What unsupervised learning tells us about language models

David M. W. Powers

Artificial Intelligence Group
Department of Computer Science
The Flinders University of South Australia

David_Powers@flinders.edu.au

Introduction

Science and Learning

It has long been recognized as a principle of scientific methodology that theory and analysis are based on assumptions. The role of science is largely to ensure the identification of assumptions, the formulation and formalization of those assumptions as hypotheses, and the elimination of those hypotheses that are demonstrably incorrect. The same considerations apply to learning. Machine Learning and Mathematical Learning Theory have supplied us with a number of results which emphasize that assumptions are always present and that every possible learning algorithm will have as many conceivable problems on which it will perform badly as problems on which it will perform well (Schaffer, 1994).

In many ways learning theory has been influenced as much by by philosophical theories of science as psychological theories of learning . We must be very careful not to underestimate the power of learning, for example in relation to language and cognition, since whatever we think we know about language has been developed using the same scientific methodologies as have inspired, and are embodied in, many of the learning systems developed within Artificial Intelligence, Automated Reasoning and Machine Learning. In fact, the formal results confirm that we must always make assumptions, and that the assumptions will in large measure determine our conclusions.

Language and Cognition

Because of the negative formal results on learning methodologies, there has been a tendency to pessimism about learning in the linguistic community. Nonetheless, all is not lost, for we are dealing not with the set of conceivable problems but with the set of actual situations, linguistic and otherwise, in which humanity is embedded. More than that, although human language and its acquisition are often viewed as separate from the natural environment, and are in a sense artifacts, we must acknowledge that language is in a sense an extension of, even an expression of, our ontology. That is language is part of the totality of what we learn from our experience of the world, and has such a deep relationship with the way in which we perceive, construe and interact with the world that is non-trivial to separate the specific features of language from the general features of a sensory-motor cognition.

Let us consider the problem of assumptions. There have always been assumptions in every branch of science, and the sign of maturity has been their making the tacit assumptions clear. In the case of language and learning there is a sense in which we are too close to the problem because the techniques we use in understanding the world are the very objects of our study. In any case, assumptions are not necessarily bad things in themselves — they just need to be exposed and controlled.

The tacit assumptions behind language and learning, which provide bias for our cognitive and linguistic theories, are only the tip of the iceberg. There are far deeper presumptions which are built into the way our perceptual and cognitive apparatus is designed to deal with the world. Some of these may themselves reflect deeper principles which relate to the scientific laws of the universe and its organization. For example, we tend to think of objects as being systems of parts which exhibit a considerable degree of spatio-temporal proximity, and this way of ‘thinking’ is very much a part of the way our perceptual and cognitive processes are organized, so that we are biased against relationships which obey more complex spatio-temporal constraints. The leaps in Physics in relation to Quantum Mechanics and Relativity relate directly to a paradigm shift away from these obvious direct proximal relationships. Special Relativity indeed gives us a new distance measure for spatio-temporal proximity.

Cognitive and Generative Linguistics

Recently, the assumption that language is independent of the rest of cognition, which has gained credence as a tenet of Generative Linguistics, has been challenged with a claim that language is an emergent property of general cognitive mechanisms — this is the foundation stone of Cognitive Linguistics. Each of these overt assumptions depends on a number other assumptions which are not usually explicitly stated. However, Hockett (1961) assembled and attacked a set of nineteen assumptions which were 'authorized' by Chomsky as underlying the generative nativist position and these have been examined criticially by Powers (1989, p138ff). The cognitivist/constructivist position (e.g. Deane, 1992) also has its implicit assumptions although, as the minority position, these do not appear to have been formalized and challenged in the literature.

In this paper we will be exploring assumptions designed to explain language learning, as opposed to acquisition by parameter setting, and we will contrast these with the assumptions of both nativist and constructivist. Our starting point is the assumptions which would allow or deny the possibility of learning of a language, or evolution of a language organ, and as we near the finishing line and discover the linguistic structure which our learning algorithms find, this will lead us to question some of our assumptions about language itself, and to postulate some counterintuitive assumptions of our own.

The Chomskian Assumptions

In this section we take the unusual step of devoting a page to reproduce (with editorial annotations removed and some items subdivided) Powers' (1989, p138ff) summary of the 19 assumptions, tacit as well as explicit, made by Chomsky, as agreed with and reported by Hocket (1961). Whilst one of the major contributions of Chomsky was the formalization of explicit hypotheses — in some cases specifically as working hypotheses which are simplifying rather than true — they have tended to fall back down to the level of tacit assumptions during the life of the generative paradigm and we do well to revaluate this list 25 years on.

