# *Formal Models of Language: Formal versus Natural Language*

*Paula Buttery*

## *Formal vs. Natural Languages*

We can define a **formal language** precisely as a set of strings over an alphabet (see the *Grammars* handout), but what is the definition of a **natural language**? A natural language can be thought of as a mutually understandable communication system that is used between members of some population. When communicating, speakers of a natural language are tacitly agreeing on what strings are allowed (i.e. which strings are ***grammatical***?[1]). Dialects and specialised languages (including e.g. the language used on social media) are all natural languages in their own right. Note that named languages that you are familiar with, such as *French*, *Chinese*, *English* etc, are usually historically, politically or geographically derived labels for populations of speakers rather than linguistic ones.[2]

***what is a natural language?***

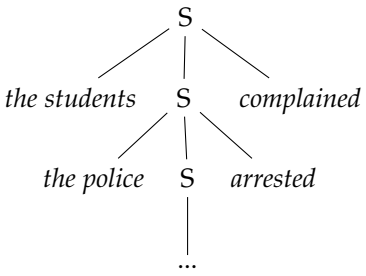
[1] Grammaticality has traditionally been considered a binary property of any given string – the string is either grammatical or it is not – however, recent work has shown that grammaticality can be gradient, with some strings found to be 'more' grammatical than others, based on native speakers' judgements (see for example `dx.doi.org/10.1111/cogs.12414`).

[2] Chinese for instance encompasses both Cantonese and Mandarin which are not mutually intelligible languages; and some of the Scandinavian languages, each of which have their own name (*Swedish*, *Danish*) might better be thought of as mutually intelligible dialects. There are various dialect continua, for example between German and Dutch, whereby geographically-juxtaposed dialects are mutually intelligible, but dialects at either 'end' of the continuum (e.g. central German and south-eastern Dutch) are not.

## *1. Language Complexity*

In the *Grammars* handout we noted a trade-off between the ***expressivity*** of a language class and the algorithmic ***running time*** for recognising a string from a language in that class. An important question then is whether all natural languages can be modelled using the class of regular grammars. This is an important question for two reasons: first, it places an upper bound on the running time of algorithms that process natural language; second, it may tell us something about human language processing and language acquisition (more on this in later sections). It turns out that regular grammars have limitations when modelling natural languages for several reasons:

*Centre Embedding* In principle, the syntax of natural languages cannot be described by a regular language due to the presence of centre-embedding; *i.e.* infinitely recursive structures described by the rule, $A \rightarrow \alpha A \beta$, which generate language examples of the form, $a^n b^n$. For instance, the sentences below have a centre-embedded structure.




1. The students the police arrested complained.

2. The luggage that the passengers checked arrived.
3. The luggage that the passengers that the storm delayed checked arrived.[3]

Intuitively, the reason that a regular language cannot describe centre-embedding is that its associated automaton has no memory of what has occurred previously in a string. In order to 'know' that $n$ verbs were required to match $n$ nominals already seen, an automaton would need to 'record' that $n$ nominals had been seen; but a DFA has no mechanism to do this. A formal proof uses the ***pumping lemma*** property to show that strings of the form $a^n b^n$ are not regular.[4] Careful here though: a regular grammar could generate constructions of the form $a^* b^*$ but not the more exclusive subset containing only $a^n b^n$ (which would represent centre embeddings). More generally the complexity of a sub-language is not necessarily the complexity of a language. If we show that the English subset of strings of the form $a^n b^n$ is not regular it does ***not*** follow that English itself is not regular. To prove something about the complexity of English, we can use the knowledge that regular languages are closed under intersection. So if we assume English is regular and intersect it with another regular language (e.g. the one generated by $/the\ a\ (that\ the\ a)^* b^*/$) we should get another regular language. However the intersection of a regular language of form $a^* b^*$ with English results in constructions of the form $a^n b^n$ (in our example case $/the\ a\ (that\ the\ a)^{n-1} b^n/$), which is not regular as it fails the pumping lemma property. The assumption that English is regular must be wrong.

