Journal of Linguistics 43 (2), 2007, 365-392.

Constraints on multiple center-embedding of clauses[1]

FRED KARLSSON

University of Helsinki

(Received 6 August 2006; revised 15 December 2006)

A common view in theoretical syntax and computational linguistics holds that there are no grammatical restrictions on multiple center-embedding of clauses. Syntax would thus be characterized by unbounded recursion.

An analysis of 119 genuine multiple clausal center-embeddings from seven 'Standard Average European' languages (English, Finnish, French, German, Latin, Swedish, Danish) uncovers usage-based regularities, constraints, that run counter to these and several other widely held views, such as that any type of multiple self-embedding (of the same clause type) would be possible, or that self-embedding would be more complex than multiple center-embedding of different clause types.

The maximal degree of center-embedding in written language is three. In spoken language, multiple center-embedding is practically absent. Typical center-embeddings of any degree involve relative clauses specifying the referent of the subject NP of the superordinate clause. Only postmodifying clauses, especially relative clauses and *that*-clauses acting as noun complements, allow central self-embedding. Double relativization of objects (*The rat the cat the dog chased killed ate the malt*) does not occur.

These corpus-based 'soft constraints' suggest that full-blown recursion creating multiple clausal center-embedding is not a central design feature of language in use. Multiple center-embedding emerged with the advent of written language, especially with Aristotle, Cicero, Livy and others in the long Greek and Latin stylistic tradition of 'periodic' sentence composition.

1. INTRODUCTION

This paper deals with constraints on multiple center-embedding of clauses in seven European V(erb) – O(bject) languages. Our CONSTRAINTS are usage-based quantitative limits and combinatory restrictions, often of the nature of tendencies, intermediate between grammatical rules and discourse regularities, closely related to the 'soft constraints' of Du Bois (1987, 2003a, b).

Center-embeddings have figured prominently for decades in linguistic argumentation, especially in theoretical syntax starting with Chomsky (1956, 1957), in psycholinguistics (e.g. Miller & Isard 1964), and in mathematical and computational linguistics (e.g. Partee et al. 1993: 477–480). A theoretical claim common to all these approaches is that there are no grammatical restrictions on the degree of multiple center-embedding.

Only few empirical corpus data have been adduced, with Dryer (1980), De Roeck et al. (1982), and Sampson (1996) as notable exceptions. In a submission to The LINGUIST List on December 4, 1995, Richard Hudson aptly remarked concerning center-embedding: 'There is no shortage of explanations of the "facts", but there does seem to be a shortage of well-established facts to be explained'. Today the situation is the same. In this paper, however, we shall strengthen the empirical foundation of the issue by adducing corpus data from several Standard Average European (SAE) languages: English, Finnish, French, German, Latin, Swedish, and Danish.

The systematically collected data were derived from the tagged machine-readable corpora Brown Corpus of American English (1 million words) and LOB (Lancaster-Oslo/Bergen) Corpus of British English (1 million words). This material was supplemented by less systematic computerized searches for similar patterns in machine-readable materials of the other languages mentioned, by naturalistic observation, and by consultation of some of the copious descriptive data accumulated over the centuries in syntactic and stylistic descriptions of SAE languages, especially Latin and older variants of German, both well-known for having reached heights of syntactic complexity.

The complexity of multiple center-embedding is important because this is the one and only syntactic factor deciding whether natural language syntax is of type 2 (context-free) or type 3 (finite-state) in the Chomsky hierarchy of languages. If and only if there is a limit on multiple center-embedding, syntax is of type 3. The theoretical issue is not conclusively resolvable by corpus-linguistic induction alone, but an empirically detected limit would at least guarantee the feasibility of finite-state parsing of running SAE text.

2. CONCEPTS

The notion EMBEDDING refers to all types of clauses occurring as subordinate parts of their superordinate clauses (which themselves may be either main or subordinate). The starting point will be the classical view of subordination as expounded in Quirk et al. (1989, Chapter 14). Typical finite SUB-CLAUSES are of three types: complement, relative, and adverbial. They are indicated by subordinators or relative pronouns, henceforth called SUB/WH-ELEMENTS.

Schema (1) covers typical English SUPERORDINATE CLAUSES which can be of three types: topmost main clause, subordinate clause, and a member clause of a coordinate sentence. The optional SUB/COORD in (1) stands for subordinators such as *because*, *if*, *that*, *when*, and coordinators like *and*, *but*. The variables X, Y denote any other superordinate clause constituents. Thus, pattern 'X Y' covers simplex main clauses, pattern 'SUB/COORD X Y' subordinate and coordinate superordinate clauses:

(1) (SUB/COORD) X Y

↑ ↑ ↑

initial embedding center-embedding final embedding

The three embedding positions, illustrated in (2), can now be defined across all types of superordinate clauses.

(2)

If	I-1	finite initial embedding, depth 1
as often happened	I-2	finite initial embedding, depth 2
she asked him	I-1	continues
to tell her about it	f-2	non-finite final embedding, depth 2
she thought	M	main clause, depth 0
that he	F-1	finite final embedding, depth 1
who had been so kind	C-2	finite center-embedding, depth 2
would understand.	F-1	continues

An INITIALLY-EMBEDDED CLAUSE (abbreviated 'I' if finite, and 'i' if non-finite) occurs either before all words of its superordinate clause (as with clause I-1 in (2), occurring before the main clause *she thought*; the number in an expression like 'I-1' indicates depth of embedding) or directly after the initial subordinator or coordinator of its superordinate clause (as with I-2 in (2), occurring after the initial subordinator *if* of its superordinate clause I-1). We follow Quirk et al. (1989: 1037) in interpreting clauses like I-2, embedded immediately after a subordinator or coordinator, as initially-embedded rather than as center-embedded (subordinators and coordinators are not as tightly integrated in their clauses as ordinary full constituents are).

CENTER-EMBEDDED CLAUSES (abbreviated 'C', 'c') have words of the superordinate clause both to their left (excluding subordinators and coordinators) and to their right, as C-2 in (2) (*he* C-2 *would understand*). SELF-EMBEDDING is multiple center-embedding of the same type of clause, e.g. two relative clauses as in (7d).

FINALLY-EMBEDDED CLAUSES (abbreviated 'F', 'f') occur after the last word of the superordinate clause, e.g. f-2 and F-1 in (2).

The DEGREE of initial, center-, or final embedding of a sequence of embedded clauses is the number of instances of that type of embedding found in the sequence. The degree of initial embedding in (2) is 2. Degrees are abbreviated by exponents: I^2 (double initial embedding), C^3 (triple center-embedding). MULTIPLE embeddings are embeddings of a degree greater than 1.

The DEPTH OF A CLAUSE is its level relative to the main clause, e.g. I-2 in (2) is a (finite initial) embedding at depth 2. The main clause is always at depth 0. In schemas like (2), progressive indentation reflects increasing depth.

3. STATE OF THE ART

The mainstream view is that there are no grammatical restrictions on clausal embedding complexity in any sentential position. This opinion has been voiced by many linguists from different camps: the comparatist Meillet (1934: 355), the generativist Chomsky (1956: 65), the historical linguist Admoni (1980: 23), the descriptive grammarians Quirk et al. (1989: 44), and writers of textbooks (Akmajian et al. 1985: 163) and overviews (Langendoen 1998: 239).

Chomsky (1956: 65) explicitly conjectured that there are no grammatical restrictions on the degree of center-embedding. This HYPOTHESIS OF UNBOUNDED CENTER-EMBEDDING COMPLEXITY is henceforth called the UCE-HYPOTHESIS. Here is how Chomsky and Miller (1963: 286–287) back it up:

> ... [(3)] is surely confusing and improbable but it is perfectly grammatical and has a clear and unambiguous meaning. To illustrate more fully the complexities that must in principle be accounted for by a real grammar of a natural language, consider [(4)], a perfectly well-formed sentence with a clear and unambiguous meaning, and a grammar of English must be able to account for it if the grammar is to have any psychological relevance.

(3) The rat the cat the dog chased killed ate the malt.

(4) Anyone who feels that if so-many more students whom we haven't actually admitted are sitting in on the course than ones we have that the room had to be changed, then probably auditors will have to be excluded, is likely to agree that the curriculum needs revision.

The UCE-hypothesis relies on the generative distinctions competence / performance and grammaticality / acceptability, where competence and grammaticality are assumed to be unaffected by any kind of processing limitations. Methodologically, the UCE-hypothesis rests on intuition-based claims concerning the grammaticality of data fabricated by the linguists themselves, like (3)–(4).

Miller & Isard (1964: 293) claimed that (5a, b) are equally grammatical; any preference for (5b) over (5a) will have a psychological rather than a linguistic explanation. Bar-Hillel (1964: 199) used (5c), a quadruple center-embedding, C^4, with four objects relativized. (5d) used by Frazier and Rayner (1988: 264) is C^3 with three relativized objects, as is Pinker's (1994: 205–206) (5e).

