Integration complexity and the order of cosisters

William Dyer
Oracle Corp william.dyer@oracle.com

Abstract

The cost of integrating dependent constituents to their heads is thought to involve the distance between dependent and head and the complexity of the integration (Gibson, 1998). The former has been convincingly addressed by Dependency Distance Minimization (DDM) (cf. Liu et al., 2017). The current study addresses the latter by proposing a novel theory of integration complexity derived from the entropy of the probability distribution of a dependent's heads. An analysis of Universal Dependency corpora provides empirical evidence regarding the preferred order of isomorphic cosisters—sister constituents of the same syntactic form on the same side of their head—such as the adjectives in *pretty blue fish*. Integration complexity, alongside DDM, allows for a general theory of constituent order based on integration cost.

1 Introduction

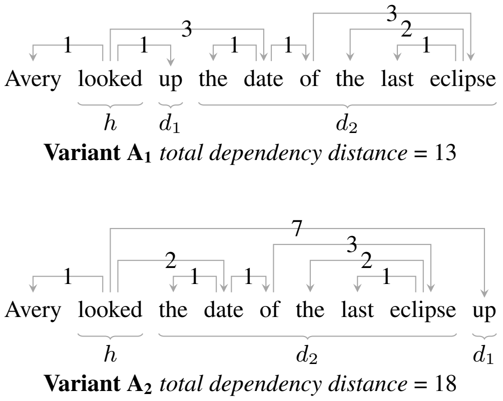
An open question in the field is why certain constituent orders are preferred to their reverse-order variants. For example, why do *pretty blue fish* or *Toni went to the store after eating lunch* seem more felicitous than *blue pretty fish* or *Toni went after eating lunch to the store*? In both sequences, two constituents of the same syntactic type depend on the same head—two 'stacked' adjectives modify *fish* and two prepositional phrases modify *went*. Yet despite their syntactic and truth-conditional equivalence, one order is preferred.

This order preference has often been treated with discrete models for each constituent type. For example, it has been proposed that stacked adjectives follow (1) a general hierarchy based on inherence (Whorf, 1945)—that is, the adjective closest to the head is more inherent to the head—discrimination (Ziff, 1960), intrinsicness (Danks and Glucksberg, 1971), temporariness (Bolinger, 1967; Larson, 2000), or subjectivity (Scontras et al., 2017); (2) a binary hierarchy based on features such as relative/absolute (Sproat and Shih, 1991), stage-/individual-level (Larson, 1998), or direct/indirect (Cinque, 2010); or (3) a multi-category hierarchy of intensional/subsective/intersective (Kamp and Partee, 1995; Partee, 2007; Truswell, 2009), reinforcer/epithet/descriptor/classifier (Feist, 2012), and perhaps most famously, semantic features such as size/shape/color/nationality (Quirk et al., 1985; Scott, 2002). Similarly, prepositional phrases and adverbials have been held to follow a hierarchy based on manner/place/time (Boisson, 1981; Cinque, 2001) or thematic roles such as evidential/temporal/locative (Schweikert, 2004). While these models may be reasonably accurate—though see Hawkins (2000); Truswell (2009); Kotowski (2016)—they seem to lack external motivation (Cinque, 2010, pp. 122-3) and explanatory power outside their specific constituent types.

A more general approach suggests that certain tendencies—constituents placed closer to their heads than their same-side sisters are more often complements than adjuncts (Culicover and Jackendoff, 2005) and are more likely to be shorter (Behaghel, 1930; Wasow and Arnold, 2003), less complex (Berlage, 2014), or have less grammatical weight (Osborne, 2007)—are the result of larger motivations such as Head Proximity (Rijkhoff, 1986, 2000), Early Immediate Constituents (Hawkins, 2004), or Minimize Domains (Hawkins, 2014). This line of inquiry seeks to explain Behaghel's (1932) observation that syntactic proximity mirrors semantic closeness, either due to iconicity or more recently as an efficiency-based aid to cognitive processing.

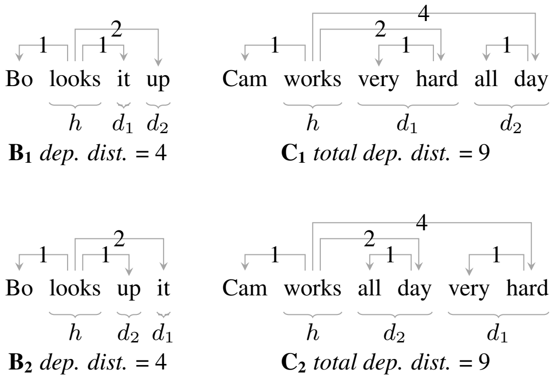
The current study sits within this latter approach of appealing to a general principle to motivate a constituent-ordering pattern.

