Minimizing Syntactic Dependency Lengths: Typological/Cognitive Universal?

David Temperley and Daniel Gildea

Eastman School of Music and Computer Science Department, University of Rochester, Rochester, New York 14627

Annu. Rev. Linguist. 2018. 4:1–15

https://doi.org/10.1146/annurev-linguistics-011817-045617

Keywords

dependency grammar, language typology, human sentence processing

Abstract

Syntactic dependencies are head/modifier relations between words in a sentence that organize sentences into a syntactic tree structure. The general principle that languages have a preference to group syntactically related words close together can be made precise as a preference for shorter dependencies. We examine evidence for this principle in the development of languages' grammars as well as in the choices made by individual speakers where syntactic variation is licensed. We survey evidence from corpus studies, computational simulations, and experiments on comprehension; altogether, this evidence makes a compelling case for DLM as an important factor in language structure and cognition.

1. INTRODUCTION

It is a basic principle of human cognition that we tend to associate objects that are close together in space and time. This is expressed most generally in the Gestalt rule of proximity, which explains many phenomena of visual and auditory perception. It is natural to suppose that this principle would apply to language as well; at the lexical level, it suggests that words that are especially closely related should be close together in the sentence. There is indeed abundant evidence for this principle from a variety of areas of language research, including language typology, corpus research, computational linguistics, and psycholinguistics. Proposals regarding the proximity of related words are most often expressed in terms of dependencies. A dependency is an asymmetrical syntactic relation between a pair of words in a sentence, known as the head and the dependent. The head of each dependency is then the dependent of another word, unless it is the head of the entire sentence; this forms a recursive structure that connects the entire sentence. Dependencies play an important role in many linguistic theories (Bresnan 1982, Dik 1989, Hudson 1991, Mel'čuk 1988, Oehrle et al. 1988, Pollard & Sag 1987, Radford 1997). Positing a dependency between two words generally implies a particularly close syntactic and semantic relationship between them. Therefore, the idea that languages tend to place closely related words close together can be expressed as dependency length minimization (DLM).

There is general agreement as to the nature of dependency structures in language. In the case of PPs in English, for example, it is usually assumed that the preposition is the head of the phrase (with the prepositional object as its dependent) and is then a dependent of the word (generally a preceding verb or noun) that the phrase conventionally modifies. In general, the head of each major constituent type (NP, VP, AP, PP) is the word after which the phrase is named, and the head of a clause is its finite verb. One controversial case involves NPs; some theories assume that NPs are headed by their main nouns (Bresnan 1982, Gibson 1998, Mel'čuk 1988, Pollard & Sag 1987), while others assume the determiner as the head (Abney 1987, Radford 1997). Coordinate structures are also a problematic case. The head of a coordinate phrase is sometimes considered to be either the first conjunct (Mel'čuk 1988) or the conjunction (Munn 1993); other theories consider all conjuncts of the phrase to be heads (forming dependencies with external words) (Hudson 1991, Pickering & Barry 1993).

A common assumption of dependency research is that dependencies may not cross one another, nor may any dependency cross over the root word of the sentence; this is known as the assumption of projectivity. In fact, this assumption is not strictly true—many languages feature occasional violations of projectivity—but it appears to hold true the vast majority of the time in nearly all languages. We address this topic in section 4.

Evidence for DLM comes from a wide range of sources. At the broadest level, this evidence can be organized as pertaining to either grammar or usage. It appears that grammars of languages have evolved in ways that reduce or minimize dependency length. Some evidence for this assertion comes from common knowledge about languages; computational methods have also been brought to bear on this issue, often comparing dependency length in corpora to random or optimal baselines. With regard to usage, the issue is the role of DLM in explaining differences among grammatical sentences within a language. These differences relate to both production and comprehension. In production, the essential concept is syntactic choice: Languages typically offer many ways of expressing a thought, and writers and speakers seem to prefer those that involve shorter dependencies. Numerous corpus studies have been done in this area. In comprehension, studies have shown that sentences with shorter dependencies are easier to process; here, experimental research has been of primary importance.

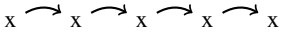
Although this three-category framework—grammar, production, and comprehension—is useful to bear in mind, in practice it is not so easy to sort DLM research into these three categories. In particular, much corpus research relates to both grammar and production (syntactic choice), not cleanly distinguishing between the two. We have found it useful to organize this review more along methodological

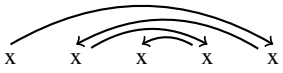
lines. We begin with a brief survey of some early (largely theoretical and typological) observations about DLM. We then turn to studies of syntactic choice, which rely mainly on corpus methods. We then examine computational studies of dependency length in natural language data. Finally, we survey experimental research on the role of DLM in comprehension.

2. TYPOLOGICAL OBSERVATIONS

The first observation of DLM is often attributed to Behaghel (1932, p. 4), who proposed four "laws" of sentence structure, the first of which is as follows: "das geistig eng Zusammengehörige [wird] auch eng zusammengestellt," or "what belongs together mentally is placed close together." In a similar vein, Givon (1991, p. 89) noted that "entities that are closer together functionally, conceptually, or cognitively will be placed closer together at the code level, i.e., temporally or spatially."

