

John Benjamins Publishing Company



This is a contribution from *International Journal of Corpus Linguistics 15:1*
© 2010. John Benjamins Publishing Company

This electronic file may not be altered in any way.

The author(s) of this article is/are permitted to use this PDF file to generate printed copies to be used by way of offprints, for their personal use only.

Permission is granted by the publishers to post this file on a closed server which is accessible to members (students and staff) only of the author's/s' institute, it is not permitted to post this PDF on the open internet.

For any other use of this material prior written permission should be obtained from the publishers or through the Copyright Clearance Center (for USA: www.copyright.com).

Please contact rights@benjamins.nl or consult our website: www.benjamins.com

Tables of Contents, abstracts and guidelines are available at www.benjamins.com

Multiple final embedding of clauses*

Fred Karlsson

University of Helsinki

There are no grammatical limits on multiple final embedding of clauses. But converging corpus data from English, Finnish, German and Swedish show that multiple final embedding is avoided at levels deeper than three levels from the main clause in syntactically simple varieties, and at levels deeper than five levels in complex varieties. The frequency of every successive level of final embedding decreases by a factor of seven down to levels 4–5. Only relative clauses allow free self-embedding, within the limits just mentioned.

These restrictions are regularities of language use, stylistic preferences related to the properties of various types of discourse. Ultimately they are explained by cognitive and other properties of the language processing mechanisms. The frequencies of final embedding depths in modern languages such as English and Finnish is not accidental. Ancient Greek had reached this profile by 300 BC, suggesting cross-linguistic generality of the preferences.

Keywords: embedding, final embedding, right-branching, syntactic complexity, depth

1. Introduction

According to a widely accepted view expressed e.g. by Meillet (1934: 355), there are no principled grammatical restrictions on clausal embedding complexity in sentences. However, in two recent papers of mine (Karlsson 2007a, b), I have shown that there are indeed absolute restrictions at least on how many times clausal embedding may be repeated in initial position (two times) and medially (center-embedding, three times).

Most often Meillet's view is illustrated with multiple final embedding, also called right-branching, which certainly allows much freedom of subordinating several clauses (henceforth **sub-clauses**) one below the other, as in sentence (1) with six successive relative clauses, where progressive indentation indicates

growing embedding depth, also indicated by growing numbers appended to “F” (short for final embedding; “M” = main clause, “&” = coordinated).

- | | |
|---|------|
| (1) The said rocker level is operated by means of a pair of opposed fingers | M |
| <i>which</i> extend from a pitman | F-1 |
| <i>that</i> is oscillated by means of a crank stud | F-2 |
| <i>which</i> extends eccentrically from a shaft | F-3 |
| <i>that</i> is rotatably mounted in a bracket | F-4 |
| and has a worm gear thereon | &F-4 |
| <i>that</i> is driven by a worm pinion | F-5 |
| <i>which</i> is mounted upon the drive shaft of | |
| the motor. | F-6 |
- (U.S. Patent; Yngve 1960:460)

The grammatical latitude of final embedding does of course not preclude the existence of usage-based regularities concerning the number and combinations of finally-embedded sub-clauses. Generative grammarians have always taken Meillet’s position as concerns the grammar of embedding complexity. On the other hand, they have readily conceded that there are performance factors inhibiting the complexity of genuine utterances and sentences (e.g. Chomsky 1965:10–15). The nature of these constraints has remained unclear.

It will be empirically shown below that there are constraints in the nature of stylistic preferences to avoid excessive final embedding in English, Finnish, German and Swedish. These preferences are similar in all these languages, and some can be detected as early as in Ancient Greek 500–300 BC, the oldest language variant on which comparable data on finally-embedded clauses are available. This suggests a more general cognitive basis for the preferences.

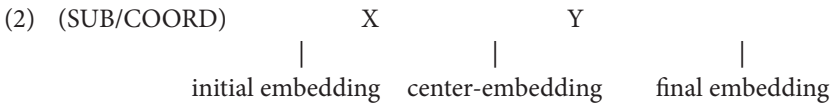
In Section 1 the central concepts are defined. The method is twofold. First, by automatic extraction and subsequent manual analysis we systematically scrutinize three well-known tagged English corpora, viz. Brown, LOB (Lancaster-Oslo/Bergen) and ICE-GB (British Component of The International Corpus of English), with the purpose of finding the most complex occurrences of multiple final embedding. These three corpora together contain an estimated 150,000 sentences which suffice for determining the basic tendencies. The method and the results are presented in Sections 2 and 3.

Second, in Section 4 an overview is provided of relevant systematic empirical investigations concerning English, Finnish, German and Swedish that have not been sufficiently considered in the literature. The overall clausal syntax of sentences in these languages is so similar that it is legitimate to compare quantitative data drawn from all of them. Several variety-dependent restrictions will be found to obtain.

In Section 5 some additional data are evaluated. Section 6 offers a discussion of the possibilities to explain the restrictions found in Sections 2–5 and a brief note on the historical origin of complex finally-embedded constructions.

