Syntactic Form Frequency: Assessing

Intermediate article

Marc Brysbaert, Royal Holloway College, Egham, UK
Don C Mitchell, University of Exeter, Exeter, UK

CONTENTS

Syntactic form frequency research aims to find out how often words and sequences of words in a given context fulfil particular syntactic roles in a sentence. This is achieved by analyzing representative samples of text or speech materials. The research is motivated by the findings that artificial grammars perform better when they take syntactic form frequencies into account, and that humans also seem to be sensitive to this kind of information.

THE NEED FOR ASSESSING SYNTACTIC FORM FREQUENCIES

Syntactic form frequency refers to the number of times words and sequences of words in a given context fulfil particular syntactic roles in a sentence. When a human or a machine tries to convey a message, it is important not only to use the correct words, but also to assign the correct roles to the different parts of the sentence (i.e. to explain 'who did what to whom'). This is equally vital when listeners or readers try to recover the meaning intended by the speaker or the writer.

The assignment of the correct syntactic structure is less straightforward than it might seem. First, the structure of a sentence is not uniquely defined by the position of the words in the sentence. Although the order subject–verb–direct object–indirect object is the basic order in English, this is by no means the only possible sequence (as shown by the examples 'Lyn gave Charles the pencil' and 'The pencil was given by Lyn to Charles'). Conversely, small differences in word order can introduce large differences in meaning (compare 'He showed her baby the pictures' with 'He showed her the baby pictures'; Frazier and Clifton, 1996). Second, many sentences include regions that allow more than one interpretation, even when the parser is able to make a distinction between the subject, the verb, and the object. For instance, in the sentence 'The cop informed the motorist that he had followed…' the final clause could either be a complement structure ('… that he had followed the instructions') or a relative clause ('…that he had followed the whole day, that …'). In principle, the number of syntactic ambiguities grows exponentially with sentence length. This became apparent when one of the first artificial grammars was applied to some example input sentences (Martin *et al*., 1983):

List the sales of products in 1973.
(3 analyses possible: the products in 1973, the sales in 1973, or the listing in 1973) (1a)

List the sales of products produced in 1973.
(10 analyses possible) (1b)

List the sales of products produced in 1973 with the products produced in 1972. (455 analyses) (1c)

Not all possible syntactic forms occur equally often, however. Some are more frequent than others. Because researchers believe those differences in frequency are important both for practical purposes (e.g. to build an artificial language device) and to understand the ways in which humans interpret the syntactic structure of sentences (see below), they have tried to get more precise estimates of the relative frequencies of different structures.

FINDING CORPORA

To assess syntactic form frequencies, one needs corpora. Usually, these are machine-readable text files that consist of written texts or transcriptions of human speech. Most of the corpora that have been collected are available on the internet and can be found relatively easily with the existing search engines. Ideally, a corpus must be very large and

contain a representative sample of general language. The larger the corpus, the more reliable the frequency estimates become and the more representative the corpus is for the kind of texts covered. In the early days of corpus research, corpora with one million words were considered huge; nowadays, owing to the massive availability of digital text sources, corpora can include up to a billion words. One much used source of corpora, for example, is the CD-ROMs made available on a yearly basis by newspaper publishers.

The problem of the representativeness of the corpus depends to some extent on the research question. With respect to syntactic structures there is, for instance, the fact that newspaper publishers have their articles text-edited before publication. This may raise problems for a researcher who is interested in actual usage within a certain language community (as opposed to use prescribed in grammar textbooks). Another limitation of newspapers and magazines is that they cover only a limited range of language registers (i.e. language use in a particular context). Therefore, a few corpora with a wider variety of texts have been collected for research purposes. Unfortunately, most of these corpora are rather limited in extent (to a few million words). Another source of corpora that is currently attracting a great deal of attention is the internet, where in discussion groups and in chat channels millions of sentences are produced weekly on a great variety of topics and without any stylistic supervision.

A final concern with the existing corpora is that most are based on written texts. This raises a number of issues. One is the extent to which the calculated syntactic form frequencies on the basis of written materials can be extended to spoken materials. Another is that written materials may tell us little about the language children are exposed to during their preschool years.

ESTIMATING FREQUENCIES

After the corpus has been chosen (or assembled), the sentences must be parsed in order to gain access to the different syntactic forms and frequencies. Depending on the research question, two different strategies are used. For some topics (for example, the development of an automatic sentence parser) the breadth of coverage is important. This means that the program must be able to handle a great variety of texts, but that it does not usually have to provide a full analysis of the sentences and may make occasional mistakes (in general, a two to four percent error rate is acceptable). For other topics (such as the resolution of specific syntactic ambiguities) the analysis must be complete and accurate. In this case, however, the amount of material that has to be handled is much more limited.

