Introduction

Sudoku puzzles such as the one in Figure 1 are (as we write this) all the rage. (If you have never worked one of these puzzles – and we expect that you have – the idea is to fill each empty square with a number between 1 and 9, inclusive, in such a way that no number is repeated in any row, any column, or any of the nine 3 × 3 squares of which the puzzle is composed.) In this book, we are, in a sense, concerned with Sudoku puzzles of a different kind. In a Sudoku puzzle of the familiar kind, the numbers provided in the puzzle's initial state give enough information to deduce all of the numbers that appear once the puzzle is successfully completed. A table of a language's inflectional paradigms is in some ways similar to a completed Sudoku puzzle: many of the forms occupying an inflectional paradigm's "squares" (we call them "cells" rather than "squares") are predictable, so that the paradigm can be given a less redundant (or redundancy-free) representation in which only certain particular cells are filled; the word forms occupying these particular cells allow one to predict all the forms that are omitted, just as the numbers in a Sudoku puzzle's initial state allow one to predict the missing numbers. Word forms that satisfy this requirement function as "principal parts" for the paradigm table. (Analogously, we might call the numbers in a Sudoku puzzle's initial state its principal parts.)

Principal parts have a long history in language pedagogy. In particular, they are helpful for learning and using languages with inflection-class systems. In such languages, lexemes belonging to the same syntactic category inflect for the same sets of morphosyntactic properties but do so with different morphology, according to their membership in one or another conjugation or declension. (Inflection-class systems might therefore be likened to books of Sudoku puzzles, with each inflection class requiring a different solution.)

But the interest of principal-part systems does not begin and end with language pedagogy. Principal parts have theoretical interest as well, because they hint at – and to an extent reify – the complex network of implicative relations affiliating the different cells in a lexeme's paradigm. Such relations help language users to learn, recognize, and produce a lexeme's full array of

Cambridge University Press
978-1-107-02924-8 - Morphological Typology: From Word to Paradigm
Gregory Stump and Raphael A. Finkel
Excerpt
More information

			8					
			7				8	
3	7							
	9	6						5
	1	7	5					
				1				
					9			
				4	1		7	
						5		

Figure 1. *A Sudoku puzzle*

forms. Principal parts "work" because they participate in implicative relations that are certainties rather than mere probabilities. There are, to be sure, relations of mere probability in morphology: for instance, if one encountered a new verb with infinitive *dring* and past participle *drung*, one would expect its past-tense form to be either *drang* (compare *sang* and *rang*) or *drung* (compare *stung* and *flung*), but one could not be sure which is right. On the other hand, one could be certain that its present participle is *dringing*: in English, if a verb with infinitive X has a present participle, it is invariably X*ing*.[1]

The networks of implicative certainties upon which principal parts depend exhibit considerable typological variation, and it is in terms of this variation that we define an inflection-class system's relative complexity. Our goal in this book is to develop this conception of complexity as an objective, measurable property of inflection-class systems. To this end, we investigate implicative relations in inflectional morphology from both formal and typological perspectives, addressing questions such as:

(a) How do languages vary with respect to their networks of implicative certainties? How do different inflection classes within a single language vary in this respect?

(b) What makes a word form predictable? What makes a word form predictive? How do predictability and predictiveness vary across inflection classes within a single language? How do they vary across languages?

[1] Rare is the verb that has an infinitive but no present participle, though there are examples, e.g. *beware* and (for some speakers) *use to* /jus tu/ (as in *I didn't use to go there*).

(c) How can the predictability and predictiveness of a lexeme's forms be measured?
(d) How are the implicative relations among a paradigm's cells to be represented? What is their theoretical status, and what are their implications for a general theory of inflectional morphology?
(e) How do these factors figure in the conception and quantification of an inflection-class system's complexity?