The conceptual 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, and they are operationally determinable by instances of relevant subordinators or relative pronouns, henceforth called sub/wh-elements, italicized in (1) as well as in all subsequent examples. Non-finite clauses are signalled by subordinators or morphological markers like *to* (+ infinitive) or the ending *-ing*.

Schema (2) 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 (2) 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 and relative clauses, pattern “SUB/COORD X Y” other subordinate clauses than relative ones, and coordinate superordinate clauses:



The three embedding positions can now be defined across all types of superordinate clauses. An initially-embedded clause (abbreviated “I” if finite and “i” if non-finite) occurs either before all words of its superordinate clause, as clause I-1 in (3), which occurs before the main clause *she thought*, or directly after the initial subordinator or coordinator of its superordinate clause, as I-2 in (3), which occurs 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. The main argument is that subordinators and coordinators are not syntactically as tightly integrated and real constituents in their clauses as ordinary full constituents and relative pronouns are.

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

An **embedding chain**, **e-chain**, is a consecutive sequence of sub-clauses embedded one below the other. An e-chain is described by a sequence of characters expressing the positions and finiteness (I, C, F, if finite; i, c, f, if non-finite) of the sub-clauses, starting at depth 1. The e-chains in (3) are II (finite initial embedding at depth 1, containing another finite initial embedding at depth 2), If (finite initial embedding at depth 1 with a non-finite final embedding at depth 2), and FC (finite final embedding at depth 1 with a finite center-embedding at depth 2). The e-chain in (1) is FFFFFFF. An e-chain is **rooted** in the superordinate clause of its topmost clause. The e-chain in (1) is thus rooted in the main clause, M.

The **degree** of embedding of an e-chain is the amount of some positional type of embedding found in the chain. The degree of initial embedding in (3) is 2. Degrees are abbreviated by exponents, e.g. I² for double initial embedding, or F⁶ for sixfold final embedding as in (1).

Multiple embeddings are of a degree greater than 1. The individual clauses in a multiple embedding are referred to by expressions like I-1, the *if*-clause initially-embedded in (3) at depth 1; or F-6, the last finally-embedded clause in (1).

2. Method and empirical data on the maximal degree of final embedding

The syntactically most complex parts of the machine-readable tagged Brown, LOB, and ICE-GB corpora (altogether three million words) were systematically searched. First, graphical sentences (delimited by full stops and their equivalents) with at least four sub/wh-elements were automatically extracted from Brown and LOB, yielding 1,126+1,134 = 2,260 very complex sentences, and these were then subjected to manual analysis. This procedure disclosed some 50 fully finite instances of F⁵ and F⁶, and just one uncontroversial instance of fully finite F⁷, example (4), where even depth 8 is reached when the non-finite clause f-5 is included. When the full array of non-finite clauses are considered as well, Brown displays an instance of F/f¹⁰, (5). Example (6) contains four infinitival clauses. These instances from the Brown corpus are likely to be about as deep as one can expect to encounter in any type of spoken or written English text (elliptical relative pronouns and complementizers in parentheses).

- | | |
|---|-----|
| (4) Or you could hope | M |
| (<i>that</i>) the parachute wouldn't open just | F-1 |
| so you could say | F-2 |
| (<i>that</i>) you saw it not open, | F-3 |
| not <i>because</i> you meant any harm to Starkey in his suit of | F-4 |
| red underwear, | |

- | | |
|--|------|
| but mainly <i>because</i> you were tired of | &F-4 |
| <i>being</i> an old maid — a thing | f-5 |
| <i>which</i> cannot admit | F-6 |
| <i>when</i> it thinks | F-7 |
| (<i>that</i>) it might be pregnant, | F-8 |
| but must stand the dizzy feeling all alone | &F-6 |
| and go on | &f-6 |
| <i>like</i> everything is all right | F-7 |
| instead of <i>being</i> able | &f-6 |
| <i>to say</i> to somebody in a normal voice: | f-7 |
| I think I'm pregnant. | |
| | |
| (5) I fled, however, not from | M |
| <i>what</i> might have been the natural fear of | F-1 |
| <i>being</i> unable | f-2 |
| <i>to disguise</i> from you | f-3 |
| <i>that</i> the things about my bridegroom | F-4 |
| — in the sense you meant the word things — | C-5 |
| <i>which</i> you had been galvanizing yourself | C-5 |
| <i>to tell</i> me as a painful part of your maternal duty | f-6 |
| were things | |
| <i>which</i> I had already insisted | F-5 |
| upon finding out for myself (despite, | f-6 |
| I may now say, | C-7 |
| the unspeakable awkwardness | |
| of making the discovery on principle, yes, | f-7 |
| on principle, and in cold blood) | |
| <i>because</i> I was resolved, as a modern woman, | F-8 |
| not <i>to be</i> a mollycoddle | f-9 |
| <i>waiting</i> for Life | f-10 |
| but <i>to seize</i> Life by the throat. | &f-9 |
| | |
| (6) [_M ... it seems ... [_{F-1} <i>that</i> ... it is unlikely [_{f-2} <i>to bear</i> much relation ... to the | |
| ... need of [_{f-3} <i>ensuring</i> ... [_{F-4} <i>that</i> there may be adequate opportunity given | |
| to the staff [_{f-5} <i>to do</i> [_{F-6} <i>what</i> they can [_{f-7} <i>to get</i> the man [_{f-8} <i>to stand</i> again | |
| ...]]]]]]]]]] | |

