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

*Paula Buttery*

*Easter 2018*

## *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 measuring 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