Cognitive Science 34 (2010) 286–310

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/j.1551-6709.2009.01073.x

Do Grammars Minimize Dependency Length?

Daniel Gildea,[a] David Temperley[b]

[a]*Computer Science Department, University of Rochester*
[b]*Eastman School of Music, University of Rochester*

Received 4 March 2008; received in revised form 22 July 2009; accepted 24 July 2009

Abstract

A well-established principle of language is that there is a preference for closely related words to be close together in the sentence. This can be expressed as a preference for dependency length minimization (DLM). In this study, we explore quantitatively the degree to which natural languages reflect DLM. We extract the dependencies from natural language text and reorder the words in such a way as to minimize dependency length. Comparing the original text with these optimal linearizations (and also with random linearizations) reveals the degree to which natural language minimizes dependency length. Tests on English data show that English shows a strong effect of DLM, with dependency length much closer to optimal than to random; the optimal English grammar also has many specific features in common with English. In German, too, dependency length is significantly less than random, but the effect is much weaker than in English. We conclude by speculating about some possible reasons for this difference between English and German.

Keywords: Syntax; Natural language processing; Word order

1. Introduction

Much recent language research has relied heavily on the concept of *dependencies*. A dependency is an asymmetrical syntactic relation between two words, the head and the dependent. The head of each dependency is then the dependent of another word (unless it is the head of the sentence), forming a recursive structure which connects all the words of the sentence. Fig. 1 shows the dependency structure for an English sentence, taken from a corpus which we describe further below. (In this and all other dependency diagrams, arrows point from the head to the dependent.) There is general, although not complete, agreement on the nature of dependency structures in English and other languages. Generally, each

Correspondence should be sent to Daniel Gildea, Computer Science Department, University of Rochester, Rochester, NY 14627. E-mail: gildea@cs.rochester.edu

```
(S (NP-SBJ (NP (NNP Pierre) (NNP Vinken)) (, ,) (ADJP (NP (CD 61) (NNS
years)) (JJ old)) (, ,)) (VP (MD will) (VP (VB join) (NP (DT the) (NN
board)) (PP-CLR (IN as) (NP (DT a) (JJ nonexecutive) (NN director)))
(NP-TMP (NNP Nov.) (CD 29)))) (. .))
```

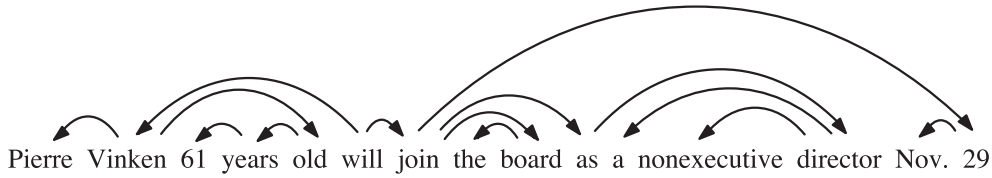


Fig. 1. The first sentence of the Wall Street Journal corpus in its original Penn Treebank notation (above) and the dependency tree produced by Collins's head-finding algorithm (below).

constituent type is headed by a word of the corresponding type—VPs by verbs, NPs by nouns, and PPs by prepositions; the head of a clause is its finite verb, and the head of a sentence is the head of its main clause.[1]

A well-established principle of language is that there is a preference for closely related words to be close together in the sentence. In dependency terms, this can be expressed as a preference for dependency length minimization (DLM). In this study, we examine the explanatory power of DLM with regard to natural languages. Our approach is to take naturally occurring language data, extract the dependencies, and then use different algorithms for reordering the words. We first consider the question, what is the optimal ordering algorithm in terms of dependency length? After answering this question, we then ask, how similar is the output of this optimal algorithm to natural language—both in terms of its dependency length, and in terms of its specific characteristics? Our reasoning is that, if word order in natural languages turns out to be similar to that produced by the optimal algorithm, then it is plausible to posit DLM as a factor in the evolution of natural languages. Note that our hypothesis is not that DLM has been the *sole* factor in the evolution of language. Clearly, grammars have been shaped by a multitude of forces, some of which we discuss below. The question is whether DLM has played a significant role in this process, alongside other factors.

We begin by reviewing some of the evidence for DLM in language. We then present tests using data from two languages, English and German. Results from English show a strong effect of DLM; results from German reflect a much weaker effect. We conclude by comparing the results from the two languages and considering how they might be explained.

2. Evidence for dependency length minimization

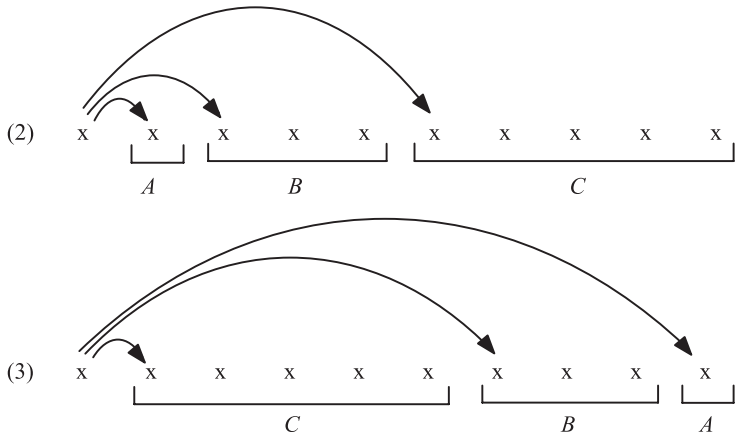
Experimental and theoretical language research has yielded a large and diverse body of evidence for DLM. Gibson (1998, 2000) argues that structures with longer dependencies are more difficult to process, and shows that this principle predicts a number of phenomena in comprehension. One example is the finding that subject-extracted relative clauses, such as

(1a), are easier to process than object-extracted relative clauses such as (1b) (King & Just, 1991). In both subject and object relatives, the verb of the relative clause (attacked) is dependent on the preceding relative pronoun (who). In subject relatives, these two words are normally adjacent, while in object relatives they are separated by the relative clause subject (the senator); thus, object relatives yield longer dependencies.

1a. The reporter who attacked the senator admitted the error.
1b. The reporter who the senator attacked admitted the error.

Phenomena of ambiguity resolution are also explained by Gibson's theory—for example, prepositional-phrase attachment decisions (Gibson & Pearlmutter, 1994; Thornton, MacDonald, & Arnold, 2000) and main-verb/reduced-relative ambiguities (Gibson, 2000). In such cases, the preferred interpretation tends to be the one with shorter dependencies.

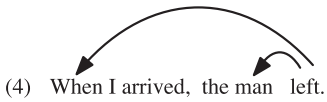
Dependency length minimization has also been put forth as an explanation for phenomena of language production. Here, we must distinguish between phenomena of grammar and phenomena of syntactic choice. Grammar refers to hard-and-fast constraints on the permissible sentences in a language; syntactic choice refers to situations where there is more than one possible way of saying the same thing. Much of the attention on DLM to date has concerned phenomena of syntactic choice. Hawkins (1994, 2004) observes that dependency length is minimized when the shorter of two dependent phrases is placed closer to the head; Temperley (2008) calls this principle ''ordered nesting.'' The two sentences below illustrate this point.



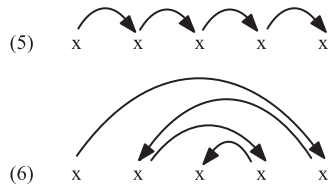
The length of a dependency is simply the number of words it spans; a dependency between adjacent words has a length of 1. The total dependency length of a sentence sums the lengths of all of its dependencies. In the case of (2) and (3), we assume that the two sentences are the same in terms of the internal structure of *A*, *B*, and *C*; thus, we need only consider only the dependencies that differ between the two sentences. The dependency length of (2) is therefore $1 + 2 + 5 = 8$, while the dependency length of (3) is $1 + 6 + 9 = 16$. Thus (2), which places the shorter dependent phrases closer to the head, has shorter dependency length.

Hawkins (1994, 2004) shows, through a series of corpus analyses, that syntactic choices generally respect the preference for ordered nesting. For example, in cases where a verb has two prepositional-phrase dependents, the shorter one tends to be placed closer to the verb. This preference is found both in head-first languages such as English, where PPs follow verbs and the shorter of two PPs tends to be placed first, and in head-last languages such as Japanese, where PPs precede verbs and the shorter one tends to be placed second (see also Yamashita & Chang, 2001). (Hawkins does not explain these patterns in terms of dependency length per se, but rather in terms of his ''Early Immediate Constituent'' [EIC] theory, which relates processing difficulty to the span of words within which the dependents of a head can be identified. The predictions of the EIC theory and the DLM theory are similar, although not identical; see Temperley, 2007, for discussion.)