Proceedings of the Second Workshop on Universal Dependencies (UDW 2018), pages 55–65
Brussels, Belgium, November 1, 2018. ©2018 Association for Computational Linguistics


Avery looked up the date of the last eclipse
h
d_1
d_2
1
1
3
1
1
3
2
1
Avery looked the date of the last eclipse up
h
d_2
d_1
1
7
2
1
1
3
2
1
Variant A$_1$ *total dependency distance* = 13

Variant A$_2$ *total dependency distance* = 18

Figure 1: DDM variants



B$_1$ *dep. dist.* = 4

C$_1$ *total dep. dist.* = 9

B$_2$ *dep. dist.* = 4

C$_2$ *total dep. dist.* = 9

Figure 2: Isomorphic cosisters

2 Dependency Distance & Isomorphic Cosisters

Dependency is a relation between words such that each word except the root depends on another word, forming a tree of dependents and heads (Tesnière, 1959; Mel'čuk, 2000). Dependency Distance Minimization[1] (DDM) holds that word orders which minimize the cumulative linear distance between dependents and their heads tend to be preferred to variants with longer total distances, where dependency distance is the count of words intervening between dependent and head (Liu et al., 2017). In Figure 1, for example, the two sentences may be semantically equivalent, but variant A$_1$ yields a total dependency distance of 13, which is smaller than that of A$_2$ at 18; thus A$_1$ is preferred according to DDM. The variants in Figure 1 hinge on whether the particle *up* appears closer to the head *looked* than the longer noun phrase *the date of the last eclipse*. DDM has been shown to be quite widespread, if not universal (Futrell et al., 2015), and rests on solid theoretical and empirical foundations from linguistics (Hudson, 1995), psycholinguistics (Futrell et al., 2017), and mathematics (Ferrer-i Cancho, 2004).

The methodology underlying DDM effectively punishes certain structures, including those in which two sister constituents are placed on the same side of their head—'cosisters' after Osborne (2007)—where the longer cosister appears closest to the head. Variant A$_2$ in Figure 1 shows such a case. One strategy for avoiding these structures is to alternate the placement of sister constituents on either side of the head (Temperley, 2008), as in many double-adjective noun phrases in Romance—the Spanish *gran globo rojo* [big balloon red] 'big red balloon'—and single- and multi-word adjective phrases in English, as in *the happy child / the child happy from playing outside*.

Another strategy for minimizing dependency distance is to place shorter cosisters closer to the head, as in Figure 1 variant A$_1$, in which the shorter dependent cosister d_1 is placed closer to the head h than its longer cosister d_2. Because the two cosisters are of differing length, DDM is able to predict that variant A$_1$ be preferred to A$_2$.

However, if the cosisters are of the same length, or more accurately if they have the same form, DDM is unable to explain the preference for one variant over another. Figure 2 shows two such structures, B and C, in which varying whether d_1 or d_2 appears closest to the head h does not yield a different total dependency distance. The cosisters d_i in B have the same structure, as do the cosisters d_i and C: the single-word *it* and *up* in B are single leaf-node dependents with no other internal structure, and the internal structure of *to LA* and *after lunch* is the same in that the first word depends on the second in both cases.

These isomorphic cosisters, or same-side sister constituents that share the same internal syntactic form, are the focus of the current study. In order to motivate a preference for one linear order over another, as in Figure 2 B$_1$ and C$_1$ over B$_2$ and C$_2$[2] we must appeal to a mechanism other than DDM.

[1] This approach is also called Dependency *Length* Minimization (DLM). Liu et al. (2017) suggests that because distance connotes a dynamic state which may vary, while 'length' is a more static feature, 'distance' is preferred. Recent literature (e.g. Ferrer-i Cancho, 2017; Futrell et al., 2017; Ouyang and Jiang, 2017) is converging on 'distance.'

[2] B$_2$ and C$_2$ are not necessarily impossible, just disfavored. When asked *Does Cam work very hard in the morning?*, the response *No, Cam works* ALL DAY *very hard*, might be marginally acceptable, especially with focus stress (Rooth, 1992). Adjective order tendencies—BLUE *pretty fish*—are also violable under similar contexts (Matthews, 2014, p. 95).

3 Integration Complexity

The cost of integrating a dependent to its head "consists of two parts: (1) a cost dependent on the complexity of the integration [... and] (2) a distance-based cost" (Gibson, 1998, p. 13). If we accept DDM as the basis for the distance-based cost and a valid motivation for preferred orders among different-length constituents (Futrell et al., 2017), a definition of integration complexity may allow the ordering preference between variant orders of isomorphic cosisters to be addressed.