Thus far, there is no easy technique that produces flawless results for the parsing of large text corpora, not even parsing by humans, unless very stringent criteria are adhered to. Reporting on their first experiences with the annotation of a corpus (i.e. the addition of markers that make it easy to retrieve and analyze information about the language), Marcus *et al*. (1993) reported that trained human annotators needed 44 minutes on average to tag 1000 words (hence requiring nearly 100 working days to go through a corpus of one million words) and showed an inter-annotator disagreement of 7.2 percent. Part of the problem is that the syntactic structure of a sentence can become quite complicated when the sentence is long and contains nested structures. Another reason is that a lot of sentences are ambiguous at a purely syntactic level and can have the ambiguity removed only by looking at the meaning of the sentence or the discourse context.

The performance of the annotators in the study of Marcus and coworkers was twice as good (in terms of both speed and accuracy) when the materials were preprocessed by an automatic parser and needed to be corrected only for the mistakes made by the algorithm. As a result of this finding, the annotation of large corpora nowadays is nearly always carried out semi-automatically (that is, the sentences are first parsed by a computer program and the output is then post-edited by humans). In this case, however, care has to be taken to avoid the possibility of annotators being biased by the suggestions of the algorithm. Luckily, owing to the efforts of previous researchers, in many cases it is not necessary to annotate a new corpus. For many research issues, one can make use of an existing corpus that has already been tagged. This is done, for instance, to test linguistic and psycholinguistic hypotheses, or to measure the performance of newly developed software and to provide the training input for these computer programs.

For other research topics, researchers do not need a fully parsed corpus. Often, their question is confined to one particular syntactic structure or to a small set of words, about which they want an in-depth analysis. For such purposes, it is usually feasible (and desirable) to do the analysis by hand. Sometimes this can be achieved simply by scanning the corpus for particular words or combinations of words. An example of this type of research concerns the issue of to which syntactic form

frequencies humans are sensitive when they are parsing sentences.

TYPE FREQUENCY VERSUS TOKEN FREQUENCY

Another issue researchers have to face when they are assessing syntactic form frequencies is how to define the different categories. In frequency counts, there is always the issue of which instances to group and which to separate, because there is rarely a full mapping of verbal forms with theoretical categories. Consider the word 'that'. It can have at least three syntactic functions, as shown below:

He showed the girl that painting. (demonstrative pronoun) (2a)

He showed the girl that he was strong. (complementizer) (2b)

He showed the girl that he had just met that he was strong. (relative pronoun) (2c)

In addition, the syntactic function of 'that' in (2a) overlaps with the function of the words 'this', 'these', and 'those', raising the question whether they have to be grouped or not. This problem is known as the issue of type versus token frequency. Roughly, types refer to theory-based distinctions, whereas tokens refer to the number of occurrences of these types in a corpus (e.g. the number of times 'that' is used as a demonstrative pronoun, as a complementizer, or as a relative pronoun). Mitchell *et al*. (1995) listed some examples of syntactic form frequency distributions that differed significantly depending on how the types had been defined (a phenomenon these authors called the grain-size problem). For instance, in the examples above, it may be that the word 'that' is used much less often as a demonstrative pronoun than as a complementizer after the sequence 'he showed the girl', but the same need not be true for the more general syntactic categories introduced by the word 'that' (i.e. a noun phrase versus a complement clause).

THE USE OF SYNTACTIC FORM FREQUENCIES

Syntactic form frequencies are useful for two purposes. First, it has been shown that automatic sentence parsing algorithms (such as those needed for artificial speech perception) perform better when they take into account not only the syntactic features of the individual words and phrases but also the frequencies of the different syntactic forms given the preceding context. So, when confronted with the sentence 'She put the dress *on the rack*', the algorithm will do a better job in interpreting the ambiguous final phrase 'on the rack' when it takes into account the probability of such a prepositional phrase following the verb 'put' versus the probability of such a phrase following the noun 'dress' (as in 'She saw the dress on the rack'). By taking this probabilistic information into account, the computer may be more likely to come to the correct attachment.

