Syntagma - How it works

The aim

Syntagma is a computer program written in C# and running on Windows computers. Its aim is to take a corpus of natural language text and to make a start on discovering the grammar of that language. It groups together words with similar distributions into classes and seeks to find groups of classes which can be considered structures.

The program is available free of charge on my web site, www.colinday.co.uk/downloads.

The method

The program runs in two phases. Phase 1 reads in the corpus and on the basis of their distribution some words are grouped into classes. Phase 1 may only be executed once.

Phase 2 may be run several times. The distribution of words is recorded with respect to the classes and structures which have been found so far, and as a result of this, some classes may have extra words assigned to them, and new classes and structures may be created.

Phase 1

First all the corpus is read and stored. The words will normally be capitalised (though this is optional), and punctuation removed except for sentence ends.

The number of times each word occurs is found together with the number of times one word is followed by another. A statistical test is made to see whether such collocations are significant. Two words may be linked if they are preceded and followed significantly by the same lists of words.

Each word is first tested to see whether it is sufficiently linked to any existing class, and if it is it is placed within that class. If it does not, it is tested against other words to see whether a new class can be started. When a new class is created, it is tested against all other classes to see whether a merger is possible.

The end result of Phase 1 is the production of a basic set of classes (the ‘seed classes’) which may provide a solid starting point for considering how words collocate with them.

Phase 2

The sentences are scanned, looking for places where items occur bracketed between two classes or structures. Such a bracket is known as a ‘context’.

If within a context only one class appears (frequently enough) then words which also are to be found (frequently enough) within that context may also be reckoned to be members of that class.

Apart from that, words occurring (frequently enough) within that context may be considered to be a new class. This new class is then compared with existing classes, to see whether they can be merged.

Besides the contexts, a test is made for classes and structures occurring side by side often enough. They may then be considered to be structures.

Classes and structures found within one run of Phase 2 may then be part of the discovery process when Phase 2 is run again.

Results

In developing the program a corpus of English text was used, so that the classes found could be examined with some idea of what they should be. English also has the advantage of being only slightly inflected, so for the most part words retain their forms. The corpus used has been the book *Frontier Ways* (1959) by Edward Everett Dale, University of Texas Press. The University of Texas Press have kindly given me permission to distribute this text along with my program. This book has 80,550 words, which has proved quite sufficient for my needs.

When Syntagma is used with this corpus, running Phase 1 and then running Phase 2 five times, the following results are obtained:
6 articles (THE HIS THEIR A AN HIS)
8 verbal auxiliaries (IS WAS WOULD MIGHT HAD MUST WERE COULD)
5 pronouns (THERE IT HE THEY SHE)
4 prepositions (IN FROM BY ON)
537 nouns
278 adjectives
12 structures
37% of the corpus classified, overall time less than 1 second.

Shortcomings

Syntagma counts the co-occurrences of words. It is not suitable for a language which is more than slightly inflected. It might be possible to use it on an inflected language if the morphemes are separated by spaces as if they were words, but I have no experience of this.

When Phase 2 is run many times, spurious classes start appearing. I do not know how to correct this.

I have no idea how Syntagma will fare when used with a corpus of a very different size, or with another language. These things remain to be seen.

Colin Day
2nd September 2018