The hypothesis that linguistic subsystems can be meaningfully said to vary in complexity has drawn considerable interest in recent work in language typology (Baerman *et al*. 2010; Dahl 2004; Hawkins 2004; Miestamo *et al*. 2008; Nichols 1992, 2007; Nichols *et al*. 2006); the dimensions and degrees of such variation are an important focus of current research. In developing the notion of linguistic complexity as it specifically relates to inflection-class systems, we compare the individual inflection classes in a language's inflectional system according to the manner and extent of their deviation from canonical ideals of maximal transparency and maximal opacity; the criteria for these comparisons in turn inform a more general comparison of whole inflection-class systems. We define the complexity of an inflection-class system as the extent to which it inhibits motivated ("certain") inferences about a lexeme's full paradigm from subsets of the forms in this paradigm.

This definition might be taken to imply that an inflection-class system that is more complex (in our sense) has a greater "cost" – that it is more difficult to learn, that it complicates the production and comprehension of utterances, and that it makes lexemes' lexical representations more intricate. Though we would be surprised if this proved not to be the case, our thesis is not that complex inflection-class systems are costly in a psycholinguistic sense. Rather, we regard relative complexity (in our sense) as an objectively observable property of inflection-class systems. The motivated inferences in terms of which we define an inflection-class system's complexity are relations of logical implication, and as such, are directly detectable by computational means. Moreover, we shall see that an inflection-class system may be complex (in our sense) for one or more of a variety of reasons. Accordingly, we will, in this course of this book, propose ten measurable correlates that allow an inflection-class system's complexity to be seen as a multifaceted but quantifiable typological variable.

Our discussion and conclusions are based on evidence from a wide range of languages with complex inflectional systems. In analyzing these systems, we make extensive use of the Principal-Parts Analyzer (PPA), a computational tool that we have devised specifically for generating principal-part analyses and for measuring patterns of predictability, predictiveness, and entropy in an inflection-class system's paradigms.

In Chapter 1, we examine the traditional notion of **principal parts** and consider the factors involved in giving this notion theoretical and typological substance. We distinguish three principal-part schemes: the lexemes in a static scheme have the same members of their paradigms as principal parts; in an adaptive scheme, a lexeme's principal parts are ordered, defining a path through a kind of flowchart determining its inflection-class membership; in a dynamic scheme, lexemes have unordered principal parts which vary in number from one inflection class to another.

In Chapter 2, we develop the notion of a **plat**, a format for modeling a language's system of inflection classes; models in this format constitute our central objects of analysis in identifying and measuring an inflection-class system's degree of complexity. As we show, the construction of plats raises a crucial representational issue: what is actually modeled by a model of a language's inflection classes? In addressing this issue, we draw a critical conceptual distinction between hearer-oriented and speaker-oriented plats.

In Chapter 3, we present a preliminary typology of principal-part systems. Traditionally, morphological typology has focused on the properties of individual word forms, involving such criteria as degree of synthesis and degree of fusion (Greenberg 1960); but in a language's inflectional system, the properties of entire paradigms are of considerable typological interest. We propose three typological criteria for a preliminary classification of inflection-class systems: (i) the number of principal parts needed to determine a lexeme's paradigm, (ii) the number of principal parts needed to determine a given cell in a lexeme's paradigm, and (iii) the extent to which certain cells enjoy a privileged status as determinants of lexemes' inflection-class membership. We apply these criteria to compare the principal-part systems of a range of languages, including Comaltepec Chinantec (Oto-Manguean; Mexico), Dakota (Siouan; USA), Fur (Nilo-Saharan; Sudan), Icelandic, Koasati (Muskogean; USA), Kwerba (Trans-New Guinea; Irian Jaya), Latin, Ngiti (Nilo-Saharan; DR Congo), Sanskrit, and Tuḷu (Dravidian; India). These criteria are observable correlates of an inflection-class system's complexity.