(5) (a) It is more likely that the man who said that a cat that the dog that the boy owns chased killed the rat is a liar than not.

(b) It is more likely than not that the man is a liar who said that the rat was killed by a cat that was chased by the dog that is owned by the boy.

(c) John whom June whom Paul whom Jean whom Dick hates adores prefers detests loves Mary.

(d) Men women children dogs bit like marry hate pets.

(e) The rapidity that the motion that the wing that the hummingbird has has has is remarkable.

The UCE-hypothesis as a presumed characteristic of grammatical competence is also found in numerous textbooks, glossaries, and encyclopedias. Crystal (2003: 407) states it using (6) as evidence.

(6) ?The dog that the cat that the man bought scratched ran away.

According to Greene (1972: 26) it would be arbitrary to say that embeddings can only be carried out three times. Pulman (1986: 204) surmised a limit of 'say, ten centre embeddings'. Most of the authors cited, Chomsky included, state that in performance the acceptability of multiple center-embeddings

degrades with each cycle of embedding (as demonstrated by Miller & Isard 1964).

The UCE-hypothesis was challenged by Reich (1969), who claimed (without empirical evidence) that the maximal degree of center-embedding in English is 1; we shall call this hypothesis C^1MAX. However, C^1max was falsified by De Roeck et al. (1982), who adduced several genuine examples of multiple center-embedding, concluding that their data support the UCE-hypothesis. Sampson (1996) provided more corpus examples and denied the existence of a clear demarcation line between C^2 and C^3.

Psycholinguistic work by Miller & Isard (1964) and Lewis (1996), as well as connectionist modeling by Christiansen and Chater (1999, 2001), point towards C^2 and/or C^3 as potential limits but no conclusive upper limit has yet been empirically motivated on purely linguistic grounds.

Here, we intend to demonstrate that there is a precise empirical constraint on the maximal degree of center-embedding: 3.

4. EMPIRICAL DATA

Five major data sources were used.

(i) The Brown and the LOB corpora were systematically searched for multiple center-embeddings. The criterion used for spotting potential extreme embedding complexity was the number of sub/wh-elements in a sentence. All sentences with four or more sub/wh-elements (N = 2260) were automatically extracted from Brown and LOB and then manually analyzed.

(ii) The thirty genuine English, German, and Latin examples of multiple center-embedding cited by De Roeck et al. (1982) and Sampson (1996) were analyzed in detail.

(iii) A manual analysis was made of 6000 sentences by three 19th century scholars known for their intricate and syntactically complex language use (Jeremy Bentham, John Stuart Mill, C. S. Peirce).

(iv) More than one hundred descriptive, stylistic, and diachronic papers, grammars, and style manuals especially of Latin and German were consulted with a view to finding data on complex center-embeddings, e.g. Kriebel (1873) and Nägelsbach (1963 [1846]) for Latin, and Blatz (1896), Engel (1922), Hoffmann-Krayer (1925), and Olzien (1933) for German.

(v) Naturalistic observation of newspapers and books over the years.

The 119 instances of multiple center-embeddings thus retrieved are

documented, analyzed in detail, and made available as Karlsson (2007c), the analyzed quantitative data as Karlsson (2007d).

In Brown and LOB the most complex center-embeddings retrieved were eighteen instances of C^2, ten in Brown and eight in LOB: examples (24), (34)–(49), (96) in Karlsson (2007c). (7) presents a handful of them, and (8) three of Jeremy Bentham's (1945: 65, 121, 197) C^2s:

(7) (a) [$_{M}$ ... the girl ... [$_{C\text{-}1}$ who was clothed in the tightest-fitting pair of slacks [$_{C\text{-}2}$ I had ever seen on a woman] and a sweater [$_{F\text{-}2}$ that showed everything [$_{F\text{-}3}$ there was]]] wanted to be sociable.] (Brown)

(b) [$_{M}$ It was not [$_{F\text{-}1}$ until he was an old man [$_{F\text{-}2}$ that one day his son, [$_{C\text{-}3}$ who, [$_{C\text{-}4}$ as was the way of the world,] had left the shamba] explained to him [$_{F\text{-}3}$ that ...]]]] (LOB)

(c) [$_{M}$ A student [$_{C\text{-}1}$ who [$_{C\text{-}2}$ while ∅ in attendance at Carleton College] participates in an athletic contest during the school year,] ... shall be permanently ineligible to ...] (Brown)

(d) [$_{M}$ And yet a widow, [$_{C\text{-}1}$ whose pension, [$_{C\text{-}2}$ for which her husband paid,] is wiped out [$_{F\text{-}2}$ because she works for a living wage,]] will now have [$_{f\text{-}1}$ to pay for her spectacles.]] (LOB)

(e) [$_{M}$ At one point in the game [$_{C\text{-}1}$ when the skinny old man in suspenders [$_{C\text{-}2}$ who was acting as umpire] got in the way of a thrown ball] [$_{\&C\text{-}1}$ and took it painfully in the kidneys,] he lay there ...] (Brown)

(f) [$_{M}$ But the idea [$_{C\text{-}1}$ that the fact [$_{C\text{-}2}$ that some pain is heading my way] gives me no special reason [$_{f\text{-}2}$ to avoid it]] seems so at odds with ...] (Internet)

(8) (a) [$_{M}$ And in particular [$_{C\text{-}1}$ when the motives [$_{C\text{-}2}$ which are applied] are of the nature of those [$_{F\text{-}2}$ which result from a change [$_{F\text{-}3}$ made in the condition of the body,]]] the power may be said ...]

(b) [$_{M}$ For an analysis of the possible modifications [$_{C\text{-}1}$ of which the pathological termination of an act [$_{C\text{-}2}$ which is not according to law] are susceptible] we have therefore ...]

(c) [$_{M}$ These are the cases [$_{F\text{-}1}$ where ... the power [$_{C\text{-}2}$ from whom the magistrate [$_{C\text{-}3}$ by whom the commands in question are issued] take their official name,] ... comes ...]]

Is C^2 the maximal degree of center-embedding? No. Older sources and naturalistic observation disclosed thirteen instances of C^3, in Latin, German, English, Swedish, and Danish, of which seven are presented here (9)–(15), and the rest in Karlsson (2007c, d) along with detailed analyses of all:

(9) [$_{M}$ Er ... war allen Gefahren ... [$_{C\text{-}1}$ welche ein jeder, [$_{C\text{-}2}$ der diese
he was all dangers which anybody who these
wilde Gegend zu jener Zeit, [$_{C\text{-}3}$ als diese Geschichte dort
wild surroundings at that early date as this story there
spielte,] durchstreifte,] gewärtig sein mußte,] gewachsen.]
took place wandered through prepared be had to up to
'He was up to all dangers that anybody had to be prepared to cope with that wandered through these wild surroundings at that early date.' (Engel 1922: 328)

(10) [$_{M}$ Indessen muß man den Mißbrauch [$_{C\text{-}1}$ der ... in der Offensive,
meanwhile must one the malpractice which in the offensive
von dem Laufschritt, [$_{C\text{-}2}$ der, [$_{C\text{-}3}$ sobald die Schützenkette
at the double quick which as soon as the riflemen
formiert ist, und selbst unter Umständen, [$_{F\text{-}4}$ wo er gar nicht
grouped are and even in circumstances where it at all not
angezeigt ist,]] vorgeschrieben zu sein scheint,] gemacht wird,]
appropriate is required to be seems done is
tadeln.]
criticize
'Meanwhile one must criticize the malpractice which in the offensive seems to be required at the double quick as soon as the riflemen have grouped themselves and even in circumstances where it is not at all appropriate.' (*Deutsche Heereszeitung* 1893; Engel 1922: 333)

(11) [$_{M}$ Der Ritter von Malzahn, [$_{C\text{-}1}$ dem der Junker sich als einen
The rider from Malzahn whom the Junker himself as a
Fremden, [$_{C\text{-}2}$ der bei seiner Durchreise den seltsamen Mann,
stranger who during his journey the strange man
[$_{C\text{-}3}$ den er mit sich führe,] in Augenschein zu nehmen
whom he with him would bring a look to have
wünschte,] vorstellte,] nötigte ihn ...]
wanted introduced urged him

'The rider from Malzahn to whom the Junker had introduced himself as s stranger who during his journey through wanted to have a look at the strange man whom he would bring with him urged him …' (H. von Kleist, *Michael Kohlhaas*; Schneider 1959: 469)