Interest in DLM increased in the late twentieth century among scholars concerned with explaining observed crosslinguistic universals or regularities. It has been noted that languages tend to be consistently head-first or head-last—in other words, they consistently place the heads either before their dependents or after them (Greenberg 1963, Vennemann 1974, Chomsky 1988) (though there are many exceptions to this, as discussed below). 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 their objects.) 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 1990). If each word in a sentence has exactly one dependency, then a consistently "same-branching" (head-first or head-last) structure (e.g., structure 1) yields shorter dependencies than one with "mixed branching" (e.g., structure 2). (In the following diagrams, arrows point from heads to dependents.)

(1) 

(2) 

Recent research suggests that the head-first/head-last principle does not actually characterize languages very well. In a study of 625 languages, Dryer (1992) found many exceptions to the principle. A more accurate generalization, he suggested, 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. Temperley (2008) noted that the pattern observed by Dryer may in fact be optimal for DLM. If a head has several dependents, placing them all on the same side of the head creates a kind of crowding effect: the closer dependents force the more distant ones to be further away from the head. Dependency length can be reduced if the dependents are balanced on either side of the head. 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. This is exactly the pattern noted by Dryer. Thus, inconsistent branching of one-word phrases, which has been empirically observed as a common crosslinguistic tendency, may serve to reduce dependency length.

Hawkins (1994, 2004) points to other crosslinguistic grammatical patterns that reflect a preference for shorter dependencies. In many cases, when two constituents are placed on the same side of the

head, grammatical rules require that the shorter one (or the one that is generally shorter) be placed closer to the head. 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.

3. DEPENDENCY LENGTH MINIMIZATION IN SYNTACTIC CHOICE

Evidence for dependency length minimization has also been found in language production. Often in language, there are multiple ways of expressing essentially the same thought; DLM has been found to be a major factor in these situations of syntactic choice. Hawkins (1994, 2004) has made important contributions to this area. Specifically, Hawkins's Early Immediate Constituent (EIC) theory states that language comprehension will be facilitated if, within each constituent, the heads of the children are all close together—within a short "window," known as the constituent recognition domain (CRD); this is advantageous because it provides the parser with "earlier and more rapid access" to the children of the larger constituent (Hawkins 1994, p. 66). In the case of a verb with two right-branching adjunct phrases (e.g., PPs), the shorter adjunct should be placed first, as in structure 3, rather than second, as in structure 4, as this placement minimizes the size of the CRD—the portion of the sentence containing the adjunct heads and the parent head:

(3) [V [P x] [P x x x x]]
| - CRD —— |

(4) [V [P x x x x] [P x]]
| ——— CRD ——— |

Examining corpora from a variety of languages, Hawkins finds evidence that, indeed, "short–long" ordering of adjuncts is generally preferred. A well-known case in point in English is "heavy-NP shift": the tendency for a direct-object NP to be placed after a PP when it is long, as in *He sold [for five dollars] [the diamond ring that his mother gave him]*. Another prediction of the EIC theory is that left-branching constituents in a predominantly right-branching language like English should tend to be short; a long left-branching constituent will tend to increase the CRD of the larger constituent. Diessel (2005) has invoked this prediction as an explanation for the greater length of final versus initial adverbial clauses. Although the EIC theory is not explicitly formulated in terms of dependencies, favoring the placement of a syntactic head and its children's heads within a short window is similar to favoring structures with short dependencies.

Temperley (2007) examines a number of syntactic choice situations in English; using corpus data from the *Wall Street Journal*, he finds a consistent preference for structures that minimize dependency length. For example, in quotation constructions like sentences 5 and 6, below, the outer subject–verb pair may be either inverted or not. There is a strong tendency to invert when the subject phrase is long, as in these two examples; sentence 5 feels much more natural than sentence 6.

(5) "I agree," [said [Jane Smith, president of Smith, Brown & Jones, a consulting firm]].
(6) "I agree," [[Jane Smith, president of Smith, Brown & Jones, a consulting firm,] said].

If the subject phrase is short (e.g., *Jane*), the preference for inversion is much weaker. This fact can be attributed to dependency length; if we assume that the quoted phrase is a dependent of the verb *said*, the uninverted construction with a long subject NP, as in sentence 6, creates a long dependency.

Temperley also proposes dependency length as a possible explanation for the greater length of object NPs versus subject NPs in *Wall Street Journal* sentences. Because (as he shows) the head of an NP tends to be near its beginning, a long subject phrase can create a long dependency to the verb. Although this is not strictly a choice between two syntactic structures equivalent in meaning, it is often possible to convey essentially the same proposition with a particular NP in either subject or object position—for example, by using a passive construction, or in other ways (*strict rules governed the meetings* = *the meetings followed strict rules*). Other explanations for the length difference between subjects and objects are also possible; in particular, objects are more likely to contain new rather than given discourse elements, and one might expect new discourse elements to be longer. Temperley shows, however, that objects tend to be longer even when the comparison is confined to indefinite NPs, which are almost always discourse-new. In addition, subject NPs are especially likely to be short when they are preceded by a premodifying adjunct phrase, whereas the presence of a premodifying adjunct has no effect on object length; this finding, too, can be attributed to DLM, because a short subject NP reduces the distance between the premodifying adjunct and the head verb of the sentence.