In Chapter 4, we discuss **inflection-class transparency**. An inflection class is transparent to the extent that a member lexeme's full inventory of inflected forms may be inferred from subsets of that inventory. In the paradigm of a lexeme belonging to a maximally transparent inflection class, each cell allows every other cell to be deduced; a paradigm of this sort needs only a single principal part, and any of its cells can serve as this principal part. The relative transparency of inflection classes that deviate from this ideal intuitively depends on three criteria: all else being equal, inflection class A with member lexeme *a* is more transparent than inflection class B with member lexeme *b* (1) if the number of principal parts required to deduce a given cell in *a*'s paradigm is, on average, lower than the number of principal parts required to deduce a given cell in *b*'s paradigm; (2) if there are more alternative principal-part analyses for *a* than for *b*; and (3) if *a*'s paradigm has fewer unpredictable cells than *b*'s paradigm. We give formal substance to these intuitive criteria by proposing precise measures of **inflection-class predictability** and **cell predictability**. We demonstrate these measures with evidence from Fur. We show that the deviation of Fur's conjugation classes from maximal transparency is irreconcilable with the No-Blur Principle (Cameron-Faulkner & Carstairs-McCarthy 2000). The proposed measures of inflection-class predictability and cell predictability afford a precise account of cross-linguistic differences in inflection-class transparency, as we demonstrate in a comparison of the conjugational system of Fur with those of Comaltepec Chinantec and Icelandic. Both of these measures are correlates of an inflection-class system's complexity.

In Chapter 5, we show that a cell's predictability and its **predictiveness** may be enhanced by supplementing its phonological realization with additional grammatical information. In particular, we show that in Sanskrit gender specifications tend to heighten both a cell's predictability and its predictiveness; information about stem delimitation also tends to heighten a cell's predictiveness (though not its predictability). Cell predictiveness is another correlate of an inflection-class system's complexity.

In Chapter 6, we discuss two phenomena that complicate the investigation of paradigms' implicative relations: *impostors* (lexemes one or more of whose realizations exhibit morphology that can be analyzed in two different ways, making them ambiguous with respect to their inflection-class membership) and *heteroclites* (lexemes that follow the patterns of different inflection classes in different parts of their paradigms). We draw a distinction between motivated and unmotivated inferences about a lexeme's paradigm of word forms; this distinction is essential to our characterization of an inflection-class system's complexity.

In the conceptual framework developed in Chapters 1 through 6, we regard a lexeme's principal parts as a subset of the cells within its paradigm. In Chapter 7, we discuss an alternative possibility: that a lexeme's principal parts are a subset of the indexed stems upon which its realizations are based. As we show, some inflectional systems favor this alternative conception of principal parts. We elaborate this idea with extensive evidence from the French verb system, whose conjugation-class distinctions are almost entirely expressed by differences in the formation and alternation of stems (Bonami & Boyé 2002). We show that the framework developed in earlier chapters is straightforwardly generalizable to accommodate stem-based principal-part systems.

In Chapter 8, we discuss the **marginal detraction hypothesis**, according to which marginal inflection classes (those with very few members) tend to detract most strongly from the predictability of other, more central inflection classes. Drawing on the evidence of Icelandic verb conjugation, we present an empirical finding that supports this hypothesis. On their own, central (i.e. nonmarginal) conjugations in Icelandic allow the morphosyntactic property sets in a verb's paradigm to be grouped into a comparatively small number of **distillations** (where a distillation is a set S of morphosyntactic property sets such that members of S are isomorphic in their realization across inflection classes); this phenomenon enhances the predictability of a verb's inflection-class membership from a relatively small number of word forms in its paradigm. Marginal conjugations, however, allow fewer morphosyntactic property sets to be grouped together, detracting from the predictability of other conjugations. This finding implies that as languages evolve historically, they are constantly subject to a kind of tension: on one hand, morphological innovations that minimize the number of distillations in a language enhance the predictability of its paradigms, and the number of distillations can be reduced by eliminating marginal inflection classes; yet, the persistence of marginal inflection classes is favored by the fact that they are inherently more predictable than central inflection classes. An inflection-class system's number of distillations is another observable correlate of its complexity.

