

Grammatical movement and gap-filling: Unproven cognitive processes

Gerard Kempen (gerard.kempen@mpi.nl)

Max Planck Institute for Psycholinguistics, Nijmegen
& Cognitive Psychology Unit, Leiden University

Abstract

In the past decade, more than twenty psycholinguistic studies have been published that claim to provide experimental evidence for the “psychological reality” of movement transformations and gap-filling during sentence production and comprehension. I argue that this interpretation is premature, and offer an alternative account of the data. It presupposes a non-transformational lexicalized grammar formalism with separate components for the hierarchical and the linear aspects of sentence structure. I sketch a grammatical processing architecture that incorporates such a grammar and whose workings are based on interactive activation and competitive optimization. In this account there is no need for movement transformations and related notions. The majority of the relevant experimental-psycholinguistic studies used measurements of cognitive processing load. These data can be accounted for in terms of the processing architecture sketched in the present paper. The remaining studies tested successfully two hypothesized effects of movement traces. However, more recent work with the experimental paradigms used in those studies suggest that accounts in terms of movement traces are very implausible. I conclude that nontransformational models of grammar and grammatical processing fit the currently available psycholinguistic data.

Keywords: generative grammar; movement transformation; filler-gap dependency; gap-filling; Performance Grammar; nontransformational grammar; competitive optimization.

Introduction

Since half a century, groups of experimental-psycholinguists have sought evidence for the psychological reality of movement transformations. The syntactic constructions they investigated belong to four types of movement distinguished in many modern grammar formalisms: (1) Argument movement (A-movement), (2) Nonargument movement (A-bar movement), (3) Head movement, and (4) Scrambling. Early studies, starting in the 1960s, aimed to prove the so-called Derivational Theory of Complexity (DTC), based on early versions of Generative Grammar. These attempts were unsuccessful (Fodor, Bever & Garrett 1974).

At the end of the 1970s, emphasis shifted to A-bar movement, in particular to Wh-fronting in relative and interrogative clauses. Mainstream Generative Grammar at the time assumed that relocated constituents leave “traces” at gaps created by movement operations. Traces are phonologically empty constituents indexing the moved constituents. One of the psycholinguistic hypotheses inspired by the assumption of movement traces holds that during sentence comprehension the parser attempts to associate traces/gaps with moved constituents—an operation similar to finding

the antecedents of pronouns. This aspect of parsing has become known as “gap-filling” (cf. Fodor, 1978).

The psycholinguistic search for evidence in support of the (neuro)cognitive reality of movement traces/gaps and gap-filling has met with considerable success and is continuing till date. Nevertheless, serious doubts remain. Alternative grammar formalisms have seen the light which explain the relevant linguistic facts satisfactorily without invoking movement transformations. This calls for a re-examination of the psycholinguistic processing facts and alternative explanations.

Given the limited space available, I sketch a nontransformational treatment of one “movement” phenomenon: Wh-fronting. To this purpose, I use the psycholinguistically motivated PERFORMANCE GRAMMAR (PG) formalism (Kempen & Harbusch, 2002, 2003; Harbusch & Kempen, 2002). Vosse & Kempen (2000, 2008/9) published two versions of a computational syntactic processing model called UNIFICATION SPACE (U-space for short). The U-space processor incorporates PG’s grammatical rules and constraints, and operates on the basis of interactive activation and competitive optimization. This enables deriving experimental predictions with respect to cognitive processing load incurred by sentences with and without grammatical movement.

I begin with elementary descriptions of PG and U-space. Then, I summarize the experimental-psycholinguistic evidence on grammatical movement and gap-filling as (neuro-)cognitively realistic processes. In doing this, I concentrate on the results obtained in the past decade (and a half). For earlier work I refer to critical reviews by Pickering & Barry (1991) and Sag & Fodor (1995). Finally, I briefly examine the evidence from the point of view of the movement-free PG/U-space architecture (see Supplemental materials) and conclude that the experimental evidence for grammatical movement and related notions is inconclusive.

Nontransformational Linguistic Treatments of Movement Phenomena

In nontransformational treatments of grammatical movement phenomena, “moved” constituents are base-generated at their surface (“noncanonical”) position. The resulting trees do not include branches terminating in empty nodes (“gaps”). Instead, the grammar rules contain provisions such that if a constituent is generated in a noncanonical position, it is annotated with features coding for its functional-grammatical role(s), in particular for the governor (subcategorizer) it is dependent on. The grammar also specifies the special conditions under which constituents end up in non-canonical surface positions (e.g. “Verb-Second” in German

and Dutch). Representatives of nontransformational grammatical frameworks are Generalized Phrase Structure Grammar (GPSG; Gazdar, Klein, Pullum & Sag (1985), Head-driven Phrase Structure Grammar (HPSG; Sag, Wasow & Bender, 2003), Lexical Functional Grammar (LFG; Bresnan, 2001; Kaplan & Zaenen, 1989), and Performance Grammar (PG; Kempen & Harbusch, 2002, 2003; Harbusch & Kempen, 2002).

