Machine Learning of Inflection

1 Introduction

The task of learning a natural language is an instance of inductive inference (making generalizations based on the past observations to predict the future ones) studied within several traditions in the sciences under such names as “machine learning,” “grammar induction,” and “computational learning theory.” In this chapter, I aim to introduce some of the concepts and approaches used in the formal study of learnability and the relevance of this field to theoretical linguistics. I then discuss issues specific to learning inflectional morphology. However, my focus is not on providing a comprehensive review of the proposals for learning inflection, but rather on the general problems and obstacles in learning inflection and on the question of what properties a computational system needs if it is to overcome these obstacles. Inflection, considered separately from other components of language, is relatively restricted in its expressive power which should make it easier to learn than syntax. However, inflectional systems are full of irregularities and mismatches between different levels of structure, and such irregularities make learning difficult. Overall, I conclude that linguistically interesting proposals for machine learning of inflection should provide explanations for the nature and extent of irregularities and for the specific patterns of language acquisition and language change. Such explanations are not usually provided by traditional theoretical models of inflection that focus on predicting what patterns are possible vs. impossible, and do not have a way of predicting what patterns should be easy to learn, or what mistakes might occur during the acquisition process.

A prominent leitmotif of this chapter is a focus on how computational learning models can serve as a valuable tool for constraining linguistic theories and for uncovering deeper explanations for a variety of linguistic phenomena. The relevance of formal learning models to linguistic theory stems from the causal connection between learning and language structure standardly assumed within the rationalist approach to language (Chomsky, 1975). On this approach, the properties we find in human languages at least to some extent reflect constraints on our cognitive capacity including constraints on the language-acquisition mechanism.

2 Goals of computational learning theory

In the machine learning literature, a language learner is a function that maps data to hypotheses about the data source (e.g., grammars) or to probability distributions over a set of such hypotheses. This learning function can be defined by an algorithm of any sort: a neural network, a stochastic process, a Boolean classifier, an inductive logic program, and so on. The field of computational learning theory as developed by Gold (1967); Blum and Blum (1975); Angluin (1980); Valiant (1984); Jain et al. (1999) and others is the study of such
functions and their behaviors within frameworks that explicitly specify both the criteria for the evaluation of learning success and various constraints on the learning situation, such as what kind of input is allowed, how it is presented, and what the resource limitations are. As in other sciences, idealizations are deliberately used to focus attention on specific aspects of the problem abstracting away from other details. Within more data-oriented traditions including much of the statistical learning literature, the research goals are somewhat different — typically, to directly model a certain aspect of human behavior or corpus distribution. However, the same mathematical results apply to these models and the models themselves can be easily translated into the computational learning theory terms (Niyogi, 2006).

Learning is understood as choosing the right or the most probable hypothesis about the data from a set of possible hypotheses (where the right hypothesis is the one that generates the target language with probability 1). The learning problem is sometimes viewed as the problem of searching for the best grammar within a space of possible grammars from some limited class. Learning algorithms are said to learn a class of languages if they can identify every language in the class given some data from this language. In terms of first language acquisition, learning a class of languages (rather than a single language) corresponds to the idea that human learners must be able to learn whatever natural language they happen to be born into. For a non-technical linguist-oriented introduction to some key concepts and controversies within the field of learnability I refer the reader to Heinz (2013).

In the next sections I highlight one of the central results of learning theory - the fact that no learning algorithm can learn an unrestricted class of languages. The restrictions on the class of languages (or on the class of grammars) to be learned constitute knowledge that is present prior to learning. These restrictions are often referred to as learning biases (which can also be viewed as constraints of UG). The fact that learning without any prior knowledge is impossible, should put to rest the debate between the nativists and the empiricists, although the related debate between the domain-general versus domain-specific constraints on language acquisition remains very vibrant (Clark and Lappin, 2011; Perfors et al., 2011; Pearl and Lidz, 2009, and others).

### 2.1 Learning without restrictions is impossible

Language learnability provides an obvious constraint on what constitutes a possible human language - such a language must be learnable. One natural starting point for investigating which classes of formal patterns are learnable is the Chomsky Hierarchy (Chomsky, 1956), which defines a set of increasingly complex classes of rewrite grammars corresponding to classes of languages generated by these grammars: finite languages, regular languages, context-free languages, context-sensitive, and recursively enumerable (or r.e.) languages.\(^1\)

The importance of the Chomsky Hierarchy is that it establishes a certain metric of complexity for languages in terms of the complexity of the mechanisms for their recognition and generation, which one might also expect to correlate with the learning difficulty. The current consensus is that at least some human languages are mildly-context sensitive in the sense of Joshi (1985), a class of languages intermediate in complexity between context-free and context-sensitive languages (Joshi, 1985)\(^2\) or, perhaps, even more complex as recently suggested based on certain kinds of copying dependencies in syntax (Stabler, 2004; Kobele,

\(^1\)For more details on the formal language theory and the Chomsky hierarchy the reader may consult Partee et al. (1990); Hopcroft et al. (2001).

\(^2\)More specifically, these are context-sensitive languages that are efficiently parsable, have limited crossing dependencies, and have constant growth (the growth of string length is linear).
However, one of the early controversial results in learnability theory due to Gold (1967) is that even relatively simple classes within the Chomsky hierarchy, such as the class of regular languages, are not learnable (or feasibly learnable, that is, learnable in a realistic amount of time) in many frameworks. For instance, regular languages are not learnable in Gold’s framework “identification in the limit from full positive texts.” This framework allows the learner an unlimited amount of time and computational resources, but requires exact convergence (that is, learning is defined as reaching a point after which no errors are ever made.) Subsequent work made clear that it is not the stringent exact convergence requirement that makes learning large classes of languages difficult. Modifying the above framework so that success is evaluated in a probabilistic fashion does not change any of its negative results (Pitt, 1985; Wiehagen et al., 1984). Similarly, none of the classes in the Chomsky hierarchy are learnable in a probabilistic framework known as Probably Approximately Correct learning (PAC) which is rooted in the principles of statistical learning with older history (Valiant, 1984; Kearns and Vazirani, 1994).

The negative results mentioned above do not indicate that the mathematical study of learnability is useless and not revealing, as is sometimes suggested. These results should be considered together with a large number of positive results showing that specific classes of languages with well-defined structure are learnable in various frameworks. Moreover, the whole class of r.e. languages becomes learnable (but not efficiently so) in certain settings that either enrich the learner’s input or restrict the types of data on which the learner is required to succeed (Gold, 1967; Horning, 1969). What these results tell us is that while learning unrestricted classes of languages is difficult or impossible, restricting assumptions about the nature of the input or about the hypotheses available to the learner dramatically improve the learning outcome. Non-trivial (i.e., generalizing) learners have been proposed for patterns similar to those found in human language (Yokomori, 2003; Heinz, 2010; Clark, 2010). These learners are defined for restricted classes of languages that do not correspond to any specific class in the Chomsky hierarchy, supporting the growing consensus that the class of human languages (i.e., languages that can be learned by humans) corresponds to some cross-section of the Chomsky hierarchy. The exact nature of restrictions on this class is still a matter of discovery – research that will surely benefit from a joint effort on the part of linguists and the computational learning community. What the latter community contributes to this research are precise formulations about what guarantees learnability in specific settings and insights about what structural properties of the learning problem drive generalization. What linguists and psychologists can bring to the table are the empirical facts that help better identify structural properties forming the core of human languages (properties that learners can exploit), as well as facts that help make learning proposals more psychologically plausible.