Temperley (2007) finds evidence for DLM in a variety of syntactic choice phenomena in written English. For example, subject NPs tend to be shorter than object NPs; as the head of an NP tends to be near its left end, a long subject NP creates a long dependency between the head of the NP and the verb, while a long object NP generally does not. The length difference between subject and object NPs cannot be attributed solely to differences in discourse function; while subjects are more often ''given'' rather than ''new'' elements in the discourse, object NPs are longer than subject NPs even when the comparison is confined to specific indefinite NPs, which are presumably new elements in all cases. Subject NPs also tend to be significantly shorter when the clause begins with an ''opener'' phrase—an adverbial phrase preceding the clause. (In [4], for example, ''When I arrived'' is an opener phrase.) By contrast, the presence of an opener phrase has no effect on the length of object NPs. Again, this can be explained by DLM. If we assume (as is customary) that an adverbial phrase forms a dependency with the verb that crosses the subject NP, as in (4), a long subject NP will lengthen this dependency; thus, there is extra pressure for the subject NP to be short. For a direct object NP, which follows the verb, no such pressure is present.

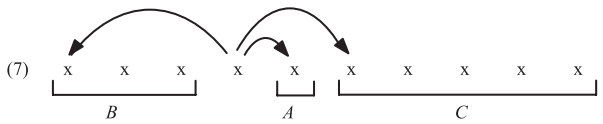


In other cases, DLM has been put forth as an explanation for actual grammatical rules. One case in point is the well-known tendency for languages to be either consistently ''head-first'' or consistently ''head-last'' (Hawkins, 1994; Radford, 1997; Vennemann, 1974). For example, languages in which the object follows the verb are predominantly prepositional—that is, just as verbs precede their objects, adpositions precede their objects. (''Adposition'' is the general name for prepositions, which precede their objects, and postpositions, which follow them.) Languages in which the object precedes the verb tend to be postpositional. Several authors have observed that a consistently head-first or head-last grammar might serve to minimize the distances between heads and dependents (Frazier, 1985; Hawkins, 1994; Rijkhoff, 1994). If each word in a sentence has exactly one dependency, it can be seen that a consistently ''same-branching'' (head-first or head-last) structure such as (5) yields shorter dependencies than one with ''mixed branching,'' such as (6).



Hawkins (1994, 2004) also shows that many cross-linguistic grammatical patterns reflect a preference for placing shorter dependents closer to the head (thus minimizing dependency length, as discussed earlier). For example, in languages in which adjectives and relative clauses are on the same side of the head noun, the adjective, which is presumably generally shorter than the relative clause, is usually required to be closer to the noun.

Temperley (2008) notes that, while a consistently head-first or head-last grammar is often desirable from the point of view of DLM, it is in fact not optimal. If a head has several dependents, placing them all on the same side of the head creates a kind of ''crowding'' effect; dependency length can be reduced if they are balanced on either side of the head. For example, whereas a same-branching construction such as (2) has a total dependency length of 8, a more balanced construction such as (7) has a dependency length of only 6.



Temperley suggests that one way to achieve such a balance is to stipulate a prevailing branching direction (e.g., right-branching) for a language but to allow some short (e.g., one-word) dependent phrases to branch in the opposite direction. Exactly this pattern has been observed empirically by Dryer (1992). In a study of 625 languages, Dryer observes that the conventional ''same-branching'' principle does not characterize actual grammars well. A more accurate generalization is that multiword phrases tend to branch consistently in a language, whereas one-word phrases are generally not consistent, sometimes branching in the prevailing direction of the language and sometimes not. Thus, inconsistent branching of one-word phrases, which appears to reduce dependency length, has also been empirically observed as a common cross-linguistic tendency.

Altogether, there is a suggestive body of evidence that DLM has influenced the evolution of natural language grammars. DLM is facilitated by same-branching constructions; the prevalence of such constructions has been widely observed. DLM is facilitated by the placing of shorter dependents closer to the head; this preference is reflected in many grammatical rules. The inconsistent branching of one-word dependent phrases, suggested by Temperley as a way of reducing dependency length, has also been empirically observed. All of these findings seem to point to DLM as a factor in the evolution of grammars. However, these findings have mostly been rather informal and qualitative in nature. In this study, we explore a quantitative way of examining the role of DLM in natural languages.

3. Dependency length minimization in English

A dependency tree can be regarded as a structure of dependencies connecting heads and dependents, such that every dependent has exactly one head, except for the ''root'' word, which is the head of the entire sentence. As such, a dependency tree corresponds to the mathematical concept of an acyclic graph.[2] Construed in this way, a dependency tree is inherently unordered, in that the words are not assigned any linear ordering; we will call such a structure an unordered dependency graph (UDG). An actual sentence entails a particular linearization of a UDG. In what follows, we extract the UDGs from natural English text, and then linearize them in the way that minimizes dependency length; we also linearize them in a completely random fashion. Comparing the dependency length of the original English to that of the optimal ordering and the random ordering gives a measure of how close English is to being optimal with regard to DLM.

The English text used in this study is from the *Wall Street Journal* portion of the Penn Treebank (hereafter the WSJ corpus), a large corpus of English text from the 1989 *Wall Street Journal* (Marcus et al., 1994). We use section 0 of the WSJ corpus, containing 1,921 sentences. The original data are annotated with conventional syntactic phrase markers: Constituents are bracketed and labeled as NP, VP, S, and the like, and terminal elements (words) are marked with part-of-speech tags (NN for singular noun, IN for preposition, etc.). Collins (1999) proposed an algorithm for recovering dependencies from Penn Treebank text, which has been widely used in computational linguistics (Charniak & Johnson, 2005; Eisner & Smith, 2005). The algorithm chooses a head from the children of each constituent (where children may be either constituents or terminal elements), using rules that list possible heads for each constituent type in order of preference; for example, the first choice for the head of a PP is a preposition (IN), the second choice is the infinitive marker ''to'' (TO), and so on. By recursively applying Collins's rules in a bottom-up fashion, a constituent representation can be converted into a dependency representation. Using Collins's algorithm, we extracted dependency trees from the WSJ corpus text; these were then treated as unordered graphs, to be linearized in different ways. We did not include nonlexical punctuation symbols such as commas and periods in the dependency trees because their dependency status (if any) is unclear. The average length of sentences in the test set was 21.2 words. Fig. 1 shows the first sentence from the test set in Penn Treebank notation and the dependency tree produced by Collins's algorithm.

An algorithm for ordering the words of a UDG will be called a dependency linearization algorithm (DLA). (See the Appendix for definitions of the terminology introduced in our study.) We explore two different kinds of DLAs. An unlabeled DLA pays no attention to syntactic categories (constituent types or parts of speech), and simply orders words based on the graph structure of the dependency tree. A labeled DLA requires consistent ordering of words in a given syntactic relation (e.g., preposition–object). Labeled DLAs may be regarded as more linguistically plausible—more like actual grammars—in that, in most languages, the ordering of words is constrained by syntactic rules (e.g., in English, prepositions must precede their objects). However, an unlabeled DLA, being unconstrained by syntactic consistency, is capable of achieving the absolute minimum dependency length, and it is interesting to see

how close natural language comes to this absolute minimum. While our main focus will be on labeled DLAs, we begin by briefly considering the case of unlabeled DLAs.

Following the usual assumptions of dependency research, we assume that dependencies may not cross and that no dependency may cross over the root word of the sentence.[3] Given these assumptions, a DLA must essentially answer two questions. One is the question of branching: on which side of the head—left or right—will a dependent be placed? The other is the question of nesting: Given that two or more dependents of a word are on the same side, how will they be ordered in relation to the head? The resolution of these two questions for each head completely determines the linear order of the words.

We begin by considering the optimal unlabeled DLA, that is, the one yielding the minimum possible dependency length for a given set of UDGs. Gildea and Temperley (2007) propose an unlabeled DLA which they prove to be optimal for DLM. The procedure is as follows. For each dependent of a word, a decision must be made as to its placement relative to the head. These decisions are made in a temporal order based on the length of the dependents' phrases (the phrase of a dependent word *w* includes *w* and all the descendants of *w*). If two phrases are of equal length, the temporal order of their placement is irrelevant. The first dependent to be placed is the one with the longest phrase; this dependent branches in the same direction as the head itself (which has already been determined, as explained below). Subsequent phrases are then placed in an alternating pattern: If the longest phrase branches to the left, the second longest branches to the right, the third longest to the left, and so on. If there is more than one dependent on the same side of the head, ordered nesting is applied, with shorter phrases placed closer to the head. Given this procedure, the words of an entire UDG can be linearized by starting with the root word, branching its longest dependent phrase in an arbitrary direction, and then branching successive dependents in an alternating fashion; the process can then be repeated recursively for the dependents of other words. Fig. 2 shows the resulting linearization for the first sentence of the WSJ corpus (shown in its original form in Fig. 1). The total dependency length for this linearization is 20—substantially lower than the original sentence, which had a dependency length of 32. Random linearizations of the corpus sentences were also produced, simply by choosing a random branching direction for each dependent of each head, and—in the case of multiple dependents on the same side—randomly ordering them in relation to the head.