Despite their widely differing rule systems, transformational and nontransformational grammar formalisms agree with each other on an important issue: Their treatments of noncanonical linear orders tends to be more complex than those of canonical ones: Noncanonical positions are licensed by a larger set of conditions—are governed by more special constraints—than canonical ones, which all need checks.

In the next section, I sketch a PG-based nontransformational account of a class of movement phenomena that has been targeted in several behavioral and neuroimaging experiments: fronting of Wh-constituents in simple clauses.

Performance Grammar: Structures

A lexicalized grammar assumes that the information needed to build grammatically correct sentences is associated with the individual lexical items making up the sentences. Retrieving a lexical item from the mental lexicon entails retrieving a LEXICAL FRAME, that is, a specification of the syntactic environment the item needs to live in. Every lexical item is head of (i.e., it “heads”) a phrase (for instance, nouns and relative pronouns are head of an NP, articles and possessive pronouns head a Determiner Phrase, verbs head a clause), and that these phrases/clauses specify grammatical functions to be fulfilled by other constituents (i.e. other phrases/clauses) in the sentence. Given the linguistic structures tested in the psycholinguistic experiments, I restrict myself to clauses and their major constituents: mainly subjects (Subj) and direct objects (DObj). Main or auxiliary verbs play the role of clausal heads (HD), and NPs function as subject or object; the NPs to be discussed are headed by nouns or relative pronouns.

For example, when parsing sentence (1), the hierarchical component of the grammar retrieves from the Mental Lexicon the lexical frames in Figure 1, and links them together by “binding” the root node of one lexical frame to a foot node of another one. The two nodes being bound should carry the same phrase label. The binding operation underlying sentence (1) yields the tree in Figure 2.

(1) Money counts

A root node can bind to a foot node only if all of their morphosyntactic features (number, person, case, etc.) *unify*—informally: are compatible. In Figure 2, successful unification is indicated by the “unification node” (U-node): the black dot in the thick line connecting the unified partners. Note that the unified partners do not merge/fuse; they only take the same feature matrix. (If unification fails, the feature matrices remain unchanged.) Comparison of Figures 1 and 2 shows that the unification yields a single value option for

the CASE attribute of the root and foot nodes of *money*, and a single option for the NUMBER, PERSON and STATUS attributes of *counts*. This illustrates how unification determines Subject-Verb agreement.

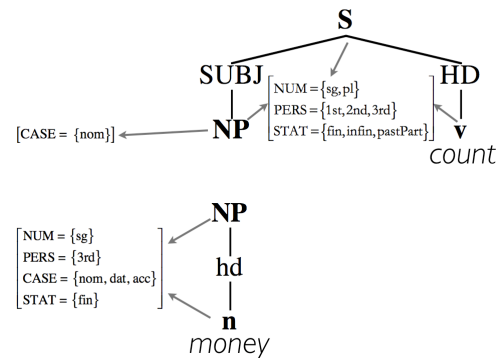


Fig. 1. Simplified lexical frames for the words of sentence (1). Only branches whose terminal leaves are involved in a binding operation or carry a lexical label, are shown. HD=head; STAT=status of a clause/verb: finite, infinitival or participial (present or past). The TENSE feature of the verb is left out.

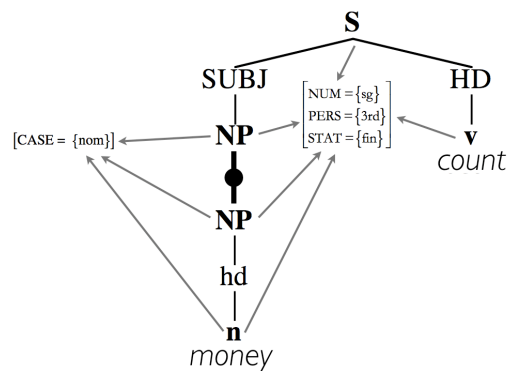


Figure 2: Syntactic tree for sentence (1) after the Subject footnote of *count* has unified with the root node of *money*.