Many have wrestled with the notion of linguistic complexity (Newmeyer and Preston, 2014) or grammatical weight (Wasow, 1997; Osborne, 2007), though a consensus has yet to emerge. Suggestions often involve number of words or phrase-structure nodes—more words or nodes equates to higher complexity—yet counterexamples to this sort of reasoning are readily found: Chomsky (1975, p. 477) notes that *they brought the man I saw in* is shorter and yet more complex than *they brought all the leaders of the riot in*. Further, isomorphic cosisters cannot be differentiated based on number of words or internal nodes, since the sister constituents in question are equal on both counts. Yet ordering preferences among this type of constituent remain; thus neither length nor syntactic structure can fully account for complexity.

We have an initial clue about relative integration complexity, inherited from the strategy used to minimize dependency distances: the shorter cosister should be placed closer to the head than the longer cosister. By analogy, we expect that the less-complex cosister should likewise be placed closer to the head than the more-complex one. For example, in Figure 2 B and C, we expect both d_1 constituents to have lower integration complexity than their d_2 cosisters; that is, because *looked it up* is preferred to *looked up it*, we infer that *looked* $\rightarrow$ *it* is a less complex integration than *looked* $\rightarrow$ *up*.

A second clue regarding integration complexity comes from nonce words, like *wug* or *tolver*, which seem to maintain order preferences when they appear as heads but not as dependents. For example, while *pretty blue wug* is preferred to *blue pretty wug* when the nonce word is a head, there is no obvious preference between *wuggy tolvic aliens* and *tolvic wuggy aliens*.

Together, these clues allow us to create two inferences: (1) integration complexity is based on a feature of dependents rather than heads, and (2) dependents with lower integration complexity tend to be placed closer to heads than their cosisters.

A plausible feature of dependents, one which could form the basis of integration complexity, is their frequency. However, a simple example shows that this cannot be the case: in *big chartreuse blanket*, the less-frequent adjective *chartreuse* is placed closest to the head, while in *miniscule white blanket* the more-frequent *white* is placed closest the head. Clearly frequency of dependent alone cannot be the force driving integration complexity.

A similar feature is the range of heads that a word can depend on. Ziff (1960) initially proposes that this 'privilege of occurrence' could be the mechanism underlying adjective order, giving the example of *little white house*, in which *little* can depend on a wider range of nouns than can *white*—*little sonnet* for example, but not *white sonnet*—suggesting that the dependent with a more narrow range of possible heads should be placed closest to the head. However, Ziff's counterexample of *intelligent old man*—"*old* has a much greater privilege of occurrence than *intelligent*" (p. 205)—suggests just the opposite, that the dependent with a wider range of heads should be placed closest to the head. Thus similar to raw frequency, the range of possible heads cannot directly define integration complexity.

Futrell et al. (2017) suggest that the mutual information of the dependent-head pair may hold the key to explaining why, "for instance, adjuncts are typically farther from their heads than arguments, if it is the case that adjuncts have lower mutual information with their heads" (p. 2). Mutual information (MI) is one of a series of information-theoretic measures based on Shannon entropy (Shannon, 1948) to gauge how knowing the value of one random variable informs us about another variable (Cover and Thomas, 1991). Pointwise mutual information (PMI) (Bouma, 2009), a version of MI, is frequently used for quantifying the relationship between words (Church and Hanks, 1989). However, PMI requires that the individual frequencies of dependent, head, and dependent-head co-occurrence be known. Nonce words by definition have no frequency, either alone or in co-occurrence with a dependent, so their PMI with a dependent is undefined. It is unclear how an integration complexity based on mutual information could deal with nonce words.

Instead of frequency, 'privilege of occurrence,' or mutual information, it seems plausible that given a dependent word, the relative predictability of its heads should correlate with integration complexity: a dependent whose set of heads is quite small or predictable should be easier to integrate, while a dependent with a wide variety of equally probable heads should be more difficult.

Therefore a measure of integration complexity should be low in the case of a word which depends predictably on very few heads and high when the word's heads are numerous or varied. Entropy (Shannon, 1948) captures this idea mathematically by measuring the 'peakedness' of a probability distribution—the more peaked a distribution, the lower its entropy (Jaynes, 1957)—and is calculated as the logarithm of the probabilities in a distribution, weighted by each probability (Cover and Thomas, 1991), as shown in Equation 1.

$$\mathrm{H}(X) = -\sum_{i=1}^{n} \mathrm{P}(x_i) * \log_b \mathrm{P}(x_i) \qquad (1)$$

A dependent whose heads form a peaked probability distribution is easier to integrate—and therefore has a lower entropy—than a dependent whose heads form a flatter distribution.

In information-theoretic terms, given a dependent with a wide variety of heads of equal probability, we expect a large amount of surprisal or information when the head is determined; this is high entropy. Conversely a dependent with a few very likely heads is expected to yield a small amount of information, captured as low entropy.

However, using the actual head-word lexemes or lemmata in our entropy calculation for dependents is problematic for a subtle reason: it would weight head words equally. Integrating a dependent to a set of heads which are themselves quite similar semantically or distributionally should not yield a large amount of surprisal. One way to more properly weight head words according to their similarity is to use syntactic categories as a basis for the probability distribution. Words of each category—nouns, verbs, adjectives, and so on—are by definition closer to each other functionally and distributionally.