(12) [$_{M}$ Der Landvogt ... fand, [$_{C\text{-}1}$ als er, [$_{C\text{-}2}$ von dem, [$_{C\text{-}3}$ was
The governor found as he about that which
vorgefallen,] benachrichtigt,] in bestürzten Märschen zurückkehrte,]
had happened had been told in forced marches returned
die Stadt in allgemeinen Aufruhr.]
the city in general turmoil
'The governor found the city in a state of general turmoil as he returned in forced marches having been informed about what had happened.' (H. von Kleist, *Michael Kohlhaas*; Hoffmann-Krayer 1925: 131)

(13) [$_{M}$ In an excellent article ... Salvini draws a parallel between the way [$_{C\text{-}1}$ in which the spoken Latin of the men [$_{C\text{-}2}$ with whom Gregory of Tours, [$_{C\text{-}3}$ whom he has no reason [$_{f\text{-}4}$ to mention,]] must have mixed] eventually became Old French ...,] and the comparable direct development of pre-Romanesque painting ...] (L. Thorpe, *Gregory of Tours: The History of the Franks*, 1974: 39; due to Geoffrey Sampson)

(14) [$_{M}$ ... the Prime Minister [$_{C\text{-}1}$ who at the height of the crisis had snapped to a junior minister [$_{C\text{-}2}$ who, [$_{C\text{-}3}$ not having seen him for some time,] had approached him in a Westminster corridor with a view to [$_{f\text{-}3}$ wishing him luck …,]] 'If you want to resign, put it in writing',] was unlikely to ...] (Patrick Cosgrave 1979; De Roeck et al. 1982: 338)

(15) [$_{M}$ A person [$_{C\text{-}1}$ who, [$_{C\text{-}2}$ when riding a cycle, [$_{C\text{-}3}$ not being a motor vehicle,] on a road or other public place,] is unfit to ride through drinks or drugs,] shall be guilty of an offence.] (*British Road Traffic Act*, 1972; Hiltunen 1984: 115)

Most of these thirteen C^3s are very convoluted and low on any scale of acceptability, e.g. beyond the acceptability limit formulated by Gibson (1998: 41) in his theory of syntactic processing complexity (rough limit: being forced to retain two or more unresolved syntactic dependencies over three or more new referents or main verbs). The simplest C^3 in our corpus is (12). This is one of the few not violating Gibson's acceptability limit because it has few new discourse referents, many pronouns, and short member clauses.

Even a superficial inspection of the internal structure of the C^3s discloses a clustering around certain clause types and configurations, in particular relative clauses; all of the six center-embedded clauses in (11) and (13) are relatives. But thirteen instances is too small a corpus for an analysis of the qualitative restrictions on multiple center-embedding. As the same tendencies are at work in C^2s, the analysis in section 5 below will be performed on the ensemble of C^2s (N = 106) and C^3s (thirteen instances, each containing two C^2s, thus N(C^2) = 26), yielding a sum total of 132 C^2s. This is a reasonable corpus for establishing basic structural tendencies over a few central syntactic variables.

Of the 132 C^2s one was found by Reich and Dell (1976), ten by De Roeck et al. (1982), eleven by Sampson (1996), two by Korthals (2001), four by Geoffrey Sampson (personal communication), one by Terttu Nevalainen (personal communication), and 103 by myself. The corpus is subdivided over language (English, German, Latin, Swedish, Finnish, French, Danish), mode (w = written, s = spoken), and time (see table 1):

Source		Eng	Ger	Lat	Swe	Fin	Fre	Dan	Total
Classical Latin	w			9					9
other pre-19th century	w	1	2						3
19th century	w	10	17		4			2	33
Brown, 20th century	w	10							10
LOB, 20th century	w	8							8
other 20th century	w	20	3		9	4	1	1	38
21st century	w	16	2		8	2			28
20th century	s	3							3
Total		68	24	9	21	6	1	3	132

Table 1.
Composition of the C^2-corpus over language, time, and mode.

All thirteen C^3s are from written language. Of the 132 C^2s only three, namely(14), (50) and (51) in Karlsson (2007c) are from spoken language. Multiple center-embeddings are extremely rare in genuine speech.

5. INCIDENCE OF CENTER-EMBEDDINGS

No data on the incidence of CEs are available. For obtaining an overview across genres of the incidence of initially-, finally-, and center-embedded clauses, finite and non-finite ones included, a systematic balanced sub-sample was generated of the Brown corpus containing every 110th graphical sentence (N = 495), of which 283 (57%) contained at least one embedded

clause. The incidence of clauses in the initial, center- and final embedding positions (IE, CE and FE, respectively) was then analyzed manually.

Embedded clauses	All		Finite		Non-finite	
	N	%	N	%	N	%
IE	55	13	44	16	11	8
CE	46	11	41	15	5	4
FE	314	76	194	70	120	88
Sum	415	100	279	101	136	100
Inserted	15					
Sum	430					
CE relative			30	65		
Other CE			16	35		
Sum			46	100		
IE relative			2	2		
CE relative			30	24		
FE relative			91	74		
Total			123	100		

Table 2.
Incidence of embedded clauses in a systematic balanced sub-sample of the Brown Corpus (495 sentences).

As is shown in table 2, 76% of all embedded clauses are FEs, as are almost 90% of the non-finite ones (most of which are infinitives). The overall share of CEs is about one tenth of all embedded clauses, or 15% if only finite clauses are counted. CE non-finite clauses such as the deepest one in (15) are uncommon, 4%.

Of the 46 CE clauses no fewer than 30 (65%) are relative. Hakulinen, Karlsson & Vilkuna (1980: 118) report that in the 10,000-clause HKV-corpus of written Finnish there were 140 finite CE sub-clauses of which 98 (70%) were relative. For obtaining data on CE in spoken English, I analyzed the Pear Story material of Chafe (1980: 301–319): of 45 CE clauses, 34 (76%) were relative. I therefore generalize across genres:

(16) The typical center-embedded clause is a relative clause.

At least in VO-languages, FEs are much more frequent than IEs, while the

latter are more frequent than CEs. This order is the same as Dryer's (1980: 126) universal Sentential NP Position Hierarchy, which defines the preferred positions for sentential NPs.

Ordinary relative clauses (and other adnominal clauses) modify their heads, i.e. are in endocentric construction, while the head functions as a constituent of the main clause. All other center-embedded clauses (*if*-, *when*-clauses, etc.) are direct main clause constituents, mainly adverbials, whereas *that*-clauses mostly are objects. Thus, adnominal clauses do not interrupt the identification of the main clause constituents as abruptly as other CE clause types do (Grosu and Thompson 1977: 144). From the main clause perspective, it is enough to determine the syntactic function of the head of the endocentric relative construction. This difference in processing ease explains the prevalence of relative clauses in CE position.

The incidence of 46 C^1s is 9% in relation to the 495 sentences. Thus, in running English text roughly every tenth sentence contains a simplex center-embedding, C^1, and of these around 70% are relative clauses. Now consider multiple center-embeddings. Our partly computerized search procedure could not spot all instances of reduced relative clauses without pronouns (*the book I BOUGHT*), and the eighteen C^2s in Brown + LOB reported in section 4 above is therefore too low. A reasonable estimate could be 25 C^2s. As the number of sentences in Brown + LOB is altogether about 110,000, the sentential incidence of C^2s would be some 0.02 %, i.e. one C^2 per 5000 sentences or 90,000 words, or twelve per one million words. De Haan (1989: 185) found the same incidence of C^2s in the Nijmegen corpus of written English (130,000 words).

Korthals (2001: 183) reports three doubly relative C^2s in the Negra Corpus of German newspaper text (355,096 words; 20,602 sentences). If German is like English in that around 40% of the C^2s are doubly relative (see table 4 below), the total number of C^2s in Negra should be eight, or twenty per one million words, almost double the share of C^2s in Brown + LOB.

To check a genre-specific and idiolectal extreme incidence of multiple center-embeddings in English, I read 100 pages (1500 sentences) of Jeremy Bentham's *The Limits of Jurisprudence Defined* (1945), known for its syntactic complexity. The most complex center-embeddings found were eight C^2s, a sentential incidence 25 times higher than that in Brown + LOB, .5% (one per 200 sentences).

Extrapolating from the 46 C^1s in 495 sentences in our Brown subcorpus, an estimate of 460 C^1s in 5000 sentences is obtained, as compared to one C^2. The rate of occurrence of C^2s in relation to C^1s would then be 1/460 or 0.02%. If increases in center-embedding complexity follow a smooth curve, 0.02% would predict one C^3 in 2,300,000 sentences and 42,000,000 words. This suggests there could be ten C^3s in the Bank of English, whose present size is 500,000,000 words.