However, examples of centre-embedding quickly become unwieldy for human processing (*n.b.* the difficulty of understanding the example sentences above). For finite $n$ we can still model the language using a DFA/regular grammar: we can design the states to capture finite levels of embedding. So are there any other reasons not to use regular grammars for modelling natural language?

*Redundancy* Grammars written using regular grammar rules alone are highly redundant: since the rules are very simple we need a great many of them to describe the language. This makes regular grammars very difficult to build and maintain.

*Useful internal structures* There are instances where a regular language[5] can recognise the strings of a language but in doing so does not provide a structure that is linguistically useful to us. The left-linear or right-linear internal structures derived by regular grammars are generally not very useful for higher level NLP applications. We need informative internal structure so that we can, for example, build up good semantic representations.[6]

In practice, regular grammars can be useful for ***partial grammars*** (i.e. when we don't need to know the syntax tree for the whole

[3] Regular languages are closed under ***homomorphism***: this means we can map all the *nouns* to $a$ and all the *verbs* to $b$ and then describe centre embeddings in 2. and 3. to be of the general form $/the\ a\ (that\ the\ a)^{n-1} b^n/$.

[4] For each $l \geq 1$, find some $w \in \mathcal{L}$ of length $\geq l$ so that no matter how $w$ is split into three, $w = u_1 v u_2$, with $|u_1 v| \leq l$ and $|v| \geq 1$, there is some $n \geq 0$ for which $u_1 v^n u_2$ is not in $\mathcal{L}$. To prove that $\mathcal{L} = \{a^n b^n | n \geq 0\}$ is not regular. For each $l \geq 1$, consider $w = a^l b^l \in \mathcal{L}$.

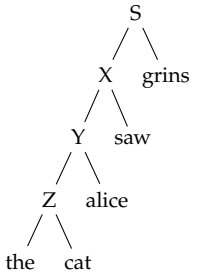
If $w = u_1 v u_2$ with $|u_1 v| \leq l$ & $|v| \geq 1$, then for some $r$ and $s$:

- $u_1 = a^r$
- $v = a^s$, with $r + s \leq l$ and $s \geq 1$
- $u_2 = a^{l-r-s} b^l$

so $u_1 v^0 u_2 = a^r \epsilon a^{l-r-s} b^l = a^{l-s} b^l$

But $a^{l-s} b^l \notin \mathcal{L}$ so by the Pumping Lemma, $\mathcal{L}$ is not a regular language

[5] Below: A left-branching tree structure derivable from some RG (ie. all rules of form $A \to Bb$ for $A, B \in \mathcal{N}$ and $b \in \Sigma$). This structure does not capture linguistic constituency.




[6] Below: a tree structure that captures linguistic constituency derived from a CFG (ie. all rules of form $A \to \alpha$ where $A \in \mathcal{N}$ and $\alpha \in (\Sigma \cup \mathcal{N})^*$). Note that $NP$ and $VP$ are single non-terminal symbols not two in a row—in linguistic terminology they represent a ***noun phrase*** (a phrased *headed* by a noun) and a ***verb phrase*** respectively.

sentence but rather just some part of it) and also when we don't care about derivational structure (i.e. when we just want a Boolean for whether a string is in a language). For example, in information extraction, we need to recognise NAMED ENTITIES. These are essentially referents *e.g.* The Computer Lab, Prof. Sir Maurice Wilkes, the Backs, Great Saint Mary's, the Gog Magog Hills, and so on. The internal structure of named entities is normally unimportant to us, we just want to recognise when we encounter them.

For instance, using rules such as:

$$\begin{aligned} NP &\rightarrow nnsb\ NP \\ NP &\rightarrow np1\ NP \\ NP &\rightarrow np1 \end{aligned}$$

where $NP$ is a non-terminal and $nnsb$ and $np1$ are terminals representing tags from the large CLAWS2 166 tag set,[7] you could match a titled name like, *Prof. Stephen William Hawking*.[8]

So the next question is whether the class of context-free grammars is expressive enough to model natural language. Or in other words, for every natural language that exists, can we find a context-free grammar to generate it?