It is the proposal of this paper that by weighting each dependent word by its integration complexity, as measured by the entropy of the probability distribution of the syntactic categories of the word's heads, the order preference between isomorphic cosisters can be modeled—specifically that the constituent with a lower integration complexity tends to be placed closer to the head. Further, cosisters with roughly equal integration complexity should not show a particularly strong order tendency, while cosisters with greatly differing integration complexity should have a strong tendency of placing the constituent with lower integration complexity closest to the head.

Formally, let the integration complexity IC of dependent d be the entropy H of the probability distribution of the syntactic categories of the heads of d. Let a head h have two isomorphic dependent constituents d_1 and d_2 appearing on the same linear side of h in the surface realization and with integration complexity $\mathrm{IC}(d_1)$ and $\mathrm{IC}(d_2)$. It is hypothesized that as the difference between the two complexities $|\mathrm{IC}(d_1) - \mathrm{IC}(d_2)|$ increases, the tendency to place the constituent with lower IC closer to the head should also increase.

4 Methodology

The Universal Dependencies (UD) project provides corpora that can be used to both calculate the integration complexity of dependent words and show a preference for one variant order over another. That is, the UD corpora can be used to formulate the probability distribution of the syntactic categories of the heads that a given word tends to depend on—training—as well as the apparent order preference for a pair of cosisters: testing.

Because one goal of Universal Dependencies is to "create cross-linguistically consistent treebank annotation for many languages within a dependency-based lexicalist framework" (Nivre et al., 2016, p. 1659), certain linguistic features are annotated in a somewhat non-intuitive way. Copula and auxiliaries are not treated as the root of a sentence, but instead depend on a predicate or main verb. Further, rather than considering adpositions as the heads of adpositional phrases, as would be common under a phrase-structure framework (cf. Stockwell, 1977), UD treats them as dependents of their associated nouns or verbs. This approach is not without controversy, and there are cross-linguistic arguments, mainly typological, to be made in favor of an adpositional-phrase treatment (Hagège, 2010). Nevertheless, because UD corpora are tagged such that copula, auxiliaries, and adpositions are dependents rather than heads, the current study uses this annotation scheme.

(1) head lemmata of *happy*
afford, always, band, birthday, camper, check, choose, customer (2), enjoy, feel (2), give, go, happy (2), holiday, hour (2), keep, make, need, safe, say (3), tell, walk, year (2)
(2) syntactic categories of lemmata
ADJ (3), ADV, NOUN (8), PROPN (2), VERB (15)
(3) probability distribution of syntactic categories
$^{3}/_{29}$, $^{1}/_{29}$, $^{8}/_{29}$, $^{2}/_{29}$, $^{15}/_{29}$
(4) entropy of probability distribution
1.78 bits

Figure 3: Calculating integration complexity of *happy*

Finally, UD version 2.2 contains multiple corpora for some languages, designed to be applied to various types of analysis. Because the current study requires as full a picture as possible for the syntactic-category tendencies for each dependent word, as well as a sufficient quantity of isomorphic-cosister sequences to test, the largest corpus for each language in the Universal Dependencies will be analyzed here.

4.1 Training

Determining the integration complexity of each dependent is done by finding the probability distribution of the syntactic categories of the heads each word depends on in the UD corpus and calculating the entropy with Equation 1.

For example, Figure 3 shows the entropy calculation for the adjective *happy*. The word appears 29 times in the English-EWT corpus as a dependent on a set of head lemmata (1), with a variety of syntactic categories (2). Those categories form a probability distribution (3) whose entropy, assuming a logarithmic base of 2, is 1.78 bits (4). For comparison, other adjectives have integration complexities such as *little* (1.56 bits), *Italian* (0.76 bits), and *chemical* (0.5 bits).

This process of finding the heads of each dependent, using the heads' syntactic categories to create a probability distribution, and calculating the entropy of that distribution, is repeated for each word in the corpus, thereby determining the integration complexity of all dependents.

4.2 Testing

The UD corpora can also be used to test the hypothesis that the lower-complexity cosister tends to be placed closest to the head. While the order of words as attested in a corpus is not a direct substitution for an order preference in all situations, the corpus order does imply that in the specific context of the sentence in the corpus, the attested order is preferred to others. In effect, we are using frequency—the sentence exists at least once in the corpus—as a logistically convenient stand-in for actual order preference (Song, 2012, pp. 14-5).