C1. The vast majority of the sentences encountered throughout life by any user of a language are encountered only once. [***uniqueness***]

C2. Any user of a language has access, in principle, to an infinite set of sentences. [***infiniteness***]

C3. The user knows the grammar of his language, though not in a sense of 'know' that would imply that he can explicitly tell others of his knowledge. [***competence***]

C4. A users performance - what he actually says and hears - reflects his competence, but is also conditioned by many other factor [***performance***]

C5. A convenient device is to imagine an ideal speaker- listener, in a perfectly homogeneous speech-community, who is unaffected by such grammatically irrelevant conditions as memory limitations, distractions, shifts of attention and interest, and errors (random or characteristic). [***ideal speaker***]

C6. Since the users competence is a mental reality, linguistics is necessarily mentalistic. [***mentalistic competence***]

C7. Probabilistic considerations pertain to performance, not to competence. [***probabilistic performance***]

C8a. The distinction between grammatical and nongrammatical sentences (whether absolute or a matter of degree) applies to competence, not performance. [***grammaticality***]

C8b. The degree of acceptability of an actually performed utterance is a different matter. [***acceptability***]

.C9. Meaningfulness, like grammaticality, pertains to competence. But these two are distinct. [***meaningfulness***]

C10. The grammar of a language is a finite system that characterizes an infinite set of (well-formed) sentences and is by definition not more powerful than an universal Turing machine. [***formal power***].

C11. At present there is no known algorithm for computing or 'discovering' the grammar of a language: knowledge of grammatical structure cannot arise by application of step-by-step inductive operations of any sort that have yet been developed within linguistics, psychology, or philosophy. [***learnability***]

C12a. It must be, therefore, that the infant brings to this task at least the following: an innate system for the production of an indefinitely large set of grammars of 'possible' human languages; and the innate ability to select, from this set, the (or a) correct grammar for the language of his community. [***universality***]

C12b. The infants innate system must include much more, eg., an algorithm for determining the structural description of an arbitrary sentence given an arbitrary grammar. [***parser***]

C13. An explicit formulation of the innate grammar-producing system just mentioned would constitute a general grammar. or general linguistic theory. [***generality]***

C14. The innate grammar-producing system is a well-defined system in the sense of C10. [***well-definition***]

C15. It is at least plausible that the grammar of a language consists of three components: (a) The syntactic component, (b) The phonological component, (c) The semantic component. [***modularity***]

C16a In searching for the grammar of a language, one may propose various explicitly formulated grammars ... for the language. ... A grammar proposed for a language is descriptively adequate to the extent that it correctly describes the intrinsic competence of the idealized native speaker. [***descriptively adequate grammar***]

C16b A descriptively adequate grammar for a language is principled to the extent that it conforms to a general linguistic theory of the type mentioned in C13. [***principled grammar***]

C17a. A proposed general linguistic theory (C13) is descriptively adequate if it makes a descriptively adequate grammar available for each natural language. [***descriptively adequate theory***]

C17b. A descriptively adequate general theory is explanatorily adequate to the extent that it approximates the innate grammar-producing system, and other innate capacities, of the infant (C12). [***explanatorily adequate theory***]

C18. The proposal in C15 says nothing about how the user employs his grammar in either the production or the reception of sentences. [***algorithms***]

C19. A linguistic change is a shift from one grammar to another (presumably similar) one. It is essential to distinguish between system-conforming and system-changing events. It is important not to mistake an awkward or inaccurate performance for one that is really symptomatic of a change in underlying competence. [***evolution***]

Some assumptions questioned

Most of the Chomskian assumptions have been questioned at one level or another by Hockett (1961) or Powers (1989). Clearly the later ones build on the former, as is clear from the cross-referencing, and in this section we will address only those for which there is obvious significance in relation to the type of linguistic theory and grammar we might develop. In some cases, notably C5, Chomsky has explicitly made simplifying assumptions by assuming an ideal speaker, and C1 and C2 also appear to fall into this class in to the extent that they are recognized as simplifying assumptions rather than empirical facts. Other assumptions are actually more like definitions, thus the fundamental competence/performance distinction of C3/C4 defines these concepts which then lead directly to the ideal language user of C5 and thus underlie most of the subsequent assumptions. But from a more cynical perspective it provides a ‘too hard’ basket by allowing convenient relegation of some language phenomena to ‘performance issues’.