2.2 What do computational learning models tell us about language

Many of the computational learning models should not be taken as modeling exactly what is going on in the brain (although some models, such as neural networks, are claiming to do exactly that). I would like to stress that this fact does not make such models uninteresting or irrelevant for linguists and psycholinguists. Computational learning models help us gain insight about what makes learning possible and answer a variety of important questions starting with really general ones, such as “what classes of languages are in principle learnable by any learning algorithm” to more specific ones such as “what effect a particular bias (say,
an expectation that no two words can have the same meaning) can have on learning.”

Additionally, restrictive learning models can help us explain and hone in on the set of properties that languages share in common (both categorical and “soft” statistical universals) by highlighting which properties of grammars or languages aid in the learning process. On the other hand, some very powerful unrestricted learning models capable of learning the entire class of r.e. languages, are not very insightful in the sense that they do not explain why the class of natural languages is vastly smaller than the class of r.e. languages. Thus, the traditional connection between linguistic universals and learning biases plays a central role in model design and selection – we assume that some of the restrictions in the data are there because they make learning easier.

For example, in statistical learning literature, simplicity considerations (discussed further in section 5.1) serve as a learning bias roughly corresponding to the idea that simpler hypotheses are preferred to more complex ones. In these models “restrictiveness” is discussed with respect to attempts to achieve a balance between simplicity and goodness of fit (the ability to account for the data as tightly as possible). The more restricted and simple the language class, the more likely it is that a simple hypothesis capturing the data also fits the data well. On the other hand, less restricted, complex classes are hard to capture with simple models without running the error of underfitting (or overgeneralizing, which is one way of giving up global goodness of fit); and vice-versa: choosing a model that fits well all the sample data leads to proposing highly complex hypotheses that run a risk of overfitting (or undergeneralizing, which is another way of giving up global goodness of fit). Thus, the best scenario for a statistical model is the one in which the language class to be learned is sufficiently restricted along the lines that accord with the model’s notion of simplicity, and, as a result, are describable with high accuracy by a simple grammar. In other learning models, simplicity biases and other linguistic restrictions can be directly build into the assumptions about the hypothesis space or the learning algorithm.

In the rest of this section I briefly address a common concern that the assumed relationship between language universals and learning biases is faulty. There are different ways in which the connection between learning biases and typological preferences can be fleshed out. One possibility (in rough terms) is that in the process of language change, which often involves co-existence of several alternatives (Weinreich et al., 1968), the alternatives for which the learner is biased are learned more faithfully and by a larger number of new learners because they require less effort. As a result these alternatives are more stable and eventually drive out the other competitors. This does not lead to complete obliteration or simplification of all irregularities, as new complications are introduced all the time by processes which may be “blind” to certain components of grammar.

One might object, however, that linguistic universals (whether categorical or statistical) might be due to non-learning factors such as historical accidents, limitations on the mode of language transmission, and functional constraints. There are several responses to this objection. The commonly suggested alternatives to learning biases in the domains of morphology and syntax (e.g., communicative and processing constraints) are only alternatives to domain-specific biases, but not necessarily to domain-general biases. Many communicative pressures can be restated as learning pressures because what hinders communication also hinders learning which relies on communication for transmission of data to the learner. Anderson (2008) also observes that it is possible for some preferences that had their origins in a functional explanation to become encoded as innate preferences in the architecture of the learning mechanism through the so-called evolutionary “Baldwin effect,” the propensity of a learned trait to become innate over a long period of time. Finally, on the empirical side,
experiments on artificial grammar learning (e.g., Moreton, 2008; Finley and Badecker, 2008; Clair et al., 2009) and on evolutionary simulations of language change (Brighton, 2002; Smith et al., 2003; Kirby et al., 2008) support the role of learning biases in emergence of linguistic universals (including biases grounded in communicative factors). Based on these facts, I conclude that it is highly plausible that some (and perhaps many) language universals are due to learning biases. Therefore, an insightful and explanatory learning model will be the one that is appropriately restrictive by incorporating such biases. To my mind, the difficulty of securely establishing the source of linguistic universals only underscores the importance of formal learning models, especially when they are embedded into broader computational models of language change. Such models can help us quantitatively test the load of different factors in explaining language universals since different factors in the model (e.g., the learning mechanism vs. the population dynamics) may contribute differently to a number of learning-related predictions, such as predictions for errors during language acquisition, rates and directions of language change, side-effects on other aspects of language structure, language shifts in bilingual populations, and so on. See Niyogi (2006) for an exploration of this topic in the context of an evolutionary model that includes factors of communicative efficiency, population dynamics, and learning algorithms inspired by linguistic proposals.

To sum up, one of the central imports of modeling language learning is similar to the major goal of linguistic research, that is, the goal of identifying restrictions on linguistic patterns and providing explanations for the source of linguistic universals.

2.3 Finer-grain predictions of learning models

The main criteria for evaluating a learning model is whether it correctly learns a desired set of data. But for models that strive for psychological plausibility there are additional criteria that have to do with how closely the model encodes the learning situation of a human child and fits children’s learning trajectory. This includes assumptions about what computational resources the model relies on, whether it has access to negative evidence, whether it can handle noise, what kinds of errors it makes, and so on. To the extent that such assumptions lead to concrete predictions about the speed of learning and the error patterns, they can be tested against acquisition data, which provides finer-grain evidence for learning models than mere facts about grammaticality or acceptability. Below I review just a few examples of the kinds of considerations that arise in evaluating models with respect to their psychological plausibility.

The first consideration concerns memory resources. Most learning proposals implement some version of memory, a repository of the data points encountered by a learner. Some models have access to every single data-point in addition to computing compressed representations of the data, some models are completely memoryless, and others are in-between these two extremes allowing for various levels of memory. Assumptions about memory resources and how they are used have consequences for the predicted pattern of errors during language acquisition. These predictions along with the psycholinguistic findings about how memory is used in language learning can be used to decide on the most realistic type of model. For example, several psycholinguistic studies based on frequency effects suggest that even regular words of relatively low frequency are stored and accessed by language users (Alegre and Gordon, 1999; Baayen et al., 2003). Such studies are taken to provide support for the memory-based models.

However, memory-based models place a heavy burden on storage and computational resources, which is why some modelers favor memoryless learners instead. A memoryless
learner is an incremental function that at each step maps a current data-point and a previous hypothesis to a new hypothesis. Several syntactic triggering learners (inspired by the Principles and Parameters framework) are of this type – after seeing a new data-point they may change some parameter setting which then produces a new grammar (Gibson and Wexler, 1994; Yang, 2002). These models provide greater storage economy which is particularly important for learning infinite or very large classes of patterns. The absence of memory, however, makes learning many problems significantly more difficult. For instance, Fodor and Sakas (2005) discuss how memoryless strategies have certain difficulties with learning languages that are in a subset relationship. Certain kinds of memoryless learners also have a problem with local maxima (discussed shortly).