The optimal and random algorithms were run on the WSJ corpus, and the average dependency length (ADL) was calculated in each case (by averaging the dependency lengths for each sentence); the ADL for the original text was also calculated. Table 1 shows the results. The optimal linearization (the third row of Table 1) of course reflects the lowest ADL, 33.5; the ADL for the original text is 47.5, while the ADL of the random linearization, 82.7, is much higher ($t[1,920] = 37.2$, $p < .0001$). The original ADL is significantly different from

Fig. 2. The first sentence of the Wall Street Journal corpus (see Fig. 1) as linearized by the optimal unlabeled DLA. The dependencies have been retained, but the words are reordered to minimize dependency length.

Table 1
Dependency length and percent correct for dependency linearization algorithms (English)

	Average Dependency Length (ADL)	Std. Dev. of Dep. Length	ADL as Percentage of Random	Percent Correct
Original text	47.5	52.1	57.4	—
Random DLA	82.7	78.8	100.0	40.5
Optimal unlabeled DLA	33.5	29.4	40.5	45.4
Extracted labeled DLA	51.4	65.0	62.2	80.1
Optimized labeled DLA	42.5	53.8	51.4	64.9

Note. Percent correct indicates the percentage of words (with one or more dependents) with the same ordering of dependents as in the original text.

DLA, dependency linearization algorithms.

both the random ADL ($t[1,920] = 37.4$, $p < .0001$) and the optimal ADL ($t[1,920] = 25.2$, $p < .0001$) but is much closer to the optimal ADL than to the random ADL. (The difference between the original and the random is much greater than the difference between the original and the optimal: $t[1,920] = 22.1$, $p < .0001$.) These results indicate a strong tendency for English to minimize dependency length. We also examined how similar the output of the optimal DLA was to English, by counting the percentage of words (with one or more dependents) whose dependents were ordered the same way as in the original text. The match of the optimal DLA to English (45.4%) was somewhat closer than that of a random algorithm (40.5%), although the difference was fairly small ($\chi^2[1] = 107.1$, $p < .0001$).

The optimal unlabeled DLA resembles natural languages in several important ways. In the first place, it reflects ordered nesting—the placement of shorter dependent phrases closer to the head—which has been observed as a consistent principle of natural grammars and syntactic choice. Like natural languages also, it reflects a tendency towards same-branching patterns, in that the longest dependent of each head must branch in the same direction as the head; if every word has exactly one dependent, the result will be a completely ''same-branching'' structure, such as (5). However, the grammar also reflects a certain amount of ''mixed'' branching, which Dryer (1992) argues is characteristic of natural languages as well. The grammar does not, however, reflect the distinction between one-word and multiword dependent phrases observed by Dryer and suggested by Temperley (2008) as a good strategy for DLM. A more fundamental problem with this algorithm as an approximation of natural languages is that it orders dependents with no regard for their syntactic type; for example, prepositional objects might be right-branching in one case and left-branching in another case (perhaps even within the same sentence). By contrast, word order in English—like many languages—is largely (though not entirely) governed by syntactic rules. Given this requirement of consistent syntactic ordering, we could hardly expect English to match the optimal dependency length, even if DLM were a powerful factor in its evolution. A more convincing test of the role of DLM in English would be to compare it with the minimal dependency length achievable by an algorithm that required consistent syntactic ordering; we call this a ''labeled'' DLA.

A labeled DLA—which could also be considered a kind of dependency grammar—is a DLA that requires each head-dependent set (a syntactic head type and a set of dependent

types) to be ordered in the same way on each occurrence. A labeled DLA can be defined in the following way. Each word in a dependency tree is given a label indicating the largest constituent that is headed by that word. The labeling is straightforward in many cases; for example, a preposition is normally labeled PP. A verb may be labeled VP, but if it is the finite verb of a clause it is the head of the entire clause and will be labeled S; a noun is usually labeled NP unless it is a noun modifier in a noun phrase, in which case it is labeled with its terminal label, that is, NN for singular noun. We used only basic constituent labels in the Penn Treebank, omitting subcategories; for example, PP-LOC and PP-TMP (representing locative and temporal PPs) are simply labeled as PP. The one exception is the SBJ (subject) tag on NPs, which we retained because we wanted to allow subjects to be distinguished from objects. For each word type, a real number line is created with the word itself at zero; each possible dependent word type is assigned a ''weight''—a position on the line, with negative indicating left-branching and positive indicating right-branching. The relative value of the weights for two dependency types determines their relative order, with weight closer to zero indicating which dependency type is placed closer to the head. Using this system, a simplified rule for ordering elements in a noun phrase in English might be stated as in Table 2 (these weights are for purposes of illustration only).

Table 2
Representative weights for children of a head noun

DT (determiner)	−1.0
JJ (adjective)	−0.7
NN (noun modifier)	−0.3
(the head noun)	0.0
PP (prepositional phrase)	0.3
SBAR (relative clause)	0.7

Such a rule can be used to linearize the dependents of a noun. The rule indicates that, for example, a determiner should be left-branching and further to the left than any adjectives or noun modifiers. Note that the exact numerical value of the weights is not significant, as only their relative ordering determines word order. If there is more than one dependent of a particular type, their ordering is chosen randomly.

Syntactic ordering in English is not completely governed by rules. For example, there is some flexibility in the ordering of the dependents of a verb; a prepositional phrase normally follows a direct object NP, but sometimes does not, especially if the NP is long—a phenomenon known as ''heavy-NP shift'' (Arnold, Wasow, Losongco, & Ginstrom, 2000; Hawkins, 1994). For this reason, it is of interest to see how well English word order can be approximated by a labeled DLA such as that proposed earlier. We did this by extracting a labeled DLA directly from the WSJ corpus, in the following way. For each dependency relation (an ordered pair of syntactic types), we assigned an integer indicating the dependent's position relative to the head (−1 for the first dependent to the left, −2 for the second, and so on); we then averaged these numbers across all occurrences of each dependency type. This ''extracted'' labeled DLA was then used to relinearize all the sentences, and the

output was compared with the original English. The results are shown in Table 1 (row 4); for 80.1% of all words (with at least one dependent), the ordering of dependents by the extracted DLA matched that of the original text. This indicates the extent to which each head-dependent set follows a single consistent ordering; thus, it provides a measure of the rigidity of English word order. The extracted DLA also provides an upper bound on how well a labeled DLA can be expected to match English word order, as it was extracted directly from the original text.

We now consider the labeled DLA that is optimal with regard to dependency length. This was found using a search procedure described in Gildea and Temperley (2007). The procedure begins with a random set of weights and adjusts each weight individually to minimize the dependency length of the corpus. When adjusting the weight for one dependency type, we take advantage of the fact that the total dependency length for the corpus changes only when the weight crosses the value of the weight of some other dependency type with which it co-occurs. We compute a set of significant intervals for the weight from these crossing points and set the weight to the midpoint of the interval that gives the smallest total dependency length for the corpus. We iterate over the weights until no further improvement is possible by adjusting any individual weight—usually three or four iterations. This optimization procedure was run on sections 2–21 of the WSJ corpus, initializing all weights to random numbers between 0 and 1. This biases the grammar towards right-branching configurations. Applying the optimized grammar to the dependency trees of section 0 of the WSJ corpus yields an ADL of 42.5 (see Table 1, row 5). Comparing the output with the original text, we find that 64.9% of words with at least one dependent have the same ordering of dependents. Thus, the optimized DLA matches English much better than a random DLA (64.9% vs. 40.5%, $\chi^2[1] = 2{,}288.8$, $p < .0001$) and achieves only a slightly lower dependency length than English (42.5 vs. 47.5, $t[1{,}920] = 37.9$, $p < .0001$).

The optimization process is not guaranteed to find the global minimum; for this reason we call it optimized, rather than optimal. In practice, however, we find that it is relatively insensitive to the initial values of the weights. The process was run 10 times; upon convergence, all runs were within 0.5 of one another in terms of ADL. This strongly suggests that the resulting grammar is close to optimal.