6. THE CONSTRAINTS

The absence of C^3s in Brown and LOB means that C^3 is practically non-existent in modern English. The thirteen instances of C^3 come from the ensemble of Western writing and philological scholarship through the ages. Given this enormous universe, the incidence of C^3 is close to zero in written language and equal to zero in spoken language. But the existence of the C^3s cannot be denied; also note that Hagège (1976) reports C^3s in the Niger-Congo language Mbum. No genuine C^4s have ever been adduced. These empirical observations underlie the quantitative constraint C^3MAX-W (where '-w' refers to written language):

(17) *C^3max-w constraint*
The maximal degree of multiple center-embedding is three in written language.

C^3max-w is a hypothesis falsifiable by contrary evidence. Engel (1922) presents a potential German counterexample of no less than C^5:

(18)

Derjenige,	M
anybody	
der denjenigen,	C-1
who the one	
der den Pfahl,	C-2
who the pole	
der an der Brücke,	C-3
which on the bridge	
die auf dem Wege,	C-4
which on the road	
der nach Worms führt,	C-5
which to Worms leads	
liegt,	
lies	
steht,	
stands	
umgeworfen,	
overthrown	
anzeigt,	
reveals	
erhält eine Belohnung.	
gets a reward	

'Anybody gets a reward who reveals the person who overthrew the pole standing on the bridge lying on the road leading to Worms.' (no source;

Engel 1922: 328)

Although Engel normally gives the source of his data, he does not in this case. (18) does look artificial. Center-embeddings typically occur at the end of the grammatical subject (see table 5 below), but the antecedents of C-2 through C-5 are objects or adverbials. The double object relativization in C-1 and C-2 violates constraint (29) below. The dubious nature of (18) is corroborated by the fact that Blatz (1896: 1274), without giving a precise source as he otherwise does, gives a version of (18) that has 'only' C^3, and Drach (1963: 46) provides a variant with C^4. (18) is just playing with language.

A sentence verging on C^4 is Cicero's (19):

(19)
Postea vos, patres conscripti, ... M
then you Senators
huic furiae, C-1
from this madman
si diutius in hac urbe, C-2
if longer in this city
quam delere I-4
which destroy
cuperet, C-3
wanted
maneret,
would stay
vox interdiceretur,
vote cancel
decrevistis ...
decided
'After this you, Senators, decided to cancel this man's right to vote if he would stay longer in this city which he wanted to destroy.' (Cicero, *De haruspicum responsis*; Nägelsbach 1963: 645)

However, the deepest clause, I-4 *quam delere*, is initially-embedded in C-3 and therefore (19) is not a genuine C^4. Nägelsbach (1963: 645) specifically remarks that (19) stretches the limits of Latin grammar to the extreme.

A good test of the tenability of C^3max-w is provided by Admoni (1980), a study of the development of complex German sentences during the period 1470-1730. Admoni's material contains some 450 sentences among which are one C^3 (id.: 198f.) and 50 C^2s.[2]

C^3 does not occur in speech. Less than a handful of spoken C^2s are on record. We therefore infer C^2MAX-S (where '-s' refers to spoken language):

(20) *C^2max-s constraint*
The maximal degree of multiple center-embedding is two in spoken language.

Of course it is to be expected that written language allows more complexity than speech,[3] manifested as one additional cycle of clausal embedding. But in ordinary language use, written C^3s and spoken C^2s are almost non-existent.

I now turn to an analysis of the qualitative composition of the constituent clauses in C^2s. I first consider the tendencies in regard to clause types (table 3).

Clause type	C-high	%	C-low	%
relative	82	62,1	83	62,9
that (compl.)	22	16,7	7	5,3
when, after, before	11	8,3	16	12,1
if	5	3,8	3	2,3
as	4	3,0	12	9,1
because	2	1,5	0	0,0
while. whereas	1	0,8	2	1,5
others	5	3,8	9	6,8
Total	132	100,0	132	100,0

Table 3.
Clause types in 132 C^2s. C-high, C-low = upper, lower clause.

Table 3 displays the same tendency as table 2: relative clauses predominate at all depths in center-embeddings. The share of relatives is over 60% both in C-high and in C-low, i.e. generalization (16) holds also for C^2s and C^3s. Clauses other than relative clauses are infrequent in multiple center-embeddings and therefore difficult to generalize over, but there seems to be a tendency for *that*-clauses (complements) to prefer C-high over C-low (which is in harmony with the results of Gibson 1998), and for *as*-clauses to prefer C-low over C-high.

The typical C^2 is indeed a pair of relative clauses, as shown in table 4:

Clause combinations	N	%	cumulative %
relative + relative	47	35,6	35,6
relative + before, when	14	10,6	46,2
relative + as	11	8,3	54,5
relative + that	1	0,8	55,3
relative + other non-relative	9	6,8	62,1
that (compl.) + that (compl.)	5	3,8	65,9

that (compl.) + relative	15	11,4	77,3
when, after, before + relative	9	6,8	84,1
as, if, whereas + relative	7	5,3	89,4
because + relative	2	1,5	90,9
other non-relative + relative	4	3,0	93,9
non-relative + non-relative	8	6,1	100,0
Total	132	100,0	100,0

Table 4.
Combinations of clause types in 132 C^2s (self-embeddings bolded).

Self-embeddings contain the same type of clause. Table 4 shows that the typical C^2 is a relative self-embedding, as in (7a, d) and (8b, c). The only other type of self-embedding encountered is a pair of *that*-clauses (noun complement), as in (7f). No C^2s are on record of two *if*- or two *because*-clauses, for example. The common feature of center-embedded relative and *that*-clauses is that they endocentrically postmodify nouns whereas *if*-clauses, *because*-clauses etc. act as independent adverbials in their superordinate clause. Even more generally, example (7g) above (with two indirect question clauses) indicates that all types of adnominal postmodifying clauses may occur in self-embeddings:

(21) *Only-postmodifying-self-embedding constraint*
Only clauses postmodifying nouns allow central self-embedding.

This constraint is explained by the same processing preference as that invoked in the analysis of (16): postmodifying center-embeddings do not interrupt the processing of the superordinate clause as clearly as center-embeddings with superordinate clause constituent status (e.g. *if*- and *because*-clauses) do. The first two lines of table 3 show that almost 80% of all C-highs and 70% of all C-lows are adnominal and postmodifying.

As (11) and (13) show, multiply self-embedded relative clauses do occur even of degree C^3, as allowed by C^3max-w and Only-postmodifying-self-embedding.

Many intuition-based claims in the literature are at variance with the data and conclusions just presented: C^2 is ill-formed or unacceptable (Newmeyer 1987: 7, McMahon 1994: 155, Givón 2001: 217); C^2 is ungrammatical and 'completely baffling' (Quirk et al. 1989: 1040) ; C^2 is unacceptable and/or self-embedding is more complex than the same amount of center-embedding without self-embedding (Miller & Chomsky 1963: 475, Lyons 1970: 102, Lewis 1996: 103); self-embedding is unacceptable (Chomsky 1965: 10, Bever 1976: 67); successive self-embedding of clauses of the same grammatical type or function is worse than embeddings of different types or

functions (Kuno 1974: 120); increasing the similarity of CE clause types increases difficulty (Lewis 1999: 105); C^2 of relative clauses does not occur (Lehmann 1984: 197); self-embedded relative clauses are ungrammatical in English and systematically avoided across languages (Hawkins 1994: 12, 5); any C^2 of relative clauses boggles the human parser (Pinker 1994: 207); any number of self-embedded reduced relative clauses are grammatical (Smith 1994: 647).

Table 4 shows that the share of C^2s with at least one relative clause is around 90%. Most C^2s that lack relative clauses turn out to have other types of adnominal postmodifying clauses, cf. (7g) above and (22):

(22) [$_{M}$ Your report today [$_{C\text{-}1}$ that any Tory constituency party [$_{f\text{-}2}$ failing [$_{f\text{-}3}$ to deselect its MP, [$_{C\text{-}4}$ should he not vote in accordance with a prime ministerial dictate,]]] might itself be disbanded,] shows ...] (*The Times* 25.11. 1994)

Of the 132 C^2s in my corpus, just a handful do not contain at least one postmodifying clause (cf. (1a), (44), (67), (93) in Karlsson 2007c), suggesting a minimal constraint on C^2s:

(23) *Minimally-one-postmodifying-clause constraint*
A double center-embedding must contain at least one postmodifying clause.

The dominance of relative clauses in center-embeddings of any degree calls for a functional interpretation. Consider table 5:

Position of center-embedding in superordinate clause	C-high	%	C-low	%
after grammatical subject	86	65,2	86	65,2
after adverbial	35	26,5	24	18,2
after others	11	8,3	22	16,7
Total	132	100,0	132	100,1
before finite verb	94	71,2	96	72,7
before grammatical subject	16	12,1	2	1,5
before adverbial	14	10,6	22	16,7
before others	8	6,1	12	9,1
Total	132	100,0	132	100,0

Table 5.
Functional-syntactic positions of C-high and C-low in 132 C^2s.