In morphology, there are proposals for dual-route learning models that implement storage for the irregular subsets of the data and compute compressed descriptions (e.g., rules) for the rest of the data (Pinker, 1999). Research by Michael Ullman and his colleagues suggests that these two types of “routes” correspond to two different types of memory humans are known to have: the declarative memory (storage of facts) and procedural memory (storage of procedures, rules) (Ullman, 2001). But experimental data from different labs is not consistent: some support the dual-route model while others do not. Thus, how much memorization and how much generalization humans perform when learning morphology remains an open empirical question.

Another important way of evaluating psychological plausibility of a learning algorithm is to look at the general trends for error patterns produced in the process of learning. One such trend is conservatism. Gibson and Wexler (ibid.) loosely define conservatism as a property learners have if they “make no large-scale reorganizations of their grammars in a single learning step” (Gibson and Wexler, 1994). They propose that conservatism is a desirable property in a psychologically plausible learning model because children are thought to learn in a very gradual piecemeal fashion.\(^3\) A widely used conservative learning method is the so-called “hill-climbing” method. This method relies on an incremental memoryless optimization strategy that proceeds by making small adjustments to a current hypothesis towards a more “optimal” solution, namely the one that better explains the current data-point (the Gradual Learning Algorithm Boersma (1997) for learning stochastic Optimality Theory is one example). However, although such learners meet the definition of conservatism above, they face a notorious problem if the hypothesis space contains local maxima.

A local maximum is a hypothesis that cannot be improved upon by the hill-climbing algorithm given any possible input, and yet it is not the most optimal global solution to the learning problem. This scenario is often illustrated through an intuitive example of actual hill-climbing. Suppose that to get to a top of a hill a climber only takes steps in the direction that will land him at a higher elevation than his current position. This strategy will guarantee that eventually the climber will reach the top of a hill, but not necessarily the highest hill since once he reaches the top of any hill, he will stop (because a step in any direction will be a step down). Therefore, hill-climbing methods can only be applied if the set of possible hypotheses for the target language class does not contain local maxima. Do grammars of human languages have local maxima? This, of course, depends on what these grammars are like. For instance, the exploration of some parametric

\(^3\)It is important to distinguish between two different kinds of conservatism: conservatism with respect to the grammar, and extensional conservatism which is conservatism with respect to the languages generated by the grammar. Children may display large jumps in their language production, suddenly producing more different types of expressions, but as long as these jumps are due to a small change in the grammar, they are still conservative in the first sense.
hypotheses spaces (in which a grammar corresponds to a set of parameters as in theories of Government and Binding (GB) or Head-Driven Phrase Structure Grammar (HPSG)) reveals that local maxima are problematic (Niyogi and Berwick, 1996). Thus, depending on the theory of grammar we adopt, a conservative hill-climbing method may not be the best learning strategy.\footnote{However, there are various methods for augmenting hill-climbing with additional assumptions (e.g., memory, restarts, stochastic modifications) which alleviate the local maxima problem.}

There are other error trends we may consider to be important for a learner to replicate. For instance, we may require that a learner be consistent. A consistent learner is a learner whose hypotheses account for all the data seen so far. A consistent learner is not necessarily conservative in the sense above, but it guarantees that the hypotheses it produces at each step agree on all the received data. There are good reasons to think that human learners are generally consistent, that is, they don’t gratuitously dismiss the data that they have seen in the past. Another example of a learning property associated with a particular error pattern is monotonicity. A monotonic learner is a learner whose current hypothesis is always a superset of the previous hypotheses. That is, a monotonic learner is a learner that never “takes back” its predictions. Such learners are mathematically nice because they make it easy to reason and to prove facts about learner’s behavior. However, it is likely that human learners are not monotonic because children occasionally overgeneralize and later correct their overgeneralization errors (Ervin, 1964; Marcus et al., 1992). This implies the existence of a mechanism that allows the learner to retreat from its previous hypothesis, producing a non-monotonic pattern of generalizations. The linguistic descriptions involving “default” and “elsewhere” rules are examples of non-monotonic reasoning systems that could be thought of as arising from a non-monotonic learning strategy. If human learners are indeed non-monotonic, this would already rule out many learning algorithms from our consideration.

Overall, this section shows that modeling is useful not only for investigating consequences of learning biases but also consequences of our assumptions about the learning situation.

3 Inflection as a formal system

Turning from general issues in machine learning, we now consider the learning of inflectional morphology in particular. What kinds of learning techniques and learning biases are appropriate for inflection? Before these questions can be answered, it is helpful to first clarify what is learned when learning inflection, and how complex this task is.

3.1 Expressive power of inflection

There is no consensus on the exact criteria separating inflection from other types of morphological marking such as derivation and cliticization (Stump, 1998), but we may roughly define the inflectional component of a language as a set of strings resulting from taking a union of wordforms of every lexeme in the language. Wordforms of the same lexeme do not differ in their core semantic meaning (i.e., the concept expressed by the word) and in their lexical category, but differ only in morpho-syntactic properties (whatever our current understanding of such properties may be). Additionally, a single wordform does not typically contain multiple instances of the same inflectional morpheme.\footnote{One possible exception to this claim is the apparent recursion known as “Suffixaufnahme” when the possessive NP carries the case and number inflection of the possessor (Plank, 1995).} To evaluate the complexity
of inflectional patterns, we can remain agnostic about whether words are derived in syntax or in a distinctly different morphological component of grammar.

Let us now consider this question: what is the simplest grammar that accounts for the set of inflectional forms of a language in an insightful way? And, relatedly, what types of learning algorithms are particularly fitted for learning this grammar? In the simplest scenario, the set of inflectional wordforms is finite (supposing finite vocabulary for the moment). The simplest kind of representation for a finite set is an unstructured list, and the simplest way to learn such a list is memorization. However, inflection cannot be plausibly learned by memorization alone. In subsection 3.1.2, I consider proposals that inflection can be fully handled by relatively simple computational devices, namely finite-state automata (which generate regular languages) and which are widely used in computational models of morphology (Sproat, 1992; Beesley and Karttunen, 2003, and others). The upshot of this section is that simple learners that can handle subsets of regular patterns are sufficient for handling inflection.

3.1.1 Against lists and memorization

Since the number of inflectional exponents in any language is finite, and since inflection is not recursive, one could in principle list all the word-forms in a language (assuming, for the sake of an argument, that speakers know only a finite number of words). Such a list could be derived in the process of simple memorization. However, most morphologists would agree that memorization alone (without generalization) is unrealistic as a model for how humans learn inflection. There are several reasons commonly cited for thinking this is the case. First, we know that speakers generalize as is evinced by children’s errors, wug-tests, and loan-word adaptations. Additionally, we know that sets of wordforms, though finite, can be extremely large in richly inflected, especially agglutinative, languages (Hankamer, 1989; Kurimo et al., 2006).

Moreover, the distribution of data within the speakers’ experience is sparse. That is, speakers mostly experience the same patterns repeated over and over again, while majority of patterns are rare. This is known as the Zipfian distribution, which was originally established for word frequencies (Zipf, 1932), but has since been found to hold for many other linguistic phenomena including inflectional morphemes (Chan, 2008). For example, as table 1 from Chan shows, many wordforms in richly inflected languages are exceedingly rare as can be seen from the fact that available corpus data does not include the full verbal paradigm for any lemma. The last column, saturation, shows the maximum percentage of verb forms attested in the corpus for a single verb lemma. For all languages in the sample, except English (which only has 6 maximally distinct verb forms), saturation is below 100%; and for some languages it is only around 50%.