Second, information about form frequencies is important to find out whether humans also make use of this kind of probabilistic information when they are parsing a sentence, and if they do, whether this information is used immediately or in a second, reanalysis stage after the initial analysis has failed. The first theories of human sentence parsing assumed such information did not play a role in the initial syntactic analysis because of the limitations in working memory capacity. More recently, researchers have argued that the build-up of the syntactic structure by the human parser cannot be understood without taking into account this type of information. Still others (e.g. Mitchell *et al*., 1995) accept an influence of syntactic form frequencies, but only at the level of the syntactic structure, not at the level of the individual words that make up the sentence. As this debate is largely based on comparisons of corpus findings with experimental reading data, assessment of syntactic form frequencies has become an important research tool in psycholinguistics as well as in (computational) linguistics.

References

Frazier L and Clifton C Jr (1996) *Construal*. Cambridge, MA: MIT Press.

Marcus MP, Santorini B and Marcinkiewicz MA (1993) Building a large annotated corpus of English: The Penn Treebank. In: Armstrong S (ed.) *Using Large Corpora*. Cambridge, MA: MIT Press.

Martin W, Church K and Patil R (1983) Preliminary analyses of a breadth-first parsing algorithm: theoretical and experimental results. In: Bolc L (ed.) *Natural Language Parsing Systems*. Berlin, Germany: Springer-Verlag.

Mitchell DC, Cuetos F, Corley MMB and Brysbaert M (1995) Exposure-based models of human parsing: evidence for the use of coarse-grained (nonlexical) statistical records. *Journal of Psycholinguistic Research*, **24**: 469–488.

Further Reading

Biber D, Conrad S and Reppen R (1998) *Corpus Linguistics: Investigating language structure and use*. Cambridge, UK: Cambridge University Press.

Jurafsky D and Martin JH (2000) *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Upper Saddle River, NJ: Prentice Hall.
Mitchell DC, Cuetos F, Corley MMB and Brysbaert M (1995) Exposure-based models of human parsing: Evidence for the use of coarse-grained (non-lexical) statistical records. *Journal of Psycholinguistic Research* **24**: 469–488.
Pickering MJ, Traxler MJ and Crocker MW (2000) Ambiguity resolution in sentence processing: evidence against frequency-based accounts. *Journal of Memory and Language* **43**: 447–475.
Tabor W, Juliano C and Tanenhaus MK (1997) Parsing in a dynamical system: an attractor-based account of the interaction of lexical and structural constraints in sentence processing. *Language and Cognitive Processes* **12**: 211–271.
Thomas J and Short M (eds) (1996) *Using Corpora for language research*. London: Longman.

Syntax

Introductory article

Colin Phillips, University of Maryland, College Park, Maryland, USA

CONTENTS

Syntactic theory aims to explain how people combine words to form sentences, and how children attain knowledge of sentence structure.

GOALS OF SYNTACTIC THEORY

Syntactic theory aims to provide an account of how people combine words to form sentences. A common feature of all human languages is that speakers draw upon a finite set of memorized words and morphemes (i.e. minimal meaning-bearing elements) to create a potentially infinite set of sentences. This property of *discrete infinity* allows speakers to express and understand countless novel sentences that have never been uttered before, and hence forms the basis of the creativity of human language. Syntactic theory is concerned with what speakers know about how to form sentences, and how speakers acquire that knowledge.

For example, speakers of English know that 'dogs chase cats' and 'cats chase dogs' are possible sentences of English, but have different meanings. Speakers know that 'chase dogs cats' is not a possible sentence of the language, and that 'cats dogs chase' is possible in specific discourse contexts, as in 'cats, dogs chase, but mice, they flee'. Speakers' knowledge of possible word combinations is often referred to as the *(mental) grammar*.

An accurate model of a speaker's knowledge of his or her language should minimally be able to generate all and only the possible sentences of the language. For this reason, syntactic theory is often known as *generative grammar*. In the 1950s, early attempts by Noam Chomsky and others to create explicit generative grammars quickly revealed that speakers' knowledge of syntax is a good deal more complex than had been anticipated. Research on syntactic theory has relied primarily upon speakers' intuitive judgments about the well-formedness ('grammaticality') of sentences of their language. Since grammaticality judgments can be gathered relatively easily, syntactic theory has amassed a large database of findings about an ever more diverse set of languages.

The complexity of syntactic knowledge sharpens the problem of how language is learned. Research on language acquisition has demonstrated that children know much of the grammar of their language before they are old enough to understand explicit instruction about grammar. Therefore, a primary challenge for syntactic theory has been to understand how a child can learn any language, relatively effortlessly, and without explicit