The tendency is clear. In 65% of both C-high and C-low a relative center-embedding occurs AFTER THE SUBJECT and BEFORE THE MAIN VERB of the superordinate clause. For comparison, in a 10,000-clause corpus of written Finnish there were 140 finite center-embedded sub-clauses, of which 98 (70%) were relative (Hakulinen et al. 1980: 118). Closer scrutiny reveals that 98 (75%) of these occurred in front of the superordinate finite verb. Danielsson (1975: 88) found that 70% of 546 center-embedded clauses in her Swedish textbook corpus modified the first constituent (especially the subject) of the superordinate clause. Similarly, two-thirds of 552 center-embedded postmodifying clauses in the Nijmegen English corpus go with the subject (De Haan 1989: 189). Thus there is ample evidence from several SAE languages for generalization (24):

(24) The typical location of a C^2 is at the end of the grammatical subject immediately before the main verb.

This is the major intrasentential break in SVO-languages, the optimal location for grounding and specification of the main clause topic, normally expressed by the grammatical subject, before new information is presented by the verb and its postverbal non-subject complements:

(25) $_S$[[Subj $C^{1\sim2\sim3}$] [V] [Obj ~ PredCompl ~ Advl]]

Non-postmodifying sub-clauses are rare in center-embeddings because the S-V break is not conducive to intricate development for example of conditional or causal reasoning. The basic discourse function of center-embedded relatives is to specify and ground the referent of the grammatical subject of the superordinate clause (Fox and Thompson 1990), an observation that dates back to Erman's (1913: 475) study of clausal embedding in Old High German. For C-high, the superordinate clause is almost always the main clause and the referent of its grammatical subject is the main topic of the whole sentence.

Prototypical C^2s with at least one and often two relative clauses are comparatively easy to process because the relative pronouns are coreferential with antecedents in their superordinate clauses, thereby reducing the information processing load. This effect is particularly clear in sentences like (26) where two CE relative pronouns are accompanied by two pronominal grammatical subjects:

(26) [$_M$ ... all the concern [$_{C\text{-}1}$ which **he** [$_{C\text{-}2}$ to whom **it** belongs by adoption] has in the matter] is the being ...] (Bentham 1945: 103)

The propensity of C^2s to have pronominal subjects is clearly seen in table 6:

Type of grammatical subject	C-high	%	cum %	C-low	%	cum %
relative pronoun	48	36,4	36,4	41	31,1	31,1
personal pronoun	8	6,1	42,5	33	25,0	56,1
it	0	0,0	42,5	7	5,3	61,4
other pronoun	13	9,8	52,3	4	3,0	64,4
zero	13	9,8	62,1	21	15,9	80,3
definite non-pronominal NP	34	25,8		18	13,6	
other textually bound non-pronominal NP	7	5,3		2	1,5	
indefinite NP	9	6,8		6	4,5	
Total	132	100,0		132	99,9	

Table 6.
Types of grammatical subjects in 132 C^2s. Cum = cumulative %.

Including pronominal zeroes, the share of pronominal subjects is over 60% in C-high and 80% in C-low, lending support to Bever (1976) and Kac (1981) in that pronominal subjects in C-low facilitate processing, and to Hudson (1996) in that a full common noun as subject in C-low hampers its processing. But still, every fifth C^2 is a counterexample, with a full NP in C-low, as in (7d) and (8c), for example. Instead, an examination of all pronouns in the 132 C-low's instead provides evidence for the somewhat weaker constraint OVERT-PRON-LOW:

(27) *Overt-pron-low constraint*
C-low must contain at least one overt pronoun, preferably as subject.

There are a few exceptions to Overt-pron-low, less than 5% (cf. (3a), (25), (44), (48), (50), (92) in Karlsson 2007c), but they are either idioms or contain repeated NPs that could (if not should) have been pronominalized. Overt-pron-low is amplified by the fact that almost half of the C-lows contain two pronouns (including zeros), as in (26) above and in instances like *den er mit sich führe*; *as he put it*; *not ∅ having seen him for some time*; *∅ I have*, etc.

Table 7 presents the preferences for combining relativized syntactic positions in the 47 self-embedded relative C^2s found (cf. table 4). The notation 'S-O', for example, is to be read 'S(ubject) is relativized in the upper clause, O(bject) is relativized in the lower clause.

Rel-Rel	N	%
S-S	13	27,7
S-O	6	12,8
S-A	4	8,5
S-PC	2	4,3
O-S	5	10,6
O-A	1	2,1
O-PC	1	2,1
PC-S	4	8,5
PC-PC	4	8,5
PC-O	3	6,4
A-S	2	4,3
GEN-S	1	2,1
GEN-PC	1	2,1
Sum	47	100,0

Table 7.
Combinations of relativized constituents in 47 self-embedded C^2s. S(ubject), O(bject), A(dverbial), P(repositional) C(omplement), G(enitive).

As is to be expected, the subject is most frequently relativized, more than 50% in both C-high and C-low. Examples of various combinations: S-S (14), S-O (7a), GEN-PC (7d), PC-S (8b), PC-PC (8c). The most striking feature of table 7 is the lack of doubly relativized objects, O-O, as in the classical example (3), here repeated for convenience as (28).

(28) The rat the cat the dog chased killed ate the malt.

Note that (28) violates Overt-pron-low, which certainly contributes to its strangeness. The low acceptability of (28) and the total absence of O-Os suggest an independent stronger constraint, the NO-MULTIPLE-OBJECT-RELATIVIZATION CONSTRAINT, abbreviated *O–O:

(29) **O–O constraint*
Direct objects must not be multiply relativized in C^2s.

Is there an explanation for *O–O? At least in SAE it is a fact that object relativization is a more resource-consuming process than subject relativization (e.g. Gibson 1998). Double object relativization in C^2s would be exceedingly costly, therefore it does not occur.

Givón (2002: 217–218), like myself, claims that O–O C^2s like (28) are in fact ungrammatical. He proposes another processing-related explanation: the coreference relation in relativization between a head and its coreferent zero

holds only across adjacent clauses. This condition is violated in (28) where C-2 interferes with the extraction site '$\__j$' of the relativized object in C-1:

(30) [$_M$ The rat$_j$ [$_{C\text{-}1}$ the cat$_k$ [$_{C\text{-}2}$ the dog$_m$ chased $\__k$] killed $\__j$] ate the malt.]

The antecedent of '$\__j$' is two clauses away and therefore (28) is ungrammatical. If Givón's condition is the appropriate generalization, *O–O needs no separate statement. This matter cannot be conclusively resolved here.

7. DISCUSSION

First, a caveat. Our data and conclusions concern complexity restrictions on embedding of clauses (which certainly is the most discussed class of embedding constructions). But there are at least two other potential loci for unbounded center-embedding complexity: multiple embedding of phrasal constituents, in particular NPs within NPs, and multiple nested nominalizations of clauses by rank-shifting them to modifier status within NPs. The literature contains few observations on the complexity limits of these constructions.

The constraints C^3max-w, C^2max-s, Only-postmodifying-self-embedding, Overt-pron-low and *O–O falsify the hypothesis of unbounded clausal embedding complexity by strongly restricting the potential clausal center-embeddings in English and several other SAE languages.

C^3max-w is a deeply entrenched constraint in many of the SAE languages because C^3s (but not C^4s) were documented in Danish, English, German, Latin, and Swedish. On the other hand, the C^3s found are so rare, and mostly so convoluted and incomprehensible, that C^3 is marginal at best. Spoken multiple center-embedding is close to non-existent in SAE languages: less than a handful of genuine (English-only) C^2s have been retrieved.

The key to understanding the SAE phenomenon of center-embedding is relative clauses, the prototype for center-embedding of all degrees and in all the SAE languages here considered. The primacy of relative clauses is manifested in the constraint Only-postmodfying-self-embedding. In consequence, the basic discourse function of SAE center-embedding is that of relative clauses and noun complement clauses: to specify the referent of a noun phrase in the superordinate clause, prototypically the subject. Given SVO-order and clausal postmodification in NPs, the result is the pattern $_S$[[Subj $C^{1\sim2\sim3}$] [V] [Obj ~ PredCompl ~ Advl]], i.e. clausal center-embedding. Many other types of clauses (e.g. *although*-, *as*-, *because*-, *if*-, *when*-clauses) may be center-embedded (table 3) but they prefer initial and/or

final embedding much more strongly than relative clauses do. For example, in a corpus extracted from Brown consisting of 894 (every eighth) *that*-clauses, 96% were finally-embedded, 3% center-embedded, and 1% initially-embedded. The corresponding shares of 772 similarly extracted *when*-clauses were 54%, 7%, and 39%.