Zipfian distribution renders the memorizing learner implausible: if most expressions are rare, one has to wait a very long time until all of them are heard. There are other, perhaps more interesting, asymmetries in the data making certain types of patterns more frequent than others. Just to mention a few examples: free-variation of morphemes appears to be rare (Kroch, 1994); syncretism is more common in the marked portions of a paradigm (Haspelmath, 2002); when allomorphy is conditioned by other material in the word, this material will almost always be closer to the root (Carstairs-McCarthy, 1987). Supposing that some of these facts are results of learning biases, a learner encoding these biases will converge on the right grammar much earlier than a memorizing learner. For instance, a learning algorithm that is biased against synonymy (e.g., by adopting the Principle of Contrast of
Table 1: Sparseness of paradigms in corpora (from Chan, p.79)

<table>
<thead>
<tr>
<th>corpus</th>
<th>millions of tokens</th>
<th># of total forms in corpus</th>
<th>max. # of forms for any lemma</th>
<th>% saturation</th>
</tr>
</thead>
<tbody>
<tr>
<td>English(Brown)</td>
<td>1.2</td>
<td>6</td>
<td>6</td>
<td>100</td>
</tr>
<tr>
<td>English(WSJ)</td>
<td>1.3</td>
<td>6</td>
<td>6</td>
<td>100</td>
</tr>
<tr>
<td>Basque</td>
<td>0.6</td>
<td>22</td>
<td>16</td>
<td>72.2</td>
</tr>
<tr>
<td>Czech</td>
<td>2.0</td>
<td>72</td>
<td>41</td>
<td>56.9</td>
</tr>
<tr>
<td>Finnish</td>
<td>2.1</td>
<td>365</td>
<td>147</td>
<td>40.3</td>
</tr>
<tr>
<td>Greek</td>
<td>2.8</td>
<td>83</td>
<td>45</td>
<td>53.2</td>
</tr>
<tr>
<td>Hungarian</td>
<td>1.2</td>
<td>76</td>
<td>48</td>
<td>63.2</td>
</tr>
<tr>
<td>Hebrew</td>
<td>2.5</td>
<td>33</td>
<td>23</td>
<td>69.7</td>
</tr>
<tr>
<td>Slovene</td>
<td>2.4</td>
<td>32</td>
<td>24</td>
<td>75.0</td>
</tr>
<tr>
<td>Spanish</td>
<td>2.6</td>
<td>51</td>
<td>34</td>
<td>66.7</td>
</tr>
<tr>
<td>Swedish</td>
<td>1.0</td>
<td>21</td>
<td>14</td>
<td>66.7</td>
</tr>
<tr>
<td>Catalan</td>
<td>1.7</td>
<td>45</td>
<td>33</td>
<td>73.3</td>
</tr>
<tr>
<td>Italian</td>
<td>1.4</td>
<td>55</td>
<td>47</td>
<td>85.5</td>
</tr>
<tr>
<td>CHILDES Spanish</td>
<td>1.4</td>
<td>55</td>
<td>46</td>
<td>83.6</td>
</tr>
<tr>
<td>CHILDES Catalan</td>
<td>0.3</td>
<td>39</td>
<td>27</td>
<td>69.2</td>
</tr>
<tr>
<td>CHILDES Italian</td>
<td>0.3</td>
<td>49</td>
<td>31</td>
<td>63.3</td>
</tr>
</tbody>
</table>

Clark (1987)), will be more efficient than a learner without such a bias since, when deciding on a meaning of a new morpheme, it will not have to entertain meanings already assigned to other morphemes (see a simulation in Taguchi et al. (2006)). Generally, whenever a finite set is structured (even if only in a probabilistic sense so that some patterns are more frequent than others), this gives a generalizing learner an advantage over a memorizing learner which can just as easily memorize unstructured data as it does structured data.

3.1.2 The upper bound on inflectional complexity

As I have already mentioned, the set of human languages is currently believed to be some subset of mildly context-sensitive languages. However, some aspects of human languages are certainly more restricted than that. Most relevantly, there are reasons to think that inflectional patterns are simpler than phrasal syntax.

First, inflection does not allow recursion. As Haspelmath (2002) points out, nothing logically rules out recursion in inflection as one can imagine hypothetical forms like cat-s-es meaning ‘sets of cats.’ But such iterated inflections are virtually unattested. Second, unlike syntactic dependencies, most word-internal inflectional dependencies tend to be local, that is, conditioned by linearly adjacent material or by structurally adjacent material with limited amount of non-local dependencies (more on this in section 5.2). Third, the order of inflectional morphemes within a word is often fixed – many languages have the so-called templatic structure where the same morphosyntactic features are realized in the same order. Fourth, languages don’t show true scrambling of inflectional affixes, although occasional restricted patterns of free variation in affix order are possible. Fifth, the semantics of inflection is relatively primitive, corresponding mostly to taking (multi)unions of inflectional features in contrast to a variety of complex semantic functions at the sentence level. And finally, at the level of the mapping between form and meaning, inflectional elements realized in the same position typically appear in complementary distribution: for instance, a nomi-
tive case marker cannot appear where dative is expected and this choice is not up to the speaker. On the other hand, “open-class” lexical items are less restricted in their distribution: languages are full of words with overlapping meanings and meanings at various levels of specificity. The speakers can choose among the lexical concepts at will (e.g. in exactly the same situation one can choose to refer to an instance of a rose as “a rose,” “a flower,” “a plant,” or “an object”).

It is because inflection, and morphology more generally, is more restricted than phrasal syntax, that finite-state computational models of morphological grammars and morphological learning, whose expressive power is relatively limited, have been so successful (Koskenniemi, 1983; Theton and Cloete, 1997; Beesley and Karttunen, 2003). These models represent the data as a finite-state automaton (FSA), a mathematical abstraction used to generate or recognize sets of expressions that require only finite memory and that correspond to the regular languages in the Chomsky hierarchy.

As exemplified in figure 1, an FSA is standardly drawn as a directed graph in which memory states (represented by circles) are connected to each other by labeled arrows. One can traverse an automaton graph by following the arrows from an initial to a final state, outputting or reading the labels on arrows along the way. Figure 1 is an example of a finite state transducer (an FSA in which symbols correspond to input-output pairs) for a small set of Russian wordforms, the nominative possessive pronouns.

FSAs can handle infinite languages (since they allow loops), but only of limited complexity – namely, those in which identity of each symbol depends on finitely many preceding symbols. It is because inflectional patterns do not require infinite or arbitrary amounts of memory (with a possible exception of reduplication) that FSAs are expressive enough for describing inflection. This contrasts with certain syntactic patterns (e.g., center-embedding) that require an arbitrary amount of memory.

Finite state methods are not limited to concatenative morphology. Computational linguists, relying on proposals of theoretical linguists such as McCarthy (1981), have developed finite-state models that can handle non-linear templatic morphology of Semitic languages. For instance, Kiraz (2000), building on a proposal by Kay (1987), presents a general finite-state framework for computing templatic patterns using multi-tape finite automata implementing the idea of multiple tiers in autosegmental phonology. Thus, morphology of Semitic languages can be learned if a learning model is built so as to expect templatic patterns and has some way of determining when a language is using concatenative or templatic method of morpheme combination.