As noted above, Hawkins's EIC theory makes similar predictions to DLM, but is not expressed in terms of dependency length. One way to distinguish the two accounts concerns heads with three right-branching dependents. In this case, the logic of the EIC theory suggests that the ordering of the first two dependents should make no difference; what matters is the length of the CRD—the distance between the parent head and the head of the last dependent—and that will be the same under either ordering of the first two dependents. By contrast, DLM predicts that the first dependent will tend to be shorter than the second, as that will minimize dependency length. Temperley (2007) examines such constructions in *Wall Street Journal* text, and finds that the first adjunct in a three-adjunct construction tends to be significantly shorter than the second. In this case, then, DLM appears to fit the evidence better than the EIC theory.

Although the tendency toward short–long ordering in right-branching languages is clearly predicted by both DLM and the EIC theory, other factors may also play a role. Arnold et al. (2000, p. 32) suggest that the preference for short–long ordering—sometimes known as end-weight—is due to constraints on language production: "When formulation is difficult, choices in constituent ordering allow speakers to postpone the long, difficult constituent while they utter the shorter, easier one." In light of this proposal, it is of particular interest to examine left-branching languages. When a head has multiple dependents to its left, DLM predicts long–short ordering, whereas the end-weight principle still predicts short–long ordering. Hawkins (1994) cites corpus evidence from Japanese, a predominantly left-branching language, that long–short orderings are indeed preferred—for example, in cases with two NP dependents of a following verb. (Although the EIC theory in its basic form does not predict long–short ordering in left-branching languages, Hawkins proposes an extension of the theory that does.) Similarly, an experimental study by Yamashita & Chang (2001) find a preference for long–short ordering of Japanese subject and object NPs; Yamashita (2002) reports the same pattern in a corpus study of written Japanese. Such phenomena seem to favor DLM over an end-weight account. In other cases, the evidence is indecisive. In English, in cases where a clause has two premodifying adjuncts, DLM predicts long–short ordering, whereas end-weight predicts short–long: Corpus data show no significant difference in length between the first and second adjunct phrases (Temperley 2007).

4. DEPENDENCY LENGTH MINIMIZATION IN GRAMMAR: COMPUTATIONAL SIMULATIONS

In this section, we attempt to evaluate quantitatively the degree to which DLM affects language structure using computational simulations. The approach taken in these simulations is to measure the

minimal dependency length possible over all linearizations of an unordered dependency structure, the dependency length observed in real language, and the dependency length of randomly selected linearizations, and to compare the results to determine whether the observed dependency length is significantly lower than that expected by chance. The simulations discussed in this section are based on text corpora annotated with syntactic dependencies, and consider other possible orders of the words in the sentences, holding the dependency relations constant, and measuring the total dependency length in each ordering. Dependency length can be affected both by grammatical contraints and by syntactic choice; some of the studies we survey attempt to isolate the effects of these two factors.

We first discuss how to determine the minimal dependency length achievable for a given structure. The problem is simplified if we consider only projective dependency structures—those in which dependencies do not cross (although this does not always lead to the minimum dependency length, as discussed below). Under this assumption, Gildea & Temperley (2010) present a simple algorithm for DLM. They show that one can determine the optimal layout of each subtree independently. (A subtree is the set of direct and indirect dependents of a given word, usually corresponding to a syntactic constituent.) Furthermore, the dependents of each word should be ordered by sorting them according to the number of words in each dependent's constituent, then arranging them from the inside out, with the shortest constituents closest to the head, and alternately placing constituents in the order of length on either side of the head. For the purposes of generating random, projective linearizations, we can, by contrast, simply choose a random assignment of each dependent to either the left or right of its head, and a random ordering within each side. Using this method, Gildea & Temperley (2007, 2010) find that English has an average per-sentence dependency length of 47.5—significantly lower than expected at random (82.7), though still quite a bit higher than the absolute minimum possible (33.5). This finding provides evidence that dependency length may have influenced the historical development of grammar of human languages.

The results discussed above assume that each sentence is arranged independently of all other sentences in the corpus. Gildea & Temperley (2007, 2010) also consider DLM in the scenario in which word order is fixed by the grammar. Here, the dependency types are taken into account: each word is labeled with the type of the largest constituent of which it is the head, and the dependency is labeled by the types of its head and dependent words. (For example, a dependency between a subject and verb would typically be labeled S $\rightarrow$ NP.) The language's grammar is then modeled as a set of rules specifying the left-to-right order of dependents, for each possible set of dependents of a given head type. This model more accurately reflects the situation in fixed-word-order languages such as English; the optimal dependency length for such a grammar may offer a more realistic comparison to determine the extent to which real English optimizes dependency length. To find the fixed grammar with the minimum dependency length, Gildea & Temperley (2007, 2010) designed a numerical optimization procedure to search over possible grammars; this procedure is guaranteed to converge but, given the computational complexity of the search problem, is not guaranteed to find the global minimum over grammars. The optimized fixed-word-order grammar achieved an average dependency length per sentence of 42.5, fairly close to the 47.5 of the original text, and much lower than the 82.7 of random linearizations.

