 1 à 5*. Presses de l'Université de Montréal, Montréal.

[Mithun, 2001] Mithun, M. (2001). Lexical forces shaping the evolution of grammar. In *Historical Linguistics 1999: Selected Papers from the 13th International Conference on Historical Linguistic, Vancouver, 9-13 August 1999*, pages 241–252.

[Morin,] Morin, Y. C. Les yods fluctuants dans la morphologie du verbe français. In Fradin, B., editor, *La raison morphologique: Hommages à la mémoire de Danièle Corbin*. John Benjamins, Amsterdam. To appear.

[Morin, 1987] Morin, Y. C. (1987). Remarques sur l'organisation de la flexion des verbes français. *ITL Review of Applied Linguistics*, 77-78:13–91.

[Morin, 1995] Morin, Y. C. (1995). De l'acquisition de la morphologie : le cas des verbes morphologiquement défectifs du français. In Shyldkrot, H. B.-Z. and Kupferman, L., editors, *Tendances récentes en linguistique française et générale : volume dédié à David Gaatone*, pages 295–310. John Benjamins, Philadelphia.

[Neuvel and Singh, 2002] Neuvel, S. and Singh, R. (2002). Vive la difference! what morphology is about. *Folia Linguistica*, 35(3-4):313–320.

[Orgun and Sprouse, 1999] Orgun, C. O. and Sprouse, R. L. (1999). From mparse to control: Deriving ungrammaticality. University of California Berkleley, Rutgers Optimality Archive 224.

[Paul, 1888] Paul, H. (1888). *Principles of the history of language*. London. Translated from the second edition of the original German text by H.A. Strong.

[Perlmutter, 1971] Perlmutter, D. M. (1971). *Deep and Surface Structure Constraints in Syntax*. Holt, Rinehart & Winston, New York.

[Pinker, 1999] Pinker, S. (1999). *Words and Rules: The ingredients of language*. Basic Books, New York.

[Pinker and Prince, 1988] Pinker, S. and Prince, A. (1988). On language and connectionism: analysis of a parallel distributed model of language acquisition. *Cognition*, 28:73–193.

[Plénat, 1981] Plénat, M. (1981). L'"autre" conjugaison. *Cahiers de grammaire*, 3. Université de Toulouse-Le Mirail.

[Pollard and Sag, 1994] Pollard, C. and Sag, I. A. (1994). *Head-driven Phrase Structure Grammar*. University of Chicago Press & CSLI Publications, Chicago & Stanford.

[Pope, 1934] Pope, M. K. (1934). *From Latin to Modern French*. Manchester University Press, Manchester.

[Prince and Smolensky, 1993] Prince, A. and Smolensky, P. (1993). Optimality theory: Constraint interaction in generative grammar. Technical Report 2, Rutgers University Center for Cognitive Science.

[Pulleyblank, 1986] Pulleyblank, D. (1986). *Tone in Lexical Phonology*. D. Reidel Publishing Company, Dordrecht, Boston, Lancaster & Tokyo.

[Pulleyblank, 1997] Pulleyblank, D. (1997). Optimality theory and features. In [Archangeli and Langendoen, 1997], pages 59–101.

[Pustejovsky, 1995] Pustejovsky, J. (1995). *The generative lexicon*. MIT Press, Cambridge MA.

[Ramscar, 2002] Ramscar, M. (2002). The role of meaning in inflection: Why the past tense does not require a rule. *Cognitive Psychology*, 45:45–94.

[Rice, 2005] Rice, C. (2005). Optimal gaps in optimal paradigms. Rutgers Optimality archive 753-0605.

[Rice, 1999] Rice, K. (1999). *Morpheme order and semantic scope : word formation in the Athapaskan verb*. Cambridge University Press, New York.

[Ross, 1967] Ross, J. R. (1967). *Constraints on variables in syntax*. Ph.D. dissertation, Massachusetts Institute of Technology.

[Rummelhart and McClelland, 1986] Rummelhart, D. E. and McClelland, J., editors (1986). *Parallel Distributed Processing: Explorations into the microstructure of cognition*. The MIT Press, Cambridge MA.