There is some evidence that natural language can contain CROSS-SERIAL DEPENDENCIES. A small number of languages exhibit strings of the form shown in Figure 1.

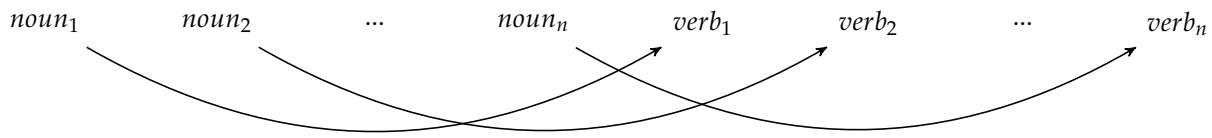



Figure 1: A schematic for cross-serial dependencies in language.

There is a Zurich dialect of Swiss German in which constructions like the following are found:

***mer d'chind em Hans es huus haend wele laa hälfe aastriiche.***
we the children Hans the house have wanted to let help paint.
*we have wanted to let the children help Hans paint the house*

Such expressions may not be derivable by a context-free grammar.[9]

If we are to use formal grammars to represent natural language, it is useful to know where they appear in the hierarchy (especially since the decision problem is intractable for languages above context-free in the hierarchy). However, notice that we can in fact divide the space of all languages any way we see fit; we are not limited to discussing language classes only in terms of the Chomsky hierarchy.

With respect to natural language, it might turn out that the set of all attested natural languages is actually as depicted in Figure 2:

[7] You can find the CLAWS2 tag set at `http://ucrel.lancs.ac.uk/claws2tags.html`. *nnsb* tags a preceding noun of style or title, abbr. (such as *Rt.* or *Hon.*); and $np1$ tags singular proper nouns (such as *London*, *Jane* or *Frederick*).

[8] Note that although noun phrases can be structurally complicated (*e.g.* the man who likes the dog which bites postmen), the relative clause is not generally part of a named entity so we don't need to capture it in the grammar (i.e. we use a partial grammar).

[9] The proof follows similarly as that for centre embeddings except that we must use the pumping lemma for context-free languages.

note the overlap with the context-sensitive languages which accounts for those languages that have cross-serial dependencies. Since the recognition problem for the class of context-sensitive languages is intractable, we don't want to have to generally use context-sensitive grammars to describe natural languages unless we really have to. What we would ideally like is a grammar that describes only the languages depicted in the set in Figure 2.

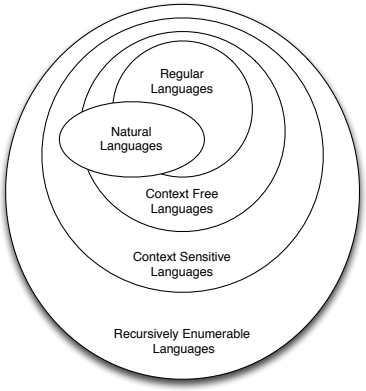



Figure 2: A Venn diagram showing the intersection of the attested natural languages with the Chomsky hierarchy

With this motivation in mind, Joshi [Joshi, 1985] defined a class of languages that is more expressive than context-free languages, less expressive than context-sensitive languages and also sits neatly within the Chomsky hierarchy (thus retaining the properties we already know about). This class of languages is known as the MILDLY CONTEXT-SENSITIVE languages. The abstract language class has the following properties:

- it includes all the context-free languages;
- members of the languages in the class may be recognised in polynomial time;
- the languages in the class account for all the constructions in natural language that context-free languages fail to account for (such as cross-serial dependencies).