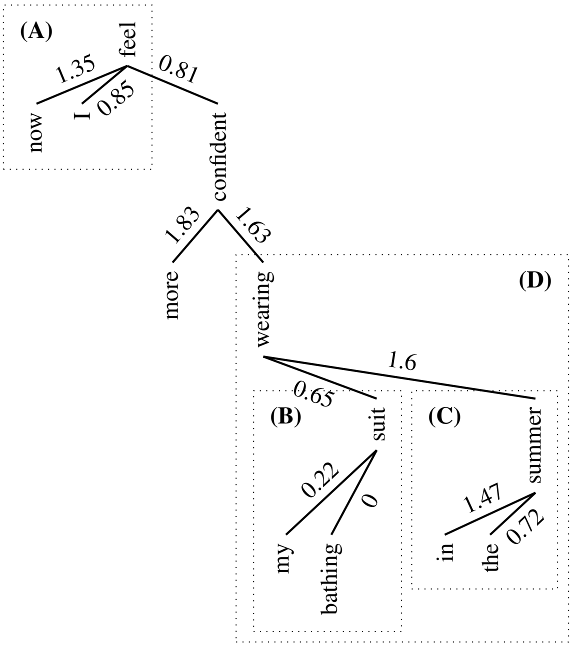


Figure 4: Integration complexity of cosisters

Figure 4 shows an example sentence from the English-EWT corpus: *now I feel more confident wearing my bathing suit in the summer*. The sentence is annotated according to the UD scheme—notably the preposition *in* is a dependent of the noun *summer*—and lists the integration complexity of each dependent word. For example, the integration complexity of *now* is 1.35 bits, calculated as the entropy of the probability distribution of the syntactic categories of the heads of the adverb *now* in the UD English-EWT corpus.

The sentence contains four instances of isomorphic cosisters and their heads: (A) *now I feel*, where *now* and *I* are cosisters of the same syntactic form—single-leaf nodes with no dependents themselves—which precede their head *feel*; (B) *my bathing suit*, where *my* and *bathing* precede their head *suit*; (C) *in the summer*, where *in* and *the* precede their head *summer*; and (D) *wearing my bathing suit in the summer*, where the multi-word *my bathing suit* and *in the summer* are isomorphic cosisters following their head *wearing*.

In the first case (A), the adverb *now* has an integration complexity of 1.35 bits, while the pronoun

I has 0.85 bits; therefore the lower-complexity cosister, *I*, has been placed closest to the head *feel*. Both (B) *my bathing suit* and (C) *in the summer* also follow this pattern—the lower-complexity *bathing* and *the* are placed closer to their heads than their cosisters *my* and *in*—thereby confirming the hypothesis for these single-word cosisters.

For the multi-word isomorphic cosisters *my bathing suit* and *in the summer*, there are at least two possible strategies. One method is to sum the integration complexity of all nodes, yielding an integration complexity of 0.87 bits for *my bathing suit*—*my* (0.22) + *bathing* (0) + *suit* (0.65)—and a summed complexity of 3.79 bits for *in the summer*: *in* (1.47) + *the* (0.72) + *summer* (1.6).

Another approach is to treat multi-word constituents according to the Dependency Distance Minimization method: a total dependency distance is created by calculating the sum of integration complexity of all words intervening between a dependent and head. This approach yields a total integration complexity of 0.99 bits for *my bathing suit*: (0.22 + 0) for *my* ← *suit*; (0) for *bathing* ← *suit*; and (0.65 + 0.22 + 0) for *wearing* → *suit*.

It is not clear which method is a better representation of the complexity of integrating multi-word constituents for a human parser. Further, given the limited scope of the structures under analysis in this study, it is not clear that one method would result in markedly different outcomes *vis-à-vis* the relative complexity of isomorphic cosisters. For simplicity, the first method of summing the integration complexity of all nodes in a constituent will be used here[3]. Thus for the isomorphic cosisters in Figure 4 (D), the complexity of *my bathing suit* is calculated as 0.87 bits, while that of *in the summer* is 3.79 bits; as such the lower-complexity cosister has been placed closer to the head.

5 Results

Table 1 shows logistic regressions for single- and multi-word isomorphic cosisters. Each language with at least 20 analyzed isomorphic cosisters is listed, along with the specific UD corpus and total number of structures analyzed. The x-axis in each graph shows the difference between the integration complexity of the two cosisters from 0 to 5 bits, and the y-axis shows the probability between 0 and 1 that the lower-complexity cosister has been placed closest to the head.

We see that of the 70 languages analyzed, 61 show a pattern that as the difference between the integration complexity increases, the lower-complexity cosister is more likely to be placed closest to the head. Croatian and Russian show a general preference for placing the less-complex cosister closest to the head, but that preference does not appear to increase as the integration complexities diverge. Japanese is indeterminate showing approximately 50% probability regardless of complexity difference. Six do not follow the hypothesized pattern: Afrikaans, Ancient Greek, Galician, North Sami, Tamil, and Vietnamese seem to prefer that the higher-complexity cosister be placed closest to the head as the difference in integration complexity increases.