Your infinity is too small

The problem with sweeping generalizations is that people tend to believe them and forget about the exceptions that disprove the rule. While we often use words like ‘all’ and ‘unique’ in a less than precise fashion, linguistics and learning theory has built heavily on C1 and C2 which formalize these words as ‘infinite’ and ‘once’, while adding hedges like ‘the vast majority’ and ‘in principle’, that are then conveniently overlooked in building up our ‘ideal’.

Note that an arbitrary context-free grammar requires an unbounded stack to produce a sentence whose length is not limited *a priori*. Both the size of our heads and the span of our lives are bounded. An arbitrary strictly context-free grammar requires an infinite stack in order to actually recognize or generate any sentence of the language. Assuming a finite vocabulary and a definite lifespan, the number of sentences in a language has an obvious upper bound.

Why split hairs? What’s the difference between ‘large’ and ‘infinite’? The point is that once we remove the assumption pertaining to the infiniteness of language, any finite language can be generated by a non-recursive grammar and is free of many of the restrictive negative results which have been proven about language and learning (Gold,1967; Chomsky,1963). A model based on analogy and copying from one level to another in a connectionist model can, for example, produce recursive-like structures but without the embarassment of having to provide arbitrary restrictions about levels of language (like subjacency). It is clearly advantageous when a range of linguistic phenomena cease to become problems to explain, but rather become emergent properties of the system or theory.

Of course, a connectionist model is just a dressed up form of statistical model, and recently there has been an increasing awareness that grammaticality judgements, by linguist and lay alike, are heavily influenced by context. Furthermore, the whole concept of subcategorization of contentives must be cast into question, since with sufficient ingenuitiy any supposed breach of subcategorization can be justified by constructing an appropriate context: consider for example that any bodily action can be used as a means of gestural communication and may thus be used with the subcategorization associated with ‘tell’ (Entwisle, 1995). Entrenchment is also implicated here (Deane, 199?). The use of a probabilistic model can actually explain as a single phenomena the two attributes of competence and performance which Chomsky postulated: a connectionist model which reflects frequency/probability/information but requires attaining a threshold to be productive allows for some ‘rules’ to be stronger than others and for grammatical judgements to be reasonably reliable whilst allowing comprehension, production and repair when the input is not as expected. This type of model also provides an explanation of the distinction between recognition and production grammars/competence/performance (unlike C18).

Once upon a time

Similarly, the assumption that we encounter each sentence only once automatically steers us to models of learning, like Gold’s (1967) identification in the limit, in which the primary issue is the order in which subjects are exposed to sentences and given an appropriate ordering language learning is possible, poverty of the stimulus notwithstanding. Most researchers either haven’t read the paper or simply overlook the fact that it is only learning from sentences in an arbitrary order which is impossible for superfinite languages (viz. infinite set of sentences) in an unsupervised paradigm (viz. without examples explicitly identified as ungrammatical).

In fact, not only is the assumption false, it is not just wrong but completely backward in that the vast majority of the sentences a child encounters occur multiple times. Our daily routine dominates and the boring repetitions far exceed the novelties. How many times in his early years does a child hear “Daddy’s home” or “Dinner’s ready” or “Clean up your room” or “Once upon a time ...”.

Sentences, words, morphs and closed classes

There is a further implicit assumption here that sentences have some sort of psychological reality: if we discard the clause periphery, or count clauses rather than sentences, the effect becomes even more marked. Indeed the minor variations are very powerful and do provide critical input to a learner. Now consider that the child’s play consists in repetition at every level. The child wants you to tell or read the same story over and over again, to watch his favourite video or sing his favourite song *ad nauseum*. The child thrives on routine and the kind of paradigmatic repetition that every teacher knows facilitates learning.

Furthermore, even granted that the content words do change, the sentence structure in terms of closed class words and affixes are highly salient. Here is another assumption. We assume that the words children *produce* relatively late are not only not *understood* but not *recognized* or *attended to* until relatively late. Computer models have demonstrated that exposure to pure untagged corpus text is sufficient to discover the closed class words and affixes and the basic sentence structures, and that these morphs can be formed into morphological, grammatical and subcategorization/semantic classes in the absence not only of explicit negative information but of any ontological input or possibility of attachment of semantics — and contra C15 the same mechanisms have been employed to discover linguistic structure relating to each of the components postulated by Chomsky (Powers, 1989, 1991, 1992, 1996). Conversely, computational constraint parsing models have been produced which can parse arbitrary corpus texts and identify ungrammatical sentences using solely this kind of closed class and affix information (Entwisle, 1994, 1995).