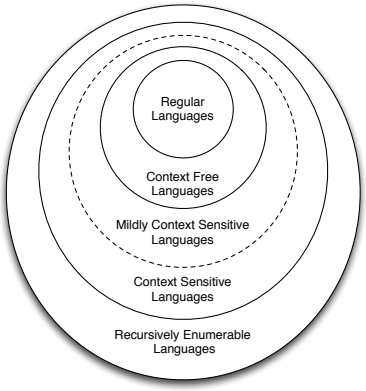



Figure 3: A Venn diagram showing the mildly context sensitive languages within the Chomsky hierarchy

The class of minimally context-sensitive languages is depicted in Figure 3. The grammar that Joshi defined to comply with these properties is called a TREE-ADJOINING GRAMMAR or TAG (see the *Grammars* handout).

For more general information on Formal Language Theory you can try Hopcroft and Ullman [1979] and Rozenberg and Salomaa [1997].

## *References*

N. Chomsky. On certain formal properties of grammars. *Information and control*, 2(2):137–167, 1959.

J.E. Hopcroft and J.D. Ullman, editors. *Introduction to automata theory, languages, and computation*. Reading, MA: Addison-Wesley, 1979.

Aravind K. Joshi. *Tree adjoining grammars: How much context-sensitivity is required to provide reasonable structural descriptions?*, volume http://dx.doi.org/10.1017/CBO9780511597855.007 of *Cambridge Books Online*, pages pp. 206–250. Cambridge University Press, 1985.

G.K. Pullum and G. Gazdar. Natural languages and context-free languages. *Linguistics and Philosophy*, 4(4):471–504, 1982.

G. Rozenberg and A. Salomaa. *Handbook of formal languages: Word, language, grammar*, volume 1. Springer Verlag, 1997.

# *Formal Models of Language: Formal versus Natural Language*

*Paula Buttery*

## *2. Human Language Processing Predictions*

The field of ***psycho-linguistics*** is concerned with how we acquire, comprehend and produce language; and consequently how we might store and process language in our brains. Questions of interest to a psycholinguist would include:

- How are words *organised* in the brain? For instance, do we *store* words in their entirety or do we store them in such a way that (abstractly speaking) they are rule generated e.g do we store the word *cat* and also *cats*, or alternatively just *cat* and use a rule that adds *s*'s to make a plural.
- What makes a sentence difficult to process? e.g. why is the sentence *the cat the dog licked ran away* easier to process than *the cat the dog the rat chased licked ran away* (we will discuss this one further below).
- Why do we prefer one particular interpretation of a sentence when there are many? e.g. for the sentence *he saw the queen with the telescope*, how do we decide who has the telescope? Do we store all the possible interpretations during processing (called ***parallelism***) or just one?
- How is the meaning of words stored in the brain? e.g. do we store the meaning of *bird* as a collection of features (such as *beak*, *feathers*, *fly*)? or do we store some representation of a prototypical bird (like a crow rather than a penguin)? or do we store the meaning of a word as an abstract statistical representation of its co-occurrence with other words?

### *Methods for measuring human response to language*

Psycholinguists use a range of methods to answer these questions. The methods we will come across in this course fall into one of two categories:

*Observations of language in the environment:* this involves gathering evidence from language after it has been produced either from wide-coverage corpora (large collections of texts built to be broadly representative of a language); or from specialised datasets (such as the language of children, second language learners, or people with specific learning impairments).

*Observations of humans in response to stimuli:* this involves measuring physiological responses to language tasks and includes measur-

ing reading times using eye tracking technology, measuring reaction times using button presses, or measuring brain responses using fMRI (which has low temporal but high spatial accuracy) or EEG/MEG (high temporal but low spatial accuracy).

*What makes a sentence complex?*

The term complexity is often used to describe the perceived human processing difficulty of a sentence: work in this area is generally referred to as computational psycholinguistics. Complexity within this domain can refer to: 1) the time and space requirements of the algorithm that your brain is posited to be executing while processing a sentence; or 2) the ***information theoretic content*** of the sentence itself in isolation from the human processor.