There does not seem to be a clear pattern to the set of languages which do not follow the study's hypothesis. Ancient Greek, and North Sami have rich inflectional systems—and resulting 'free' word order —but so do Basque, Estonian, Latin, Old Church Slavonic, and Turkish, which conform to the study's hypothesis.

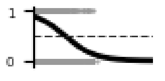
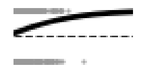
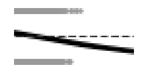
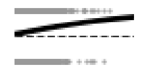
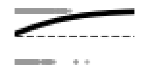
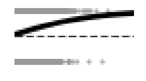
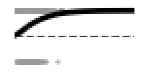
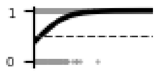
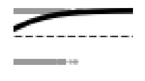
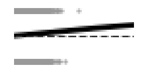
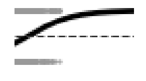
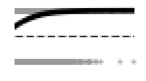
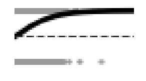
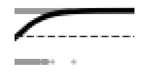
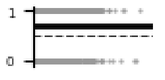
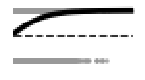
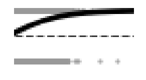
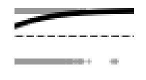
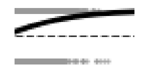
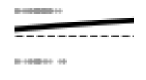
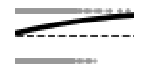
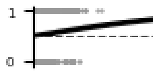
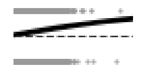
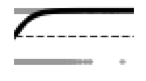
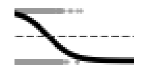
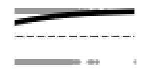
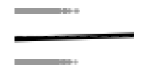
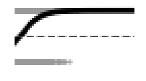
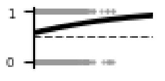
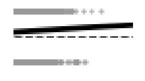
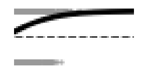
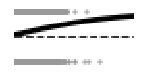
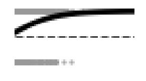
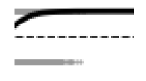
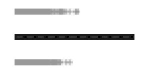
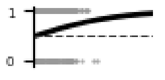
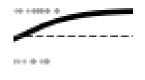
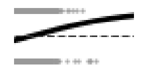
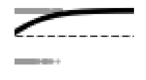
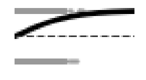
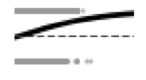
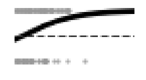
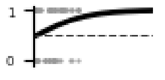
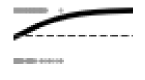
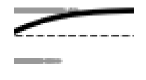
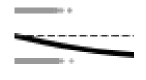
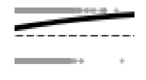
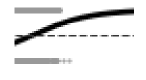
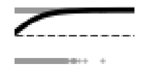
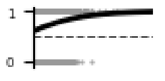
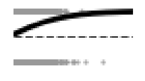
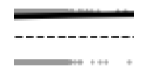
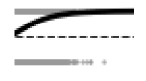
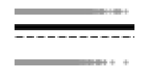
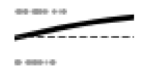
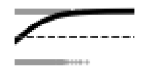
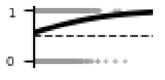
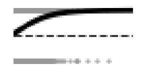
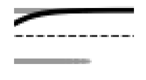
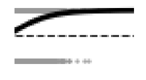
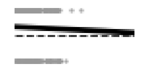
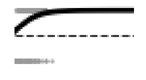
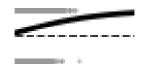
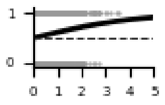
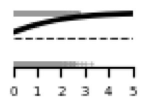
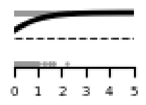
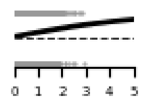
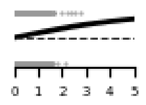
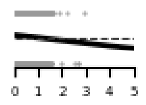
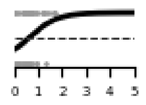
Nor do language families seem to play a role in these non-conforming languages. Afrikaans and Gothic are outweighed by the many other Germanic languages—Danish, Dutch, English, and so on—which do follow the hypothesis; likewise the conformity of Catalan, French, Italian, Latin, Old French, Portuguese, Romanian, and Spanish to the hypothesized pattern discounts Romance as an explanation for Galician's non-conformity. North Sami is countered by its Uralic cousins of Estonian, Finnish, Hungarian, and Komi Zyrian.

Data sparsity is a possibility—North Sami and Vietnamese both contain fewer than 1,000 structures analyzed—but Ancient Greek and Galician seem to have sufficient data, and other corpora with few structures conform to the hypothesis: Armenian (398), Belarusian (267), and so on.

Instead, a likely cause is noise from language-specific tagging and lemmatization in the UD corpora, amplified by the calculation of integration complexity, especially in multi-word cosisters. However, that noise actually makes the overall success rate—61 of 70, or 87.1% of languages—more impressive, as it suggests that a real structural regularity can be found in the data.