The morphologically and syntactically tagged ICE-GB comes with a search system offering full integration of lexical, morphological and syntactic information. This corpus provides one instance each of fully finite F⁵, F⁶, and F⁸. The F⁸ deserves presentation as it ends with a further non-finite f-9 and even contains a double center-embedding C² before F/f⁹, (7).

- (9) [Then came the Holy One and slew the Angel of Death [*that* killed the chef [*that* cooked the bread [*that* inspired the meal [*that* ... (+ 90 relative that-clauses) [*that* flavored the goat [(*that*) my father bought for two zebeks] — one little goat!...]...]]]]] (<http://www.pastemob.org/trap/042.html>, accessed October 2006)

Do (8) and (9) prove (i) that unlimited final embedding exists, and (ii) that unlimited final embedding is an important design feature of syntax? The answer to (i) is a conditional yes, to (ii) no. The basic motivation for these answers is that (8) is a cumulative folk tale, a conventionalized genre based on language play and utilizing relative clauses only — similar folk tales (or other genres) employing only multiple complement or adverbial clauses do not occur. Sentence (8) was first published in 1755 and later became modified in countless versions: *The World That Jack Built*, *The Picture That Mom Drew*, *The House That Drac Built*, *The World That God Made*, etc. Jack-sentences of type (8) are obvious stereotypes not to be found in ordinary discourse. Sentence (9) is artificially modelled on the same idea — its author promised to deliver unlimited successive heaps of relative clauses against a modest fee.

Repeated instances of embedding of the same type of clause (several relative clauses, several *if*-clauses, several *that* COMPL-clauses, etc.) constitute the special type of embedding called **self-embedding**. A soft constraint, i.e. a tendency governing language use, can now be formulated concerning final embedding of clauses: **relative clauses alone allow unlimited final (self-)embedding, but only in stereotypical discourse.**

There is no corpus evidence of final self-embeddings of the magnitudes under discussion that would consist of e.g. twelve *that* COMPL-clauses or six *because*-clauses. Note that the maximally complex instances of multiple final embedding found in Brown, LOB, and ICE-GB (4–7) all contain an assortment of various clause types. Therefore the obviously invented F⁷ with multiple *that* COMPL-clauses presented by Akmajian et al. (1985:163) does not count as counterevidence falsifying the prohibition of other self-embeddings than relative clauses: here, the linguists' intuitions go astray.

Thus, Jack-sentences instantiating the constraint on final self-embedding represent stereotypical, functionally and structurally restricted language use, unheard and unseen in everyday discourse. Jack-sentences have an alternative based on juxtaposition, as in *The Old Woman and Her Pig* (10):

- (10) As soon as the cat had lapped up the milk, the cat began to kill the rat; the rat began to gnaw the rope; the rope began to hang the butcher; the butcher began to ...

Juxtaposition, coordination and subordination are related types of chaining. Relative pronouns are equivalent to sequences of coordinating conjunctions and anaphoric pronouns, e.g. *who* = *and (s)he*, *which* = *and it* (Chafe 1988:21). Jack-sentences are convertible to juxtaposed or coordinated clauses. For example, ... *sowing the corn, that kept ...*, as in (8), can be converted to a sequence of coordinated clauses, either ... *sowing the corn; the corn kept...* or ... *sowing the corn and it kept...* Repeated final embedding (just like repeated initial embedding as in (3)) is **tail-recursion**, formally equivalent to **iteration** at the same depth, like chains of coordinated clauses (also cf. Karlsson 2010). Jack-sentences with F^{12} are not perceived to be more complex than juxtaposed or coordinated variants retaining depth 1. Rather, Jack-sentences are simpler than juxtaposed or coordinated alternatives because they have less repetition of full NPs. Cumulative folk tales are the simplest of all stories and so much enjoyed by children because of their maximal rhythm and minimal plot where “episodes follow each other neatly and logically in a pattern of cadenced repetition” (Sutherland & Arbuthnot 1977:146).

4. Comparison to data drawn from the literature

Sections 2 and 3 presented the results of our systematic search for maximally complex multiple final embeddings. Section 4 checks how these results compare to previous systematic studies of clausal embedding complexity.

Several sources estimate the ratio of finally-embedded clauses to all embedded clauses to be around 0.8 in non-complex language use, i.e. four out of five embedded clauses are finally-embedded. For instance, Saari (1975) reports 82% for spoken Swedish, Varantola (1984) 82% for English periodicals and Sunday newspapers, Hakulinen et al. (1980) 81% for Finnish newspapers, periodicals and encyclopaedias, and Karlsson (2007b) 76% for a systematic balanced sub-sample of the Brown Corpus (including both finite and non-finite clauses; 70% if finite clauses alone are counted).