*Sentence complexity for the human processor:* Work in this area has looked mainly at parsing algorithms to discover whether they exhibit properties that correlate with measurable predictors of complexity in human linguistic behaviour. Two general assumptions are made in this work:

1. Sentences will take longer to process if they are more complicated for the human parser. Processing time is usually measured as the time it takes to read a sentence (often done with eye-tracking machines). These also identify whether the subject re-read any parts of a sentence.
2. Sentences will not occur frequently in the spoken language if they are complicated to produce or comprehend. Frequencies are calculated by counting constructions of interest in spoken language corpora.

The assumption then is that one (or both) of the two measurements of perceived complexity above will correlate with time and space requirements of the parsing algorithm. For instance, Yngve[1] suggested that human processing is limited by memory and that the size of the stack formed during processing will correlate with measures of perceived complexity. He predicted that sentences which required many items to be placed on the stack would be difficult to process and also less frequent in the language. He also predicted that when multiple parses are possible we should prefer the one with the minimised stack.

[1] V.H. Yngve. A model and hypothesis for language structure. In *Proceedings of the American Philosophical Association*, number 104, pages 444–466, 1960

*Information theoretic content of the sentence:* This work is concerned with the amount of information conveyed by each word or structure in a sentence. The general assumption made in this work is that the more we expect a certain type of structure, the more difficult it is to hypothesise an alternative structure. According to this model, a sentence is more complex when it is unexpected. Again, evidence for these theories is found in correlations with reading times or corpus frequencies. An example of this work

would be Hale[2] who uses a probabilistic Earley parser as a psycholinguistic model. Hale's paper predicts that the cognitive effort associated with integrating the next word into a sentence is related to the word's conditional probability (that is, the word's probability given the partial trees hypothesised for the words already heard).

[2] John Hale. A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the 2nd Conference of the North American Chapter of the Association for Computational Linguistics*, Pittsburgh, PA., 2001

### *Spoken versus written language*

Speech is very different in nature to written language.[3,4,5,6] The most obvious difference is the mode of transmission: the phonetics (sounds) and prosody (manner) of producing speech versus the characters and orthography (spellings) of writing systems. Other distinctive features of speech include intonation and co-speech gestures to convey meaning, and turn-taking, overlap and co-construction in dialogue interaction. Intonation refers to the way speakers' pitch rises and falls in line with words and phrases, to signal a question, for example. Co-speech gestures involve parts of the body which move in coordination with what a speaker is saying, to emphasise, disambiguate or otherwise (sometimes these are cultural practices).

[3] David Brazil. *A grammar of speech*. Oxford: Oxford University Press, 1995

[4] Douglas Biber, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. *Longman Grammar of Spoken and Written English*. London: Longman, 1999

[5] Geoffrey Leech. Grammars of spoken English: new outcomes of corpus-oriented research. *Language Learning*, 50:675–724, 2000

[6] Ronald Carter and Michael McCarthy. Spoken Grammar: where are we and where are we going? *Applied Linguistics*, 38:1–20, 2017

Turn-taking is the way that dialogue is constructed: speakers usually take it in turns to speak, and there are unspoken ways of ceding and holding 'the floor' (rules which can be broken of course, sometimes leading to offence). Overlap occurs when two or more speakers talk at the same time – pay attention to some conversations in the next few days: it happens surprisingly often without causing a problem! Similarly, co-construction occurs when one speaker finishes what another speaker is saying (couples and close friends do this a lot).

A fundamental characteristic of speech is the lack of the sentence unit used by convention in writing, delimited by a capital letter and full stop (period). Indeed it has been said that, "such a unit does not realistically exist in conversation" [7]. Instead in spoken language we refer to 'speech-units' (SUs)– token sequences which are usually coherent units from the point of view of syntax, semantics, prosody, or some combination of the three. Thus we are able to model SU boundaries probabilistically,[8] and also improve parses of the SUs using extra-linguistic information, such as the prosody.[9]

[7] Douglas Biber, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. *Longman Grammar of Spoken and Written English*. London: Longman, 1999

[8] Ann Lee and James Glass. Sentence detection using multiple annotations. In *Proceedings of INTERSPEECH 2012*. International Speech Communication Association, 2012

[9] E.J. Briscoe and P.J. Buttery. *The Influence of Prosody and Ambiguity on English Relativization Strategies*. Conference on the Interdisciplinary Approaches to Relative Clauses, Research Centre for English and Applied Linguistics, 2007

Other well-known characteristics of speech are disfluencies such as hesitations (1), repetitions (2) and false starts (3):

1. um he's a closet yuppie is what he is.
2. I played, I played against um.
3. You're happy to – welcome to include it.

