Constraints on multiple initial embedding of clauses*

Fred Karlsson
University of Helsinki

The received view is that there are no constraints on clausal embedding complexity in sentences. This hypothesis will be challenged here on empirical grounds from the viewpoint of multiple initial embedding of clauses. The data come from the British National Corpus, Brown, LOB, and philological scholarship. The results extend to several other 'Standard Average European' (SAE) languages like Finnish, German, Latin, and Swedish. There is a precise quantitative constraint on the degree of initial clausal embedding, and that limit is two. In double initial embeddings, a qualitative constraint prescribes that typically the highest embedded clause is an *if*-clause. The lower embedded clause should be the sentential subject of the *if*-clause. Here is a real example of a maximally complex, prototypical, initial clausal embedding in mainstream SAE: [$_{\text{Main}}$ [$_{\text{Init–1}}$ *If* [$_{\text{Init–2}}$ *what is tantamount to dictatorship …*] *continues in a union*] *it can …*] (LOB). Multiple initial self-embeddings are prohibited.

Keywords: constraint, clause, embedding, initial embedding, clausal embedding, multiple embedding, self-embedding, syntactic complexity, recursion

1. Introduction

This paper is about constraints on initial embedding of clauses in sentences. **Constraints** are quantitative limits and combinatory restrictions, often in the nature of tendencies. The empirical data are mostly English but the constraints detected seem more generally valid in much of 'Standard Average European languages' (SAE), e.g. Finnish, French, German, Latin, and Swedish.

The data is mainly derived by systematic examination of the tagged machine-readable corpora British National Corpus (BNC, 100 million words),

International Journal of Corpus Linguistics 12:1 (2007), 107–118.
ISSN 1384–6655 / E-ISSN 1569–9811 © John Benjamins Publishing Company

Brown Corpus of American English (Brown, 1 million words), and Lancaster-Oslo/Bergen Corpus of British English (LOB, 1 million words). This was supplemented by less systematic computerized searches of similar patterns in machine-readable materials of the other languages mentioned, naturalistic observation, and consultation of some of the copious descriptive data accumulated over the centuries in syntactic and stylistic descriptions of several SAE languages, especially Latin (e.g. Nägelsbach 1963 [1846]) and older variants of German (e.g. Admoni 1980), both languages well-known for having reached heights of syntactic complexity.

There is a precise quantitative constraint on the degree of initial clausal embedding and that limit is two. Furthermore, there are specific qualitative constraints governing the composition of the members in double initial embeddings. For example, the subjunction of the upper clause must be *if*.

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). Without further discussion of the floating border between subordination and coordination (Haiman & Thompson 1988), the starting point will be the classical view of subordination (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* and *but*. The variables X, Y denote any other superordinate clause constituents. Thus, the pattern "X Y" covers simplex main clauses and relative clauses, the pattern "SUB/COORD X Y" other subordinate clauses than relative ones, and coordinate superordinate clauses:

```
(1)  (SUB/COORD)          X                         Y
                     |                     |                     |
              initial embedding     center-embedding     final embedding
```

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 (2), occurring before the main clause *she thought*, or directly after the initial subordinator or coordinator of its superordinate clause, as I-2 in (2), occurring after the initial subordinator *if* of its superordinate clause I-1. Note that 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.

(2)	If	I-1	finite initial embedding	depth 1
	as often happened	I-2	finite initial embedding	depth 2
	she asked him	I-1	the bulk of I-1	depth 1
	to tell her about it	f-2	non-finite final emb.	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	depth 1

Occasionally an I-2 may be altogether preposed before the subjunction of I-1:

(3) [$_{\text{M}}$ [$_{\text{I-1}}$ [$_{\text{I-2}}$ When the president of the US is rocking along …,] if I were the leader of the opposition party] I might…] (*Newsweek*, October 2, 1989, p. 12)

The *when*-clause is initially embedded in the *if*-clause but feels markedly preposed and almost coordinated with the *if*-clause. A more natural order would be for the *when*-clause to occur after the *if*-clause, i.e. to be finally embedded in it. The *when*-clause could even occur initially embedded immediately after *if*, cf. (6c) (Section 4). Initial linearizations like (3) above will not be treated further in this paper.