One possible challenge for finite-state models of morphology is reduplication since it is best implemented as a copying operation. Unbounded copying, however, is beyond the power of finite-state/regular grammars because it requires keeping track of potentially unlimited number of crossing dependencies (as shown below).

Whole-word reduplication is particularly problematic because the model has to work for arbitrarily long words. (Of course, we can build an FSA that allows reduplication of words up to some sufficiently large number of segments $n$, but such an automaton would be extremely complex and redundant.) Interestingly, whole-word reduplication intuitively seems to be the least problematic type of reduplication for human learners since the constraint on the size of the reduplicant is very simple. This suggests that either representations of higher complexity (i.e., those that can describe context-sensitive patterns) are involved in learning some aspects of morphology, or that reduplication is best handled not by copying of phonological segments, but by doubling of a single substring within a word.

To sum up, while we can in principle account for inflection by listing all inflectional forms,
The initial state is an ellipse, the final state - a double circle. This FST captures the fact that 3rd person possessive pronouns behave differently from other pronouns by not taking gender agreement suffixes.

the more elegant, succinct grammars of inflection capturing regularities in the data are more complex, but likely no more complex than regular (or finite-state) grammars. Since finite-state languages are relatively simple and well-studied, there are a number of efficient and robust methods for learning subclasses of such languages (for examples of learning proposals in morphology that rely on finite-state representations see Kiraz (2000); Clark (2001)).

3.1.3 Modularity within language

Lurking behind the above discussion is the issue of modularity. The issue is whether the human acquisition device is “modularized” so that more restricted learning procedures are used to learn more restricted aspects of grammar. If inflection can be described with less expressive machinery than syntax, could it be that it is also learned by simpler, more restricted learners?

Remember that we have not committed ourselves to the view that inflection in fact corresponds to a separate grammar module. It could very well be that it does not, in which case we can assume that the same learning algorithm that is used to learn more complex syntactic patterns is also used to learn inflection (or both inflection and derivation). This possibility is parsimonious in the sense that it reduces the number of different learning
strategies one has to assume.

On the other hand, modularity can be argued to be beneficial for learners: namely, it could be more advantageous to have different learning algorithms for aspects of grammar that differ in complexity, just as it is advantageous for a golfer to carry different golf clubs, each tailored for a specific situation. An advantage of a modularized learner over an all-purpose learner is that such a learner can learn certain simpler aspects of grammar before the more complex ones, something that can simplify learning at higher levels of structure. This advantage exists because learning is more efficient within more constrained hypothesis spaces as was discussed in the first half of this chapter.

4 Challenges for inflectional learning models: ambiguity and irregularity

Although I suggested that learning the set of inflectional wordforms of a language may be relatively simple, it is still far from trivial. One of the main obstacles for this task is ambiguity. Ambiguity involves a non one-to-one mapping or a mismatch between the two components. Inflectional mismatches exist at various levels of representation: at the level of the mapping between phonological form and grammatical function (e.g., null morphemes, syncretism, allomorphy), at the level of mapping between word-positions and features typically realized in these positions (e.g., portmanteau morphemes, variable affix ordering), at the level of mapping between morphological form and syntactic/semantic function (e.g., depenency, heteroclisis), and at the level of mapping morphological constituents to syntactic or semantic constituents (the so-called “bracketing paradoxes”).

Why are mismatches problematic for learning? In short, because they enlarge the learning space (the possible hypotheses the learner has to entertain) and because in some cases these hypotheses become hard to tell apart from each other. At one extreme, mismatches may make learning extremely difficult or impossible. For instance, imagine learning morpheme meanings in a language in which every morpheme is homonymous or has a million synonyms. On the other hand, when there is no ambiguity at all, learning something about one component of the grammar automatically tells you something about the other “matched” component, and this fact can be exploited by relatively simple generalization techniques. In other words, one-to-one mappings allow the learner to “bootstrap” when learning interlocking components.

Perhaps the most common and widely discussed inflectional mismatch is allomorphy, the phenomenon in which the same morpho-semantic features are realized differently. Particularly problematic is global allomorphy, which leads to seemingly arbitrary inflectional classes, that is, classes whose membership is not fully predictable from any of the mem-
bers’ properties. The allomorphy in the English past tense exemplifies largely arbitrary lexical classes corresponding to different realizations of the feature [past]. The learning of English past tenses is probably the most studied question in machine learning of inflection, a topic that sparked the debate between the connectionist vs. symbolic approaches to learning, as well as the debate between the dual route (symbolic rules for regulars + associative memorization for irregulars) vs. single route (associative learning or rule learning for both regulars and irregulars) mechanisms for learning and storing morphological patterns (Rumelhart and McClelland, 1996; Pinker and Prince, 1988; Ling and Marinov, 1993, and others). The prominent feature of lexical allomorphy is that, while it is not fully predictable, many allomorph classes display patterns of semi-regularity. That is, the lexical items that select the same allomorphs may have a Wittgensteinian “family resemblance” structure with overlapping and crossing patterns of similarity without sharing a single set of necessary and sufficient features. An artificial language learning study reported in Brooks et al. (1993) found that fully arbitrary gender-like inflectional classes were not learned by adults and children, unlike classes in which a subset of nouns shared a set of phonological features. These facts suggest that there are statistical restrictions on the amount and arbitrariness of lexical allomorphy.

The same can be said about another prominent example of ambiguity - syncretism. Syncretism refers to a situation in which different morpho-syntactic features are realized by the same inflectional exponent. The simplest examples of syncretism are structured and can be described with feature underspecification, but the more complicated cases exhibit the same semi-regularity as allomorphy: the cells that are syncretic do not share a set of necessary and sufficient properties, yet most of the time they also do not comprise a completely random set. For instance, most patterns of syncretism that cannot be captured by underspecification alone, can be captured by underspecification and blocking (Pertsova, 2011). There are also cross-paradigmatic regularities involving syncretism: in many cases, the same cells tend to be syncretic in different paradigms within the same language revealing that something systematic is going on.

These two examples of mismatches reveal a general pattern: while ambiguity is pervasive, it is still restricted. In a finite domain with few categorical universals, such restrictions become the defining properties of the system. Therefore, the main goal of modeling inflectional learning is to predict restrictions on ambiguity (and irregularity) and explain their underlying causes.

5 Learning biases in acquisition of inflection

In this section I discuss some learning biases that are used in many models of inflectional learning. These biases reflect the researchers’ hypotheses about what regularities in the data can be exploited by learners. We will see that some such biases allow learners to successfully generalize in the presence of ambiguity and at the same time explain why ambiguity exists.

5.1 Simplicity

The first bias I discuss is based on a general idea of simplicity or economy implementing the principle of Occam’s razor, familiar throughout science. Most learning models, as well

6Syncretism is often assumed to arise for principled reasons, in the words of (Carstairs-McCarthy, 1987) “in the sense of being more than mere accidental by-products of phonological processes or morphological ‘spell-out’ rules”. Namely, forms that are realized syncretically are also expected to have related function.
as theories of inflection, invoke some version of simplicity or economy constraints in their architecture. In this section, I first introduce how simplicity may be understood and implemented, and then discuss how this notion is used in models for learning inflection. I will also discuss how trade-offs in simplicity at different levels of grammar may introduce the sort of ambiguity we must account for.

