

ВОРОНЕЖСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ  
ФАКУЛЬТЕТ РОМАНО-GERMANСКОЙ ФИЛОЛОГИИ  
НАУЧНО-МЕТОДИЧЕСКИЙ ЦЕНТР  
ПО КОМПЬЮТЕРНОЙ ЛИНГВИСТИКЕ

**ПРОБЛЕМЫ  
КОМПЬЮТЕРНОЙ  
ЛИНГВИСТИКИ**

Выпуск 3

Воронеж – 2008

Fred Karlsson  
(Helsinki, Finland)

### Implications for computational syntax of restrictions on clausal embedding complexity

Complex sentences are due to repeated embedding of clauses in initial, center-, or final embedding position. Initially embedded clauses (IEs) have either nothing to their left, or at the most a subjunction or conjunction of the superordinate (S-)clause. Center-embedded clauses (CEs) have S-clause constituents to their left and right. Final embeddings (FEs) have no material rightwards that immediately belongs to their S-clause.

The mainstream view holds that clausal embedding is grammar-wise unconstrained and fully recursive (e.g. Noam Chomsky, Randolph Quirk et al., Steven Pinker). Thus, (1) with triple IE and (2) with quadruple CE would be grammatical even if not acceptable.

(1) *If if if it rains it pours I get depressed I should get help.*  
(Pinker)

(2) *John whom June whom Paul whom Jean whom Dick hates adores prefers detests loves Mary.* (Bar-Hillel)

In two recent papers (Karlsson 2007a, b) I have demonstrated that there are indeed CONSTRAINTS on clausal embedding complexity, i.e. quantitative limits and other restrictions.

My data were: (i) complex sentences in the British National Corpus (BNC), the Brown corpus, and the LOB corpus. (ii) Computerized extraction of embedding patterns in Finnish, German, and Swedish. (iii) Consultation of 100+ corpus-based syntactic and stylistic descriptions of European languages, especially Latin and older variants of German, both well-known for heights of syntactic complexity.

Here are some constraints operational in 'Standard Average European' (SAE) languages like English, Finnish, German, Latin, and Swedish:

(3)  $I^2_{MAX}$ : the maximal degree of IE is 2 (100 instances found, as in (8)).

(4)  $QUALITATIVE I^2$ -CONSTRAINTS:  $I^2$  strongly prefers a) written language, b) an if-clause as I-high, c) a sentential subject, i.e. a what-clause as I-low, d) finiteness. Cf. (8)).

(5)  $C^3_{MAX-W}$ : in written language, the maximal (utterly rare) degree of multiple CE is 3 (13 instances retrieved, (9)).

(6)  $C^2_{MAX-S}$ : in spoken language, the maximal (utterly rare) degree of multiple CE is 2 (less than 5 instances retrieved, (10)).

(7)  $ONLY$ - $ADNOMINAL$ - $SELF$ - $EMBEDDING$ : only adnominal clauses (i.e. typically relative clauses, marginally postmodifying that-clauses) allow self-embedding (9, 10), where 'self-embedding' means repeated embedding of the same clause type.

(8) [<sub>M</sub> [<sub>I-1</sub> If [<sub>I-2</sub> what he saw through security] did not impress him] Tammuz ... ] (BNC)

(9) [<sub>M</sub> Der Ritter von Malzahn, [<sub>C-1</sub> dem der Junker sich als einen Fremden, [<sub>C-2</sub> der bei seiner Durchreise den seltsamen Mann, [<sub>C-3</sub> den er mit sich führe,] in Augenschein zu nehmen wünschte,] vorstellte,] nötigte ihn ...] (von Kleist, Michael Kohlhaas)

(10) [<sub>M</sub> A lot of the housing [<sub>C-1</sub> that the people [<sub>C-2</sub> that worked in New Haven] lived in] was back that way.]

No genuine  $I^3$ 's nor  $C^4$ 's are on record; (1, 2) were obviously fabricated. Only some 130 double CEs were retrieved among the tens of millions of words checked. The overall rarity and constrainedness of multiple IE and CE prompt conclusion (11):

(11) Multiple clausal initial and center-embedding are not fully recursive in SAE.

FE is less constrained but this does not overturn (11). FE is  $TAIL$ - $RECURSION$  which is always convertible to  $ITERATION$ . (11) is

amplified by the fact that it is absent from spoken language.

(12) Multiple center-embedding is not fully recursive in SAE. It was unimportant in the past.

Rather, they developed in written language and especially in rhetoric and stylistics. Their followers in written language are bound to the notion PERIODICITY, of which is clausal center-embedding brought to completion. Center-embedding is important in written language.

These restrictions have important implications for syntax. It is well known that there are no recursive formal power series in the formal power series must be at least countable. The need of formal languages of unbounded center-embedding is the need of unlimited recursion.

If and only if there is no center-embedding (and a limited formal power series) on the Chomsky hierarchy is simpler than mainstream syntactic parsers. More than the formal power series makes possible the construction of a parser that by far surpasses the current state of the art.

1. Karlsson, F. Center-embedding of clausal

amplified by the fact that multiple IE and CE are almost totally absent from spoken language. Therefore:

(12) *Multiple IE and CE (unconstrained clausal recursion) were unimportant when spoken language emerged.*

Rather, they developed along with the advent of written language and especially with the development of Greek and Latin rhetoric and stylistics, as influenced by Aristotle, Cicero, Livy and their followers in Western writing traditions. In particular, this led to the notion PERIODIC SENTENCE STRUCTURE the central concept of which is clausal center-embedding. A well-formed 'period' is brought to completion when the main clause is resumed after a center-embedding has been completed.

These restrictions on clausal embedding complexity have very important implications for interpreting the nature of computational syntax. It is well known that if the mainstream view would be true (that there are no restrictions on clausal embedding complexity), the formal power of a computational-syntactic model of syntax must be at least context-free (level 2) on the Chomsky hierarchy of formal languages. The main reason is the purported existence of unbounded center-embedding, which leads to the theoretical need of unlimited memory resources.

*If and only if there is a clear restriction on clausal center-embedding (and above I have demonstrated that there is one), the formal power of computational syntax is finite-state* (level 1 on the Chomsky hierarchy). This means that syntax is formally simpler than mainstream theorizing holds, and it also implies that syntactic parsers can be built and implemented that require no more than the formal power of finite-state devices. This also makes possible the use of well understood, efficient algorithms that by far surpass the parsers of context-free types.

#### References

1. Karlsson, Fred. (2007a). Constraints on multiple initial embedding of clauses.