Disfluencies are pervasive in speech: of an annotated 767k token

subset of the Switchboard Corpus of telephone conversations, 17% are disfluent tokens of some kind. Furthermore they are known to cause problems in natural language processing, as they must be incorporated in the parse tree or somehow removed. Indeed an 'edit' transition has been proposed specifically to deal with automatically identified disfluencies, by removing them from the parse tree constructed up to that point along with any associated grammatical relations.[10]

[10] Matthew Honnibal and Mark Johnson. Joint incremental disfluency detection and dependency parsing. *Transactions of the Association for Computational Linguistics*, 2:131–142, 2014

# *Formal Models of Language: Formal versus Natural Language*

*Paula Buttery*

## *3. Language Learnability*

If we define a ***grammatical system***, $(\mathcal{H}, \Omega, \mathcal{L})$ as:

- $\mathcal{H}$ a hypothesis space of language descriptions (e.g. all possible grammars)
- $\Omega$ a sample space (e.g. all possible strings)
- $\mathcal{L}$ a function that maps from a member of $\mathcal{H}$ to a subset of $\Omega$

Then a ***learning function***, $F$, maps from a subset of $\Omega$ to a member of $\mathcal{H}$.[1]

[1] For example, if we have $(\mathcal{H}_{cfg}, \Sigma^*, \mathcal{L})$ (that is, the grammatical system of all context-free languages over $\Sigma$) then for some $G \in \mathcal{H}_{cfg}$:

- $\mathcal{L}(G) = \{s_a, s_b, s_c ...\} \subseteq \Sigma*$
- and $F(\{s_a, s_b, s_c ...\}) = G$ for some $\{s_a, s_b, s_c ...\} \subseteq \Sigma*$

***learnability***

***Learnability*** is a property of a language class and occurs when $F$ is surjective (when we can learn every grammar in the hypothesis space using the learning function). The learning function manifests as an algorithm for ***grammar induction*** (and is often referred to as the ***learner***).

Gold's paper on learnability[2] introduced a number of ***learning paradigms*** one of which has been extremely influential in Linguistics, the details are as follows:

[2] E Mark Gold. Language identification in the limit. *Information and Control*, 10 (5):447 – 474, 1967. ISSN 0019-9958. DOI: https://doi.org/10.1016/S0019-9958(67)91165-5. URL `http://www.sciencedirect.com/science/article/pii/S0019995867911655`

For a grammatical system $(\mathcal{G}, \Sigma^*, \mathcal{L})$ –

- An $L \in \mathcal{L}$ is selected as the ***target language*** (i.e. the language that the learner is attempting to learn).
- All samples from $L$ (i.e. all $s_i$ such that $s_i \in L$) are presented to the learner one at a time, $s_1, s_2, ...$, in an infinite sequence.[3]
- After receiving each sample, the learner produces a hypothesis, $G_i \in \mathcal{G}$.[4]
- Learning is *successful* when $G$ has been ***identified in the limit***: that is, there is some number $N$ such that for all $i > N$, the hypothesised grammar $G_i = G_N$, and $\mathcal{L}(G_N) = L$ (the target language).[5]

[3] Note that the learner receives only ***positive evidence*** (as opposed to ***negative evidence*** which would be where strings not in $L$ were also presented to the learner but specifically flagged as errors). Also note that the evidence is exhaustive (i.e. every $s \in L$ will eventually be presented in the sequence.)

[4] So, after seeing the sequence $s_1, ...s_n$, the learner produces $G_n$.