The feasibility of deducing all of the cells in a lexeme's paradigm from a small subset of these cells raises an important question about the definition of a language's inflectional morphology: should it be defined by means of rules of exponence (which deduce the realization of a given cell K in the paradigm of a lexeme L from L's stem(s) together with the morphosyntactic property set associated with K), or should it instead be defined by means of implicative rules (which deduce the realization of a given cell in the paradigm of a lexeme L from the realizations of one or more other cells in L's paradigm)? In Chapter 9, we compare these approaches – the **exponence-based approach** (Anderson 1992; Matthews 1972; Stump 2001; Zwicky 1985) and the **implicative approach** (Ackerman *et al.* 2009; Blevins 2006) – arguing that they differ in their strengths and that they are not mutually exclusive. We propose a hybrid approach in which implicative rules are derived as theorems of an inflectional system's exponence-based definition; we illustrate with a fragment of Sanskrit declensional morphology.

Recent research on morphological complexity (Ackerman *et al.* 2009; Milin *et al.* 2009; Moscoso del Prado Martín *et al.* 2004) has employed the information-theoretic measures of **entropy** and **conditional entropy** (Shannon 1951) as a way of quantifying the degree to which cells in a lexeme's paradigm are predictable. Principal parts are those cells in a paradigm that reduce the conditional entropy of every remaining cell to zero. In Chapter 10, we compare entropy measures with our measures of predictability and predictiveness. In our discussion, we focus particular attention on the relevance of type frequency in applying all these measures. We demonstrate that measures of entropy, predictability, and predictiveness reveal different patterns and are therefore complementary as elucidations of an inflection-class system's complexity.

In Chapter 11, we discuss the general program of investigating complexity as a dimension of typological contrast, and we situate our approach to the complexity of inflection-class systems within this general program. We draw together the various correlates of inflection-class system's complexity:

(a) the number of distillations the system has
(b) the size of the system's optimal static principal-part sets
(c) the density of the system's optimal static principal-part sets (given (a) and (b))
(d) the average size of optimal dynamic principal-part sets for the system's inflection classes
(e) the average ratio of actual to possible optimal dynamic principal-part sets for the system's inflection classes
(f) the average number of principal parts required to deduce the realization of a given cell in a lexeme's paradigm
(g) a cell's average predictiveness
(h) the average inflection-class predictability of the system's inflection classes
(i) the average cell predictability of the system's inflection classes
(j) what we call the system's *m*-system entropy.

As we show, these measures quantify subtly different aspects of a paradigm's implicative structure; all are therefore informative as components of an inflection-class system's complexity.

In Chapter 12, we devote additional discussion to the technical task of constructing plats for investigating principal parts, implicative relations, and the complexity of inflection-class systems. We revisit the choices introduced in Chapter 2 and describe how these choices affect the measures introduced in Chapters 4 and 5.

The research on which this book rests is informed by a range of computational algorithms, all of which are embodied in the Principal-Parts Analyzer (PPA). In Chapter 13, we present the formal details of the PPA: form of input, form of output and algorithms employed. We have made the PPA freely available at the following website, along with the plats employed in our research: www.cambridge.org/stump_finkel.

During the preparation of this book, we have presented our work on the typology of inflection-class systems in a number of places; see Finkel & Stump (2006a,b; 2007; 2008a,b; 2009; 2010; 2011a,b; 2012) and Stump (2010). We wish to thank the organizers of the following conferences for inviting us to present our work in progress:

- 12th International Morphology Meeting, Budapest, 2006.
- Conference on Analogy in Grammar: Form and Acquisition, Max Planck Institute for Evolutionary Anthropology, 2006.
- Southeast Morphology Meeting, University of Surrey, 2008.
- Décembrettes 6: Colloque International de Morphologie – Morphologie et classes flexionnelles, Université de Bordeaux, 2008.
- Workshop on Morphological Complexity: Implications for the Theory of Language, Harvard University, January 22, 2010.
- Workshop on Morphology and Formal Grammar, Université Paris Diderot, 2010.
- Workshop on Quantitative Measures in Morphology and Morphological Development, Center for Human Development, UC San Diego, 2011.
- Workshop on the Challenges of Complex Morphology to Morphological Theory, Linguistic Society of America Summer Institute, Boulder, 2011.
- Conference on Morphological Complexity, British Academy, 2012.