The linear order of constituents within a clause is determined by a mechanism that works with a data structure called *topology*. Associated with every clause is a one-dimensional array of slots that serve as placeholders for constituents belonging to that clause (and sometimes for constituents lower in the hierarchical structure of the sentence). Clauses in English, Dutch and German are associated with “clausal topologies” consisting of nine slots, divided into three groups called Forefield, Midfield, and Endfield (Figure 3). Placement rules such as those in Table 1 assign individual constituents a position in one of the slots. Notice that the foremost slot F1 is exclusively accessible to constituents with special properties: those being topical, focused, or Wh-constituent. Endowed with such properties, they do not end up at their normal/default/canonical position but are “fronted.” Crucially, this placement proceeds in one step only. There is no earlier step during which the constituent occupies its canonical position.

English									
Forefield			Midfield				Endfield		
F1	F2	F3	M1	M2	M3	M4	E1	E2	

Dutch and German									
Forefield		Midfield				Endfield			
F1		M1	M2	M3	M4	M5	M6	E1	E2

Figure 3: Clausal topologies in English, Dutch and German.

Table 1. Examples of topology slot fillers for English clauses. Precedence between constituents landing in the same slot is marked by "<".

Slot	Filler
F1	<i>Declarative main clause:</i> Focus, Topic <i>Interrogative main clause:</i> Wh-constituent <i>Complement/relative clause:</i> Wh-constituent
F2	<i>Complement clause:</i> Complementizer <i>that</i>
F3	Subject (iff non-Wh-constituent)
M1	Head verb
M2	<i>Interrogative main clause:</i> Subject (iff non-Wh-constituent)
M3	Indirect Object < Direct Object (both non-Wh-constituent)
M4	Verb particle
E1	Non-finite Complement clause of Auxiliary verb
E2	Finite Complement clause, Adverbial clause

Processing Assumptions: Interactive Activation and Competitive Optimization

The U-space syntactic processor includes a network of nodes that code for grammatical notions and options. They have an activation level varying between zero and unity. Consider the diagrams in Figure 4, which illustrates the flow of activation and inhibition in a small part of the network. Excitatory connections (single or double arrows) transmit activation; inhibitory connections (lines with black dots at one or both ends) transmit inhibition (i.e., activation with a negative sign). The amount of activation (or inhibition) transmitted by a node to other nodes it is linked to, depends on its current activation level. For present purposes, I simply assume that all nodes have the same low initial (resting) level of activation, except for a few nodes representing “default” (unmarked, standard, canonical) options. They are initialized at an above-standard level of activation: PREACTIVATION. Incoming activation (or inhibition), modulated by connection weights, is added to (or subtracted from) the node’s current activation level.

Nodes have a lower and an upper activation threshold called “excitation threshold” and “selection threshold”, respectively. As long as the net activation level is below the excitation threshold, the node is “dormant” and does not transmit any activation or inhibition to other nodes. As soon as the activation level reaches the selection threshold, the grammatical property the node codes for, is “selected”, i.e. is supposed to be present in the currently processed struc-

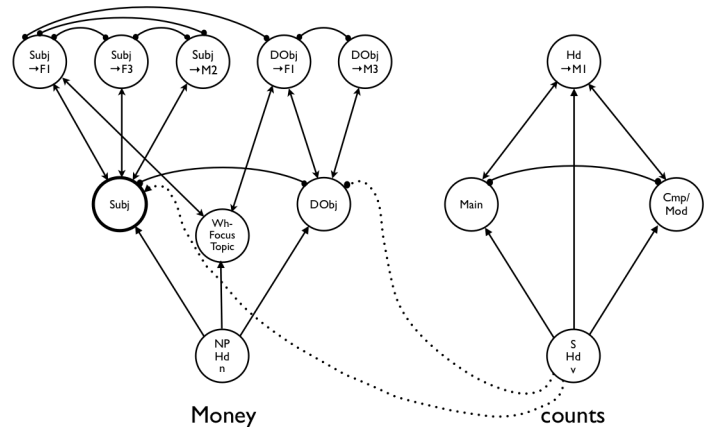


Figure 4: Nodes involved in processing example (1). The current activation level of a node is graphically rendered by the thickness of the circle line.

ture. Given adequate settings for connection weights and selection thresholds, a network can represent conjunctive constraints. For instance, U-nodes (in Figure 4: the nodes labeled Subj and DObj) have a relatively high selection threshold that only can be surpassed if they receive activation from both a foot node and a root node (in case of sentence (1): from the root node of NP *money* and the from the Subject foot node of *count*; Figure 2). Another conjunctive constraint illustrated in Figure 4 concerns the selection of the F1 slot. The nodes labeled “Subj→F1” and “DObj→F1” can be activated above the selection threshold only by receiving sufficient activation from both a U-node and the node labeled “Wh- Topic Focus”. The network thus licenses Fronting of subjects and direct objects if and only if they are a topical, focused, or Wh-constituent.