It is also of interest to compare the optimized labeled DLA to English in more detail. First, we examine the DLA's tendency to distinguish between one-word and multiword phrases. Recall that Dryer (1992) noted a tendency for many languages to reflect consistent ''same-branching'' of multiword phrases but inconsistent branching of one-word phrases; Temperley (2008) noted that consistent ''opposite-branching'' of one-word phrases is a good strategy for reducing dependency length. English shows a strong tendency for opposite-branching of one-word phrases. On the WSJ test set, 79.4% of left-branching phrases are one-word compared with only 19.4% of right-branching phrases ($\chi^2[1] = 13{,}971.3$, $p < .0001$). The optimized labeled DLA also reflects this pattern, although somewhat less strongly; 75.5% of left-branching phrases are oneword versus 36.7% of right-branching phrases ($\chi^2[1] = 5{,}007.4$, $p < .0001$).

We can also compare the optimized DLA with English with regard to specific rules. As explained earlier, the optimized DLA's rules are expressed in the form of weights assigned

to each relation, with positive weights indicating right-branching placement. Table 3 shows rules for some of the most frequent dependency types. The middle column shows the syntactic situation in which the relation normally occurs. We see, first of all, that object NPs are to the right of the verb and subject NPs are to the left, just as in English. PPs are also to the right of verbs; the fact that the weight is greater than for object NPs indicates that they are placed further to the right, as they normally are in English. Turning to the internal structure of noun phrases, we see that determiners are to the left of both object and subject nouns; PPs are to the right of both object and subject nouns. We also find some differences with English, however. Clause modifiers of nouns (these are mostly relative clauses) are to the right of object nouns, as in English, but to the left of subject nouns; adjectives are to the left of subject nouns, as in English, but to the right of object nouns. Of course, these differences partly arise from the fact that we treat NP and NP-SBJ as distinct, whereas English does not (with regard to their internal structure).

In some cases, as already noted, ordering choices in English are underdetermined by syntactic rules. For example, a manner adverb may be placed either before the verb or after (''He ran quickly/he quickly ran''). Here the optimized DLA requires a consistent ordering, while English does not. One might suppose that such syntactic choices in English are guided at least partly by DLM, and indeed there is evidence for this. In the case of heavy-NP shift, for example, placing the object NP after the PP when the NP is long keeps the shorter dependent phrase closer to the verb, thus reducing dependency length (Hawkins, 1994). In this connection, it is interesting to consider the DLA extracted directly from the corpus (Table 1 row 4). Recall that this is the labeled DLA that achieves the best possible match to English while requiring consistent ordering of each syntactic relation. This DLA yields a dependency length of 51.4, slightly higher than that of the original text, 47.5 ($t[1,920] = 9.4$, $p < .0001$). This suggests that DLM in English results not only from consistent word order patterns governed by syntactic rules, but is also affected, at least to a small degree, by patterns of syntactic choice.

Table 3
Sample weights from optimized dependency linearization algorithms

Label	Interpretation	Weight
S → NP	verb–object NP	0.037
S → NP-SBJ	verb–subject NP	−0.022
S → PP	verb–PP	0.193
NP → DT	object noun–determiner	−0.070
NP-SBJ → DT	subject noun–determiner	−0.052
NP → PP	obj noun–PP	0.625
NP-SBJ → PP	subj noun–PP	0.254
NP → SBAR	obj noun–rel. clause	0.858
NP-SBJ → SBAR	subject noun–rel. clause	−0.110
NP → JJ	obj noun–adjective	0.198
NP-SBJ → JJ	subj noun–adjective	−0.052

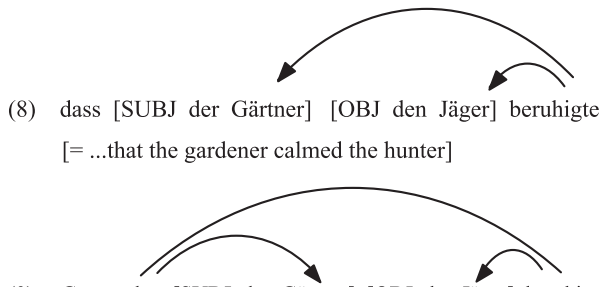
Note. Negatively weighted dependents appear to the left of their head.

The tests reported here give strong evidence for the role of DLM in English. The dependency length of English is much closer to that of an optimal linearization algorithm than it is to a random linearization. And the optimized dependency grammar for English contains many striking similarities to English grammar. Given the strong body of evidence for DLM in language generally, as described at the beginning of this article, it seems reasonable to suppose that the presence of DLM in English grammar is not mere coincidence but reflects pressures towards dependency minimization in the evolution of the language. To further explore this issue, it seemed prudent to examine DLM in other languages.

4. Dependency length minimization in German

We chose to investigate DLM in German. Two considerations motivated this choice. First, large syntactically annotated corpora are available in German, as well as rules for extracting dependencies from them. Second, German offers an interesting comparison with English as it has somewhat more freedom of word order; its word order is sometimes described as being ''semi-free'' or free within certain constraints (Kempen & Harbusch, 2008; Webelhuth, 1992). A well-known example of this freedom is the ordering of arguments of the verb (subject, object, and indirect object), which is highly variable in German (Heylen, 2005; Kempen & Harbusch, 2008). One might suppose that syntactic word order rules, while they may in themselves somewhat favor DLM, limit the extent to which dependency length can be minimized. In English, for example, the optimized labeled DLA has a somewhat higher dependency length than the optimal unlabeled DLA (see Table 1). Thus, in a language with free (or semi-free) word order, one might expect to see greater evidence of DLM.

While dependency length effects have not been studied in German to the same extent as in English, they have been examined in several studies. Bornkessel, Schlesewsky, and Friederici (2002) compare the ordering of subject and direct object in verb-final constructions such as (8) and constructions with the finite verb in ''second'' position such as (9).



In the first case, the finite verb (beruhigte) is clause-final, and both the subject and object are left-branching; DLM thus expresses no preference as to their ordering. In the second

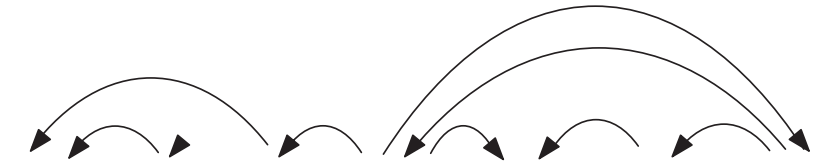
case, the subject connects to the finite verb (hat), while the object connects to the participle (beruhigt); thus, the subject-first ordering yields shorter dependencies than the object-first ordering (the object-first ordering would also involve crossing dependencies). Bornkessel et al. found that, while subject-first ordering is preferred in both cases, the preference is far stronger in the second case, just as DLM predicts. In another study by Konieczny (2000), subjects rated the complexity of sentences such as (10), in which the relative clause may be extraposed, as in (a), or not, as in (b) (the relative clause is in brackets):

10a. Er hat das Buch auf den Tisch gelegt [das Lisa gestern gekauft hatte].
10b. Er hat das Buch [das Lisa gestern gekauft hatte], auf den Tisch gelegt.
[= He lay the book that Lisa had bought yesterday on the table]

Konieczny found that the relative complexity of the extraposed version of the sentence (10a) was lower if the relative clause (das Lisa gestern gekauft hatte) was long, and if the main-clause verb phrase (auf den Tisch gelegt) was short. A corpus analysis reinforced this pattern, showing that structures judged to be more complex tend to be avoided. These findings are just as predicted by DLM. A long relative clause in a nonextraposed construction such as (10b) increases the distance from the auxiliary verb ''hat'' to the participle ''gelegt,'' whereas in an extraposed construction such as (10a) it does not have this effect; thus, a long relative clause favors extraposition. Similarly, a short verb phrase in an extraposed construction such as (10a) reduces the distance of the relative clause to its antecedent noun (Buch), but has no such effect in (10b), again favoring extraposition. Hawkins (1994) also notes that with some kinds of verb complements that are especially long and complex, such as full clauses with ''dass,'' extraposition is mandatory (''Er hat gesagt, dass…''; ''*Er hat dass…gesagt''). Again, this ordering is well predicted by DLM; a long complement will greatly lengthen the dependency from the clause-final verb to its preceding head.