This also points to a further assumption, that the conventions which each language has adopted about where to write its spaces has some relationship to a meaningful grammatical unit, which we call a word. In the computer models just discussed, the units which play a role are functional units which may be words or morphs.

The chicken and the egg

Another implict assumption (C10-C14,C16,C17) is that there is a Language Acquisition Device (LAD) whose job is to learn the mother's language. The learning theoretical problems arise because the unrestricted grammars permitted by the usual assumptions are difficult to learn because there are always many (an infinite number) of larger grammars which include them. Let us take stock for a moment and consider 'which comes first' and propose an alternate hypothesis that there is a Language Inventing Device (LID) which makes use of all the relevant cognitive, linguistic and physical abilities of the child as well as the richness of his environment in order to construct a vehicle for communication.

It's not that the LAD is the way it is because it has to handle all human languages, but rather that all human languages are the way they are because they have been invented by the LID!

And while we are on the subject of the origin of language, the argument for an innate language organ has been built on a number of fallacies, one of which is that evolution is somehow more powerful than learning. In fact, evolutionary techniques are just part of the arsenal of techniques used by Machine Learning researchers, and resent breakthroughs in Neurology, Psychology and Immunology show that mechanisms which had previously been thought to be restricted to evolution and genetics are part and parcel of the daily function of our bodies — for example, antibodies are produced by a kind of natural selection.

Thus stating that a capacity could not have arisen by learning, or involve general cognitive abilities, simply prejudges and prejudices research at understanding why language is the way it is. Sure, descriptive linguists may not want to address these scientific issues, but it is unprofessional to make claims which go beyond their own competence and the evidence of their own research. In fact, if it could not have arisen by any learning process, then neither could it have arisen by means of the far more limited evolutionary processes. The theoretical results don't hang on any particular algorithm, but rather depend on assumptions about paradigm — the nature of the learning situation and the type of input the learning receives. The fact that we don't know something (C11) doesn't mean it is impossible, and indeed there are experimental results that demonstrate that learning of at least some linguistic constructs is possible and that the specialization of the language organ is substantially overstated.

Claims that language, or certain principles, are innate have three problems: first, they are tendencious as language is clearly an innate and highly characteristic human ability; second they are unscientific and do not remove the requirement of explaining how and why the universals came about; third they are unparsimonious insofar as they specifically deny the sharing of cognitive mechanisms beyond the pale.

References

Noam Chomsky (1963), *Formal Properties of Grammar*, 323-418, and with Miller, G.A., *Introduction to the Formal Analysis of Natural Language*, 269-321 and *Finitary Models of Language Users, 419-491,* **Handbook of Mathematical Psychology**, ed. R.A. Luce, R.R. Luce and E Galanter.

Paul D. Deane (1992), **Grammar in Mind and Brain: Explorations in Cognitive Syntax**, Mouton

J. Entwisle and Groves, M. (1994), *A method of parsing English based on sentence form,* **Proceedings of the International Conference on New Methods in Language Processing**, 116-122

J. Entwisle (1995), **A parser for English using constraints on surface sentence form**, Ph.D thesis, Department of Computer Science, The Flinders University of South Australia

E.M. Gold(1967), *Language Identification in the Limit,* **Information and Control 10**:447-474

Charles F. Hockett (1961), *Grammar for the Hearer,* **Proceedings of the Symposia in Applied Mathematics XII**, p3ff.

D. M. W. Powers (1989), **Machine Learning of Natural Language**, Springer Verlag.

D. M. W. Powers (1991) *How far can self-organization go? Results in unsupervised language learning,* in **Proc. AAAI Spring Symp. on Machine Learning of Natural Language & Ontology**, 131-7

D. M. W. Powers (1992) *On the Significance of Closed Classes and Boundary Conditions: Experiments in Lexical and Syntactic Learning,* **Background and Experiments in Machine Learning of Natural Language: First SHOE Workshop on Extraction of Hierarchical Structure**, 245-266

D. M. W. Powers (1996) "Unsupervised learning of linguistic structure: An empirical evaluation", **Int'l Journal of Corpus Linguistics 1**#2

Cullen Schaffer (1994), *A conservation law of generalization performance,* **Proceedings of the 1994 International Machine Learning Conference**, 259-266.