[Russel, 1997] Russel, K. (1997). Optimality theory and morphology. In [Archangeli and Langendoen, 1997].

[Schuchardt, 1885] Schuchardt, H. (1885). *Über die Lautgesetze: Gegen die Junggrammatiker*. Berlin.

[Siegel, 1977] Siegel, D. (1977). The adjacency constraint and the theory of morphology. In *Proceedings of the North Eastern Linguistic Society*, number 8, pages 189–197.

[Singh, 1987] Singh, R. (1987). Well-formedness conditions and phonological theory. In [Dressler et al., 1987].

[Singh, 1990] Singh, R. (1990). Do we really want a syntax-organ? a note on phonology, morphology and the autonomy thesis. *Indian Linguistics*, 49:109–114.

[Singh, 1996] Singh, R., editor (1996). *Trubetzkoy's Orphan—Proceedings of the Montréal Roundtable "Morphonology: Contemporary Responses"*. John Benjamins Publishing Company, Amsterdam & Philadelphia. Edited with the collaboration of Richard Desrochers.

[Singh and Starosta, 2003] Singh, R. and Starosta, S. (2003). *Explorations in Seemless Morphology*. Sage, New Delhi.

[Skousen, 1989] Skousen, R. (1989). *Analogical modeling of language*. Kluwer Academic Publishers, Dordrecht & Boston.

[Stampe, 1973] Stampe, D. K. (1973). *A Dissertation on Natural Phonology*. Indiana University Linguistics Club, Bloomington.

[Stampe, 1987] Stampe, D. K. (1987). On phonological representation. In [Dressler et al., 1987].

[Stanley, 1967] Stanley, R. (1967). Redundancy rules in phonology. *Language*, 43(2):393–436.

[Stump, 1990] Stump, G. (1990). Breton inflection and the split morphology hypothesis. In Hendrick, R., editor, *The syntax of the modern Celtic languages*, pages 97–119. Academic Press, San Diego.

[Stump, 2001] Stump, G. (2001). *Inflectional morphology: A theory of paradigm structure*. Cambridge University Press, Cambridge.

[Tesar and Smolensky, 1998] Tesar, B. and Smolensky, P. (1998). Learnability in optimality theory. *Linguistic Inquiry*, 29(2):229–268.

[Thogmartin, 1970] Thogmartin, C. O. (1970). *The French Dialect of Old Mines, Missouri*. Ph.D. dissertation, University of Michigan.

[Trask, 1996] Trask, R. L. (1996). *Historical linguistics*. Arnold, London & New York.

[Vaux, 1998] Vaux, B. (1998). *The Phonology of Armenian*. The Phonology of the World's Languages. Clarendon Press, Oxford.

[Vaux, 2003] Vaux, B. (2003). Syllabification in armenian, universal grammar, and the lexicon. *Linguistic Inquiry*, 34(1):91–125.

[Weinreich et al., 1968] Weinreich, U., Labov, W., and Herzog, M. I. (1968). Empirical foundations for a theory of language change. In Lehmann, W. P. and Malkiel, Y., editors, *Directions for historical linguistics; a symposium*, pages 95–195. University of Texas Press, Austin.

[Williams, 1976] Williams, E. S. (1976). Underlying tone in margi and igbo. *Linguistic Inquiry*, 7(3):463–484.

[Williams, 1981] Williams, E. S. (1981). On the notions 'lexically related' and 'head of a word'. *Linguistic Inquiry*, 12(2):245–274.

[Zalizniak, 1977] Zalizniak, A. A. (1977). *Grammaticheskii slovar' russkogo iazyka : slovoizmenenie : okolo 1000 000 slov*. Russkii iazyk, Moscow.

[Zwicky, 1975] Zwicky, A. M. (1975). Settling on an underlying form: The english inflectional endings. In Cohen, D. and Wirth, J. R., editors, *Testing Linguistic Hypotheses*, pages 129–185. Hemisphere, Washington.

[Zwicky, 1992] Zwicky, A. M. (1992). Some choices in the theory of morphology. In Levine, R., editor, *Formal grammar: Theory and implementation*, pages 327–371. Oxford University Press, Oxford.