Our German tests use the Tiger corpus, a corpus of about 50,000 sentences of text drawn from the German newspaper *Frankfurter Rundshau* (Brants, Dipper, Hansen, Lezius, & Smith, 2002).[4] We extracted every 10th sentence of the corpus as a test set (5,046 sentences) and used the rest for training. The average sentence length in the corpus is 15.6 words, somewhat lower than that for our English corpus (21.2 words). The sentences are marked with constituent structures; a set of head-finding rules proposed by Dubey (2004) was used to extract the dependencies. As with the English tests, each dependency was labeled with the constituent type of the head and that of the dependent. The constituent bracketing of the Tiger corpus involves somewhat different conventions from the Penn Treebank: noun phrases within prepositional phrases are not identified, and subordinate clauses are also coded differently. We recoded the sentences in the Tiger corpus and modified the head-finding rules to resolve these differences, so as to make the dependency structures as similar as possible to those of the Penn Treebank. There are also a number of differences in constituent names between the two corpora. For example, the Tiger corpus distinguishes between common noun phrases (NP) and proper ones (PN), while the Penn Treebank does not; the top-level constituent of each sentence in the Tiger corpus is labeled as ''VROOT,'' not ''S'' as in the Penn Treebank. We did not attempt to remove these differences. (The labels of

```
(VROOT (NP-SBJ (ART Die) (MTA (NE New) (ADJA Yorker)) (NN Aktienbörse))
(VAFIN hat) (VP (PP (APPRART am) (NN Freitag)) (NP (PPOSAT ihre) (NN
Rekordfahrt)) (VVPP fortgesetzt)))
```



Die New Yorker Aktienbörse hat am Freitag ihre Rekordfahrt fortgesetzt .

(= The New York Stock Exchange continued its record-setting journey on Friday)

Fig. 3. A sentence from the Tiger corpus, showing the constituent structure (above) and the dependency structure recovered by Dubey's head-finding algorithm (below).

constituents do not effect the dependency length and correctness results for the random DLA and the unlabeled DLA; they do affect these results for the labeled DLAs, however.) We did, however, introduce a distinction between subject and non-subject NPs, allowing the labeled DLA to choose different orderings for subject and object NPs. Fig. 3 shows a sentence from the Tiger corpus with the constituent representation (as modified by us) and the resulting dependency representation.

We ran the same tests on the German corpus that were run on the English corpus. We first identified the dependency length of the original German text. We then applied the algorithms described earlier to generate the optimal unlabeled ordering and optimized labeled ordering, and also generated a random ordering. For each ordering, we examined the ADL and also the match to the original text.

The results are shown in Table 4. Note first of all that the ADL of the original German text, 46.0, is significantly lower than that of a random ordering, 55.0 ($t[5{,}045] = 25.9, p < .0001$). The optimal unlabeled DLA yields an ADL of 24.4, significantly lower than that of the original text ($t[5{,}045] = 53.4, p < .0001$). The original ADL is somewhat closer to the random ordering than to the optimal unlabeled DLA; this result is different from English, where the original text was much closer to the optimal ordering (see Table 1). This suggests right away

Table 4
Dependency length and percent correct for dependency linearization algorithms (German)

	Average Dependency Length (ADL)	Std. Dev. of Dep. Length	ADL as Percentage of Random	Percent Correct
Original text	46.0	54.4	83.6	—
Random DLA	55.0	59.7	100.0	40.0
Optimal unlabeled DLA	24.4	30.4	44.4	39.6
Extracted labeled DLA	56.2	60.2	102.2	73.9
Optimized labeled DLA	32.1	36.5	58.4	34.9

Note. Percent correct indicates the percentage of words (with one or more dependents) with the same ordering of dependents as in the original text.

DLA, dependency linearization algorithms.

that the effect of DLM is not as strong in German as it is in English. Given this result, we would not expect an optimal DLA to have a match close to German with regard to the ordering of dependents, and indeed it does not. The optimal unlabeled DLA matches only 39.6% of the words in the original corpus, about the same as a random ordering, which matches 40.0% of words ($\chi^2[1] = 1.8$, $p = .18$); an optimized labeled DLA matches only 34.9% of words, significantly worse than the random DLA ($\chi^2[1] = 229.4$, $p < .0001$).

Inspection of the dependency trees in the German corpus suggests several factors that may contribute to the greater dependency length of German compared with English. Of particular importance is the fact that verbs in German clauses are often clause-final (all participles, as well as finite verbs in dependent clauses). A participle may be a great distance from the preceding auxiliary verb, with NPs, PPs, and other things in between; in English, by contrast, auxiliaries and participles are almost always adjacent. In a dependent clause, similarly, the clause-final head verb of the clause may be far from its head, which most often precedes the clause—for example, the relative pronoun of a relative clause. (Sentence [10] offers a case in point: The relative clause verb ''hatte'' is quite far from its head, the relative pronoun ''das.'') The placement of the verb at the end of the clause also means that a crowding effect may occur, with several dependent phrases all on the same side; in English, by contrast, one dependent—the subject NP—normally branches to the left, while others branch to the right, creating a more ''balanced'' configuration. These points are illustrated by the sentence in Fig. 3; the clause-final participle ''fortgesetzt'' is quite far from its head, the auxiliary ''hat;'' on the same side as its head are two dependent phrases, the prepositional phrase ''am Freitag'' and the object phrase ''ihre Rekordfahrt.'' Thus, both the parent head and the two dependents are on the same side of the verb, creating a situation of *maximal* dependency length. Further exacerbating this situation is the fact that, when one of the noun arguments is a pronoun, it is normally placed first (Kempen & Harbusch, 2008). In a verb-final construction, this means that the shortest dependent phrase (the pronoun) is likely to be furthest from the head; again, this ordering maximizes dependency length.

We quantitatively investigated our supposition that the greater dependency length of German was largely due to its verb-final structure. This is in fact a complex matter. By the logic presented in the previous paragraph, finite verbs will tend to have long dependencies to both their heads and dependents when they are clause-final; participles will also tend to have long dependencies to both heads and dependents—although the dependency from a participle to its head (the finite verb) will generally be long only if the finite verb is *not* clause-final (as in Fig. 3). To a first approximation, most of the long-dependency situations described here involve finite verbs, either as head or dependent. Thus, we simply examined the length of dependencies involving finite verbs, either as head or dependent, in both English and German. The data are shown in Table 5. It can be seen that, indeed, dependencies involving finite verbs (which account for roughly one-third of all dependencies in both English and German) are much longer in German than in English ($t[331{,}732] = 106.6$, $p < .0001$); for other dependencies, German reflects slightly greater length, but the difference is much smaller (although still significant: $t[761{,}563] = 87.2$, $p < .0001$). This supports our hypothesis that the greater dependency length of German is largely due to dependencies involving finite verbs. We also compared the length of dependencies from the

Table 5
Dependency length in English and German

	Count	Average Length
English		
Dependencies involving finite verbs	227,811	3.20
Other dependencies	565,283	1.94
All dependencies	793,094	2.30
German		
Dependencies involving finite verbs	187,323	4.83
Other dependencies	472,822	2.37
All dependencies	660,145	3.07

verb to its subject with the length of all other dependencies, in both English and German. In both languages, the length of subject–verb dependencies is almost the same as the length of other dependencies: in German, 3.06 (for subject–verb dependencies) versus 3.07 (for all others) (t[45,661] = 0.8, n.s.); in English, 2.28 (for subject–verb dependencies) versus 2.31 (for all others) (t[78,293] = 1.9, n.s.). Thus, while verb position seems to be a contributor to German's longer dependencies, it is not specifically the distance from subject to verb that results in this effect.

The poor match of our optimized labeled DLA to German deserves comment. It is, of course, partly due to the fact that in some respects German simply does not minimize dependency length. Another reason for the poor match is that the optimized labeled DLA is not fine-grained enough to capture certain syntactic rules: For example, the dependency type ''VROOT → NP-SBJ'' (a subject NP attaching to the root verb) may be either left-branching or right-branching, depending on whether there is an ''opener'' adjunct phrase, a condition that our grammar cannot capture. Yet another reason for the poor fit between the optimized DLA and German is the tendency of German towards somewhat free word order patterns than cannot be captured by any consistent rule; we return to this point below. It is surprising, however, that the match of the optimized labeled DLA to German is substantially worse than random because, overall, the actual dependency length of German is somewhat lower than random. It may be that there are certain common dependency types whose branching direction has little effect on overall dependency length, on which the labeled DLA happened to make the ''wrong'' choice (choosing left-branching while German has right-branching or vice versa); a random ordering would get such dependencies right by chance half the time, but the labeled DLA would get them wrong all the time.