Center-embedded clauses (abbreviated C if finite and c if non-finite) 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**, also called **nesting**, is multiple center-embedding of the same type of clause, e.g. two relative clauses or, hypothetically, two *if*-clauses.

Finally embedded clauses (abbreviated F if finite and f if non-finite) occur after the last word of the superordinate clause, as f-2 and F-1 in (2).

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 (I, C, F; i, c, f) of the sub-clauses, starting at depth 1. The e-chains in (2) are II (a finite initial embedding containing another finite initial embedding), If (finite initial embedding with a non-finite final embedding), and FC (finite final embedding with a finite center-embedding).

The **degree** of initial, center-, or final embedding of an e-chain is the number of that type of clauses in the chain. The degree of initial embedding in (2) is two. Degrees are abbreviated by exponents: I^2 (double initial embedding), C^3 (triple center-embedding). **Multiple embeddings** are of a degree greater than one.

The **depth of a clause** is its level relative to the main clause. I-2 in (2) is at depth 2. The main clause is always at depth 0. In (2), progressive indentation reflects increasing depth.

An e-chain is **rooted** in the main clause, as II in (2) and FIIf in (6h). A partial e-chain is rooted lower, as IIf in (6h) rooted in F. Most initial embeddings are rooted in the main clause. But there are I^2s rooted at least one level lower (6h) (see Section 4), with I-2 and I-3. I-high denotes the higher, I-low the lower member in an I^2. In (2), I-high is I-1 and I-low is I-2, in (6h) I-high is I-2 and I-low is I-3.

3. State of the art

The mainstream view is that there are no restrictions whatsoever on clausal embedding complexity in any sentential position. This opinion has been explicitly voiced by many linguists from different camps: the comparatist Meillet (1934:355), the generativist Chomsky (1956:65, 1957:21), 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). This **hypothesis of unbounded clausal embedding complexity** will here be challenged with regard to initial clausal embedding.

The first attempt at a theory of clausal embedding complexity was Victor Yngve's papers on the Depth Hypothesis (1960, 1961). He discussed among other things hypothetical multiple initial embeddings and found (4b) awkward, (4a,c) ungrammatical, but (4d) plausible because it contains different types of clauses.

(4) a. *Because because because she wouldn't give him any candy, he called her names, she hit him, he cried.
b. ?That that it is true is obvious isn't clear.
c. *That that that they are both isosceles is true is obvious isn't clear.
d. ?That what the poem the woman he knows wrote implies, is obscure, is obvious.

Yngve (1960:461) surmised that the limit of initial embedding is three or four clauses. Given that I means initial embedding and that the superscripted index in I^n expresses the quantitative limit, Yngve's conjecture can be expressed as $I^{3\sim4}$max. Henceforth, expressions like I^2 will also be used to refer to the complexity level of individual sentences: (4a,c) are I^3s and (4b,d) I^2s.

Pinker (1994:205) judged the obviously made-up I^3s (5a,b) grammatical. According to him the sentences check out perfectly.

(5) a. If if if it rains it pours I get depressed I should get help.
b. That that that he left is apparent is clear is obvious.

Note that Yngve and Pinker disagree on the grammaticality of triple initial *that*-embeddings (4c, 5b), both basing their judgements on intuition alone.

Quirk et al. (1972:793) and Langendoen (1975:549) hypothesized that only one cycle of initial embedding is possible: I^1max. Quirk et al. (1989:1039) concluded that I^2 is awkward at best. This hypothesis we call $I^{1\sim2}$max.

Thus, there are four different claims in the literature concerning the quantitative grammaticality limit of multiple initial clausal embedding. Grouped according to increasing complexity they are: I^1max (Quirk et al. 1972, Langendoen), $I^{1\sim2}$max (Quirk et al. 1989, with I^2 marginal), $I^{3\sim4}$max (Yngve); I^{∞} (no limit: Pinker and others).