It is often remarked that simplicity is a subjective and vague concept. Correspondingly, to say that the model is driven by simplicity is not to say much - one needs to specify how this notion is to be understood. Statistical learning models get certain simplicity biases automatically from the kinds of methods they presuppose. For example, models that rely on the Minimum Description Length (MDL) principle (Rissanen, 1978) are based on the idea that simplicity can be measured as the size of the hypothesis and the size of the data described with this hypothesis (within a fixed description language). Shorter hypotheses are preferred matching the fact that sets with regularities can be compressed and hence have shorter descriptions. Other statistical models are based on the idea of selecting not the shortest but the most probable hypothesis. However, in many cases this amounts to the same thing since there is a close correspondence between the length of a hypothesis and its probability (for instance, see Vitányi and Li (2000) for conditions under which Bayesianism and MDL converge). While we can’t really know whether these are the right measures of simplicity, the success of statistical models that maximize probability of the data or minimize the hypothesis length on some linguistic tasks, such as morpheme and word segmentation, supports the relevance of the simplicity measures they adopt.

In non-statistical models the simplicity bias has to be explicitly built in. Some researchers prefer building-in simplicity bias “from scratch” as it allows them greater control in interpreting and implementing this bias given the particulars of the data.

As one specific example of how a simplicity bias may aid in learning inflection, consider a Bayesian classification model proposed in Frank et al. (2008). The model in question is a general categorization model that the authors apply to the learning of artificial languages in the lab and to the learning of inflectional rules like the English past tense. The rules in this model correspond to conjunctions of properties specifying the restrictions on a pattern (e.g., a rule like “add -ed to the end of a word” would correspond to a conjunction [−ed and position:word-final]). The goal of the model is to find the best set of rules that explains the data (call it a ‘grammar’) given the set of all potential rules. The set of potential rules is defined as a set of all apriori possible conjunctions of different phoneme sequences and word-positions. A grammar then is a partition of lexical items into the categories defined by a set of rules. A lexical item “fits” a category or a rule if it contains all the properties specified by that rule. For example, the word “smelled” fits the rule [−ed and position:word-final]. (Notice that a single item can fit several different rules, so that different partitions are possible.) The choice of the “best” grammar is determined by Bayesian inference which involves calculating probabilities of all rival grammars\footnote{In practice, because the hypothesis space is so huge, approximation methods have to be used for these calculations.} and choosing a grammar that maximizes the likelihood of the data according to Bayes’ rule. The Bayes rule applied to grammars (H) and data (D) is given below. It states that the probability of a grammar for some set of data is equal to the probability of the data being generated by this grammar times the (prior) probability of the grammar divided by the cumulative probability of the data given any possible grammar.
Frank et al. (2008)'s model implements a simplicity bias by preferring fewer number of rules (this preference is regulated by a parameter in the model that can be adjusted to make the preference weaker or stronger). An additional bias that this model assumes is a preference for more specific rules which give a tighter fit to the data. (The authors refer to this bias as a principle of “minimum generalization”.) In other words, this model implements a typical tension between simplicity and goodness of fit. The authors acknowledge that their model has a serious limitation due to their choice of whole segments (phonemes) as features. If they were to adopt phonological features instead, their model would also be able to capture rules that refer to natural classes of segments (e.g., voiceless obstruent). Nevertheless, this model was able to predict the past tenses of 88.5% of wordforms in the testing phase.8

Why was this model able to do so well even without using a realistic set of features? Because many verbs that fall into the same irregular class in English are phonemically similar. So, when the model is tested on a verb it hasn’t seen before, chances are this verb is similar enough to an already discovered category of verbs and so, fits a particular learned rule. If the model did not prefer hypotheses with smaller number of rules, a word-specific rule for each word would be adopted as the best rule. In other words, there would be no generalization. Thus, we see how a simplicity bias (preferring less rules or shorter grammars) drives generalization by taking advantage of structure in the data (phonological similarities among verbs), even if this structure is best captured by statistical tendencies rather than absolute constraints.

In contrast to the above model, several non-statistical rule-based models achieve similar results without striving for representational economy (preference for shorter grammars). For instance, the morphological learners described in Albright and Hayes (2002) and in Molnar (2001) both define a generalization procedure that works by iteratively merging properties shared by words undergoing the same structural change. This type of learner first stores word-specific rules (e.g., “to get a past tense of poke add -t to the end of the stem”), and then gradually collapses contexts of similar rules. For instance, after a learner has seen the pairs poke – poked and smoke – smoked and formulated word-specific rules for these pairs, it notices the fact that the contexts of the specific rules share a common substring and proposes a more general rule, such as “to get a past tense of a stem that ends in [ok] add -t to the end of the stem”. This method leads to the discovery of rules, but it does not entail representational economy. For instance, Albright and Hayes assume that the learner stores all the rules discovered in the learning process, including word-specific rules. This is clearly not the most economical option, but their model ends up closely matching human performance on a past tense wug test.

Given that some learners can generalize successfully without minimizing storage space, one might wonder to what extent the storage economy bias is active in human learning, and in learning of inflection in particular. Perhaps it is not representation size that humans try to minimize, or not only representation size. It is well known that what may be simple or economical for one component of language may turn out to be complex for some other component. Homonymy presents a simple example: on the one hand, homonymy makes processing less economical since it requires extra time and resources to resolve the

---

8The predictions for novel items presented during testing were generated by taking the most likely a-posteriori inflected form given the most likely hypothesis chosen by the model.
ambiguity in the form-meaning mapping (e.g., homonymy presents a disadvantage in lexical categorization tasks); on the other hand, homonymy is economical since it minimizes the number of distinct phonological forms in the lexicon. Thus, the question is: what type of simplicity or economy should the learning procedure try to maximize?

Some psycholinguists emphasize the trade-offs between minimizing storage vs. minimizing processing economy, arguing that the best model of human behavior is the one that achieves a balance between these two competing aims (this view is advocated in the dual-route processing models of Caramazza et al. (1988); Baayen and Schreuder (1968)). The conflicting pressures arising from the effects of economy at different levels of grammar may explain why certain types of ambiguity exist in the first place: ambiguity may arise because it is beneficial for some component of language, but at the same time it will be curtailed because it creates a burden for a different component. This kind of economy tug-of-war explanation of statistical tendencies in languages has been proposed for several phenomena ranging from Zipfian distributions of word frequencies to patterns of language change (Zipf, 1949; Kirby, 2001; Haspelmath, 2008). It could be that statistical regularities in morphology are particularly likely to be a product of conflicting economy biases given that morphology serves as an interface connecting several components of grammar, and given that it involves lexical storage introducing additional considerations of lexical access.