The optimal unlabeled DLA also matches German relatively poorly in comparison with English. However, it is interesting to note that it matches German better than the labeled DLA (39.6% vs. 34.9%; $\chi^2[1] = 190.7$, $p < .0001$); this is in contrast to English, where the labeled DLA yields a much better match. This suggests that German involves some phenomena of DLM which are not captured by consistent grammatical rules—that is, phenomena of usage or syntactic choice. One possible example is the fact, discussed earlier, that relative clauses are more likely to be extraposed when they are long (Konieczny, 2000). Further evidence for the role of syntactic choice in DLM in German is the finding that the

DLA extracted directly from the corpus—which indicates the best match to German that can be achieved by any consistent grammar—yields a considerably higher dependency length than the original text (56.2 vs. 46.0; $t[5{,}045] = 40.5$, $p < .0001$). (This pattern is also seen in English, but the difference there between the extracted DLA and the original text is smaller.) However, the finding that the match between the unlabeled DLA and German is still quite poor (no better than random) suggests that DLM does not have a large effect on German word order choices. One might then ask what factors do determine word order choices in German. Various factors have been posited with regard to the ordering of verb arguments; these include animacy, discourse accessibility, and thematic role (Kempen & Harbusch, 2008).

In explaining our results on German, we have referred several times to the relative freedom of German word order. We thought it advisable to examine empirically the extent to which word order in our German corpus is truly free. One way to do this is by considering the labeled DLA extracted from the corpus. In the German case, this DLA matches the original text on 73.9% of words, whereas in English it achieved a somewhat higher match of 80.1% ($\chi^2[1] = 308.4$, $p < .0001$). This suggests that German does indeed reflect freer word order than English. As another test, we extracted from the corpus the most frequent ordering of each head-dependent set and examined the percentage of instances that matched that ordering. This yielded a value of 91.3% on English and 84.9% on German ($\chi^2[1] = 539.1$, $p < .0001$); this result, too, suggests that German reflects somewhat freer word order than English. In both of these tests, however, the difference between English and German is relatively small, suggesting that there is not a large difference in freedom of word order between the two languages.

5. Two other corpora

The aforementioned results seem to point to a striking difference between English and German: English appears to reflect a much higher degree of DLM than German. Before proceeding further, we must address an important question: Do our results truly reflect a difference between the two languages, or are they merely artifacts of the two corpora being studied? Different corpora within the same language may differ quite significantly with regard to syntactic patterns (Biber, 1993; Meyer, 2002), and it seems possible that they may vary widely in dependency length as well. To resolve this issue, we examined dependency length in two other corpora.

For English, we used the Brown corpus (Francis & Kučera, 1964). The Brown corpus is a varied corpus of written English, drawn from a wide variety of sources, including newspapers, magazines, scholarly writing, and fiction. The Penn Treebank includes a portion of the Brown corpus along with syntactic analyses, using the same constituent notation and conventions as the WSJ corpus. For German, we used the Tüba-D/Z corpus (Telljohann, Hinrichs, & Kübler, 2004). Like the Tiger corpus, the Tüba-D/Z corpus is drawn from a German newspaper, *Die Tageszeitung*. However, *Die Tageszeitung* and *Frankfurter Rundschau* (the source of the Tiger corpus) are quite different in style; a quantitative comparison of the two corpora (Rehbein & van Genabith, 2007) found the Tüba-D/Z corpus to contain

Table 6
Dependency length in four corpora

	Mean Sentence Length	Original ADL	Random ADL	Optimal Unlabeled ADL	Original as % of Random
WSJ corpus (English)	21.2	47.5	82.7	33.5	57.4
Brown corpus (English)	16.2	34.2	59.2	24.1	57.7
Tiger corpus (German)	15.6	46.0	55.0	24.4	83.6
Tüba corpus (German)	15.2	41.8	51.4	23.7	81.4

APL, average dependency length.

many more interjections, personal pronouns, and other kinds of elements indicative of an informal, personal style. The Tüba-D/Z data include syntactic analyses with dependencies explicitly marked, reflecting conventions very similar to those of the Tiger corpus; as with the Tiger corpus, we slightly modified the dependencies to bring them into accordance with the Penn Treebank conventions.

Our present aim is to determine whether the difference in DLM between English and German, reflected in the WSJ and Tiger corpora, is also reflected in the Brown and Tüba-D/Z corpora. To determine this, we need to consider three measures: the original ADL of the corpus, the optimal unlabeled ADL (the absolute minimum that could be achieved), and the random ADL (the dependency length we would expect to find if there were no tendency towards DLM at all). These figures are shown in Table 6 for the Brown and Tüba-D/Z corpora, with figures for the WSJ and Tiger corpora shown again for comparison.

Table 6 confirms the difference between English and German observed earlier. For the Brown corpus, as for the WSJ corpus, the original ADL is closer to optimal than to random; for the Tüba corpus, as for the Tiger corpus, the original ADL is closer to random than to optimal. Represented as a percentage of random (see the rightmost columns), the two German corpora reflect a markedly higher observed ADL than the two English corpora. These differences are difficult to examine statistically, due to uncontrolled differences between the corpora. What we wish to measure is the observed ADL of each corpus in relation to both the optimal and random ADL. To compare this precisely across corpora would require corpora that were matched, not only in sentence length, but in optimal and random ADL. To construct matched corpora of this kind would be an interesting exercise but was not attempted here. As an approximation, however, we may compare the Brown and Tiger corpora, which are very close in average sentence length (16.2 vs. 15.6) and also in optimal ADL (59.2 vs. 55.0) and random ADL (24.1 vs. 24.4). Comparing the two corpora, we find that the difference in observed ADL (34.1 for the Brown corpus and 46.0 for the Tiger corpus) is highly significant ($t[6{,}713] = 11.1$, $p < .0001$).

6. Discussion

In this paper, we have presented studies of DLM in both English and German. The results of our tests on English show a strong effect of DLM. The dependency length of English is

much closer to the optimal arrangement than it is to a random arrangement; and the optimized labeled DLA for English has many specific features in common with English. Of particular interest is the finding that the optimized DLA is not purely ''same-branching'' but favors more ''balanced'' configurations in ways that accord remarkably well with English: for example, the placing of subject and object on opposite sides of the verb. The match of the optimized DLA to English is far from perfect; thus, it is clear that English grammar has been shaped by other factors besides DLM. However, in light of the strong independent evidence for DLM in other aspects of language processing and usage, it seems likely that the strong effect of DLM observed in English is not merely due to chance but represents an important causal factor in the evolution of English grammar.

This line of reasoning suggests that we should expect to find similar effects of DLM in other languages. German seems to cast some doubt on this view, however. The dependency length of German is somewhat less than that of a random arrangement, but it is much closer to random than to the optimal arrangement; an optimized labeled DLA also bears little similarity to German. It appears, then, that DLM is reflected much less strongly in German than in English. We have cited several specific phenomena of German that may contribute to this, notably the fact that many verbs in German are clause-final, with their heads and all of their dependents to the left. However, there remains a deeper question: Given that—according to widespread agreement—short dependencies facilitate processing, and given the strong presence of DLM in English, why is it observed so much less strongly in German? In the following discussion we explore some possible answers to this question.

The approach of the current study may be described as ''functional,'' in that it attempts to explain the evolution of languages in terms of general principles of information processing and cognition. While much research on language change has avoided functional explanations, or even explicitly rejected this approach (e.g., Chomsky, 1975, 2005), in recent years functional explanations of language evolution have attracted increasing attention (Bybee, 2007; Christiansen & Chater, 2008; Kirby, 1999; Levinson, 2000). Essentially, the approach of this study has been to see how much of English and German grammar can be explained in terms of a single functional principle: DLM. Clearly, the results have been more successful in English than in German. Given the overall functional approach of this study, it is natural to ask whether the aspects of German that seem to conflict with DLM can also be explained from a functional perspective.

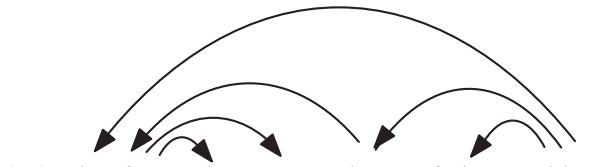
One general cross-linguistic tendency that has been cited is a preference for ''short-long'' ordering of constituents—sometimes known as ''end-weight'' (Wasow, 1997). While this is a complex phenomenon, it appears to be due at least partly to the preference for placing ''given'' discourse elements before ''new'' ones in the sentence—the assumption being that expressions referring to given elements will generally be shorter (Arnold et al., 2000; Gundel, 1988). It has also been observed cross-linguistically that subjects are more likely than objects to reflect ''given'' rather than ''new'' discourse elements (Givon, 1983). In combination with the preference for given–new ordering, we might then expect a cross-linguistic preference for orderings in which the object follows the subject, and this has in fact been observed: The three most common orderings of subject/verb/object—V–S–O, S–O–V, and S–V–O—all place the object after the subject (Dryer, 2005). As noted earlier, it has also

been observed that object phrases tend to be longer than subject phrases, at least in English (Temperley, 2007); no doubt this is partly due to the much stronger tendency of subject phrases to be given rather than new.