4. Empirical data on multiple initial embedding

Disregarding type (3), multiple initial embeddings are easy to spot in corpora because they start with several instances drawn from the closed class of sub/*wh*-elements, e.g. *if because what*, *when although if*, *after where*, where each sub/*wh*-element must start a clause in order to qualify as a relevant example. Such patterns may be searched for even in untagged corpora or on the Internet just by asking for character strings.

Of course, a conclusive search key reaching 100% recall is simple to formulate and apply if part-of-speech tagged corpora are available. A search was first made of BNC, Brown, and LOB for any third-degree (I^3) or second-degree (I^2) initial embeddings, with the search key "subordinator (+ subordinator) + sub/*wh*-element". Significantly, not a single instance of I^3 was found in the English corpora mentioned, nor were any found in Finnish or Swedish corpora exceeding three million words, nor have I found any genuine instances in any other corpora or sources, in any of the languages or grammar books, style manuals, philological studies etc. that have been consulted (the total number of which exceeds 100).

To assess the significance of this total absence of I^3 in more than 100 million words of diverse running text in three languages, consider the fact that in LOB, for instance, there are 50 subordinators alone with an overall token frequency of 19,439: *that* 7118, *as* 3328, *if* 2209, etc. The theoretical number of distinct initial subordinator triples like *if because although* is $50^3 = 125{,}000$; none occur. This absence can be nothing but the consequence of a strong constraint. The number 125,000 would be bigger if relatives like *what*, *whatever*, *where*, *whom* would be included that are possible as I-3s.

Further corroboration of the claim that the total empirical absence of I^3 follows from a strong constraint comes from Admoni's (1980) study of New High German sentence structure (1470–1730). Admoni analyzed 450 sentences from the viewpoint of clausal embeddding complexity. This material discloses no instances of I^3.

We now examine the possibility of double initial embedding, I^2. The BNC turned out some fifty instances of I^2s starting with *if what,* and some five each of *if* + *after* ~ *as* ~ *because* ~ *whatever* ~ *when* ~ *while*, almost all of them from written registers. Brown had two instances of *if as* and one of *when what*, LOB one of *whereas what*. The Internet provides a few examples of *although when*.

(6) a. [$_{\text{M}}$ [$_{\text{I-1}}$ If [$_{\text{I-2}}$ what he saw through security] did not impress him] Tammuz …] (BNC)

b. [$_{\text{M}}$ [$_{\text{I-1}}$ If [$_{\text{I-2}}$ what is tantamount to dictatorship …] continues in a union] it can …] (LOB)

c. [$_{\text{M}}$ [$_{\text{I-1}}$ If [$_{\text{I-2}}$ when I'm 38] Metallica ends] I don't think …] (BNC)

d. [$_{\text{M}}$ [$_{\text{I-1}}$ If [$_{\text{I-2}}$ after the investment is made] the value … rises] intervention …] (BNC)

e. [$_{\text{M}}$ [$_{\text{I-1}}$ If [$_{\text{I-2}}$ because risk had passed to him] the buyer bears loss] it …] (BNC)

f. [$_{\text{M}}$ [$_{\text{I-1}}$ If, [$_{\text{I-2}}$ as often happened,] he had to repeat [$_{\text{F-2}}$ because he had spoken too softly,]] he would …] (Brown)

g. [$_{\text{M}}$ [$_{\text{I-1}}$ But when [$_{\text{I-2}}$ what is new in a particular context] is also fairly obvious,] there is normally…] (Brown)

h. [$_{M}$ It will be appreciated [$_{F\text{-}1}$ that [$_{I\text{-}2}$ whereas [$_{I\text{-}3}$ what I am about [$_{f\text{-}4}$ to relate]] passed in a series of flashes] it seemed …]] (LOB)

i. [$_{M}$ [$_{I\text{-}1}$ Although [$_{I\text{-}2}$ when I print the sentences] they look fine,] I get …] (Internet)

The sentences (6a–i) are grammatical and acceptable, thereby showing that I^{2} (contrary to I^{3}) is a real structural option. Quirk et al. (1989:1039) overgeneralized when claiming that all I^{2}s would be marginal at best.[1] The homogeneous composition of (6a–i) indicates that additional constraints confine the domain of possible I^{2}s.