Yang (2005) is an example of a proposal for how to implement a trade-off between processing and representational economy. Yang presupposes a rule-based morphological system and a serial search processing model in which words and rules are accessed in the order of their frequency. He then proposes a particular metric for the cost of exceptions to rules based on minimizing processing time. The basic idea is that a rule that has exceptions is costly because an application of such a rule requires first “scanning” all of its exceptions to make sure they don’t apply in a given case. Yang suggests that this processing-based cost metric can be used to determine whether a partially regular pattern is most efficiently stored as a rule with a number of exceptions (which themselves could correspond to more narrow rules) or simply as a list of words. Notice that there is a trade-off between the number and frequency of exceptions of a rule and the number of items it explains. A pattern that covers a lot of data with few exceptions will be stored as a rule leading to compression of representation size. On the other hand, an irregular pattern with many exceptions may be better stored as a list of words, especially if these words are frequent and hence easily accessed, leading to savings in processing time. Similar ideas related to defining processing costs can be found in earlier proposals such as Jackendoff (1975).

In short, the notion of simplicity is very prominent in computational models of inflection and in learning models in general. Interestingly, the fact that simplicity considerations are active in both production/generation and perception/processing can potentially explain why patterns of ambiguity and irregularity arise in languages and also why they remain restricted. However, it is far from settled what the most helpful measure of simplicity is, and what is the most appropriate way to specify trade-offs between different types of simplicity. Learning models provide a rigorous way of exploring and testing these issues.

5.2 Locality

Another prominent bias in morphological and morpho-phonological learning models regards the notion of locality, the idea that dependencies between morphemes tend to be in some sense adjacent.

The locality bias is well established in phonology as an empirical constraint on phono-
tactics and phonological alternations (Jensen, 1974; Gafos, 1996). Physiological and articulatory factors can explain many instances of locality patterns in phonology (i.e., those patterns that are due to co-articulation of adjacent segments or segments of the same kind). So we may wonder whether in other domains, in which such factors are irrelevant, locality effects are less prominent. This does not seem to be the case in morphology.

For instance, artificial grammar learning experiments in which subjects learn morphotactic dependencies between nonsense syllables or words show a bias for adjacent dependencies over non-adjacent ones (Gómez, 2002; Newport and Aslin, 2004; Onnis et al., 2003). Non-adjacent dependencies in such experiments have been shown to be learnable only under certain conditions, either when the amount of training is significantly increased (Vuong et al., forthcoming), or when the variation in the material intervening between the two dependents is significantly increased (Gómez, 2002).

The above findings raise the question of whether natural language patterns (e.g., non-phonological, contextual allomorphy and morphotactics) exhibit the same locality bias. While a large scale empirical investigation into this topic is yet to be undertaken, several theoretical proposals (based on small samples of languages) hypothesize specific locality restrictions on contextual allomorphy (Siegel, 1978; Bobaljik, 2000; Carstairs-McCarthy, 2001; Embick, 2010). In these proposals conditions on the distance between the dependent morphemes are typically stated in terms of hierarchical structure of words. For instance, Adger et al. (2003) propose that only sister nodes can be in a conditioning relation. Embick (2010) working from within the framework of Distributed Morphology puts forth a “hybrid” theory of restrictions on allomorphy that involves both the linear notion of locality and the hierarchical cyclic locality (a Minimalism notion).

Many learning models of inflection implement an all-or-nothing bias for some type of local dependencies. One interesting case in point is Albright and Hayes’ “Minimal Generalization Learner” for learning morpho-phonological alternations. Their original learner is biased towards local alternations insofar as its generalization procedure looks for the largest shared material immediately adjacent to the locus of the morpho-phonological change (Albright and Hayes, 2002). In their subsequent work the authors extend this learning procedure to handle non-local dependencies. Interestingly, in doing so they run into a problem. The extended learner applied to Navajo data was able to learn the non-local sibilant harmony generalization (showing up as allomorphy in the perfective prefix), but it also learned 89 other complex generalizations none of which are productive in Navajo, and in fact violate the data not present in the training set (see Albright and Hayes (2006) for discussion and some efforts to fix this problem). This is a typical example of overfitting which suggests that the hypothesis space should be more tightly constrained to exclude most non-local dependencies, supporting the relevance of the locality bias in learning morphology.

The simplest statistical models that incorporate a strong locality bias are the n-gram models. These models rely on frequencies of sequences of n adjacent elements to predict the frequencies of larger strings. Bigram models, a special case of n-gram models where n = 2, are frequently applied to the task of learning sequencing patterns. A recent proposal by Ryan (2010) incorporates bigram constraints (sequences of two immediately adjacent morphs) within a statistical Maximum Entropy framework (Hayes and Wilson, 2008) to learn morphotactics. The goal of Ryan’s learner is to model patterns of free variation in morpheme order which are not due to phonological or semantic factors (the proposal is illustrated using data from Tagalog). This proposal is interesting in that the restrictiveness that comes from using bigram constraints (that is, the assumption that the learner only pays attention to sequencing of immediately adjacent morphemes) proves to derive exactly the typologically
attested patterns of free variation such as local context sensitive variation (XYA [*XYB] but YXB [*YXA]), non-transitive patterns (XY, YW, WX), and fails to generate the non-attested ones such as non-local context sensitivity. Moreover, Ryan’s proposal provides an explanation for why free variation might arise in the first place: he shows that when his learner is fed categorical data it can sometimes produce a grammar that generates free variation of the kind that occurs in natural languages. This is, therefore, another example in which a learning bias explains how certain restricted patterns of ambiguity (here a mismatch between the position within a word and morpho-syntactic features typically realized in this position) may arise and be tolerated in languages as a result of learning biases.

A locality bias may also be implemented in a gradient rather than categorical fashion. One possibility would be to adopt the hypothesis of Gómez (2002) that learners by default pay attention to adjacent dependencies, but can shift attention to non-adjacent dependencies when adjacent dependencies fail to be sufficiently predictive. This proposal can be implemented in a statistical learning model like the one discussed in the previous section. That is, we can design a model with a trade-off between the goodness of fit and the distance between dependent elements within a rule, so that if the local rules fit the data poorly, competing non-local rules become more probable.

The successes of learning models that incorporate a locality bias (and failures of those that do not) support the presence of this bias in learning morphological patterns. We have considered a few proposals that explicitly build in a locality bias into the learner. However, it is possible that this bias may be ultimately reduced to a preference for patterns that are easier to process. The correlation between processing difficulties (in particular, memory-related difficulties) and long-distance dependencies has been well documented in the literature on syntactic processing (e.g., processing of wh-questions, Phillips et al. (2005); Fiebach et al. (2002)). Thus, it is possible that a locality bias could be a particular instance of a simplicity bias discussed earlier.

5.3 Conjunctive bias

We have seen that learning allomorphy can be modeled as grouping lexical items into categories (one for each distinct set of allomorphs) based on the properties of these lexemes. In fact, many learning problems can be stated as such categorization problems, in which the task of the learner is to group the stimuli into distinct classes based on their behavior. Categorization problems have been studied extensively in machine learning and in psychology. In the 50s and 60s, psychologists have identified a number of interesting biases when testing how people learn artificially created categories defined by discrete features such as shape, color, size, and so on. One of the most robust biases they have discovered is a conjunctive bias (Bruner et al., 1956; Haygood and Bourne, 1965), or a bias for categories defined in terms of the joint presence of several feature values. This bias is also apparent in grammatical categories of natural languages.