This approach to the study of DLM has also been applied to other languages. Liu (2008) examines 20 languages, and finds that the average dependency length per word is much lower than expected by chance in all cases, and generally in the range of two to three words, with Chinese having the highest average dependency length among the languages studied, at 3.66. Futrell et al. (2015) considered an even wider set of languages; they compared the observed dependency length with random linearizations using syntactically annotated corpora from 37 languages from a variety of language families around the world. The authors found very consistent evidence of DLM, with observed dependency

lengths significantly lower than expected by chance in all 37 languages.

Could DLM explain the rarity of nonprojective structures? While the experiments described above assume projectivity, projectivity itself is often correlated with short dependencies. This is because, when two dependencies cross, the total dependency length of the sentence can often be improved by switching the position of the two inner words of the four words involved, which will both remove the crossing and bring each of the inner words closer to the outer word to which it is linked. Nonetheless, it is interesting that for a given tree the optimal layout is not always projective (see sentence 13 in the next section). Still, if projective trees are generally lower in dependency length than nonprojective ones, one might suppose that the general rarity of nonprojective trees might itself be a result of DLM.

The effect of nonprojectivity on dependency length in natural languages was examined by Ferrer i Cancho (2004), who compared observed, random, and optimal layouts for nonprojective dependency trees in Romanian and Czech. He found that dependencies are much shorter than expected by chance, and not very far from optimal in this setting. Ferrer i Cancho (2006) also examined the number of crossing dependencies in random linearizations and linearizations chosen to minimize dependency length for both randomly generated tree structures and sentences from a Romanian treebank. Layouts with small dependency length have far fewer crossings. This finding aligns with the fact that crossing dependencies are relatively rare in natural language, and suggests that this state of affairs may result from a general preference toward shorter dependencies. In another study, Park & Levy (2009) generated structures that were not projective, but with the restriction that each subtree in the dependency structure consists of at most two continuous spans with a single gap between them. This relaxation of the projectivity constraint handles a very large number of the nonprojective trees observed in linguistically annotated treebanks, and is sufficient to capture structures produced by Tree Adjoining Grammar (Joshi 1985, Abeillé & Rambow 2001), a mildly context-sensitive formalism that has been posited to handle many of the non-context-free phenomena observed in human language. Park & Levy (2009) found that allowing such nonprojective structures yielded a relatively small decrease in optimal dependency length, for example, from 34.1 to 32.7 in English. This finding provides evidence that projectivity is consistent with DLM, and that a general principle of DLM could help explain that preponderance of projective structures in language.

We now consider the possibility that the DLM minimization found in these simulations might be caused by the confounding factor of lexical predictability. Lexical predictability, in particular as measured by an n-gram language model, which computes a probability of each word in the context of $n-1$ preceding words, is another factor that has been found to have a large impact on processing difficulty. Gildea & Jaeger (2015) investigate the interaction between dependency length and n-gram probability, using corpora from English, German, Czech, Chinese, and Arabic. They evaluate a large number of randomly generated, fixed-word-order grammars in terms of both the average per-sentence dependency length and the average log probability of the next word given the previous two words. They find some correlation between these factors, which is to be expected because, if a word's dependents are close by, they are more likely to be within the window considered by the n-gram model, and hence will be more accurately predicted by an n-gram model. Nevertheless, the correlation is only moderate, indicating that these two factors can be optimized independently of one another. The observed languages are consistently both lower than expected by chance among fixed-word-order grammars in terms of dependency length and higher than expected in terms of average log probability. Similar trends are observed for German and Czech, despite the fact that they are widely considered to be free-word-order languages. In fact, both languages have strongly dominant orderings for a given set of dependency types. Thus, even in the case of such languages, modeling grammars as having fixed word order is more realistic than the models discussed above that do not take dependency type into account and order each sentence independently.

Gildea & Jaeger (2015) also perform optimization experiments, searching over possible grammars optimized for dependency length, average log probability, or some weighted combination of the two. The authors find that optimizing one criterion does not generally lead to good performance on the other, but that the observed languages are generally closer to some optimal frontier that trades off dependency length against average log probability. This finding provides evidence that both of these measures of processing difficulty have independently contributed to the development of the languages studied.

The studies discussed in this section are based on text corpora, holding the dependency structures constant over different orderings. It is important to bear in mind that the observed data are the result of both the language's grammar and syntactic choices available to the speakers. Thus, it is possible that some of the observed tendency toward shorter dependencies is due to these choices at the time of the production of the sentences. Temperley & Gildea (2010) explore this possibility by creating a fixed-word-order grammar that matches the word order in the corpus as closely as possible. (For each combination of a head and a set of dependents, the most common ordering of that set is applied to all cases of the set.) This grammar has an average per-sentence dependency length of 51.4, compared with the actual dependency length of 47.5. This provides a rough estimate of the extent to which syntactic choice, rather than grammar, reduces dependency length.

5. DEPENDENCY LENGTH MINIMIZATION IN COMPREHENSION

DLM is also a factor in sentence comprehension. Particularly important in this regard has been research by Gibson (1998, 2000) showing that the dependency structure of a sentence plays a major role in its processing difficulty. According to Gibson's Dependency Locality Theory (DLT), the syntactic complexity of a sentence can be predicted by two factors: storage cost, the cost of maintaining in memory the syntactic predictions or requirements of previous words, and integration cost, the cost of syntactically connecting a word to previous words with which it has dependent relations. The integration cost for a word increases with the distance to the previous words with which it is connected, on the reasoning that the activation of words decays as they recede in time, making integration more difficult.