We first infer a clear-cut constraint on the degree of initial embedding which shall be called I^{2}max:

(7) I^{2}max: the maximal degree of initial embedding is two.

Being an empirical generalization over several SAE languages and large corpora, I^{2}max is more solidly founded than the conjectures I^{1}max, I$^{1\sim2}$max, I$^{3\sim4}$max and I$^{\infty}$ as a hypothesis of the 'real limit' of initial clausal embedding complexity. Further support for the reality of I^{2}max will be presented shortly in the form of a functional explanation of typical I^{2}s.

The occurring I^{2}s obey additional qualitative restrictions. I^{2} occurs mainly in written language. Of the one hundred or so instances retrieved, the vast majority contain an *if*-clause as I-high. Marginally, at least *when*, *whereas*, *although* also occur as subordinators in I-high (6g,h,i). Furthermore, there is a clear tendency for I-low to be the sentential subject of I-high. Thus, the prototypical instantiation of I^{2} is a clause complex starting with *if what* (6a,b). The propensity for I-low to be a sentential subject is further corroborated by (6g,h) with *when*-, *whereas*-clauses as I-high and sentential subjects as I-low. The two clauses entering I^{2} must not be of the same type (**if if*, **when when*), i.e. initial self-embedding is prohibited. Nor should both clauses be non-finite. English requires both to be finite but Swedish allows e-chain iI, i.e. an initial infinitive construction can contain a finite sub-clause immediately after the infinitival complementizer *att* 'to':

(8) [$_{M}$ [$_{i\text{-}1}$ Att [$_{I\text{-}2}$ som regeringen gör] förbise riskerna] är …] '(M) (i-1) To (I-2) as the government does (i-1) neglect the risks (M) is … ' (Internet)

I-low must always be finite in I^{2}s, in English I-high must also be finite. It is highly uncommon for I^{2} not to be rooted in the main clause but (6h) shows that I^{2} may be rooted in a final embedding, e-chain FII. All these qualitative I^{2}-constraints are summarized in (9).

(9) **Qualitative** I^2**-constraints:** I^2 strongly prefers (a) written language, (b) an *if*-clause as I-high, (c) a sentential subject, i.e. a *what*-clause as I-low; (d) finiteness (especially in I-low), and (e) rooting in the main clause.

The absence of multiple initial self-embedding is a joint effect of (9b,c) and therefore needs no separate statement.

The preferred absolute initial position of *if*-clauses has a pragmatic explanation. It is related to the natural flow of argumentation, formulated by Greenberg (1963:84) as Universal 14: the conditional clause precedes the conclusion as the normal order in all languages. The propensity of the prototypical I^2 to have an I-high with an initially embedded sentential subject as I-low is explicable in similar terms. Sentential subjects express a whole proposition and are therefore a complex construction, likely to occur in complex discourse such as conditional reasoning.

The prototypical instantiation of I^2 is *if what*, i.e. an *if*-clause with a sentential subject (6a,b). As *which* is equivalent to *what* in certain types of relative clauses, the combination *if which* is also predicted and it does indeed occur (10).

(10) [$_M$ [$_{I\text{-}1}$ If [$_{I\text{-}2}$ (which is in the present climate a realistic possibility)] a student seeks to allege a breach by the College of its obligations towards him/her,] the College will …] (Internet)

The net effect of the qualitative I^2-constraints is not quite as dramatic as the effect of I^2max which prohibits at least 125,000 theoretically possible I^3s. The theoretically possible I^2s in LOB amount to 50 * 50 = 2,500 if only subordinators are counted but the qualitative I^2-constraints reduce the occurring types to a handful.

Even if I^2 is a real structural option, its occurrences are rare indeed. The incidence in the BNC is one or two per one million words. Brown and LOB with a handful of I^2s indicate a similar overall frequency. As there are 54,544 sentences in the LOB corpus (Johansson & Hofland 1988), another rough estimate would be one or two I^2s per 50,000 sentences.

