Evolutionary fitness, homophony and disambiguation through sequential processes

Caroline Lyon, Chrystopher L. Nehaniv, Sandra Warren, Jean Baillie
C.M.Lyon@herts.ac.uk C.L.Nehaniv@herts.ac.uk

School of Computer Science, University of Hertfordshire, UK

Abstract

Human language may have evolved through a stage when words were combined into structured linear segments, before these segments were used as building blocks for a hierarchical grammar. Experiments using information theoretic metrics show that such a stage could have its own evolutionary advantage, before the benefits of a full grammar are obtained.

This hypothesis is approached by examining the apparently ubiquitous prevalence of homophones. It shows how, perhaps contrary to expectation, communicative capacity does not seem to be adversely affected by them, and they are routinely used without confusion. This is principally explained by disambiguation through syntactic processing of short word sequences. It indicates that local sequential processing plays an underlying role in language production and perception, a hypothesis that is supported by evidence that small children engage in this process as soon as they acquire words.

Experiments on a corpus of spoken English calculated the entropy for sequences of syntactically labelled words. They show there is a measurable advantage in decoding word strings when they are taken in short sequences, rather than as individual items. This suggests that grammatical fragments of speech, could have been a stepping stone to a full grammar.

Introduction

The mapping of speech sounds onto meaning lies at the core of the human ability to communicate by language, and the limited range of sounds that other creatures can make contrasts markedly with the much wider range and combinatorial use of phonetic elements in human speech. The physiological changes to the vocal tract that were necessary to enable the production of speech sounds has concomitant disadvantages, but the value of the mechanisms exapted or adapted to support language appears to have outweighed these problems [1].

If human language had been designed to a teleological programme, we might have expected that there would be an optimum number of phonemes that provided the basis for speech. However, we find that the number of phonemes varies from about 12 to well over 100 [2]. There is massive redundancy. Some phonetic elements that can serve as particularly salient distinguishing features, such as clicks or ejectives, only occur in a subset of human languages. We see here not survival of the fittest, but survival of the many, varied, fit.

We might also have expected a one-to-one mapping between sounds and meanings. Indeed, recent mathematical models showing how language might have evolved take this approach and show how a limited number of phonemes can be combined to produce an indefinitely large number of unambiguous words [3, 4]. Nowak asserts that “ambiguity … is the loss of communicative capacity that arises if individual sounds are linked to more than one meaning” [3, p 613], that absence of word ambiguity is a mark of evolutionary fitness, and that word formation provides an exponential increase in fitness with length.

However, these models do not reflect language in the real world. Seemingly ubiquitous homophony is common in English as in other languages, though it is certainly not the case that a shortage of phonetic elements leads to a need for the same sounds to have multiple meanings. Many of the most frequently used words are ambiguous homophones (for example: *to, too, two; there, their; I, eye*) [5]. In spite of the theoretical possibilities of exploiting combinatorial properties of a set of phonemes, this does not in practice necessarily occur, yet communicative capacity does not seem to be adversely affected. We find homophones in the speech of small children [5], and observe the slippage of language into forms with more homophones [6;7,p 5].

Analysis of homophones

We can analyse homophones in two groups: those in which the homophonous forms are the same grammatical parts-of-speech, and those in which they are different. In English, and other languages [8], the second class is much the larger. Taking the smaller class first, semantic information may be necessary to distinguish these words. They may be distinct concepts spelt differently, such as *hair* and *hare*; or distinct concepts spelt the same such as *(river) bank* and *(money) bank.* They may have common ancestry, and been subject to a gradual semantic shift. For instance *to stamp* can have the distinct meanings to stamp a foot, or to stamp a letter. Linking these two meanings was a stage when letters were sealed with a heavy stamp. Homophonous forms may also be variations on a theme, as in the example from Wittgenstein of the word *game* [9, sections 66-76]. He points out that there is nothing common to all meanings of the word, but rather “a complicated network of similarities, overlapping and criss-crossing”. This class of homophones with the same parts-of-speech has been the subject of mathematical modelling, for example by Wang et al. [7], where a word refers to “an association between a meaning and an utterance”, and there seems to be an implicit assumption that they are *content words*.