For instance, an allomorph class or an inflectional class is typically described as a conjunction rather than a disjunction of properties that condition the allomorphy (e.g., “masculine nouns that end in a consonant” vs. “masculine nouns or nouns that end in a consonant”). Similarly, learning of inflectional paradigms can be viewed as learning a number of mostly conjunctive categories. Namely, each affix is a pairing of a phonological string and a featural description of the environments in which this affix is inserted. The featural description most commonly takes the form of a conjunction of features (e.g., [2p. dual masculine]) and much less frequently as disjunction (e.g., [2p. or dual or masculine], equivalent to three
homophonous affixes). According to the typological data by Baerman and Brown (2005) from the World Atlas of Language Structures, from a sample of 140 languages that mark subject and person agreement on the verbs, 80 (or 57%) have no syncretism or homonymy in any of the paradigms, which guarantees that every phonologically distinct form can be mapped to a single conjunction of features values. Other typological work on person number marking found that in paradigms with syncretism the most common type of syncretism is the one that can be described by underspecification (Cysouw, 2003; Pertsova, 2011), which is also a conjunctive type of syncretism.

Interestingly, similar results are found in phonology. In fact, for a long time phonologists took for granted that phonological processes affect “natural classes” of segments, that is segments that can be described by a single conjunction of phonological features (e.g., [+stop, +labial]). Mielke’s dissertation, based on a survey of 561 languages, showed that this was not always the case (Mielke, 2004). Yet, he still found that conjunctive phonological classes are typologically more frequent, accounting for about 76%.

The preference for conjunctive categories is not obviously reducible to some notion of simplicity. For instance, simplicity metrics that count the number of symbols in feature representations are unable to distinguish between the two operations. From a logical point of view, both conjunction and disjunction are also similar: their truth tables contain three instances of one Boolean value, and one instance of the other. One difference between the two types of operations is that conjunction has fewer number of positive instances in its truth-table compared to disjunction (see table 2). One might think that perhaps the fewer different positive instances of the category exist, the easier it is to learn it. However, this hypothesis is not consistent with a robust finding that inclusive disjunction (three positive instances) is easier to learn than exclusive disjunction (two positive instances) (Haygood and Bourne, 1965; Gottwald, 1971). The truth tables for these categories in the order of their subjective difficulty appear in table 2.

Table 2: Conjunction, inclusive and exclusive disjunction

<table>
<thead>
<tr>
<th>Conjunction</th>
<th>Disjunction</th>
<th>Excl. disjunction (XOR)</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1 F2 F1 ∨ F2</td>
<td>F1 F2 F1 ∨ F2</td>
<td>F1 F2 F1 ⊕ F2</td>
</tr>
<tr>
<td>T T T</td>
<td>T T T</td>
<td>T T F</td>
</tr>
<tr>
<td>F T F</td>
<td>F T T</td>
<td>F T T</td>
</tr>
<tr>
<td>T F F</td>
<td>T F T</td>
<td>T F T</td>
</tr>
<tr>
<td>F F F</td>
<td>F F F</td>
<td>F F F</td>
</tr>
</tbody>
</table>

The formal learning theory also does not tell us why simple disjunction should be more difficult than conjunction, although it suggests that conjunctive categories should be easier than categories that can only be represented with complex formulas involving both conjunctions and disjunctions (e.g., disjunctions of conjunctions, or DNF formulas). Specifically, conjunctions are efficiently PAC learnable by a simple algorithm (Valiant, 1984), while arbitrary DNF formulas are not (Pitt and Valiant, 1988). Similarly, many psychological models of categorization are unable to predict the difference between simple conjunction and disjunction because these models simply look for a way to separate positive examples from the negative ones. This strategy cannot distinguish between the two types of concepts since disjunctive concepts can be derived from conjunctive ones by simply switching positive examples to negative ones and vice versa (assuming all features are binary).

An explanation for the conjunctive bias may turn out to lie in the particulars of the
generalization strategy used by human learners. Certain linguistic models of morphology are actually consistent with the conjunctive bias because these models assume that the child learns from positive examples alone. There is a very simple learning strategy that can learn conjunctions from positive examples alone, while disjunction is best learned from negative examples. This strategy involves looking for features shared among all the instances belonging to the same category (Siskind, 1996; Molnar, 2001; Pertsova, 2011). Such a procedure automatically favors conjunctive categories because every positive instance of a conjunctive category includes all the properties that are shared by the members of the category at large (possibly together with some irrelevant properties).

For instance, if we are learning the category masculine and singular, than all examples of this category will have the property masculine and the property singular, so keeping track of the properties that never change from one example to the next is sufficient to identify all the conjuncts and to rule out the irrelevant properties in the process. On the other hand, when learning a category like masculine or singular, each example may only have one of the two properties along with many other irrelevant properties. In this latter case, the identification of relevant features becomes much more difficult. Thus, a simple bottom-up memoryless learner whose initial hypothesis is an empty conjunction, and whose subsequent hypotheses upon receiving new data $x$ is derived by intersecting the previous hypothesis with the properties of $x$, would learn conjunctive categories, but not disjunctive ones. To learn disjunctive categories, the learner has to do some extra work, predicting that such categories should be harder to learn. The learner described in Pertsova (2011) introduces another bias on top of the conjunctive bias, namely a bias towards concepts that can be succinctly described as complements of conjunctive concepts (making a prediction that “elsewhere” type patterns should be the second easiest to learn after the conjunctive patterns).

The literature on concept learning includes other kinds of biases related to the category structure that I will not discuss here. The relevance of such biases to learning of inflection and other linguistic categories is yet to be systematically investigated (for some steps in this direction see (Moreton and Pertsova, 2012)). This investigation would help us establish the appropriate set of biases for human category learning, and to test whether domain-general biases alone are sufficient to explain the specific instance of category learning, grammatical learning.

6 Concluding remarks

The central question of any learning theory is generalization - how can any intelligent system go beyond the available data to make correct predictions about the future? In this chapter, I have attempted to show how machine learning provides insights into this question, and how it touches on several issues of inherent interest to linguists. Perhaps the most central issue is the nature of constraints on the human language and the human mind. Machine learning provides a particular angle on this question by allowing us to rigorously explore what constraints on the learner can produce the patterns we find in languages.

In terms of acquiring inflection, learning biases help us account for patterns of semi-regularity and limited ambiguity (e.g., allomorphy, syncretism, free variation in affix order). Such learning proposals are particularly valuable in the finite domain of inflection because they make empirical predictions which go beyond the predictions of purely generative or descriptive theories. Generative theories of inflection account for what patterns are accept-
able or possible, and typically make very similar or exactly the same predictions because in a finite domain one can always resort to listing exceptions (something that most theories allow in one way or another). On the other hand, learning proposals make predictions for a wider range of phenomena (language acquisition, typological preferences, trends in language change) and hence are easier to distinguish from one another. These predictions can be additionally tested using the artificial grammar learning experiments, assuming that such experiments do in fact reflect something about language acquisition.

Overall, the main problems that arise for learning inflection, irregularity and mismatches, are in principle solvable (for instance, by the memorizing learner), but linguistically interesting solutions are those that capture the linguistic regularities (categorical or statistical universals) and that match the trajectory of language acquisition and language change. Such solutions will provide a deeper explanation for the inflectional component of language.

References


25