5. Discussion

I^2max and the qualitative I^2-constraints are empirically motivated, clear-cut, functionally motivated restrictions on the potential complexity of initial clausal embeddings in English. Many other SAE languages essentially seem to function like English in this regard. In Karlsson (to appear), I demonstrate that several quantitative and qualitative constraints also restrict center-embedding of clauses in SAE. For example, the maximal degree of clausal center-embedding is three in written language and two in spoken language. Furthermore, Karlsson (2002) showed that final embedding (right-branching) of clauses below degree four is extremely uncommon in SAE (also cf. Karlsson & Sinnemäki, to appear, for more details). In view of all this empirical evidence, the hypothesis of unbounded clausal embedding complexity must be considered falsified.

I^2max is an absolute limit on initial embedding complexity. Even I^2s are so rare that, for practical purposes and especially as concerns spoken language, it is a reasonable approximation to say that multiple initial embedding of clauses does not occur. The only clearly identifiable construction invoking I^2 is an initially embedded *if*-clause with an initial sentential subject for which we have provided a functional explanation drawn from known universal properties of conditional reasoning.

I^2max and the qualitative I^2-constraints rule out Yngve's and Pinker's made-up sentences (4, 5). In particular, Pinker's explicit claim of full grammaticality for the triple initial embeddings (5a,b) violates I^2max. Yngve's intuitions about the grammaticality of (4b,d) seem too subjective in view of the qualitative I^2-constraints. The total absence of I^3s in language use demonstrates the fallibility of linguists' grammatical intuitions when it comes to assessment of non-clear instances. It is not rigorous methodology to declare controversial fabricated structures fully grammatical if they are never used. As Wittgenstein puts it (1978 [1953]:19): when native intuition is confronted with weird made-up examples, "language goes on holiday", i.e. the grammaticality/acceptability status of such sentences is indeterminate as they lack backing in real usage.

Our constraints were initially defined as quantitative limits and combinatory restrictions, often in 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, (11) is a morphological rule of English:

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

(12) *Sue kissed he.

Sentence (12) violates rule (11) and is therefore blatantly ungrammatical, conflicting with the natural norm expressed by (11).

Fabricated sentences like (4, 5) do feel strange and this strangeness can be seen as an indication of norm breach, albeit of a somewhat weaker kind than the breach in (12). **If if if*, **if because what*, **that when as* etc. violate basic tendencies of information flow. The sentence beginning is for pointing out a topic or giving a reason for something that is to be said subsequently, not for amassing a jungle of conditions, reasons, or subsidiary statements.

Constraints like I^{2}max and the qualitative I^2-constraints are much like the **soft constraints** for Preferred Argument Structure discussed by Du Bois (2003a, b) which also express discourse-related quantitative tendencies, e.g. '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.

Our constraints are somewhat stronger and more normative than Du Bois' soft constraints. Sentences violating I^{2}max and the qualitative I^2-constraints do feel markedly strange and therefore normativeness, i.e. rule-likeness is invoked. These constraints are less arbitrary than typical basic-level morphological and syntactic rules but still have some normative force. They populate a cline between grammatical rules and behavioural language-related regularities.

Constraints of the type discussed here 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 the dependence of discourse management on short-term memory limitations (Lewis 1996; Gibson 1998).

Notes

* I am grateful for insightful comments and valuable help provided by John W. Du Bois, Risto Hiltunen, Arja Nurmi, and Geoffrey Sampson. Financial support was provided by the Academy of Finland under grant 201601.

[1] Quirk et al.'s example sentence sounds fabricated and unacceptable: [[That [if you could] you would help me] is of small comfort]. It does indeed violate two of our qualitative I^2-constraints (9b,c): *that*-clauses do not occur as I-high, nor do *if*-clauses as I-low.

References

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

Akmajian, A., Demers, R. A. & Harnish, R. M. (1985). *Linguistics: an Introduction to Language and Communication*. Second edition. Cambridge, MA: MIT Press.