However, the much larger class of homophones that are different parts of speech raise significant issues and deserve further scrutiny. Homophonous forms are frequently *function words*, and the fact that we can disambiguate them with such facility provides clues to our underlying syntactic abilities. For example, the words *to / too / two* are used and understood correctly by children very early on. We see that disambiguation must be through contextual processing, and this contextual processing seems to be mainly based on relations with adjacent words (for example *me too, two sweets, to the swing)*. The subconscious use of grammatical categories can explain how the appropriate lexical item is selected. Without invoking a full grammar, short word sequences, grammatical fragments, can be acceptable or not.

Perception and production of syntactically correct phrases

There is an ongoing debate as to how children acquire syntactic knowledge [10, 11, 12], but there is a general consensus that children from a very young age are aware of syntactic categories. Infants are aware of prosodic clues to syntactic elements, and can exploit them in the processing of speech [13, 14]. For instance, in English children use correct word order as soon as two words are produced [15]. This helps to explain how young children can understand phrases and sentences with homophonous terms: local syntactic constraints are employed as soon as words are acquired. Older speakers as much as infants are implicitly aware of syntactic categories. The fact that many could not explicitly define these categories does not detract from the proposition. In the same way, we can estimate the distance to a remote object implicitly using optical rules that we cannot explicitly formulate.

If we accept this proposition, then we can see that the disambiguation of homophones will often be based on the admissibility or otherwise of neighbouring parts-of-speech. For instance, consider *their / there. “their” is* a possessive pronoun typically followed by a noun or noun phrase. “*there”* is not usually followed by a noun or noun phrase, but typically by a verb, adverb or preposition:

Their adventures made a good story.
Their thrilling exploits amazed us.

There are many more to come.
They went there quickly

Figure 1

In Figure 1, the alternative forms *their / there* cannot be confused, because of local syntactic disambiguation. For homophonic *function words* like these, there is little or no content in them to aid disambiguation, nor is it necessary.

Experiments on the efficient decoding of word strings

The observations made so far suggest that processing short, syntactically labelled word sequences could play an underlying role in speech production and perception. To test this hypothesis, we carried out experiments to see if there was an advantage in processing words as short strings rather than as individual items.

Using Information Theoretic tools we have investigated the efficiency of decoding word sequences segmented in different ways. The concept on which these experiments are based is that we can measure the entropy of a sequence, and a decline in entropy is associated with an increase in predictability, an improvement in the efficiency of decoding and comprehensibility [16]. For a simple introduction to this concept see [17]. A standard reference is [18].

Taking the proposition that we are implicitly aware of syntactic categories or part-of-speech tags, we investigate whether tag strings are more easily decoded if they are taken in short sequences rather than as single items. In the rest of this paper we take the term “tag” to mean “part-of-speech tag”.

If we find that entropy declines as we take tags in pairs and triples, this would indicate that processing of short sequences is likely to have developed with improved understanding of speech. In turn, this would help explain how homophonous words are routinely used without confusion: they are disambiguated by being taken in conjunction with neighbouring words.

For our experiments, we take the Machine Readable Spoken English Corpus – MARSEC, organized by Arnfield [19, 20]. About 26,000 words are used. MARSEC includes prosodic annotation, which we are not using in the current experiments. The corpus includes unscripted news commentary, scripted news and lectures. This can be considered “well formed” language, not like informal conversation. Experiments are planned on other types of spoken language.

The first step in the experiment is to map words onto part-of-speech tags. This was done using a version of the CLAWS tagger (supplied by the University of Lancaster) described by Garside [21]. The CLAWS tagset was mapped onto a smaller customised tagset consisting of 26 part-of-speech tags (Appendix A).

The next stage is to measure the entropy in four cases: with no statistical information, then with information on single tags, tag pairs and tag triples. Taking the symbol H as entropy, H is the average number of bits needed to determine a symbol (tag). We need to find:

H_0 : entropy with no statistical information, symbols equi-probable.