Gibson shows that the DLT predicts a number of phenomena in comprehension. One example is the greater complexity of object-extracted relative clauses, such as as sentence 7, versus subject-extracted relative clauses, such as sentence 8 (King & Just 1991).

(7) The reporter who the senator attacked admitted the error.
(8) The reporter who attacked the senator admitted the error.

In both object relatives (sentence 7) and subject relatives (sentence 8), the verb of the relative clause (*attacked*) is dependent on the preceding relative pronoun (*who*); in subject relatives, these two words are normally adjacent, but in object relatives they are separated by the relative clause subject (*the senator*), which yields a higher integration cost for object relatives. According to the DLT, the length of a dependency is not measured by the sheer number of words it spans, but rather by the number of discourse referents it crosses; these include nouns and tensed verbs, but not pronouns and other function words. For this reason, object relative clauses are easier to process if the relative clause subject is a pronoun, as in *The reporter who I attacked admitted the error*. In cases where one object relative clause is embedded inside another—so-called center-embedding constructions, like *The mouse the cat the dog bit chased ran*—the theory correctly predicts that processing difficulty will be especially acute.

The DLT also successfully predicts more complex distinctions in sentence processing, such as the greater difficulty of sentence 9 over sentence 10:

(9) The executive who the fact that the employee stole office supplies worried hired the manager.

(10) The fact that the employee who the manager hired stole office supplies worried the executive.

In Gibson's theory, the difficulty of sentence 9 results from the high integration cost at the word *worried*, due to the two long dependencies to previous words (the relative pronoun *who* and the relative clause subject *fact*). In sentence 10, no word has such a high integration cost. Gibson and colleagues have also applied this theory to phenomena in other languages, including relative clauses in Chinese (Hsiao & Gibson 2003) and Russian (Levy et al. 2013).

The DLT has stimulated a large amount of further research, testing and refining the theory and proposing alternatives. In one study, Demberg & Keller (2008) tested the theory's ability to predict data from the Dundee corpus, a collection of reading time measurements on newspaper text. The study found the DLT to be limited as a predictor of reading times in general because it generates predictions only for nouns and verbs. When nouns and verbs were analyzed separately, the DLT was a significant though fairly small predictor of reading times. The authors recommend some modifications of the theory to increase its predictive power. They suggest that the cost of integrating a preceding NP with a verb may depend in a gradient way on the "givenness" of the NP, so that, for example, indefinite NPs would have higher cost than definite NPs; this idea has also been explored by Warren & Gibson (2002).

A challenging phenomenon for the DLT consists of so-called antilocality effects. In some cases, the processing difficulty of a word (measured by reading time) seems to decrease, not increase, as it gets further from its dependents. In a study by Levy & Keller (2013), German subjects read sentences such as these:

(11) Hans hat den Fußball versteckt.
Hans has the football hidden.

(12) Hans hat dem Sohn den Fußball versteckt.
Hans has (from) the son the football hidden.

The difference between the sentences is that, in the second case, a dative argument *dem Sohn* is added. This additional argument inserts a discourse referent between the object *den Fußball* and the verb *versteckt*, which should increase the processing difficulty of the verb, according to the DLT. However, the verb was read more quickly when the dative argument was included. Levy & Keller (2013) argue (following Konieczny 2000 and others) that this is because the additional argument makes the verb more predictable. The more arguments have occurred, the more likely it is that the verb will follow, and the more information we have as to what it will be. However, a second experiment showed evidence for locality effects: In relative clause constructions, the verb was more difficult to process when both an adjunct and a dative phrase were included than when only one or the other was included. Levy & Keller (2013) suggest that this may be due to the relative clause context; locality effects may arise under such conditions because they involve a high memory load.

Another interesting phenomenon with regard to dependency length is extraposition: the extraction of a constituent out of its normal position. In English, extraposition sometimes occurs when a phrase

modifying a direct object (such as a relative clause or prepositional phrase) is moved after a phrase modifying the preceding verb, which can result in crossing dependencies:

(13) We spoke to a man at the party who says he knows you.

It has been observed that this strategy can sometimes be used to reduce dependency length. In sentence 13, for example, the dependency length of the extraposed construction is shorter than that of the "canonical" version below:

(14) We spoke to a man who says he knows you at the party.

If extraposition is a strategy to reduce dependency length, then it should be especially advantageous when the extraposed constituent is long. Francis (2010) finds that, indeed, sentences with extraposed relative clauses are processed more easily (relative to canonical versions of the same sentences) when the relative clause is long, and are also rated more highly in acceptability judgments. However, dependency length does not explain all aspects of extraposition processing. Levy et al. (2013) find that, even when dependency length is controlled, a sentence with an extraposed relative clause, like sentence 15, causes more processing difficulty than one in which the relative clause is merely nonadjacent to the NP, like sentence 16:

(15) The reporter interviewed the star about the movie who was married to the famous model.

(16) The reporter interviewed the star of the movie who was married to the famous model.