In verb-final languages, placement of object after subject combined with the greater length of objects creates a situation in which the longer of the two dependents is closer to the verb. This is exactly the situation in German, and this may be part of the explanation for the relatively high ADL of German. This suggests that verb-final ordering would be inherently disadvantageous, and it remains to be explained why it would ever be preferred; no functional explanation has been offered for this, to our knowledge. We should note, however, that finite verbs in German are only clause-final in subordinate clauses; this serves a clear informational function of distinguishing main clauses from subordinate clauses.

One specific aspect of German verb placement that often creates long dependencies is the separation of auxiliaries and participles in main clauses. This too is difficult to explain from a functional perspective. Here, however, we should note that this syntactic pattern is quite rare. In most other languages with separate auxiliaries and participles, the auxiliary and participle are adjacent, either after the subject (as in English) or at the end of the clause (as in Japanese).[5]

Another difference between English and German that may be relevant from a functional perspective is the much greater use of syntactic inflections in German, particularly case and gender marking. That languages with more case inflections (and other kinds) allow greater freedom of word order seems to be fairly well established (Keenan, 1978; Smith, 1996); perhaps such languages allow longer dependencies as well. For example, the fact that both articles and relative pronouns are gender specific in German means that they may be separated from their noun heads by a considerable distance with much less risk of ambiguity than would occur in English. Consider the noun phrase below (from the Tiger corpus):



(= The for-end-of-1992-planned free elections)

In this case the noun ‘‘Wahlen’’ is separated from its article ‘‘der’’ by a distance of six words, something that would rarely occur in English. Perhaps the gender specificity of nouns and articles facilitates the parsing of such long-distance dependencies. If so, it may be that DLM exerts less pressure in highly inflected languages such as German than it does in English.

One might wonder if the literature on historical language change could shed any light on the role of DLM in English and German. Let us consider just the aspect of word order that

has been most widely studied, namely the ordering of subject/verb/object. From the point of view of DLM, it would appear that the most favorable orderings are those with the verb in the middle—S–V–O or O–V–S—because that balances the two dependents on either side. (As described earlier, the preference for given-new ordering and the tendency for subjects to be given rather than new may account for the preference of S–V–O over O–V–S.) In the case of English, fixed S–V–O order seems to have arisen out of a situation of fairly free word order as a way of expressing case information given the decline of inflectional case markers (Pyles, 1971; Smith, 1996); DLM might well be posited as an explanation for why the specific ordering of S–V–O gained prevalence. In German, similarly, the evidence points to a gradual ''rigidification'' of word order, and to a shift from S–O–V to S–V–O ordering in main clauses (Ebert, 1978; Hopper, 1975). Thus, at least in this one respect, dependency length in both English and German may have tended to decrease over time. However, the mystery remains as to why S–O–V order in German evolved in the first place, and why it persists in subordinate clauses. Further examination of the historical evidence would certainly be of interest and might provide valuable insights into the role of DLM in language evolution.

This discussion has focused on issues of grammar, but we should remember that dependency length is also affected by phenomena of syntactic choice. Indeed, one possible explanation for the difference between German and English relates to the differing amount of syntactic choice—word order freedom—in the two languages. Grammatical rules may, to some extent, evolve to minimize dependency length; but when word order is relatively unconstrained by syntactic rules, as it is in German, it may be that considerations other than DLM tend to dominate word order choices. However, this argument does not seem wholly convincing. For one thing, the difference in word order freedom between English and German appears to be rather small, as shown earlier. Secondly, many studies have found evidence of DLM in patterns of syntactic choice, as discussed in Section 2. And our own tests support this view: In particular, the finding that the German text reflects substantially lower dependency length than the grammar extracted from the text suggests that there are indeed patterns of syntactic choice in German that contribute to DLM.

Overall, our study points to a positive but cautious verdict regarding the role of DLM in the shaping of grammars. While our results make a strong case for the role of DLM in English, the results from German are much less conclusive. Clearly, however, there are limits on what we can conclude from only two languages. To perform similar experiments on other languages would be of interest in a number of respects. For example, our reasoning suggests that S–O–V languages may inherently tend to have longer dependency length than S–V–O languages, due to the greater length of object NPs; it would be interesting to see whether this generalization holds true. In addition, the relatively unusual nature of German syntax (e.g., the separation of auxiliaries and participles) may make it hazardous to draw general conclusions from it. In future work, we intend to examine the role of DLM in other languages and to seek a general explanation for why DLM appears to operate in some languages much more than others.

Notes

1. One controversial issue concerns the heads of NPs; while most research in psycholinguistics and computational linguistics considers the main noun to be the head (Collins, 1999; Gibson, 1998), some theoretical linguists assume the determiner to be the head (Abney, 1987; Radford, 1997). We adopt the ''noun-headed'' view here.
2. A graph is a mathematical structure consisting of vertices (corresponding to words in dependency trees) connected by arcs (corresponding to dependencies). To our knowledge, dependency trees do not correspond exactly to any standard type of graph. Dependency graphs are similar to directed acyclic graphs: They are ''directed,'' in that each dependency points from the head to the dependent, and they are ''acylic,'' in that it is not possible to trace a directed path that begins and ends at the same word. However, a directed acyclic graph allows a dependent to have more than one head, whereas a dependency tree does not.
3. See Melcuk (1998); these assumptions are also implicit in the algorithm of Collins (1999). Crossing (nonprojective) dependencies are rare cross-linguistically (Steedman, 1985), although there are well-known examples in certain languages such as Dutch (Bresnan, Kaplan, Peters, & Zaenen, 1982); German also features some nonprojectivity (see note 4).
4. The corpus is available from the website http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERCorpus/. One complication with German is that it involves a significant degree of crossing dependencies or ''nonprojectivity.'' The Tiger corpus includes two sets of dependency annotations (as described below, we did not use these annotations but extracted the dependencies ourselves): One set allows nonprojectivity and the other does not. Comparing them, we found that about 2.3% of dependencies were nonprojective (that is, the projective and nonprojective annotations differed in about 2.3% of dependencies). By extracting dependencies from constituent structures, as we do here, we essentially guarantee that dependencies will never cross. However, it would also be interesting to examine DLM in cases where nonprojectivity is allowed; we intend to do this in future work.
5. We are indebted to Holger Diessel (personal communication) for this point. Quantitative typological data regarding the separation of auxiliaries and participles are apparently not available. However, discussions of auxiliary-participle separation in specific languages often imply that this is an unusual phenomenon (see Dryer, 2005; Gensler, 1994).

Acknowledgments

We are grateful to Florian Jaeger, Holger Diessel, and three anonymous reviewers for helpful comments on this work. The authors were supported by NSF grants IIS-0546554 and IIS-0325646.

References

Abney, S. (1987). *The noun phrase in its sentential aspect*. Unpublished doctoral dissertation. Cambridge, MA: Massachusetts Institute of Technology.

Arnold, J. E., Wasow, T., Losongco, T., & Ginstrom, R. (2000). Heaviness vs. newness: The effects of structural complexity and discourse status on constituent ordering. *Language*, *76*, 28–55.

Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing*, *8*, 243–257.

Bornkessel, L., Schlesewsky, M., & Friederici, A. D. (2002). Grammar overrides frequency: Evidence form the online processing of flexible word over. *Cognition*, *85*, B21–B30.

Brants, S., Dipper, S., Hansen, S., Lezius, W., & Smith, G. (2002). The TIGER Treebank. In E. Hinrichs & K. Simov (Eds.), *Proceedings of the workshop on treebanks and linguistic theories* (pp. 24–42). Sozopol, Bulgaria: Bulgaria Academy of Sciences.

Bresnan, J., Kaplan, R., Peters, S., & Zaenen, A. (1982). Cross-serial dependencies in Dutch. *Linguistic Inquiry*, *13*, 613–635.

Bybee, J. (2007). *Frequency of use and the organization of language*. Oxford, England: Oxford University Press.

Charniak, E., & Johnson, M. (2005). Coarse-to-fine *n*-best parsing and maxent discriminative reranking. In *Proceedings of the Association for Computational Linguistics* (pp. 173–180). Ann Arbor, MI: Association for Computation Linguistics.

