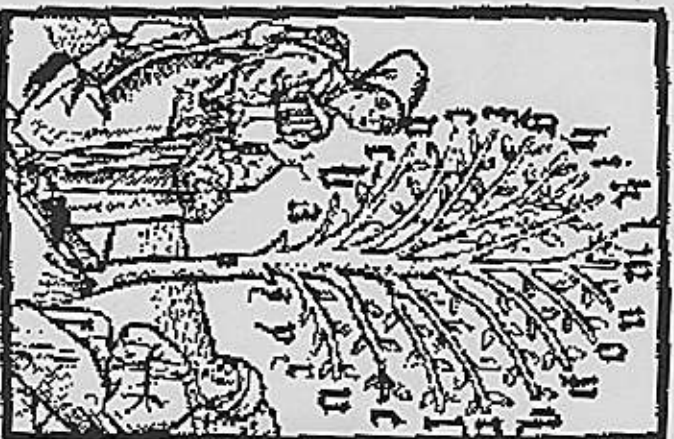


Tartu Ülikooli üldkeeleeaduse õppetooli toimetised 3
Publications of the Department of General Linguistics 3
University of Tartu

Tähendusepüüdja Catcher of the Meaning



Tartu 2002

Is there an upper limit to right-branching embedding of clauses?

Fred Karlsson

University of Helsinki

A common belief especially in generative linguistics is that a sentence is a sequence of clauses with no essential upper bounds. Surprisingly little empirical work has been done in 20th-century linguistics to check the truth of this assumption even if research on sentence composition has venerable traditions in philology and stylistics.

Clauses may be embedded in their matrix clauses in three different positions: initial, medial, and final. Consequently we may talk about initial, center-, and final embedding. Repeated initial embedding creates left-branching structures, repeated center-embedding self-embedded structures, and repeated final embedding right-branching structures. My topic in this paper, written to celebrate the 60th birthday of my old friend Haldur Öim, is whether there are restrictions on right-branching.

Linguistic intuition is fragile in other than so-called clear cases and cannot be expected to yield a reliable answer. I shall approach the problem in two other ways: by consulting available literature and existing large machine-readable corpora.

A terminological note is in order concerning the notion 'clause'. By clause I mean 'finite clause'. At the first stage of the analysis, it is important to keep finite and non-finite constructions apart because the claims will be different if all infinitival and participial constructions are included along with the finite ones. Thus, note the following example taken from the British part of the International Corpus of English (ICE-GB, one million words). Progressive levels of indentation indicate deeper levels of embedding:

	ICE-GB	FK
Selling is	depth 0	depth 0
what you do	depth 1	depth 1
to persuade people	depth 2	
to buy today	depth 3	
what you've got	depth 4	depth 2
to offer to them today.	depth 5	

Given this restrictive finite-based definition of depth we now proceed to discuss some data. Alvar Ellegård (1978) used a sub-corpus of 128,000 words from the Brown Corpus in his analysis of the syntactic structure of English texts. The genres he studied were popular fiction, journalism, literary essays, and science. As clauses he accepted also non-finite verbs not governed by auxiliaries, which of course increases the syntactic depths reported. The total number of (finite and non-finite) clauses was 17,900. The maximum depth observed was 5 of which Ellegård (1978: 27-8) found around 50 instances, with the relative frequency 0.1-0.6% in the four registers being investigated. Depths 3 and 4 were collapsed in his statistics. Their relative frequencies ranged between 4% and 13%. As the 50 occurrences of depth 5 also contain non-finite clauses, whose share was almost 25% of all clauses, it can be safely concluded that pure sequences of five finite final embeddings must have been very rare in his sub-corpus if occurring at all.

Ikola *et al* (1989) investigated the syntax of spoken Finnish dialects and written standard Finnish using most of the machine-readable tagged corpus of the Finnish Syntax Archives (*Lauseopin arkisto*), University of Turku. The spoken corpus contains some 54,300 sentences (166,000 finite clauses; 885,000 words), the written part 15,600 sentences (27,300 finite clauses; 191,000 words). The material is not split up on initial, center- and final embedding, but even so it is equally relevant for assessing the maximal depth of embedding. Of course, the majority of the embeddings in this (and any) corpus are final.

Table 1. Frequency of depth of sub-clauses in spoken Finnish dialects and written standard Finnish (Ikola *et al* 1989: 18)

	Spoken		Written	
	N	%	N	%
1	42,864	84.5	5,863	85
2	6,884	13.6	877	13
3	858	1.7	94	1.4
4	109	0.2	12	0.2
5	11	(11)		
6	2	(5)		
7	3	(3)		
Total	50731	100	6865	100

The embedding depths of finite clauses in spoken and written Finnish are almost identical, which is interesting. Written language is not more complex in this regard, as many would expect. Embeddings occur down to depth 4, where a few instances may be encountered but they are extremely rare (0.2%). The corpus contains a couple of instances at depths 5-7 but the authors are hesitant about their interpretation and they state (p. 19) that these examples might be coding errors (therefore they are put in parentheses in table 1). All the data provided by Ellegård and Ikola et al. strongly indicate that 4 is the practical upper limit of final embedding of finite clauses, the 'usus maximum'. Instances of final embedding at 5-7 or even lower depths are extremely rare.

Similar results are easy to find in many other stylistic and philological studies of various languages. Final embeddings below depth 4 are extremely uncommon - although they might occasionally occur. An interesting example of depth 6 of final embedding was provided by Victor Yngve (1960: 460) in his classical paper on grammatical depth:

The said rocker level is operated by means of a pair of opposed fingers which extend from a pitman that is oscillated by means of a crank stud which extends eccentrically from a shaft that is rotatably mounted in a bracket and has a worm gear thereon that is driven by a worm pinion which is mounted upon the drive shaft of the motor.
(U. S. Patent)

The ICE-GB corpus is syntactically coded and therefore it is easy to spot structures according to desired search keys. In this corpus there is one instance of final embedding of depth 5, one of depth 6, and one of depth 8, which we quote here (capital 'F' indicates finite clauses, lower-case 'f' non-finite construction):

The reason is
that I think
this subject of symbolic representations of Christ has importance F-2
because you know F-3
that this was a problem F-4
if that's the word F-5
that assailed Byzantium in the twenties F-6
when they convinced themselves F-7

that we shouldn't have portraits of Christ
shown in human form.

F-8
f-9

Note the reduced non-finite clause extending down to depth 9. Of course, instances like this are utterly rare, even if some more may be found in nursery rhymes and folklore. Still, they do not (in my mind) falsify the generalization established above, that the *usus maximum* of final embedding is 4.

References

- Ellegård, Alvar. 1978. The Syntactic Structure of English Texts. A Computer-based Study of Four Kinds of Text in the Brown University Corpus. Göteborg, Sweden: Acta Universitatis Gothoburgensis. Gothenburg Studies in English, 43.
- Ikola, Osmo; Palomäki, Ulla; Koitto, Anna-Kaisa 1989. Suomen murteiden lauseoppia ja tekstikielipopia. Suomalaisen Kirjallisuuden Seuran Toimituksia 511. Helsinki: SKS.
- Yngve, Victor 1960. A model and an hypothesis for language structure. - Proceedings of the American Philosophical Society 104, 444-466.