An alternative proposal that has some aspects in common with the DLT is the ACT-R (Adaptive Control of Thought–Rational) model of Lewis & Vasishth (2005, p. 398). Like the DLT, the ACT-R theory predicts processing difficulty when a word has to be integrated with a relatively distant previous word; like the DLT, this is due to the decrease in activation for the previous word. In the case of the ACT-R theory, however, this decrease in activation depends primarily on the time elapsed since the previous word, rather than on the structural "distance" between them, as in the DLT. The ACT-R theory posits additional effects as well. In particular, the ACT-R theory asserts that the difficulty of processing a dependency will depend in part on the interference caused by intervening constructions that are syntactically similar to the one being processed. For example, in the sentence *The assistant forgot [that] the student . . . was standing*, the difficulty of processing *was standing* is increased if it is separated from its subject by a relative clause, and increased still further if an embedded clause intervenes as well; because these constituents involve subject–verb relations, they create "structural interference" with the higher-level subject–verb dependency.

Other modifications of the DLT have also been proposed, relating particularly to the factors affecting the integration cost of a dependency. Gibson & Warren (2004) point to the ease of processing sentence 17 as opposed to sentence 18:

(17) The manager who$_i$ the consultant claimed t_i that the new proposal had pleased t_i will hire five workers tomorrow.

(18) The manager$_{ii}$ who$_i$ the consultant's claim about the new proposal had pleased t_i will hire five workers tomorrow.

They note that in the first sentence, the long dependency between *who* and *pleased* crosses a trace of the relative pronoun; this trace may boost the relative pronoun's activation, in effect facilitating its integration with the following verb and making the sentence easier to process. Alexopoulou & Keller (2007) offer yet another modification of the theory, in order to explain the difference in acceptability between sentences such as these:

(19) Who does Mary claim that we will fire?

(20) Who does Mary wonder whether we will fire?

These authors propose that the processing difficulty of the second sentence arises from the fact that (by general agreement) it spans an additional syntactic head (a C node) but that the first sentence does not. They suggest that the calculation of integration cost due to a dependency should be modified to take into account the number of syntactic heads that it spans.

Another type of evidence for the role of dependency length in sentence processing comes from ambiguity resolution. When two interpretations of a sentence are possible, there is often a preference for the one that minimizes dependency length. Consider this sentence, discussed by Gibson (1998):

(21) The bartender told the detective that the suspect left the country yesterday.

Yesterday could modify either the first verb *told* or the second verb *left*, but there is a strong preference to attach it to the second verb. Gibson suggests that this is because the latter attachment incurs lower integration cost with *yesterday*, because it is more recent than the first verb and thus more activated. Gibson argues that structural factors (integration and storage cost) interact with semantic plausibility and other factors to determine the preferred interpretation.

Although the idea that DLM might play a role in ambiguity resolution is intriguing, complexities arise when we consider carefully how it might be implemented. One possibility is that multiple parses of a sentence are fully constructed, and dependency length is used as a criterion (along with plausibility and other factors) to decide the preferred interpretation. This appears to be Gibson's assumption with sentence 21, above. This proposal seems doubtful to us, for several reasons. First, it seems intuitively unlikely that improbable interpretations such as the *told...yesterday* interpretation of sentence 21 are fully constructed. This becomes even clearer if we consider long sentences; imagine a 30-word sentence in which *told* was the third word and *yesterday* was the thirtieth word (such a sentence could easily arise in written text). A dependency between the two words would be extraordinarily long and would undoubtedly cause great processing difficulty, but surely we do not have to actually construct this dependency in order to reject the associated interpretation (especially when we consider that long sentences can contain a large number of possible adjunct connections of this type; Church & Patil 1982). In addition, if the dependency structures for both interpretations of sentence 21 are being constructed, then presumably the overall processing time for the sentence (assumed to be linear or at least monotonic with processing complexity; Gibson 1998, pp. 15–16) would reflect the processing time of both interpretations—either their sum (if processing is serial) or the maximum (if processing is parallel). Thus, we would predict high processing complexity for any sentence that has at least one high-complexity interpretation.

Another possibility is that parsing proceeds in a "greedy" fashion: Some incomplete structures (with unfilled dependencies) are eliminated from further consideration, while others are retained for further processing. At each point, dependency length is a factor in deciding which interpretations are

eliminated or retained. This reasoning is evident in Gibson's handling of reduced-relative/main-verb ambiguities, such as the following:

(22) The witness who the evidence examined by the lawyer implicated seemed to be very nervous.

The word *examined* could be the main verb of a relative clause with *evidence* as the subject (though this is semantically implausible) or part of a reduced-relative construction (i.e., *the evidence that was examined*); experiments show that readers generally strongly favor the main-verb interpretation, and then encounter great difficulty when it proves incorrect (at *by the lawyer*), suggesting that they are being forced to reconsider a previously abandoned analysis. Gibson suggests that, at *examined*, the reduced-relative interpretation is abandoned due to its high memory cost (the syntactic predictions generated by the matrix subject and the relative pronoun). Note that if the search procedure is greedy, then it is not guaranteed to find the overall interpretation that might actually be preferable, once plausibility and other factors are considered (or even just in terms of structural complexity). We suspect that this strategy may be difficult to implement in detail, given the human parser's remarkable ability, in most cases, to find the most plausible overall interpretation of a sentence. In any case, the point we wish to emphasize is that there are two possible ways that dependency length might be brought to bear on ambiguity resolution: in evaluating fully constructed, self-sufficient dependency structures, and in choosing particular incomplete structures to be retained or abandoned.