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

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

Du Bois, J. W. (2003a). Argument structure: Grammar in use. In J. W. Du Bois, L. E. Kumpf & W. J. Ashby (Eds.), *Preferred Argument Structure: Grammar as architecture for function* (pp. 11–60). Amsterdam: John Benjamins.

Du Bois, J. W. (2003b). Discourse and grammar. In M. Tomasello (Ed.), *The New Psychology of Language* (pp. 43–87). Mahwah, NJ: Erlbaum.

Gibson, E. (1998). Linguistic complexity: locality of syntactic dependencies. *Cognition* 68, 1–76.

Greenberg, J. (1963). Some universals of grammar with particular reference to the order of meaningful elements. In J. Greenberg (Ed.), *Universals of Language* (pp. 73–113). Cambridge, MA: MIT Press.

Haiman, J. & Thompson, S. A. (Eds.) (1988). *Clause Combining in Grammar and Discourse*. Amsterdam: John Benjamins.

Johansson, S. & Hofland, K. (1988). *Frequency Analysis of English Vocabulary and Grammar*. Oxford: OUP.

Karlsson, F. (2002). Is there an upper limit to right-branching embedding of clauses? In R. Pajusalu & T. Hennoste (Eds.), *Tähendusepüüdja. Catcher of the Meaning. Festschrift for Professor Haldur Õim on the Occasion of His 60th Birthday* (pp. 196–199). Tartu: Publications of the Department of General Linguistics, University of Tartu, 3.

Karlsson, F. (to appear). Constraints on multiple center-embedding of clauses. *Journal of Linguistics.*

Karlsson, F. & Sinnemäki, K. (to appear). Constraints on multiple final embedding of clauses.

Langendoen, D. T. (1998). Linguistic theory. In W. Bechtel & G. Graham (Eds.), *A Companion to Cognitive Science* (pp. 235–244). Oxford: Blackwell.

Langendoen, D. T. (1975). Finite state parsing of phrase structure languages and the status of readjustment rules in grammar. *Linguistic Inquiry* 6, 533–554.

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.

Meillet, A. (1934). *Introduction a l'étude des langues indo-européennes*. Paris: Librairie Hachette.

Nägelsbach, K. F. (1963 [1846]). *Lateinische Stilistik*. Darmstadt: Wissenschaftliche Buchgesellschaft.

Pinker, S. (1994). *The Language Instinct*. Harmondsworth: Penguin Books.

Quirk, R., Greenbaum, S., Leech, G. & Svartvik, J. (1989). *A Comprehensive Grammar of the English Language*. London: Longman.

Quirk, R., Greenbaum, S., Leech, G. & Svartvik, J. (1972). *A Grammar of Contemporary English*. London: Longman.

Wittgenstein, L. (1978 [1953]). *Philosophical Investigations*. Oxford: Basil Blackwell.

Yngve, V. H. (1961). The depth hypothesis. In *Proceedings of Symposia in Applied Mathematics* 12 (pp. 130–138). Providence, RI: American Mathematical Society.

Yngve, V. H. (1960). A model and an hypothesis for language structure. *Proceedings of the American Philosophical Society* 104, 444–466.

Corpora

British National Corpus. Oxford University Computing Services. http://www.hcu.ox.ac.uk/BNC/

Brown Corpus. A Standard Corpus of Present-Day Edited American English, for use with Digital Computers. The Brown Corpus Manual by W. N. Francis and H. Kučera, 1971, revised and amplified 1979. http://www.hit.uib.no/icame/brown/bcm.html.

LOB Corpus. The Tagged LOB Corpus. User's Manual by Stig Johansson in collaboration with Eric Atwell, Roger Garside, Geoffrey Leech, 1986. http://www.hit.uib.no/icame/lobman/lob-cont.html.

Author's address

Fred Karlsson
Department of General Linguistics
P.O. Box 9, FI-00014 University of Helsinki, Finland

fgk@ling.helsinki.fi
www.ling.helsinki.fi/~fkarlsso