H_1 : entropy from information on the probability of single symbols occurring.

H_2 : entropy from information on the probability of 2 symbols occurring in sequence.

H_3 : entropy from information on the probability of 3 symbols occurring in sequence

Figure 2

Description of entropy

Intuitively, we are looking at how much extra information about the part-of-speech tag we have when we take contextual information into account. Take for example the partial sentence:

We	see	a	complicated	network ….
<pronoun>	<verb>	<article>	<adjective>	<noun>

Figure 3

Consider the prediction of the tag of the word “complicated”. With no statistical information, we can only say that all tags are equally probable. With information on the probability of single tags occurring we can make a better estimate. If we have information on tag pairs this is improved further, and we can better predict the tag of the word “complicated” if we know it is preceded by the tag <article>. With information on tag triples we can again make a further improvement in prediction, also knowing that the tag of “complicated” is followed by a noun.

By calculating the entropy, we have a metric to quantify our intuitive understanding.

Formula for entropy

In mathematical terms let T be a tagset, and X be a discrete random variable taking values x in T. The probability that X takes symbol x as its value is $p(x)$

$$p(x) = probability\ (X = x)$$

$$H(x) = - \sum p(x) * log_2 (p(x)) \quad \text{over} \quad x \varepsilon T$$

Since $p(x)$ is a probability, $p(x) \leq 1$, so $log_2 (p(x))$ is negative (or zero if $p(x) = 1$).
The minus sign at the start of the formula cancels this out.

The derivations of the formulae for H_0 H_1, H_2, and H_3 can be found in [16; 17, appendix A].

Results

Shannon showed that the entropy of a sequence can decline as more of the statistics are taken into account. The n-gram entropy H_n measures entropy with information extending over n adjacent symbols and: $H_n \leq H_{n-1}$

Applying this analysis to the tagged MARSEC corpus we get the following results :

Speech representation	H_0	H_1	H_2	H_3
Entropy	4.70	4.10	3.31	2.99

Table 1

The figure for H_3 in Table 1 means that having identified the part-of-speech for two consecutive words in an utterance, there are on average about 3 bits of uncertainty in the next tag. Additional information, e.g. from the phonetic stream, would further reduce this uncertainty in the disambiguation of homophones.

We find a decline in entropy between H_0 and H_3 As information is taken over more adjacent tags the uncertainty decreases, comprehensibility increases. This suggests that processing of short sequences was likely to emerge in evolutionary changes, as it would be correlated with an improvement in communicative success.

Conclusion

We have analysed the level of uncertainty in processing short word sequences from a corpus of transcribed English speech. The results lend credence to the theory that sequential processing plays a role in the perception of language. This can help to account for the fact that we do not seem to have any difficulty in disambiguating homophones from different syntactic categories: the possible interpretations of the homophonic form are limited, usually to one, by syntactic constraints.

One of the purposes of this paper is to open up a discussion on the distribution of homophones in other languages, and to investigate whether syntactic processing of short word sequences is universally advantageous, or just a phenomenon found in a limited group of languages.

At a low level, sequential processing by primitive neural elements plays a key role in the production of human speech [1], and similar processes also operate at higher levels of speech production and comprehension. Observed combinations of phonemes are controlled by sequential regulators, phonotactic rules. Then groups of phonemes are combined into syllables, and syllables into words subject to morphophonemic constraints. The focus of this paper has been the next level, where short word sequences are processed. We find that there is a disambiguation role for syntactic categories before we move up to a full, sentence based hierarchical grammar. This is consistent with the hypothesis that language may have evolved through an intermediate stage of structured linear segments, before these segments were themselves used as building blocks for a hierarchical grammar [17].