Continuous lines code for activation or inhibition spreading WITHIN the subnetwork recruited by a lexical item. For instance, the two arrows ascending from the NP node *money* in Figure 4 code for the fact that this NP can fulfill two grammatical functions in a clause—subject or direct object. The inhibition links between the two functional options ensure that only one of them is selected at any particular time. Dotted lines stand for activation or inhibition flowing BETWEEN lexical items. The two dotted lines represent the subcategorization properties of the intransitive verb *count*: It needs a subject (i.e., sends activation to nodes coding for Subject function of other currently active phrases) but refuses a direct object (i.e., sends inhibition to nodes representing direct object roles played by other phrases).

An additional assumption concerns DECAY of activation over time. With respect to nodes coding for unification between NPs and verbs, this implies that with larger linear distance between the NP (dependent) and verb (governor) the node needs more time to reach the selection threshold: delayed unification.

Each constituent originally activates one subnetwork, but the activation soon spreads to other subnetworks. Due to inhibitory links between nodes that code for incompatible structural options, competitions are launched, in particular

competitions for grammatical functions and linear positions. The processing load incurred by a sentence at a particular point in time is a function of (among other things) the AMOUNT OF UNRESOLVED COMPETITION raging at that time. With respect to the unification process within clauses, this amount correlates positively with the number of U-nodes that are currently active (above the excitation threshold) but remain in limbo, being unable as yet to reach an amount of activation above the selection threshold. This number includes all active U-nodes that have not yet found a unification partner.

When a U-node receives sufficient activation from a foot node and a root node from two different subnetworks (and is not experiencing competition), its activation can quickly rise above the selection threshold (short “rise time”). Every successful unification entails a reduction of the processing load because it removes one U-node from the set of “in limbo” nodes (active but unselected), and possibly more (if the selected node succeeds in bringing its competitor(s) into the dormant state). However, if the unification process is delayed (longer rise time), the reduction of processing load materializes at a lower speed; hence, the average load per time unit is higher.

An important reason for a unification attempt to incur a delay is inhibition from relatively powerful competitors. This may even result in reversals of the “balance of power” between alternative unification options. For instance, consider example (2b), from Chen, Gibson & Wolf (2005). The initially most promising (but unselected, hence provisional) function sought by the relative pronouns *who* is that of subject. However, the NPs immediately following the pronouns contest this provisional assignment by sending inhibition to the U-node coding for *who*-as-subject. The result (to be accounted for in a much more detailed version of the model than described here) is that the pronouns are pushed out of their provisional subject role. This reversal is one of the factors contributing to the high processing capacity demanded by center-embedded clauses like in (2b) relative to right-branching counterparts such as (2a). Another factor is the large amount of unresolved competition due to a sequence of five NPs that all cause subthreshold activation levels in competing U-nodes, followed by three verbs all competing for the same dependents. In (2a), every NP can select its definitive grammatical function without experiencing fierce competition.

- (2) a. Mary met the senator who attacked the reporter who ignored the president
 b. The reporter who the senator who Mary met attacked ignored the president

The PG/U-space syntactic processor resembles Gibson’s (1998, 2000) DLT model of sentence comprehension with respect to the predicted cognitive processing capacity for input sentences with vs. without movement. DLT distinguishes two factors contributing to processing difficulty: the cost of STORING grammatical predictions (e.g., a nominative input NP raises the expectation that a finite verb will follow), and the cost of INTEGRATING the syntactic and concep-

tual properties of new input into the structure built thus far (e.g. verifying the agreement properties of a verb against those of a nominative NP processed earlier). These cost factors roughly correspond to the abovementioned aspects of unresolved competition: the number of active but unselected U-nodes, and the rise time of activation levels at unification. The DLT model differs from the U-space model in that it works with a transformational grammar. For instance, each token of *who* in example (2) is supposed to elicit the prediction/expectation of a gap at the canonical postverbal position of the direct object.

Reinterpreting the Experimental Evidence

In the psycholinguistic literature of the past 10-15 years, I identified more than twenty publications with experimental evidence explicitly interpreted by the authors as evidence for the psychological—or (neuro)cognitive—reality of grammatical movement. I categorized them as follows:

- A. Comprehension tasks with measurements of the cognitive processing load imposed by clauses with vs. without filler-gap dependencies (through self-paced reading or neuroimaging techniques; 9 studies)
- B. Comprehension tasks with measurements of reactivation of the meaning of the antecedent of movement traces (through cross-modal lexical priming; 3 studies)
- C. Production tasks assessing the difficulty of producing clauses with canonical vs. noncanonical word orders (all types of grammatical movement, mostly in Dutch, some studies with agrammatic speakers; 8 studies)
- D. Production tasks assessing the difficulty of producing correct subject-verb agreement in structures with displaced constituents (attraction errors with nonintervening attractors) (2 studies).