In syntactically complex varieties the shares of initial and especially of center-embeddings grow: Hiltunen (1984) found only 60% finally-embedded clauses in English legal language.

The most comprehensive statistics on embedding depth known to us is provided by Ikola et al. (1989) in their analysis of the syntax of spoken Finnish dialects (for which they examined 54,300 sentences, 166,000 finite clauses; 885,000 words) and written standard Finnish (for which they examined 15,600 sentences, 27,300 finite clauses; 191,000 words). The three embedding positions (initial, center-, final) were not distinguished in this study but even so Table 1 certainly shows the essentials of the most frequent final embedding position.

Table 1. Depth of finite clausal embedding in spoken and written Finnish (from Ikola et al. 1989: 18)

Depth	Spoken		Written	
	N	%	N	%
1	42,864	84.5	5,863	85.4
2	6,884	13.6	877	12.8
3	858	1.7	94	1.4
4	109	0.2	12	0.2
5	11		11	
6	2		5	
7	3		3	
Sum	50,731	100.0	6,865	99.8

Each depth decreases by a factor of roughly seven down to depths 4–5. The distribution of depths is almost identical in spoken and written Finnish, which (somewhat surprisingly) are equally complex in this regard. Here it must be added that the variety of spoken Finnish analyzed by Ikola et al. was interview speech of the type much used in classical dialectology where the informants answer questions about times long gone. In such speech there are many reporting sub-clauses (“then she said that ...”) which decisively contributes to the surprisingly high incidence of embeddings.

Ikola et al. (1989: 19) consider depth 4 to be a breakpoint for multiple embedding of finite clauses and are suspicious of the coding and analysis of all embeddings deeper than 4, especially those occurring in spoken language. Had Ikola et al. considered also non-finite clauses in their study, the breakpoint can rather be estimated to have been 5 (taking into account the proportions between finite and non-finite embeddings, cf. Karlsson 2007a).

Syntactically complex varieties can be expected to contain more embeddings below level 1 than simple varieties do. This conjecture is confirmed by Hiltunen’s (1984) results for English “legalese” (Table 2), frequently considered to be the syntactically most complex variety of all.

The column ‘ratio’ (between the current depth and the next lower one) shows that each depth lower than 1 is much more frequent in legal language than in Ikola et al.’s non-complex varieties where the ratios tend to be 6–7 throughout. Especially the incidence of embeddings at depths 2–4 is markedly higher in legal language, and so is the incidence of clauses embedded at levels deeper than 5 even if the absolute numbers are small. The extreme “legalese” depths conform to our results in Sections 2–3.

Table 2. Depth of final clausal (finite and non-finite) embedding in English legal language (from Hiltunen 1984: 118)

Depth	N	%	ratio
1	403	52.1	2.3
2	176	22.7	1.5
3	121	15.6	3.3
4	37	4.8	2.2
5	19	2.5	1.5
6	13	1.7	6.5
7	2	0.3	1
8	2	0.3	2
9	1	0.1	
Sum	774	100.1	

Note. Ratio = relation between any embedding depth and the next lower one

Ellegård's (1978: 20–26) English sub-sample from Brown contained some 150 (finite or non-finite) clauses per 1,000 words; depending upon text type .1%–.4% reached depth 6, which was the maximum depth encountered. This would roughly mean that depth 6 is reached once in 5,000–10,000 words.

Admoni's (1987) corpus of 19th-century German encyclopaedic articles and belles lettres (36,400 words) reached depth 4 only thrice. Hodler (1969: 645) found that depth 3 was hardly ever reached in the syntax of the Bern German vernacular. Similarly, when I analyzed the clausal embedding complexity of the twenty *Pearl Story* speakers (data from Chafe 1980: 301–319), the deepest embeddings found were four F-3s. Laury & Ono's (2010) Finnish and Japanese conversation data (1,000 randomly picked clausal units each) displayed depth 4 three times (Finnish) and six times (Japanese).

Jørgensen (1978) analyzed subordination in spoken Swedish (6,808 sentences, 4,303 sub-clauses, non-finite ones included, four discourse types, men and women, academics and factory workers). Of 1,994 *att* "that"-clauses, 75 reached F-3, which was represented in all eight sub-corpora. F-4 occurred 18 times but was unrepresented in two sub-corpora. There were eight instances of F-5 + F-6 + F-7, but in very few text types.

Westman (1968) found no clauses below F-3 in fourteen Swedish textbooks from a 100-year period (100,000 words). Danielsson (1975) spotted one F-4 in Swedish secondary school textbooks (7,600 sentences). In a Swedish official regulation (187 sentences) Gunnarsson (1982) found 341 sub-clauses of which eight were F-3s. The Swedish projects "Spoken Syntax" (300,000 words) and "Written

Syntax” (175,000 words) uncovered “very few” sub-clauses at depths below 4 (Teleman 1974: 33).

