# Reducing the Complexity of Parsing by a Method of Decomposition

**Caroline Lyon** and **Bob Dickerson***
School of Information Sciences, University of Hertfordshire,UK

A method of automatically locating the subject of a sentence has been developed, and we may be able to take advantage of this to reduce the complexity of parsing English text. Declarative sentences can almost always be segmented into three concatenated sections: *pre-subject - subject - predicate.* The pre-subject segment may be empty; for imperative sentences the subject section is empty. Other constituents, such as clauses, phrases, noun groups, are contained within these segments, but do not normally cross the boundaries between them. Though a constituent in one section may have dependent links to elements in other sections, such as agreement between the head of the subject and the main verb, once the three sections have been located, they can then be partially processed separately, in parallel.

The tripartite segmentation can be produced automatically, using the ALPINE parser. This is a hybrid processor in which neural networks operate within a rule based framework. Readers are invited to access a prototype via telnet, to use on their own text (for details contact the authors).

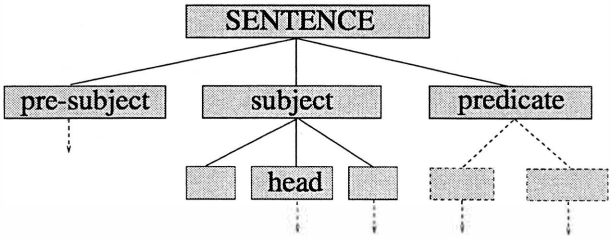

Figure 1: Decomposition of the sentence into syntactic constituents

A sentence is represented like this:

If a cooler is fitted to the gearbox, [ the pipe connections of the cooler ] must be regularly checked for corrosion.

ALPINE has been trained and tested on text of technical manuals ("Perkins"), and also run on other technical manuals ("Dynix" and "Trados"), see Table1. The first parsing step applies the neural processor to tag sentences and place the hypertags marking the subject. For more information see [Lyon and Frank, 1997] and other papers at `ftp://www.cs.herts.ac.uk/pub/caroline`.

This tripartite decomposition of sentences is supported by an argument from information theory [Lyon and Brown, 1997], derived from Shannon's original work with letter sequences [Shannon, 1951]. His ideas can be extended to other linguistic entities. Shannon showed that the entropy of letter sequences declines, the degree of predictability is increased, as information from more adjacent letters is taken into account.

The entropy can also be reduced if an extra character representing a space between words is introduced, producing a 27 letter alphabet. "A word is a cohesive group of letters with strong internal statistical influences" (Shannon), and the introduction of the space captures some of the structure of the letter sequence. A similar technique can be applied to text mapped onto part-of-speech tags.

* email: {C.M.Lyon,R.G.Dickerson}@herts.ac.uk Tel: +44 (0)1707 284266/4355

| Text | Number of sentences | % tags + hypertags correct | % hypertags correct |
|---|---|---|---|
| Perkins | 42 | 92.9 | 100 |
| | 59 | 91.4 | 100 |
| | 63 | 89.2 | 100 |
| | 67 | 84.4 | 95.5 |
| Dynix | 114 | - | 92.1 |
| Trados | 134 | - | 85.1 |

Table 1: Results on declarative sentences in technical manuals. From test data in Perkins corpus (2% sentences omitted) used for development, and from other texts.

Now, language can be represented at a primary level as a regular grammar, and we can apply Shannon's analysis to tag sequences. However, this is an inadequate representation: the patterns of tag sequences may be disrupted at clause and phrase boundaries. "Unlikely" tag combinations such as *noun - pronoun* and *verb - auxiliary verb* may occur at constituent boundaries in sentences like *The shirt he wants is in the wash.*

In a similar manner to the insertion of a space between words, the embedded clause is delimited by inserting hypertags, like virtual punctuation marks. The sentence is represented as

The shirt [ he wants ] is in the wash.

The part-of-speech tags have relationships with adjacent hypertags in the same way that they do with each other. Using this representation, one level of embedding has been modelled. We can thus represent sequences higher in the Chomsky hierarchy than regular grammars, though not fully context free.

Since the boundaries of clauses and phrases often coincide with the boundary of the subject, we expect that the insertion of hypertags to demarcate the subject will lower the entropy. If their insertion has captured some of the structure of language the bipos and tripos entropy should be reduced. This is indeed what was found on the data from the Perkins corpus. See Table 2. A tagset of 32 was used, including the hypertags.

| | $H_0$ | $H_1$ | $H_2$ | $H_3$ |
|---|---|---|---|---|
| plain tag string | 5.0 | 3.962 | 2.659 | 2.132 |
| tags + subject markers | 5.0 | 4.135 | 2.472 | 1.997 |

Table 2: Entropy measures for text with and without subject boundary markers. $H_n$ is the entropy when information is taken from $n$ adjacent tags.

This type of system could be used as a pre-processor to facilitate the processing of longer sentences by other NLP methods. Though there are arbitrary limits on the length of constituents that can be processed by ALPINE (15 words in the pre-subject, 12 words in the subject), these bounds are comparatively wide, and of course we plan to extend them. In the Trados data 9 sentences out of 134 fell outside these limits.

One of the advantages of this approach to parsing is that it lends itself to the extraction of predicate/argument structure. After the subject has been located the main verb will be found in the predicate, and then the object or complement. With the head of the subject found, we then have the raw material from which we can begin to extract the predicate/argument structure.

# References

[Lyon and Brown, 1997] C Lyon and S Brown. Evaluating Parsing Schemes with Entropy Indicators. In *MOL5, 5th Meeting on the Mathematics of Language*, August 1997.

[Lyon and Frank, 1997] C Lyon and R Frank. Using Single Layer Networks for Discrete, Sequential Data: an Example from Natural Language Processing. *Neural Computing Applications*, 5 (4), 1997. To appear.

[Shannon, 1951] C E Shannon. Prediction and Entropy of Printed English. *Bell System Technical Journal*, pages 50–64, 1951.