Constraints were initially defined as 'quantitative limits and combinatory restrictions, often of the nature of tendencies'. Typical grammatical rules are different, i.e. well-defined and categorical. Violations of rules are perceived as deviant because they breach the normativeness of the rules. For example, (31) is a morphological rule of English:

(31) The object form of *he* is *him*.

(32) *Sue kissed he.

Sentence (32) violates rule (31) and is ungrammatical, conflicting with the natural norm expressed by (31). The strangeness of fabricated sentences like (3) and (5a–e) indicates norm breach of a weaker kind than in (32).

Constraints like C^3max-w, C^2max-s, Only-postmodifying-self-embedding, Overt-pron-low and *O–O are much like the SOFT CONSTRAINTS for Preferred Argument Structure discussed by Du Bois (1987, 2003a, b), which also express quantitative and other tendencies discernible in language use − for example, 'avoid more than one lexical core argument', 'avoid more than one new core argument', 'avoid lexical NPs for subjects of transitive verbs'. Such constraints are universal regularities of discourse, recurrent patterns of language use that cannot be reduced to prototypical grammatical rules even if they are formulated using grammatical concepts. When a soft constraint is overstepped, e.g. when a transitive verb occurs with two lexical core arguments, the result is not ungrammatical nor does it need to result in processing failure.

Of our constraints, Minimally-one-postmodifying-clause and Overt-pron-low seem to be most like Du Bois' 'soft constraints'. The others are somewhat stronger and more normative, and violations of them often lead to processing difficulties. Sentences like (5c) violating C^3max-w do feel markedly strange and therefore some amount of normativeness, i.e. rule-likeness is invoked. The same is true of (3) and (6), which violate *O–O. These constraints are less arbitrary than typical basic-level morphological and syntactic rules but still have some normative force. They occupy a continuum between grammatical rules and behavioral language-related regularities. Note, in passing, that Givón (2001: 218) and Jackendoff (2002: 32) also conclude that more aspects of competence (i.e. grammar) are involved in multiple center-embedding than Chomsky and his followers have been assuming.

Our soft constraints can be linguistically interpreted and functionally explained by known facts of discourse management, especially referent specification by use of relative clauses. On the other hand, the constraints have their ultimate basis in the material language-processing resources and limitations of the human organism. In this sense the constraints are epiphenomenal consequences of more basic cognitive properties, especially short-term memory limitations.

An important property of the constraints C^3max-w and C^2max-s is that they (in contradistinction to the hypothesis of unbounded center-embedding) are falsifiable. The message of C^3max-w and C^2max-s is that running SAE text is in fact syntactically analyzable by finite-state methods because there are indeed systematic quantitative and qualitative constraints on the center-embedding complexity to be dealt with (also cf. Kornai 1985).

Embedding constraints determined on the basis of finite corpora do not suffice to conclusively falsify the hypothesis of unbounded embedding complexity: 'absence of evidence is not evidence of absence'. However, taken together, our observations do warrant a weaker conclusion:

(33) Recursive clausal center-embedding is not a highly important design feature of SAE in actual use.

Only one cycle of center-embedding is in really productive use. Much more important than recursive center-embedding ('nested recursion', in the terminology of Parker 2006)) are the processes of TAIL-RECURSION, i.e. recursive left-branching and right-branching, e.g. in English genitive constructions (34), but above all in PP chains (35) and sequences of finally-embedded finite and/or non-finite clauses as in (36)–(37).[4]

(34) [[[Hilary's] lawyer's] … secretary] (LOB)

(35) [The season will open [at [[the new [Hall [of Flowers]]] [in [[Golden Gate] Park]]]] [on [November 20]] [at [8:30 p.m.]] [with [[a concert] [by [the [Mills [Chamber Players.]]]]]]] (Brown)

(36) [This is [to confirm [that I would like [to enquire [whether it would be possible [to employ a First Aider for a series of days … [as we will be using several examination halls [which will be too far away … [for us to provide ... first aid service to them.]]]]]]]]]

(37) [This is the farmer [sowing the corn, that kept the cock, [that crowed in the morn, [that waked the priest all shaven and shorn, [that married the man all tattered and torn, [that kissed the maiden all forlorn, [that milked the cow with the crumbled horn, [that tossed the dog, [that

worried the cat, [that chased the rat, [that ate the malt, [that lay in the house [that Jack built.]]]]]]]]]]]]

The most productive of these constructions is repeated final embedding (right-branching) of clauses, as in (36) which reaches depth 8 with an assortment of non-finite and finite clauses, especially infinitival ones. But the excessive depth is rather illusory; (36) is not felt to be extremely complex. A realistic interpretation is that its eight finally-embedded clauses are, rather, CONCATENATED one after another by iteration in a way which does not create more complexity than a sequence of coordinated NPs or clauses. (36) exhibits tail-recursion, which is always convertible to iteration (Aho, Sethi & Ullman 1986: 53). Thus, (36) is not essentially 'recursion downwards'. The depth layers are mainly relevant for semantic reasons (Parker 2006).

Sentence type (37) is especially interesting. This construction with multiply self-embedded relative is truly open-ended (in certain restricted types of discourse). But again, (37) is not perceived as particularly complex. Here an incisive observation by Chafe (1988: 21) is relevant: relative pronouns are equivalent to linking sequences of coordinating conjunctions and anaphoric pronouns, e.g. *who = and she/he, which = and it*. That is, the F^{12} in (37) could be rewritten as a shallow string of conjuncts:

(38) This is the farmer and he sowed the corn, and he kept the cock, and it crowed …

These remarks suggest that final clausal embedding is not a phenomenon where (tail-)recursion would generate great complexity. In Karlsson (2007a, b) I demonstrated that there is also an absolute quantitative constraint on the degree of initial clausal embedding complexity, 2 (I^2MAX), and that in the rarely occurring I^2s the subordinating conjunction of the upper clause almost always is *if*. The presumed unbounded clausal syntactic recursion is thus heavily constrained in all three embedding positions.

The conclusion that no empirical facts support competence grammar with unbounded recursive center-embedding was also arrived at by Christiansen and Chater (1999, 2001) in a setting of connectionist modeling of recursion in human linguistic performance. My purely linguistic analysis of corpus data from several languages agrees with these processing-related results.

Independent evidence from child language acquisition also supports these conclusions. Menyuk (1969: 16) reported that center-embedding is the last embedding position picked up in the course of first language acquisition, often later than age 7. Similarly, Diessel and Tomasello (2005: 884) found that center-embedded relative clauses attached to the main clause grammatical subject, i.e. the prototype of adult written center-embeddings (table 5), 'are essentially absent from naturally occurring child speech'. In

consequence, one cannot credibly assume such non-occurring constructions and the putative mechanisms producing them to be of prime importance, at the heart of the language faculty, determining the essence of syntactic embedding complexity in spoken language.

Thus, embedding complexity is LEFTWARDS AND RIGHTWARDS ITERATIVE, or concatenative, as in (34–37), than recursively center-oriented. This finding is relevant also for the rapidly developing research on language origin in the human species. One of its basic postulates is that major gains in syntactic recursion were the decisive breakthrough in the phylogeny of language (Bolinger 1975: 308–310). Today the alleged importance of unbounded syntactic recursion is even more strongly emphasized in language origin research, as witnessed e.g. by Berwick (1998: 322) in his analysis of language evolution in the framework of the minimalist program; Nowak, Komarova & Niyogi (2001: 117f.) in their mathematical model of the conditions under which natural selection favored the emergence of rule-based grammars; Hurford (2000: 329) in his model of how social transmission favors linguistic generalizations; Kirby and Hurford (2002: 130), who have a sub-chapter entitled 'From proto-language to recursive syntax'; Bickerton (1996) in his theory of the rapid late emergence of full-blown syntactic language; Corballis (2002: 60–61) who holds that the extended use of recursion is what distinguishes humans from chimpanzees, macaques, and capuchins; and Hauser, Chomsky & Fitch (2002: 1577), who in their theory of the evolution of the faculty of language surmise that recursion is the core component of the FLN ('faculty of language in the narrow sense') and maintain that unbounded center-embedding is the crucial property of the FLN.

Li (2002: 209–212) is one of the few language origins researchers to emphasize that recursion is not a unique design feature of natural language syntax because such phenomena (in practice, tail-recursion) are also found in the communicative behavior of humpback whales and mockingbirds. Our findings, in the same vein, do indicate that the assumed central role of recursive center-embedding should also be downplayed in language origins research. What really matters in the evolution of syntactic arrangement is the emergence of repeated concatenation in combination with modest embedding depth. This scenario agrees with reports that syntactic complexity is modest in several aboriginal languages thriving in oral cultures: Inuktitut is now in the process of obtaining clausal hypotaxis along with the development of the native press (Kalmár 1985); Pirahã is claimed to have no syntactic embedding (Everett 2005).