Chomsky, N. (1975). *Reflections on language*. New York: Pantheon Books.

Chomsky, N. (2005). Three factors in language design. *Linguistic Inquiry*, *36*, 1–22.

Christiansen, M. H., & Chater, N. (2008). Language as shaped by the brain. *Behavioral and Brain Sciences*, *31*(5), 489–509.

Collins, M. J. (1999). *Head-driven statistical models for natural language parsing*. Unpublished doctoral dissertation. Philadelphia, PA: University of Pennsylvania.

Dryer, M. (1992). The Greenbergain word order correlations. *Language*, *68*, 81–138.

Dryer, M. (2005). Order of subject, object, and verb. In M. Haslpelmath, M. Dryer, D. Gil, & B. Comrie (Eds.), *World atlas of language structures* (pp. 330–331). Oxford, England: Oxford University Press.

Dubey, A. (2004). *Statistical parsing for German: Modeling syntactic properties and annotation differences*. Unpublished doctoral dissertation. Saarland University, Germany.

Ebert, R. P. (1978). *Historische syntax des deutschen*. Stuttgart, Germany: Sammlung Metzler.

Eisner, J., & Smith, N. A. (2005). Parsing with soft and hard constraints on dependency length. In *Proceedings of the international workshop on parsing technologies (IWPT)* (pp. 30–41). Vancouver, BC.: Association for Computational Linguistics.

Francis, W. N., & Kučera, H. (1964). *Manual of information to accompany a standard corpus of present-day edited American English, for use with digital computers*. Providence, RI: Department of Linguistics, Brown University.

Frazier, L. (1985). Syntactic complexity. In D. Dowty, L. Karttunen, & A. Zwicky (Eds.), *Natural language parsing: Psychological, computational, and theoretical perspectives* (pp. 129–189). Cambridge, England: Cambridge University Press.

Gensler, O. (1994). On reconstructing the syntagm S–AUX–O–V–OTHER to Proto–Niger–Congo. In Kevin Moore (Ed.), *Proceeding of the Berkeley Linguistics Society*, Vol. 20 (pp. 1–20). Berkeley, CA: Berkeley Linguistics Society.

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 W. O. A. Marantz & Y. Miyashita (Eds.), *Image, language, brain* (pp. 95–126). Cambridge, MA: MIT Press.

Gibson, E. & Pearlmutter, N. (1994). A corpus-based analysis of psycholinguistic constraints on PP attachment. In C. Clifton, L. Frazier, & K. Rayner (Eds.), *Perspectives on sentence processing* (pp. 181–198). Hillsdale, NJ: Erlbaum.

Gildea, D., & Temperley, D. (2007). Optimizing grammars for minimum dependency length. In *Proceedings of the Association for Computational Linguistics* (pp. 184–191). Prague: Association for Computational Linguistics.

Givon, T. (Ed.) (1983). *Topic continuity in discourse*. Amsterdam: Benjamins.

Gundel, J. (1988). Universals of topic-comment structure. In M. Hammond, E. Moravcsik, & J. Wirth (Eds.), *Studies in syntactic typology* (pp. 209–239). Amsterdam: Benjamins.

Hawkins, J. (1994). *A Performance theory of order and constituency*. Cambridge, England: Cambridge University Press.

Hawkins, J. (2004). *Efficiency and complexity in grammars*. Oxford, England: Oxford University Press.

Heylen, K. (2005). A quantitative corpus study of German word order variation. In S. Kepser & M. Reis (Eds.), *Linguistic evidence: Empirical, theoretical and computational perspectives* (pp. 241–264). Berlin: DeGruyter.

Hopper, P. J. (1975). *The syntax of the simple sentence in proto-germanic*. The Hague, The Netherlands: Mouton.

Keenan, E. K. (1978). Language variation and the logical structure of Universal Grammar. In H. Seiler (Ed.), *Language universals* (pp. 89–124). Tubingen, Germany: Gunter Narr Verlag.

Kempen, G., & Harbusch, K. (2008). Comparing linguistic judgments and corpus frequencies as windows on grammatical competence: A study of argument linearization in German clauses. In A. Steube (Ed.), *Sentence and context* (pp. 179–192). Berlin: DeGruyter.

King, J., & Just, M. A. (1991). Individual differences in syntactic processing: The role of working memory. *Journal of Memory and Language*, *30*, 580–602.

Kirby, S. (1999). *Function, selection, and innateness: The emergence of universals*. Oxford, England: Oxford University Press.

Konieczny, L. (2000). Locality and parsing complexity. *Journal of Psycholinguistic Research*, *29*, 627–645.

Levinson, S. (2000). *Presumptive meanings: The theory of generalized conversational implicature*. Cambridge, MA: MIT Press.

Marcus, M. P., Kim, G., Marcinkiewicz, M. A., MacIntyre, R., Bies, A., Ferguson, M., Katz, K. & Schasberger, B. (1994). The Penn Treebank: Annotating predicate argument structure. In *ARPA human language technology workshop* (pp. 114–119). Plainsboro, NJ: Morgan Kaufmann.

Melcuk, I. (1998). *Dependency syntax: Theory and practice*. Albany, NY: State University of New York Press.

Meyer, C. (2002). *English corpus linguistics: An introduction*. Cambridge, England: Cambridge University Press.

Pyles, T. (1971). *The origins and development of the English language*. New York: Harcourt Brace Jovanovich.

Radford, A. (1997). *Syntactic theory and the structure of English*. Cambridge, England: Cambridge University Press.

Rehbein, I., & van Genabith J. (2007). Why is it so difficult to compare treebanks? Tiger and tüba-d/z revisited. In K. de Smedt, J. Hajič, & S. Kübler (Eds.), *Proceedings of the sixth international workshop on treebanks and linguistic theories* (pp. 115–126). NEALT.

Rijkhoff, J. (1994). Explaining word order in the noun phrase. *Linguistics*, *28*, 5–42.

Smith, J. (1996). *An historical study of English: Function, form, and change*. London: Routledge.

Steedman, M. (1985). Dependency and coordination in Dutch and English. *Language*, *61*, 523–568.

Telljohann, H., Hinrichs, E., & Kübler, S. (2004). The Tüba-d/z treebank: Annotating German with a context-free backbone. In *Proceedings of the fourth international conference on language resources and evaluation* (pp. 2229–2235). Lisbon, Portugal: European Language Resources Association.

Temperley, D. (2007). Minimization of dependency length in written English. *Cognition*, *105*, 300–333.

Temperley, D. (2008). Dependency length minimization in natural and artificial grammars. *Journal of Quantitative Linguistics*, *15*, 256–282.

Thornton, R., MacDonald, M. C., & Arnold, J. E. (2000). The concomitant effects of phrase length and informational content in sentence comprehension. *Journal of Psycholinguistic Research*, *29*, 195–203.

Vennemann, T. J. (1974). Topics, subjects, and word order: From SXV to SVX via TVX. In J. M. Anderson & C. Jones (Eds.), *Historical linguistics I (proceedings of the first international conference on historical linguistics)* (pp. 339–376). Hillsdale, NJ: Erlbaum.

Wasow, T. (1997). Remarks on grammatical weight. *Language Variation and Change*, *9*, 81–105.

Webelhuth, G. (1992). *Principles and parameters of syntactic saturation*. Oxford, England: Oxford University Press.

Yamashita, H., & Chang, F. (2001). ''Long before short'' preference in the production of a head-final language. *Cognition*, *81*, B45–B55.

Appendix: Terminology

Term	Definition
UDG	Unordered dependency graph: A graph connecting the words of a sentence, without any linear ordering of the words
DLA	Dependency linearization algorithm: A procedure for arranging the words of a UDG in a linear order (without crossing dependencies or crossing over the root word)
Head-dependent set	A syntactic head type (e.g., ''VP'') and a set of dependent types (for example, ''NP, PP'')
Unlabeled DLA	A DLA that does not require syntactically consistent ordering (different instances of the same head-dependent set may be ordered differently)
Labeled DLA	A DLA that requires every instance of a head-dependent set to be ordered the same way
Random DLA	A DLA that orders the words of each UDG randomly
Extracted labeled DLA	A labeled DLA generated directly from a corpus: For each head-dependent set, the most common ordering in the corpus is applied to every instance of the set
Optimal unlabeled DLA	The unlabeled DLA that yields the absolute minimum dependency length for a UDG
Optimized labeled DLA	A DLA produced by an algorithm which searches for the labeled DLA that minimizes dependency length in the corpus (not guaranteed to find the absolute optimum)