[3] Entropy is additive for independent systems (Wehrl, 1978). Because the integration of each dependent to its head is treated as a separate event—the integration of dependent A to head B is independent the integration of B to its head C—summing integration complexity should be sound.

Table 1: Results

Afrikaans AfriBooms (3001) 	**Amharic** ATT (969) 	**Anc. Greek** Perseus (7969) 	**Arabic** PADT (6271) 	**Armenian** ArmTDP (399) 	**Basque** BDT (2593) 	**Belarusian** HSE (268) 
Breton KEB (727) 	**Bulgarian** BTB (8989) 	**Buryat** BDT (314) 	**Cantonese** HK (301) 	**Catalan** AnCora (36146) 	**Chinese** GSD (3723) 	**Coptic** Scriptorium (1103) 
Croatian SET (9086) 	**Czech** PDT (45147) 	**Danish** DDT (5234) 	**Dutch** Alpino (15100) 	**English** EWT (15347) 	**Erzya** JR (68) 	**Estonian** EDT (12087) 
Faroese OFT (535) 	**Finnish** TDT (7241) 	**French** GSD (37556) 	**Galician** CTG (8170) 	**German** GSD (20174) 	**Gothic** PROIEL (1839) 	**Greek** GDT (4120) 
Hebrew HTB (8961) 	**Hindi** HDTB (11782) 	**Hungarian** Szeged (1065) 	**Indonesian** GSD (3756) 	**Irish** IDT (752) 	**Italian** ISDT (28351) 	**Japanese** BCCWJ (24091) 
Kazakh KTB (302) 	**Komi Zyrian** IKDP (28) 	**Korean** GSD (1548) 	**Kurmanji** MG (365) 	**Latin** ITTB (13229) 	**Latvian** LVTB (4402) 	**Lithuanian** HSE (124) 
Maltese MUDT (120) 	**Marathi** UFAL (120) 	**Naija** NSC (784) 	**N. Sami** Giella (996) 	**Norwegian** Bokmaal (17446) 	**Old Ch. Slav.** PROIEL (1899) 	**Old French** SRCMF (11606) 
Persian Seraji (4445) 	**Polish** SZ (2929) 	**Portuguese** Bosque (21094) 	**Romanian** RRT (9468) 	**Russian** SynTagRus (47769) 	**Sanskrit** UFAL (63) 	**Serbian** SET (3970) 
Slovak SNK (4897) 	**Slovenian** SSJ (6778) 	**Spanish** AnCora (39467) 	**Swedish** Talbanken (5033) 	**Tamil** TTB (285) 	**Telugu** MTG (272) 	**Thai** PUD (866) 
Turkish IMST (1475) 	**Ukrainian** IU (4142) 	**Up. Sorbian** UFAL (584) 	**Urdu** UDTB (4871) 	**Uyghur** UDT (550) 	**Vietnamese** VTB (870) 	**Yoruba** YTB (140) 

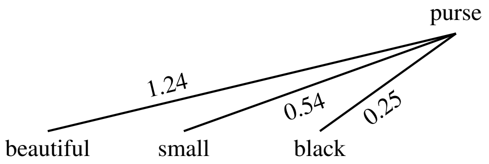


Figure 5: Hierarchical adjective order restrictions

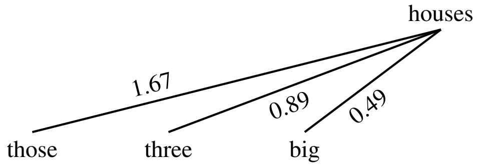


Figure 6: Greenberg's Universal 20

6 Discussion

The findings of this study reveal a widespread cross-linguistic tendency to order isomorphic cosisters such that those placed nearest to the head have the lowest integration complexity (IC). Because this tendency seems to occur across all constituent types, many fine-grained models previously proposed for specific constituent types can be subsumed by an IC approach. Further, by combining IC with DDM, a general theory of constituent ordering based on integration cost begins to take shape.

6.1 Subsuming previous models

Previous constituent-specific models of ordering can be reformulated in terms of the larger insight of ordering based on integration complexity. For example, rather than appeal to an arbitrary adjective-specific hierarchy of features such as subjective comment, size, and color to explain the order of *beautiful small black purse*—preferred to other permutations (Teodorescu, 2006)—the order can be attributed to integration complexity and the pattern that cosisters with lower IC tend to be placed closest to the head. Figure 5 shows the IC of each adjective, and indeed they follow the pattern: *beautiful* (1.24 bits[4]), *small* (0.54 bits), and *black* (0.25 bits).

As to why adjectives of size or color should congregate with regard to their placement around the noun, because the distribution of size- or color-type adjectives is likely quite similar—the set of heads that *black* depends on is presumably similar to the set that *white* or *yellow* depend on as well—their IC is likely much the same. As such, the hierarchy reveals itself as an epiphenomenon resulting from the distributional similarity of classes of adjectives.