At these conferences and elsewhere, we have benefited from the comments and suggestions of a number of people; we wish to convey our particular thanks to Farrell Ackerman, Adam Albright, Matthew Baerman, Jim Blevins, Olivier Bonami, Gilles Boyé, Dunstan Brown, Greville Corbett, Andrew Hippisley, Rob Malouf, Fermín Moscoso del Prado Martín, and Andrea Sims. Olivier Bonami also kindly supplied a database of French conjugations of which we have made extensive use.

1 *Principal parts*

We regard the complexity of an inflection-class system as the extent to which it inhibits motivated inferences about a lexeme's full paradigm of forms from subsets of those forms. Because principal parts are a crystallization of the implicative relations among different cells in a lexeme's paradigm, they are a good starting point for an examination of inflection-class systems' complexity. In this chapter, we consider principal parts both in their traditional pedagogical function (§1.1) and in the broader context of linguistic theory and typology (§1.2). In §1.3, we expound some preliminary assumptions about the nature of principal parts and implicative relations.[1]

1.1 The traditional notion of principal parts

The notion of principal parts depends on the logically prior notion of an inflectional paradigm. This concept has been defined in different ways by different people. For complete clarity, we define the **paradigm** of a lexeme L as a complete set of cells for L, where each **cell** is the pairing of L with a complete and coherent **morphosyntactic property set** (MPS) for which L is inflectable.[2] Given any such MPS σ, we represent the pairing of L with σ as ⟨L, σ⟩. The cell

[1] Portions of this chapter first appeared, in somewhat different form, in Finkel & Stump (2007), © Springer Science+Business Media B.V. 2007. Used with kind permission from Springer Science +Business Media B.V. Other portions of this chapter first appeared, in somewhat different form, in R. Finkel & G. Stump (2009), used with kind permission of Oxford University Press.

The data sets that we have employed in this chapter are:

principal.A	principal.C	principal.E	principal.ngiti
principal.B	principal.D	principal.F	principal.norse

These are available at the *Morphological Typology* website www.cambridge.org/stump_finkel.

[2] All boldface words are technical terms, some of our own invention. We gather their definitions in the glossary. See pp. xxiii–xxiv for a complete list of the abbreviations that we employ.

⟨L, σ⟩ is expressed morphologically as a word form *w*; *w* is in this context the **realization** of L, of σ, and of ⟨L, σ⟩. For example, the paradigm of the English verb BE[3] is a set containing such cells as these:[4]

⟨BE, {1sg pres ind}⟩
⟨BE, {3sg pres ind}⟩
⟨BE, {3sg past ind}⟩
⟨BE, {3sg irrealis}⟩
⟨BE, {past ptcp}⟩.

These cells have the respective realizations *am*, *is*, *was*, *were*, and *been*. A lexeme's **realized cells** are pairings of its realizations with the MPSs that they realize. The realized cells of the lexeme BE include these pairs:

⟨*am*, {1sg pres ind}⟩
⟨*is*, {3sg pres ind}⟩
⟨*was*, {3sg past ind}⟩
⟨*were*, {3sg irrealis}⟩
⟨*been*, {past ptcp}⟩.

Thus, we also say that ⟨*am*, {1sg pres ind}⟩ realizes ⟨BE, {1sg pres ind}⟩, and so on. The complete set of a lexeme's realized cells constitutes its **realized paradigm**. We find it useful to define the related notion of a syntactic category's **paradigm schema**: the set of complete and coherent MPSs realized by the paradigms of specific lexemes belonging to that syntactic category. For instance, the paradigm schema of a Latin noun is this set:

{{nom sg}	{nom pl}
{voc sg}	{voc pl}
{gen sg}	{gen pl}
{dat sg}	{dat pl}
{acc sg}	{acc pl}
{abl sg}	{abl pl}
{loc sg}	{loc pl}}.

3 Here and throughout, we follow the conventional practice of representing lexemes in small capital letters.

4 A **morphosyntactic property** is the specification of an inflectional category by one of its permissible values. For example, the morphosyntactic property 'NUMBER: singular' is a specification of the inflectional category of number. Where there is no risk of ambiguity, we abbreviate the morphosyntactic property C: *v* (where C is an inflectional category and *v* is one of C's permissible values) as *v*.