6. CONCLUSIONS

The research surveyed in this review makes a compelling case for DLM as an important factor in language structure and cognition. We see evidence of DLM in syntactic choice: In cases where a language permits different orderings of constituents, writers and speakers seem to favor constructions that reduce dependency length. DLM is also apparent in large-scale corpus studies. The actual dependency length of natural language text seems to be significantly lower than random, although still significantly higher than the optimal length. Such studies reflect the effect of DLM on both grammar and syntactic choice. And finally, a number of experimental studies suggest that DLM facilitates language comprehension, and may also be a factor in ambiguity resolution.

Clearly, this area holds great potential for further research. One interesting possibility is that these empirical explorations of dependencies might shed light on theoretical issues regarding the nature of dependencies in language. The criteria for determining "headhood"—what is a dependent of what—are somewhat unclear. Zwicky (1985, 1993) has argued that the term "head" has been used to capture three quite different properties—a semantic, "kind of" relation between a phrase and its head (a *red bird* is a kind of bird; *makes a box* is a kind of making), a syntactic property of obligatoriness (*a bird* is grammatical, whereas *a red* is not), and a syntactic similarity or transfer of features between a word and its larger phrase (*makes a box* inherits the properties of *makes*). [In response, Hudson (1987) argued that these apparently inconsistent criteria can in fact be reconciled with one another.] As noted above, there are also specific cases—such as NPs and coordinate phrases—for which headhood is a controversial issue. The various empirical approaches discussed here—corpus studies, computational simulations, and psycholinguistic experiments—could potentially be useful in this regard. That is to say, given the strong evidence for DLM, a finding that one syntactic analysis of (say) coordinate phrases resulted in shorter dependencies than another analysis (across a variety of languages), or better predicted processing complexity, could be regarded as a strong point in its favor.

An issue not addressed above is the ultimate cause of DLM. Two possibilities, not mutually exclusive, are that it arises from constraints on either production or comprehension. Hawkins's EIC theory

(which, as demonstrated above, is not equivalent to DLM but closely related to it) adopts a strongly comprehension-based view, suggesting that language producers place dependents close to their heads to facilitate parsing. Studies by Gibson and others, showing that sentences with longer dependencies create processing difficulty, might also be taken to suggest that a comprehension-based explanation for DLM is more plausible. However, evidence has also shown that comprehension difficulty is greatly affected by expectation, which in turn is shaped by the statistics of the linguistic input—the frequency of different syntactic structures and configurations (Levy 2008). If (for example) language production favors a short–long ordering of adjuncts, then it would not be surprising if this ordering was easier to process, simply because it is more often encountered and therefore more expected. Perhaps the most likely possibility is that DLM is inherently advantageous for both production and comprehension, and that each one is further shaped by the other in a mutually reinforcing process.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

LITERATURE CITED

Abeillé A, Rambow O, ed. 2001. *Tree Adjoining Grammars: Formalisms, Linguistic Analysis and Processing*. Stanford, CA: Cent. Study Lang. Inf.

Abney SP. 1987. *The English noun phrase in its sentential aspect*. PhD thesis, MIT, Cambridge, MA

Alexopoulou T, Keller F. 2007. Locality, cyclicity, and resumption: at the interface between the grammar and the human sentence processor. *Language* 83:110–60

Arnold JE, Wasow T, Losongco T, Ginstrom R. 2000. Heaviness versus newness: the effects of structural complexity and discourse status on constituent ordering. *Language* 76:28–55

Behaghel O. 1932. *Deutsche Syntax: eine geschichtliche Darstellung*, Bd. 4: *Wortstellung; Periodenbau*. Heidelberg, Ger.: C. Winter

Bresnan J, ed. 1982. *The Mental Representation of Grammatical Relations*. Cambridge, MA: MIT Press

Chomsky N. 1988. *Language and Problems of Knowledge: The Managua Lectures*. Cambridge, MA: MIT Press

Chung FRK. 1984. On optimal linear arrangements of trees. *Comput. Math. Appl.* 10:43–60

Church KW, Patil R. 1982. Coping with syntactic ambiguity. *Am. J. Comput. Linguist.* 8:139–49

Demberg V, Keller F. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition* 109:193–210

Diessel H. 2005. Competing motivations for the ordering of main and adverbial clauses. *Linguistics* 43:449–70

Dik SC. 1989. *The Theory of Functional Grammar*, part 1: *The Structure of the Clause*. Dordrecht, Neth.: Foris

Dryer M. 1992. The Greenbergian word order correlations. *Language* 68:81–138

Ferrer i Cancho R. 2004. Euclidean distance between syntactically linked words. *Phys. Rev. E* 70:1–5

Ferrer i Cancho R. 2006. Why do syntactic links not cross? *Europhys. Lett.* 76:1228–34

Francis EJ. 2010. Grammatical weight and relative clause extraposition in English. *Cogn. Linguist.* 21:35–74

Frazier L. 1985. Syntactic complexity. In *Natural Language Parsing: Psychological, Computational, and Theoretical Perspectives*, ed. DR Dowty, L Karttunen, A Zwicky, pp. 129–89. Cambridge, UK: Cambridge Univ. Press