Sinnemäki (2004) found no e-chains containing only finite clauses beyond depth 4 in a subcorpus of 12.6 million words of journalistic writing and 400,000 words of prose from the Finnish Language Bank. However, outside that data set, the first 55,000 words of the markedly artistic novel *Alastalon salissa* (“In the living room of Alastalo”, by Volter Kilpi) contained seven F-5s, two F-6s, and one F-7.

To pinpoint the effect of non-finite clauses on final embedding complexity, all 1,337 sentences in the Brown corpus with at least two *to*-infinitive clauses were analyzed with equal consideration of finite and non-finite clauses. One instance was found of degree 8, two of 7, 13 of 6, and 39 of degree 5, for a summed incidence of 4%.

Table 3 sums up this review of data available in the relevant literature. By ‘maximal productive depth’ is meant the maximal depth observed in the sources mentioned, disregarding extremely rare depths instantiated only once or twice.

Table 3. Maximal productive depth (MPD) of final embedding

MPD	Sources
3	Hodler (1969), Pear Story speakers, Westman (1968), Danielsson (1975), Gunnarsson (1982)
4	Admoni (1987), Jörgensen (1978), Teleman (1974), Sinnemäki (2004), Laury & Ono (2010)
5	Ellegård (1978), Ikola et al. (1989)
6	Hiltunen (1984), Brown Corpus

All these data indicate a variety-dependent stylistic tendency for English, Finnish, German and Swedish to constrain the degree of final embedding. We call this tendency ‘ $F^{3-5}\text{max}$ ’:

$F^{3-5}\text{max}$: Syntactically simple varieties (such as everyday conversation, textbooks) avoid multiple final clausal embedding in excess of degree 3, complex varieties (i.e. most of written language) in excess of degree 5.

$F^{3-5}\text{max}$ is true especially for sequences of finally-embedded finite clauses. All our sources are not fully clear about whether non-finite clauses have been included in the counts. Even when non-finite clauses are included, depth 5 is transcended in less than one sentence out of 100 (as shown by our Brown Corpus data), and depth 6 in one or two sentences out of 1,000.

Of course, grammatical sentences transcending $F^{3-5}\text{max}$ do occur, especially in highly complex varieties like legal language, but overall they are utterly rare. Admoni (1980: 44–49) pointed out an extreme instance of this type in his detailed

study of the development of written German sentence structure, especially that of “bureaucratese”, during the period 1470–1730. A monstrous German 790-word sentence from 1411 with 42 sub-clauses occurred in an official letter from the Archbishop of Trier to the City of Frankfurt. Thirty-nine of its sub-clauses precede the main clause and the deepest embedding reaches F-15.

5. Interplay of final with initial and center-embedding

Normally an e-chain of finally-embedded clauses is rooted in the main clause, as in (1). But it may also be more deeply rooted. In (11) F/f⁵ is rooted in I-1:

- (11) [_M [_{I-1} *If* his circumspection in regard to Philip’s sensibilities went so far [_{F-2} *that* he even refused [_{f-3} *to grant* a dispensation for the marriage of Amadee’s daughter, Agnes, to the son of the Dauphin of Vienne — a truly peacemaking move according to thirteenth-century ideas, [_{F-4} *for* Savoy and Dauphine were as usual fighting on opposite sides — for fear [_{F-5} *that* he might seem [_{f-6} *to be* favoring the anti-French coalition,]]]]]] he would certainly never take ...] (Brown)

An indication that this structure is approaching some crucial limit of processing complexity is the fact that Mark Liberman inaugurated his (intentionally amusing!) “Trent Reznor prize for tricky embedding, to be awarded intermittently” for an instance precisely of e-chain IF/f⁵ (12):

- (12) [_M [_{I-1} *When* I look at people [_{F-2} *that* I would like [_{f-2} *to feel* [_{F-3} *have been* a mentor or an inspiring kind of archetype of [_{F-4} *what* I’d love [_{f-5} *to see* my career [_{f-6} *eventually be mentioned* as a footnote for in the same paragraph,]]]]]] it would be, like, Bowie.] (<http://itre.cis.upenn.edu/~myl/languageelog/archives/002621.html>)

In (13), from De Roeck et al. (1982), an instance of F/f⁶ is rooted one level further down and in an exceptionally complicated way in IC, starting at depth 3, where the upper C-2 simultaneously contains a C-3, yielding a double center-embedding before the sextuple F/f⁶ starts:

- | | | |
|------|---|------|
| (13) | <i>When</i> he asked the very unpromiscuous Noël Coward | I-1 |
| | (<i>whose</i> ... financially promising “Life” he was engaged upon ... | C-2 |
| | and <i>whom</i> ... he had accurately discerned as “one of the kindest, | &C-2 |
| | <i>as</i> he is without doubt the wittiest, | C-3 |
| | of human beings”; one | |
| | <i>who</i> incidentally shared with James the rare quality, | F-3 |
| | strongly emerging in this book, | c-4 |

of		
	being able	f-4
	to be extremely funny about love and sex,	f-5
	both men having plenty of time for women's beauty	f-6
	and brains	
	if little for their bodies	f-7
	even when androgynous)	f-8
	whether he thought it "disorderly" of him	F-2
	to be housing a Horsegard,	f-3
	he was urbanely told: ... (<i>The Spectator</i> 25.4.1981)	M