References

1. Lieberman, P, *On the Nature and Evolution of the Neural Bases of Human Language* Yearbook of Physical Anthropology, 2002

2. Maddieson, I, *Patterns of sounds* Cambridge University Press, 1984

3. Nowak, M A; Komaraova, N L and Niyogi, P, *Computational and evolutionary aspects of language,* Nature, vol 417*:* 611-617, June 2002

4. Plotkin, J B and Nowak, M A, *Language Evolution and Information Theory* Journal of Theoretical Biology, vol 205: 147-159, 2000

5. Warren, S, *Phonological Acquisition and Ambient language: a Corpus Based Cross-Linguistic Exploration*, PhD thesis, University of Hertfordshire, UK, 2001.

6. Warren, P; Rae, M and Hay, J, *Goldilocks and the Three Beers: Word Recognition and Sound Merger,* Proceedings 9th Australian International Conference on Speech Science and Technology, 2002.

7. Wang, William; Ke, Jinyun and Minett, James, *Computational Studies of Language Evolution* Proceedings of COLING, 2002.

8. Ke, Jinyun; Wang, Feng and Coupé, Christophe, *The Rise and Fall of Homophones: a Window to Language Evolution,* Proceedings of 4th International Conference on the Evolution of Language, 2002.

9. Wittgenstein, L *Philosophical Investigations,* translated by G.E.M. Anscombe, Blackwell, 1953.

10. Tomasello, M and Brooks, P, *Early Syntactic Development: A construction grammar account* in Barrett, M (ed) *The Development of Language,* Hove, UK: Psychology Press*,*1999

11. Fisher, C, *The role of abstract syntactic knowledge in language acquisition: A reply to Tomasello,*Cognition 82 (3) : 259-278, 2002

12. Waxman, S and Markow, D, *Words as invitations to form categories: Evidence from 12- to 13-month-old-infants* Cognitive Psychology 29 (3), 1995

13. Morgan, J and Demuth, K, *Signal to Syntax,* Lawrence Erlbaum, 1996

14. Eimas, P *The Perception and Representation of Speech by Infants* in Morgan, J, and Demuth, K (Eds) *Signal to Syntax,* Lawrence Erlbaum, 1996.

15. Mazuka, R *Can a grammatical parameter be set before the first word? Prosodic contributions to early setting of a grammatical parameter* in Morgan, J, and Demuth, K (Eds) *Signal to Syntax,* Lawrence Erlbaum, 1996.

16. Shannon, C, *Prediction and entropy of printed English,* 1951*,* in Sloane, N. J. A, and Wyner, A (Eds) *Shannon: collected papers,* IEEE Press, 1993

17. Lyon, C; Dickerson, B and Nehaniv, C, *The segmentation of speech and its implications for the emergence of language structure* Evolution of Communication Vol 4, No 2: 161-182, 2003

18. Cover, T and Thomas, J.A. *Elements of Information Theory,* John Wiley, 1991

19. MARSEC: http://www.rdg.ac.uk/AcaDepts/ll/speechlab/marsec/

20. Arnfield, S *Prosody and Syntax in Corpus Based Analysis of Spoken English*, PhD thesis, University of Leeds, UK, 1994.

21. Garside, R *The Claws word-tagging system* in Garside, R, Leech, G, and Sampson, G (Eds) The Computational Analysis of English: a corpus based approach, Longman, 1987.

Appendix A

Description of the Tagset

The tagset used in these experiments is derived from CLAWS4, mapped onto a smaller set of 26 classes. They are as follows:

- article - singular e.g. “a”
- determiner - singular or plural “the”
- predeterminer e.g. “all”
- pronomial determiner e.g. “some”
- pronomial determiner - singular e.g. “this”
- proper noun
- noun - singular
- noun - plural
- pronoun - singular
- pronoun - plural
- relative pronoun
- possessive pronoun
- verb - singular
- verb - plural
- auxiliary verb - singular
- auxiliary verb - plural
- existential “here” or “there”
- present participle
- past participle
- infinitive “to”
- preposition
- conjunction
- adjective
- singular number “one”
- adverb
- exceptions

The tagging process includes the identification of common phrases or idioms, which are then treated as single lexical items. For instance, “of course” is tagged as an adverb.