[5] Note that $N$ is finite but there are no constraints placed on the computation time of the learning function.

In this paradigm the class of languages, $G$, is learnable if every language in the class can be identified in the limit, no matter what order the samples appear in. A well known result of Gold's work is that suprafinite classes of languages[6] are not learnable.

[6] A suprafinite class of languages is one that contains all possible finite languages and at least one infinite language—all the language classes in the Chomsky hierarchy are suprafinite.

### *Child language acquisition versus Gold*

Gold provides us with a framework for a thought experiment in which specific details must be fleshed out; in particular the definition of the hypothesis-space, $\mathcal{H}$, and the learning function, $F$.

Some linguists (sometimes called ***nativists***), believe in innate linguistic knowledge or a specific language faculty in the brain.[7] From the point of view of these linguists the hypothesis-space of grammars is relatively small, constrained by the innate knowledge.[8] Learning functions in this scenario tend to be algorithmic, analysing an input string and moving systematically from one grammar to the next within the small hypothesis-space.

***Empirical*** or ***usage-based*** linguists, on the other hand, believe that language may be acquired without the aid of an innate language faculty. These linguists have suggested that learning can be modelled as a statistical competition between all the grammars within the hypothesis-space. For these linguists, the hypothesis-space is unconstrained and could consequently be very large. A statistical learning function returns a probability distribution over the possible grammars. The distribution represents each grammar's fitness to describe the sentences encountered so far. In this scenario the current hypothesised grammar, $G_i$, could be selected according to the distribution. Note that under this model of learning, there needs to be a modified definition for success: for example, we could say that $F$ ***converges*** to $G \in \mathcal{H}$ if there exists a finite $N$ such that for all $i > N$, $F$ is defined on $\{s_1 ... s_i\}$ and returns a distribution over $\mathcal{H}$ such that $G$ is most likely.

Notice there are several points of difference between Gold's learning paradigm and language acquisition in children:

1 Gold's paradigm requires convergence in a finite number of steps (i.e after a finite number of hypothesised grammars). The amount of data the learner sees, however, is unbounded and the learner can use unbounded amounts of computation.

- In child language acquisition a child only sees a limited amount of data, and has only limited computational resources.

2 Gold's paradigm doesn't tell us anything about a learner's state at any particular time. In fact, at any particular time, it is not possible to tell whether learning has been successful (identified in the limit), since the learner may always guess a new grammar when presented with the next sentence.

- In reality children learn progressively and could perhaps be considered to be converging towards a target language (as is described above for the statistical learning models).

3 The learner hypothesises a grammar after every presentation of a string—this includes presentations that have been chosen by an adversary with knowledge of the internal state of the learner.

- It is arguable that actual input distributions received by children are in some way helpful (referred to as ***parentese***) and that children might even receive helpful negative evidence (as opposed to positive evidence only).[9] It has also been suggested that children

[7] This is referred to by Chomsky as ***Universal Grammar***.

[8] Nativists might argue that the hypothesis space *must* be constrained due to Gold's result that *none* of the classes in the Chomsky hierarchy are learnable—whether the learnability of Chomskyan classes is relevant to a human learner is a matter of debate.

[9] Note that linguists do not agree on these points.

only attend selectively to evidence—that is, they notice only the strings that are *just right* for them to learn from (this is referred to as the *Goldilocks effect*).

4 Within Gold's paradigm the target language is static and the learner is required to exactly identify the target language.

- Natural languages are dynamic not static. Also some linguists claim that we can observe differences in word choices and grammaticality judgments between adults speakers from quite similar backgrounds (that is, they do not appear to have a common target language). It is also not without argument that we ever converge on a single stable grammar within our lifetimes.

## *References*

E Mark Gold. Language identification in the limit. *Information and Control*, 10(5):447 – 474, 1967. ISSN 0019-9958. DOI: https://doi.org/10.1016/S0019-9958(67)91165-5. URL http://www.sciencedirect.com/science/article/pii/S0019995867911655.