The deeper any embedding is, the less likely it is to contain further embedded clauses. At any depth, the most likely further embedding is a finally-embedded clause. But occasionally initially- and center-embedded clauses may occur in final embeddings, even at deeper levels. Sentence (5) contained two center-embedded clauses, one at depth 5 and one at depth 7, in the middle of an e-chain of final embeddings reaching depth 10. In (13) F-3 contains the non-finite center-embedding c-4. In (14) F-2 contains an initial embedding I-3 which contains a further center-embedding C-4, e-chain fFIC. Sentence (15) displays an initially-embedded sentential subject in e-chain FFI.

(14)	To do this,	i-1
	it is sufficient	M
	to point out	f-1
	that	F-2
	if the principle	I-3
	in terms of which alternatives are to be conceived	C-4
	is such	
	as to exclude more than two,	F-4
	then the question of a third possibility is a meaningless question.	
	(Brown)	
(15)	People say	M
	(that) we need religion	F-1
	when	F-2
	what they really mean	I-3
	is	
	we need police	F-3
	(H.L. Mencken)	

Again, it is instructive to refer to one of Mark Liberman's "Trent Reznor Prizes for Tricky Embedding", this time awarded to Andrew Ilachinsky, for having produced sentence (16):¹

- | | |
|---|------|
| (16) It is nonetheless tempting | M |
| <i>to speculate</i> about | f-1 |
| <i>whether</i> there exists () — | F-2 |
| and, | &F-2 |
| if so, | I-3 |
| what the properties are, of — a universal grammar of combat. | |
| (http://itre.cis.upenn.edu/~myl/languagelog/archives/005531.html) | |

Like (15), (16) contains an initial embedding in the lower clause of a double final embedding, but it is further complicated by the long cataphoric pronominal reference from () to *a universal grammar of combat*, which crosses the initially-embedded pronominalized sentence I-3, for which the antecedent of *so* in turn is found backwards beyond the preceding trace ().

4. Discussion and conclusion

Even if multiple final embedding of clauses is grammatically unlimited, we have observed a clear preference in language use to constrain the number of embeddings as prescribed by F^{3-5} max. In simple varieties such as everyday speech and textbooks, final clausal embeddings at depths below 3 are extremely rare, and so are embeddings deeper than 5 in more complex (especially written) varieties. If an explanation of these tendencies is possible, it must obviously be sought among human cognitive properties, in particular short-term memory (STM) restrictions on discourse management both in language production and language understanding.

Jarvella (1971, 1979) demonstrated the importance of clauses and their boundaries for short-term retention of discourse. A sequence of spoken sentences was interrupted and the subjects had to recall as many words as possible. They usually recalled the last clause, sometimes the one before that, with clear breaks in recall performance occurring at clause boundaries. Clauses were used for structuring discourse to be kept in short-term memory, one or two clauses being retained verbatim. Glanzer & Razel (1974) demonstrated that simplex sentences functioned as units. The recall results deteriorate from one to two clauses but the decisive break is between two and three. Pawley & Syder (1983:564–565) had a limit similar to that of Jarvella in mind when they proposed the one-clause-at-a-time constraint for spoken language processing. It refers to a single integrated sequence of encoding actions, a single span of attention focus to plan word for word the content of a novel clause of up to about ten words. Chafe's (1985:106) seven-word idea unit is similar. Kimball's (1973) principle of two clauses allows primary two-clause

chunks. The common denominator of all these proposals is that 1–2 clauses is the basic syntactic chunk of spoken language processing, the second of these clauses typically being a short relative clause. Closely related to these structural principles are the constraints on the normal amount of new information per spoken message, e.g. Chafe's (1994: 109) principle of one new idea per intonation unit.

In "The magical number 4 in short-term memory", Cowan (2000) reviewed the large body of research conducted since Miller (1956) presented his famous theory of STM working on 7 ± 2 chunks. Cowan arrived at a processing span of maximally four chunks where one chunk might be an isolated datum or a scene integrated as one whole. Chunking takes place at least on the levels of syntax and discourse representation. Chunk sizes might differ from one level to another. As clauses are chunks with sub-constituents, only a few clauses can be successfully processed at a time.

The limit F^3 max which has been detected in simple varieties is in perfect harmony with these experimental results, especially in view of the fact that even depth 3 is very rare in ordinary conversation. Recall that in the Pear Stories there were only four instances of depth 3, and Hodler (1969) reported that it was hardly ever reached in the Bern German vernacular.