The existing clausal center-embeddings in SAE languages are moderately recursive and almost totally confined to written language. This suggests that the origin of SAE multiple center-embedding should be sought within a time span not longer than the advent of written language 5000 years ago.

Fundamentally, SAE clausal center-embedding is a byproduct of the development of Latin stylistics. Clausal embedding below depth 1 was not established in Latin before 100 BC (Lindskog 1896). Center-embedding of relative clauses next to the antecedent was consolidated by Cicero, who treated this construction from a stylistic and rhetorical point of view in *De oratore* (55 BC). By laying down rules for sentential composition he completed the doctrine of PERIODS, a cornerstone of Western rhetoric and stylistics initiated by Aristotle. By definition, a periodic sentence contains at least one center-embedding, a detour from the overriding sentence scheme, brought to completion by the latter part of the superordinate clause when it is resumed. The master of the use of periodic sentences was Livy (59/64 BC – AD 13), along with Cicero a stylistic icon for centuries. From the Renaissance to the 1900s there were hundreds of SAE grammars, stylistic manuals, and scholarly monographs treating periodic sentence structure − Boivie's (1834: 99–102) Swedish and Becker's (1870: 418–423) German grammars, for example This is the historical source of the fairly uniform patterns of clausal subordination found in present-day SAE (Blatt 1957).

The upshot of this brief diachronic note is that all SAE variants of clausal center-embedding, as well as the recursive mechanisms underlying these structures, are young phenomena related to the emergence of written language and therefore situated within the historical time span. They cannot have played an important role in the emergence of SAE spoken language complexity much earlier. This argumentation supports Johansson's (2005: 235) conclusion that of the four most central grammatical design principles (Structured, Hierarchical, Recursive, Flexible), the prime candidate for being a late evolutionary addition to human grammar is Recursive.

Our data come from seven SAE languages only and therefore care must be exercised in pondering whether the constraints here inferred are valid for language in general. Insofar as the constraints derive from material processing limitations of the human organism, especially short-term memory management, it nevertheless seems reasonable to assume a more general validity.

REFERENCES

Admoni, W. G. (1980). *Zur Ausbildung der Norm der deutschen Literatursprache im Bereich des neuhochdeutschen Satzgefüges (1470–1730)*. Berlin: Akademie-Verlag.

Aho, A. V., Sethi, R. & Ullman, J. D. (1986). *Compilers.* Reading, Mass.: Addison-Wesley Publishing Company.

Akmajian, A., Demers, R. A. & Harnish, R. M. (1985). *Linguistics.* (2nd edn.). Cambridge, MA: MIT Press.

Bar-Hillel, Y. (1964). *Language and information.* Reading, Mass.: Addison Wesley Publishing Company, Inc.

Becker, K. F. (1870). *Ausführliche deutsche Grammatik als Kommentar der Schulgrammatik.* Zweiter Band. (2nd edn.). Prag: Verlag von Tempsky.

Bentham, J. (1945). *The limits of jurisprudence defined.* Ed. C. Everett. New York: Columbia University Press.

Berwick, R. C. (1998). Language evolution and the minimalist program. In Hurford, J. R., Studdert-Kennedy, M. & Knight, C. (eds.), *Approaches to the evolution of language.* Cambridge: Cambridge University Press. 320–340.

Bever, T. G. (1976). The influence of speech performance on linguistic structure. In Bever, T. G., Katz, J. J. & Langendoen, D. T. (eds.), *An integrated theory of linguistic ability*. New York: Crowell. 65–88.

Bickerton, D. (1996). *Language and human behaviour*. London: UCL Press.

Blatt, F. (1957). Latin influence on European syntax. Copenhagen: *Travaux du Cercle Linguistique de Copenhague* **11**. 33–69.

Blatz, F. (1896). *Neuhochdeutsche Grammatik mit Berücksichtigung der historischen Entwickelung der deutschen Sprache*. Satzlehre (Syntax). (3rd edn.). Karlsruhe: J. Lang's Buchdruckerei.

Bolinger, D. (1975). *Aspects of language*. (2nd edn.). New York: Harcourt Brace Jovanovich, Inc.

Boivie, P. G. (1834). *Försök till en svensk språklära jemte inledning, innehållande allmänna grammatikan*. Upsala: Palmblad & C.

Chafe, W. (ed.) (1980). *The Pear Stories: cognitive, cultural, and linguistic aspects of narrative production*. Norwood, NJ: Ablex Publishing Corp.

Chafe, W. (1988). Linking intonation units in spoken English. In Haiman & Thompson (eds.), 1-27.

Chomsky, N. (1956). On the limits of finite-state description. *MIT Research Laboratory for Electronics, Quarterly Progress Report* **41**. 64–65.

Chomsky, N. (1957). *Syntactic structures*. The Hague: Mouton.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, Mass.: The MIT Press.

Chomsky, N. & Miller, G. A. (1963). Introduction to the formal analysis of natural languages. In Luce et al. (eds.), 269–321.

Christiansen, M. H. & Chater, N. (1999). Toward a connectionist model of

recursion in human linguistic performance. *Cognitive Science* **23**. 157–205.
Christiansen, M. H. & Chater, N. (2001). Finite models of infinite language: a connectionist approach to recursion. In Christiansen, M.H. & Chater, N. (eds.), *Connectionist psycholinguistics*. Westport, Connecticut: Ablex Publishing. 138–176.
Corballis, M. C. (2002). *From hand to mouth. The origins of language*. Princeton and Oxford: Princeton University Press.
Crystal, D. (2003). *A dictionary of linguistics & phonetics*. (5th edn.). Oxford: Blackwell Publishers Ltd.
Danielsson, S. (1975). *Lärobokssprå*k. Umeå: Acta Univ. Umensis 4.
De Haan, P. (1989). *Postmodifying clauses in the English noun phrase. A corpus-based study.* Amsterdam: Rodopi.
De Roeck, A., Johnson, R., King, M., Rosner, M., Sampson, G. & Varile, N. (1982). A myth about centre-embedding. *Lingua* **58**. 327–340.
Diessel, H. & Tomasello, M. (2005). A new look at the acquisition of relative clauses. *Language* **81**. 882–906.
Drach, E. (1963). *Grundgedanken der deutschen Satzlehre*. Darmstadt: Wissenschaftliche Buchgesellschaft.
Dryer, M. S. (1980). The positional tendencies of sentential noun phrases in universal grammar. *The Canadian Journal of Linguistics* **25**. 123–196.
Du Bois, J. W. (1987). The discourse basis of ergativity. *Language* **63**. 805-855.
Du Bois, J. W. (2003a). Argument structure: grammar in use. In Du Bois, J. W., Kumpf, L. E. & Ashby, W. J. (eds.), *Preferred argument structure: grammar as architecture for function*. Amsterdam: John Benjamins. 11–60.
Du Bois, J. W. (2003b). Discourse and grammar. In Tomasello, M. (ed.), *The new psychology of language*. Mahwah, NJ: Erlbaum. 43–87.
Engel, E. (1922). *Deutsche Stilkunst*. (30th edn.). Wien / Leipzig: Hölder-Pichler-Tempsky A-G.
Erman, K. B. (1913). Beziehungen zwischen Stellung und Funktion der Nebensätze mehrfacher Unterordnung im Ahd. *Zeitschrift für deutsche Philologie* **45**. 1–46, 153–216, 426–484.
Everett, D. L. (2005). Cultural constraints on grammar and cognition in Pirahã. Another look at the design features of human language. *Current Anthropology* **46**. 621-646.
Fox, B. A. & Thompson, S. A. (1990). A discourse explanation of the grammar of relative clauses in English conversation. *Language* **66**. 297–316.
Frazier, L. & Rayner, K. (1988). Parameterizing the language processing system: left- vs. right-branching within and across languages. In Hawkins, J. A. (ed.), *Explaining language universals*. Oxford: Blackwell. 246–279.