Futrell R, Mahowald K, Gibson E. 2015. Large-scale evidence of dependency length minimization in 37 languages. *PNAS* 112:10336–41

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 *Image, Language, Brain: Papers from the 1st Mind Articulation Symposium*, ed. A Marantz, Y Miyashita, W O'Neil, pp. 95–126. Cambridge, MA: MIT Press

Gibson E, Warren T. 2004. Reading-time evidence for intermediate linguistic structure in long-distance dependencies. *Syntax* 7:55–78

Gildea D, Jaeger TF. 2015. Human languages order information efficiently. arXiv:1510.02823 [cs.CL]

Gildea D, Temperley D. 2007. Optimizing grammars for minimum dependency length. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, ed. A Zaenen, A van den Bosch, pp. 184–91. Stroudsburg, PA: Assoc. Comput. Linguist.

Gildea D, Temperley D. 2010. Do grammars minimize dependency length? *Cogn. Sci.* 34:286–310

Givón T. 1991. Isomorphism in the grammatical code: cognitive and biological considerations. *Stud. Lang.* 15:85–114

Greenberg J. 1963. Some universals of grammar with particular reference to the order of meaningful elements. *Universals of Language* 2:73–113

Hawkins J. 1994. *A Performance Theory of Order and Constituency*. Cambridge, UK: Cambridge Univ. Press

Hawkins J. 2004. *Efficiency and complexity in grammars*. Oxford, UK: Oxford Univ. Press

Hsiao F, Gibson E. 2003. Processing relative clauses in Chinese. *Cognition* 90:3–27

Hudson RA. 1987. Zwicky on heads. *J. Linguist.* 23:109–32

Hudson RA. 1991. *English Word Grammar*. Oxford, UK: Blackwell

Joshi AK. 1985. Tree adjoining grammars: How much context-sensitivity is required to provide reasonable structural descriptions? In *Natural Language Parsing: Psychological, Computational, and Theoretical Perspectives*, ed. DR Dowty, L Karttunen, A Zwicky, pp. 206–50. Cambridge, UK: Cambridge Univ. Press

King J, Just MA. 1991. Individual differences in syntactic processing: the role of working memory. *J. Mem. Lang.* 30:580–602

Konieczny L. 2000. Locality and parsing complexity. *J. Psycholinguist. Res.* 29:627–45

Levy R. 2008. Expectation-based syntactic comprehension. *Cognition* 106:1126–77

Levy R, Fedorenko E, Gibson E. 2013. The syntactic complexity of Russian relative clauses. *J. Mem. Lang.* 69:461–95

Levy RP, Keller F. 2013. Expectation and locality effects in German verb-final structures. *J. Mem. Lang.* 68:199–222

Lewis RL, Vasishth S. 2005. An activation-based model of sentence processing as skilled memory retrieval. *Cogn. Sci.* 29:375–419

Liu H. 2008. Dependency distance as a metric of language comprehension difficulty. *J. Cogn. Sci.* 9:159–91

Mel'čuk IA. 1988. *Dependency Syntax: Theory and Practice*. Albany, NY: SUNY Press

Munn AB. 1993. *Topics in the syntax and semantics of coordinate structures*. PhD thesis, Univ. Md., College Park

Oehrle RT, Bach E, Wheeler D, ed. 1988. *Categorial Grammars and Natural Language Structures*. Dordrecht, Neth.: Reidel

Park YA, Levy R. 2009. Minimal-length linearizations for mildly context-sensitive dependency trees. In *Proceedings of the 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 335–43. Stroudsburg, PA: Assoc. Comput. Linguist.

Pickering M, Barry G. 1993. Dependency categorial grammar and coordination. *Linguistics* 31:855–902

Pollard C, Sag IA. 1987. *Information-Based Syntax and Semantics*, vol. 1: *Fundamentals*. Chicago: Univ. Chicago Press

Radford A. 1997. *Syntactic Theory and the Structure of English: A Minimalist Approach*. Cambridge, UK: Cambridge Univ. Press

Rijkhoff J. 1990. Explaining word order in the noun phrase. *Linguistics* 28:5–42

Temperley D. 2007. Minimization of dependency length in written English. *Cognition* 105:300–33

Temperley D. 2008. Dependency-length minimization in natural and artificial languages. *J. Quant. Linguist.* 15:256–82

Vennemann T. 1974. Topics, subjects and word order: from SXV to SVX via TVX In *Historical Linguistics I*, ed. J Andersen, C Jones, pp. 339–376. Amsterdam: North Holland

Warren T, Gibson E. 2002. The influence of referential processing on sentence complexity. *Cognition* 85:79–112

Yamashita H. 2002. Scrambled sentences in Japanese: linguistic properties and motivations for production. *Text* 22:597–634

Yamashita H, Chang F. 2001. “Long before short” preference in the production of a head-final language. *Cognition* 81:45–55

Zwicky AM. 1985. Heads. *J. Linguist.* 21:1–29

Zwicky AM. 1993. Heads, bases and functors. In *Heads in Grammatical Theory*, ed. GC Corbett, NM Fraser, S McGlashan, pp. 292–315. Cambridge, UK: Cambridge Univ. Press