It is to be expected that many written varieties are more complex. This is captured by the tendency to comply with F^5 max. For both the reader and the writer, written language offers the obvious possibility of returning to the beginning of a complex sentence to update discourse information that was overshadowed or forgotten due to a multitude of material in later more deeply embedded sub-clauses. But "returning to the beginning" is an uneconomical method both of writing and of reading comprehension, compared to smooth simultaneous production and understanding of simpler sentence constructions with less embeddings. Trying to write a very complex sentence structure easily leads to grammatical mistakes (e.g. improper agreement), anacolutha etc. Therefore it is rational for the central functions of speaking, hearing, writing and reading to use constructions the complexity of which does not transcend the limits of the processing mechanism.

As for the historical emergence of multiple final embedding in written language, the earliest relevant data known to us concern Ancient Greek.² Webster (1941) studied the development of sentence complexity in Ancient Greek, especially of final embedding. F^3 was used already by Homer around 700 BC but he has only one instance of it. This shallow syntactic depth of the *Iliad* and *Odyssey* is in good keeping with current theories of their origin in oral narrative. F^4 was first used by Herodotus, Sophocles, and Thucydides around 450 BC, and F^5 first by Xenophon a century later. Demosthenes and Plato both have one F^6 and, as an extreme, Dinarchus (360–292 BC) one F^9 . Overall, even F^3 was rare in Ancient Greek (Schwyzer 1950: 710). These historical data and the relative incidences of

the various embedding depths are in perfect keeping with the data from present-day languages reviewed above (see Karlsson 2009 for a more detailed treatment of the origin of clausal embedding complexity, including initially- and center-embedded clauses).

One may conclude that the tendency to comply with the preference F^{3-5} max of final embedding in written language was reached in Ancient Greek in less than 500 years and that it has remained as such ever since. This is natural because F^{3-5} max is ultimately rooted in cognitive constraints on short-term memory management and discourse processing. There is no evidence that these would have changed over the past 2000–3000 years.

Notes

* I am grateful for constructive comments provided by two anonymous reviewers and by Kaius Sinnemäki. This paper is an elaboration and revision of views first presented briefly in Karlsson (2002).

1. The example appeared in the paper “Exploring self-organized emergence in an agent-based synthetic warfare lab”, *Kybernetes*, 32 (1/2): 38–76, 2003.
2. I am not aware of any relevant historical studies for English, Finnish, German or Swedish concerning this particular aspect of the emergence of syntactic complexity.

References

- Admoni, W. G. 1980. *Zur Ausbildung der Norm der Deutschen Literatursprache im Bereich des Neuhochochdeutschen Satzgefüges (1470–1730)*. Berlin: Akademie-Verlag.
- Admoni, W. G. 1987. *Die Entwicklung des Satzbaus der Deutschen Literatursprache im 19. und 20. Jahrhundert*. Berlin: Akademie-Verlag.
- Akmajian, A., Demers, R. A. & Harnish, R. M. 1985. *Linguistics. An Introduction to Language and Communication*. 2nd ed. Cambridge, MA: MIT Press.
- Chafe, W. (Ed.) 1980. *The Pear Stories: Cognitive, Cultural, and Linguistic Aspects of Narrative Production*. Norwood, NJ: Ablex Publishing Corp.
- Chafe, W. 1985. “Differences between speaking and writing”. In D. R. Olson, N. Torrance & A. Hildyard (Eds.), *Literacy, Language and Learning. The Nature and Consequences of Reading and Writing*. Cambridge: Cambridge University Press, 105–123.
- Chafe, W. 1988. “Linking intonation units in spoken English”. In J. Haiman & S. A. Thompson (Eds.), *Clause Combining in Grammar and Discourse*. Amsterdam/Philadelphia: John Benjamins, 1–27.
- Chafe, W. 1994. *Discourse, Consciousness, and Time*. Chicago: The University of Chicago Press.
- Chomsky, N. 1965. *Aspects of the Theory of Syntax*. Cambridge, Mass.: The MIT Press.