Gibson, E. (1998). Linguistic complexity: locality of syntactic dependencies. *Cognition* **68**. 1–76.
Givón, T. (2001). *Syntax.* Vol. II. Amsterdam: John Benjamins.
Greene, J. (1972). *Psycholinguistics.* Harmondsworth: Penguin Books Ltd.
Grosu, A. & Thompson, S. A. (1977). Constraints on the distribution of NP clauses. *Language* **53**. 104–151.
Hagège, C. (1976). Relative clause center-embedding and comprehensibility. *Linguistic Inquiry* **7**. 198–201.
Haiman, J. & Thompson, S. A. (eds.) (1988). *Clause combining in grammar and discourse.* Amsterdam: John Benjamins.
Hakulinen, A., Karlsson, F. & Vilkuna, M. (1980). *Suomen tekstilauseiden piirteitä: kvantitatiivinen tutkimus.* Helsinki: Publications of the Department of General Linguistics, University of Helsinki, No. 6.
Halliday, M. A. K. (1963). Class in relation to the axes of chain and choice in language. *Linguistics* **2**. 5-15.
Hauser, M. D., Chomsky, N. & Fitch, W. T. (2002). The faculty of language: what is it, who has it, and how did it evolve? *Science* **298**. 1569–1579.
Hawkins, J. W. (1994). *A performance theory of order and constituency*. Cambridge: Cambridge University Press.
Hiltunen, R. (1984). The type and structure of clausal embedding in legal English. *Text* **4**. 107–121.
Hoffmann-Krayer, E. (1925). *Geschichte des deutschen Stils in Einzelbildern*. Leipzig: Verlag von Quelle & Meyer.
Hudson, R. (1996). The difficulty of (so-called) self-embedded structures. University College London: *Working papers in linguistics* **8**. 283–314.
Hurford, J. R. (2000). Social transmission favours linguistic generalisation. In Knight, C., Studdert-Kennedy, M. & Hurford, J. R. (eds.), *The evolutionary emergence of language*. Cambridge: Cambridge University Press. 324–352.
Jackendoff, R. (2002). *Foundations of language. Brain, meaning, grammar, evolution.* Oxford: Oxford University Press.
Kac, M. (1981). Center-embedding revisited. *Proceedings of the third annual conference of the Cognitive Science Society, August 1981*. Hillsdale: Lawrence Erlbaum. 123–124.
Kalmár, I. (1985). Are there really no primitive languages? In Olson, D. R., Torrance, N. & Hildyard, A. (eds.), *Literacy, language and learning. The nature and consequences of reading and writing*. Cambridge: Cambridge University Press. 148-166.
Karlsson, F. (2007a). Constraints on multiple initial embedding of clauses. In print: *International Journal of Corpus Linguistics* **12(1)**.
Karlsson, F. (2007b). Genuine data on multiple initial embedding of clauses. [Available at www.ling.helsinki.fi/~fkarlsso/ieb_data.pdf]
Karlsson, F. (2007c). Genuine data on multiple center-embedding of clauses.

[Available at www.ling.helsinki.fi/~fkarlsso/ceb_data.pdf]
Karlsson, F. (2007d). Quantitative analysis of multiple center-embedding of clauses. [Available at www.ling.helsinki.fi/~fkarlsso/ceb_quant.xls]
Karlsson, F. (2007e). Constraints on phrase-level recursion. Ms., University of Helsinki, Helsinki.
Kirby, S. & Hurford, J. R. (2002). The emergence of linguistic structure. In Cangelosi, A. & Parisi, D. (eds.), *Simulating the evolution of language*. London: Springer. 121–147.
Kornai, A. (1985). Natural languages and the Chomsky hierarchy. In King, M. (ed.), *Proceedings of the 2nd European conference of the Association for Computational Linguistics*. 1–7.
Korthals, C. (2001). Self embedded relative clauses in a corpus of German newspaper texts. In Striegnitz, K. (ed.), *6th ESSLLI student session, August 13–24, 2001*. Helsinki, Finland. 179–190.
Kriebel, W. (1873). Der Periodenbau bei Cicero und Livius. Rostock, Universität Rostock PhD dissertation. Prenzlau: A. Mieck.
Kuno, S. (1974). The position of relative clauses and conjunctions. *Linguistic Inquiry* **5**. 117–136.
Langendoen, D. T. (1998). Linguistic theory. In Bechtel, W. & Graham, G. (eds.), *A companion to cognitive science*. Oxford: Blackwell. 235–244.
Lehmann, C. (1984). *Der Relativsatz.* Tübingen: Gunter Narr Verlag.
Lewis, R. L. (1996). Interference in short-term memory: the magical number two (or three) in sentence processing. *Journal of Psycholinguistic Research* **25**. 193–215.
Lewis, R. L. (1999). Accounting for the fine structure of syntactic working memory: similarity-based interference as a unifying principle. *Behavioral and Brain Sciences* **22**. 105–106.
Li, C. N. (2002). Some issues concerning the origin of language. In Bybee, J. & Noonan, M. (eds.), *Complex sentences in grammar and discourse. Essays in honor of Sandra A. Thompson*. Amsterdam/Philadelphia: John Benjamins Publishing Company. 203–221.
Lindskog, C. (1896). Beiträge zur Geschichte der Satzstellung im Latein. Lund: *Acta Universitatis Lundensis* **32**. 1–60.
Luce, R. D., Bush, R. R. & Galanter, E. (eds.) (1963). *Handbook of mathematical psychology, II*. New York: John Wiley and Sons, Inc.
Lyons, J. (1970). *Noam Chomsky*. New York: The Viking Press.
McMahon, A. (1994). *Understanding language change*. Cambridge: Cambridge University Press.
Meillet, A. (1934). *Introduction a l'étude des langues indo-européennes.* Paris: Librairie Hachette.
Menyuk, P. (1969). *Sentences children use*. Cambridge, Mass.: The MIT Press.
Miller, G. A. & Chomsky, N. (1963). Finitary models of language users. In

Luce et al. (eds.), 419–491.
Miller, G. A. & Isard, S. (1964). Free recall of self-embedded English sentences. *Information and Control* **7**. 292–303.
Nägelsbach, K. F. (1963 [1846]). *Lateinische Stilistik*. Darmstadt: Wissenschaftliche Buchgesellschaft.
Newmeyer, F. J. (1987). Extension and implications of linguistic theory: an overview. In Newmeyer, F. (ed.), *Linguistics: the Cambridge survey*, Vol. II. Cambridge: Cambridge University Press. 1–14.
Nowak, M. A., Komarova, N. L. & Niyogi, P. (2001). Evolution of universal grammar. *Science* **291**. 114–118.
Olzien, O. H. (1933). *Der Satzbau in "Wilhelm Meisters Lehrjahren"*. Leipzig: J. J. Weber.
Parker, A. R. (2006). Evolving the narrow language faculty: was recursion the pivotal step? In: Cangelosi, A., A. D. M. Smith & K. Smith (eds.), *The Evolution of Language: Proceedings of the 6th International Conference (Evolang 6)*. London: World Scientific. 239–246.
Partee, B. H., ter Meulen, A. & Wall, R. E. (1993). *Mathematical methods in linguistics*. Dordrecht etc.: Kluwer Academic Publishers.
Pinker, S. (1994). *The language instinct*. Harmondsworth: Penguin Books.
Pulman, S. G. (1986). Grammars, parsers, and memory limitations. *Language and Cognitive Processes* **1**. 197–225.
Quirk, R., Greenbaum, S., Leech, G. & Svartvik, J. (1989). *A comprehensive grammar of the English language*. London: Longman.
Reich, P. A. (1969). The finiteness of natural language. *Language* **45**. 831–843.
Reich, P. & Dell, G. S. (1976). Finiteness and embedding. In Di Pietro, R. J. & Blansitt Jr., E. L. (eds.), *The third LACUS forum 1976*. Columbia, South Carolina: Hornbeam Press. 438–447.
Sampson, G. (1996). From central embedding to corpus linguistics. In Thomas, J. & Short, M. (eds.), *Using corpora for language research*. London: Longman. 14–26.
Schneider, W. (1959). *Stilistische deutsche Grammatik*. Freiburg im Breisgau: Verlag Herder KG.
Smith, N. V. (1994). Competence and performance. In Asher, R. E. (ed.), *The encyclopedia of language and linguistic*s. Oxford: Pergamon Press. 645–648.

Author's address: General Linguistics
P.B. 24
FI-00014 University of Helsinki
Finland
E-mail: fgk at ling dot helsinki dot fi
home page: www.ling.helsinki.fi/~fkarlsso

FOOTNOTES

1. Several colleagues have offered valuable help and criticism which is gratefully acknowledged: Andrew Chesterman, Guy Deutscher, John W. DuBois, Marita Gustafsson, Risto Hiltunen, Richard Hudson, Jarmo Korhonen, Kimmo Koskenniemi, Heikki Mattila, Terttu Nevalainen, Martti Nyman, Simo Parpola, and especially Geoffrey Sampson. Sincere thanks for constructive criticism are due also to two JL referees. The work reported here was supported by the Academy of Finland under grant 201601.

2. Admoni (1980) was encountered when the corpus work of this paper was almost completed. His valuable material is at times difficult to interpret due to variable interpunctuation and structural vagueness. Insofar as I was able to analyze it, it seemed to conform to our generalizations and was therefore not included in our corpus but rather used as a test-bench. Admoni did not specifically address the degree restrictions on center-embedding.

3. Halliday (1963: 12) suggested that spoken English, and perhaps language generally, would tolerate greater depth in recursion than written English does. This hypothesis finds no support in our center-embedding data.

4. The nature of the phrase-level recursion in (34)–(35) cannot be addressed here (cf. Karlsson 2007e).