The evidence in categories A and C is readily interpretable nontransformationally in terms of the PG/U-space model. In the comprehension studies, the crucial factor is the larger amount of competition spawned by clauses with, than by clauses without, displaced constituents. In the production studies, the smaller proportion of errors in producing canonical than noncanonical word orders can be attributed to preactivation of the network nodes coding for canonical orders, causing a bias in favor of the latter. (In addition, producing noncanonical orders requires checking a larger number and/or more specialized conditions than producing canonical orders.) The studies under B and D tested two hypothesized effects of movement TRACES. However, the results of more recent work with the experimental paradigms used in those studies cast serious doubts on the plausibility of accounts in terms of movement traces. I conclude that non-transformational models of grammar and grammatical processing fit the currently available psycholinguistic data.

Supplemental Materials

In this “long abstract” there is no space for details concerning the relevant experimental papers and my reinterpretation of the results. This information is available on my website: www.gerardkempen.nl/Downloadables.

References

- Bresnan, J. (2001). *Lexical-Functional Syntax*. Oxford, UK: Blackwell.
- Chen, E., Gibson, E., & Wolf, F. (2005). Online syntactic storage costs in sentence comprehension. *Journal of Memory and Language*, 52, 144–169.
- Fodor, J.A., Bever, T.G., & Garrett, M.F. (1974). *The psychology of language*. New York: McGraw-Hill.
- Fodor, J.D. (1978). Parsing strategies and constraints on transformations. *Linguistic Inquiry*, 9, 427-473.
- Fodor, J.D. (1989). Empty categories in sentence processing. *Language and Cognitive Processes*, 4, 155-209.
- Gazdar, G., Klein, E., Pullum, G.K., & Sag, I.A. (1985). *Generalized phrase structure grammar*. Oxford UK: Blackwell.
- Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68, 1–76.
- Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. In Miyashita, Y., Marantz, A., & O'Neil, W. (Eds.), *Image, language, brain*. Cambridge MA: MIT Press.
- Harbusch, K., & Kempen, G. (2002). A quantitative model of word order and movement in English, Dutch and German complement constructions. In: *Proceedings of the 19th International Conference on Computational Linguistics (COLING-2002)*, Taipei (Taiwan). San Francisco: Morgan Kaufmann.
- Kaplan, R.M., & Zaenen, A. (1989). Long-distance dependencies, constituent structure, and functional uncertainty. In Baltin, M.R., & Kroch, A.S. (Eds.), *Alternative conceptions of phrase structure*. Chicago: Chicago University Press.
- Kempen, G., & Harbusch, K. (2002). Performance Grammar: A declarative definition. In: Nijholt, A., Theune, M., & Hondorp, H. (Eds.). *Computational Linguistics in the Netherlands 2001*. Amsterdam: Rodopi.
- Kempen, G., & Harbusch, K. (2003). Dutch and German verb constructions in Performance Grammar. In: Seuren, P.A.M., & Kempen, G. (Eds.), *Verb constructions in German and Dutch*. Amsterdam: Benjamins.
- Pickering, M., & Barry, G. (1991). Sentence processing without empty categories. *Language and Cognitive Processes*, 6, 229-259.
- Sag, I.A., & Fodor, J.D. (1995). Extraction without traces. In Aranovich, R., Byrne, W., Preuss, S., & Senturia, M. (Eds.), *Proceedings of the Thirteenth West Coast Conference on Formal Linguistics, Vol. 13*. Stanford CA: CSLI Publications.
- Sag, I.A., Wasow, T., & Bender, E.M. (2003). *Syntactic theory: A formal introduction*. Stanford CA: CSLI Publications.
- Vosse, T., & Kempen, G. (2000). Syntactic structure assembly in human parsing: A computational model based on competitive inhibition and lexicalist grammar. *Cognition*, 75, 105-143.
- Vosse, T., & Kempen, G. (2008). Parsing verb-final clauses in German: Garden-path and ERP effects modeled by a parallel dynamic parser. In: Love, B.C., McRae, K., & Sloutsky, V.M. (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (Washington DC, July 2008). Austin, TX: Cognitive Science Society.
- Vosse, T., & Kempen, G. (2009). The Unification Space implemented as a localist neural net: Predictions and error-tolerance in a constraint-based parser. *Cognitive Neurodynamics*, 3, 331-346.