- Cowan, N. 2000. "The magical number 4 in short-term memory: A reconsideration of mental storage capacity". *Behavioral and Brain Sciences*, 24 (1), 87–114.
- Danielsson, S. 1975. *Läroboksspråk*. Umeå: Acta Universitatis Umensis 4.
- De Roeck, A., Johnson, R., King, M., Rosner, M., Sampson, G. & Varile, N. 1982. "A myth about centre-embedding". *Lingua*, 58 (3), 327–340.
- Ellegård, A. 1978. *Syntactic Structure of English Texts*. Göteborg: Gothenburg Studies in English 43.
- Glanzer, M. & Razel, M. 1974. "The size of the unit in short-term storage". *Journal of Verbal Learning and Verbal Behavior*, 13 (1), 114–131.
- Gunnarsson, B.-L. 1982. *Lagtexters begriplighet*. Lund: LiberFörlag.
- Hakulinen, A., Karlsson, F. & Vilkuina, M. 1980. *Suomen tekstilauseiden piirteitä: kvantitatiivinen tutkimus*. Helsinki: Publications of the Department of General Linguistics, University of Helsinki, No. 6.
- Hiltunen, R. 1984. "The type and structure of clausal embedding in legal English". *Text*, 4–1 (3), 107–121.
- Hodler, W. 1969. *Berndeutsche Syntax*. Bern: Franke Verlag.
- Ikola, O., Palomäki, U. & Koitto, A.-K. 1989. *Suomen murteiden lauseoppia ja tekstikielioppia*. Helsinki: The Finnish Literature Society.
- Jarvella, R. J. 1971. "Syntactic processing of connected speech". *Journal of Verbal Learning and Verbal Behavior*, 10 (3), 409–416.
- Jarvella, R. J. 1979. "Immediate memory and discourse processing". In G. H. Bower (Ed.), *The Psychology of Learning and Motivation*, Vol. 13. New York: Academic Press, 379–421.
- Jörgensen, N. 1978. *Underordnade satser och fraser i talad svenska*. Lund: Walter Ekstrand Bokförlag.
- Karlsson, F. 2002. "Is there an upper limit to right-branching embedding of clauses?". In R. Pajusalu & T. Hennoste (Eds.), *Tähendusepüüdjä. Catcher of the Meaning. Festschrift for Professor Haldur Öim on the Occasion of His 60th Birthday*. Tartu: Publications of the Department of General Linguistics, University of Tartu, 196–199.
- Karlsson, F. 2007a. "Constraints on multiple initial embedding of clauses". *International Journal of Corpus Linguistics*, 12 (1), 107–118.
- Karlsson, F. 2007b. "Constraints on multiple center-embedding of clauses". *Journal of Linguistics*, 43 (2), 365–392.
- Karlsson, F. 2009. "Origin and maintenance of clausal embedding complexity". In G. Sampson, D. Gil & P. Trudgill (Eds.), *Language Complexity as an Evolving Variable*. Oxford: Oxford University Press, 192–202.
- Karlsson, F. 2010. "Recursion and iteration". In H. van der Hulst (Ed.), *Recursion and Human Language*. Berlin/New York: Mouton de Gruyter, 43–67.
- Kimball, J. 1973. "Seven principles of surface structure parsing". *Cognition*, 2 (1), 15–47.
- Laury, R. & Ono, T. 2010. "Recursion in conversation: What speakers of Finnish and Japanese know how to do". In H. van der Hulst (Ed.), *Recursion and Human Language*. Berlin/New York: Mouton de Gruyter, 69–92.
- Meillet, A. 1934. *Introduction a l'Étude des Langues Indo-européennes*. Paris: Librairie Hachette.
- Miller, G. 1956. "Human memory and the storage of information". *I. R. E. Transaction on Information Theory IT-2*, 129–137.

- Pawley, A. & Syder, F. G. 1983. "Natural selection in syntax: Notes on adaptive variation and change in vernacular and literary grammar". *Journal of Pragmatics*, 7 (4), 551–579.
- Quirk, R., Greenbaum, S., Leech, G. & Svartvik, J. 1989. *A Comprehensive Grammar of the English Language*. London: Longman.
- Saari, M. 1975. *Talsvenska*. Helsingfors: Svenska Litteratursällskapet.
- Schwyzler, E. 1950. *Griechische Grammatik auf der Grundlage von Karl Brugmanns Griechischer Grammatik. Zweiter Band. Syntax und Syntaktische Stilistik*. München: C. H. Beck'sche Verlagsbuchhandlung.
- Sinnemäki, K. 2004. "Complex right-branching clauses". Unpublished MA thesis, Department of General Linguistics, University of Helsinki, Helsinki.
- Sutherland, Z. & Arbuthnot, M. H. 1977. *Children and Books*. Glenview, Ill.: Scott, Foresman & Co.
- Teleman, U. 1974. *Manual för grammatisk beskrivning av talad och skriven svenska*. Lund: Studentlitteratur.
- Varantola, K. 1984. *On Noun Phrase Structures in Engineering English*. Turku: Annales Universitatis Turkuensis B:168.
- Webster, T. B. L. 1941. "A study of Greek sentence construction". *American Journal of Philology*, 62 (3), 385–415.
- Westman, M. 1968. "Meningar och satsfogning i historieläroboksspråk från de senaste hundra åren". Unpublished licentiate's thesis. Stockholm: University of Stockholm.
- Yngve, V. H. 1960. "A model and an hypothesis for language structure". *Proceedings of the American Philosophical Society* 104, 444–466.

Corpora

- Brown Corpus. A Standard Corpus of Present-Day Edited American English, for use with Digital Computers. *The Brown Corpus Manual* by W. N. Francis & H. Kučera, 1971, revised and amplified 1979. Available at: <http://www.hit.uib.no/icame/brown/bcm.html>.
- LOB Corpus. The Tagged LOB Corpus. User's Manual by S. Johansson in collaboration with E. Atwell, R. Garside & G. Leech, 1986. Available at: <http://www.hit.uib.no/icame/lobman/lob-cont.html>.

Author's address

Fred Karlsson
 Department of Modern Languages
 P.O. Box 24
 FI-00014 University of Helsinki
 Finland
 fgk@ling.helsinki.fi