Other patterns of noun modifiers also seem to yield to an integration-complexity explanation. In Universal 20, Greenberg (1963) observes that "When any or all of the items (demonstrative, numeral, and descriptive adjective) precede the noun, they are always in that order. If they follow, the order is either the same or its opposite." Dryer (2009) further refines the formulation based on a set of languages larger than Greenberg's, confirming the prenominal order as near-universal and showing that postnominal orders are vastly more likely to be the mirror order. However, why this pattern might be appears to be an open question.

Adopting an integration-complexity approach, we see in Figure 6 that the IC of the demonstrative *those* (1.67 bits) is larger than than of the numeral *three* (0.89 bits), which is itself larger than the adjective *big* (0.49 bits)[5]. Thus the IC of the noun modifiers[6] in *these three big houses* follows the established pattern that constituents placed closest to the head tend to have lower IC.

Other phenomena, such as heavy noun phrase shift, dative shift or alternation, and particle movement or placement (Gries, 1999; Wasow and Arnold, 2003), largely deal with deviations from the supposedly canonical verb-complement-adjunct order. However, both the canonical order and its deviations can be reformulated as an effect of a strategy based on integration complexity: because both complements and constituents with lower IC tend to be placed closest to the head, complements likely have lower IC than adjuncts. Similarly, deviations tend to occur when the adjunct has a lower IC than the complement.

Integration complexity is the more inclusive mechanism, able to account for preferred orders of adjectives, noun modifiers, and both the canonical order of complements and adjuncts as well as deviations from that order.

[4] Here and throughout this section, integration complexity is calculated from the UD-English-EWT corpus.

[5] UD marks demonstratives as "PronType=Dem" and cardinal numerals as "NumType=Card." Descriptive adjectives are not differentiated from modals or intensionals like *possible* or *former* by UD.

[6] There is an ongoing debate over whether demonstratives or determiners in general modify nouns and are therefore part of the noun phrase, or if nouns instead are the dependents of a larger determiner phrase (cf. Szabolsci, 1983; Abney, 1987; Hudson, 2004; Matthews, 2014). The current study follows UD and treats determiners as syntactic dependents of nouns.

6.2 Integration Cost

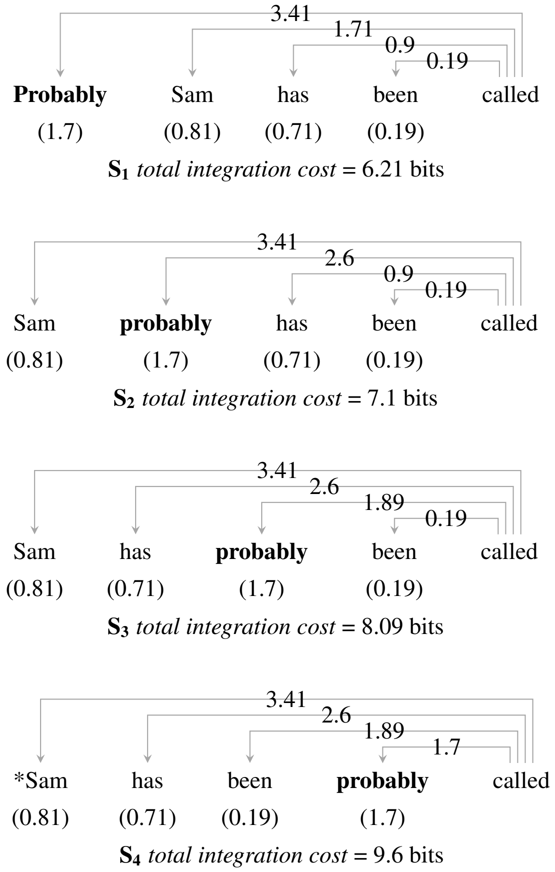
DDM measures the distance between a word and its head as the count of words intervening between the two (Liu et al., 2017). This count quantifies the distance-based cost of integrating dependents to their heads (Gibson, 1998, 2000). By introducing integration complexity as formulated in the current study as a sort of weight for each word, we are able to capture both the distance- and complexity-based parts of the cost of integration. Integration cost is therefore the sum of the integration complexity of a dependent and that of any words intervening between the dependent and its head.

Integration cost as so defined allows us to address another constituent-ordering phenomenon: English adverb placement. For example, Potsdam (1998), citing Jackendoff (1980), suggests that inserting the adverb *probably* into *Sam has been called* is possible in three preverbal positions but disfavored in a fourth. As the examples in Figure 7 show, it may appear clause-initially (S_1); immediately after the subject (S_2); immediately after a modal or finite auxiliary (S_3); but is disfavored immediately after a non-finite auxiliary (S_4).

Figure 7 also shows the integration complexity and cost of each dependent and the total integration cost for each variant. For example, *probably* has a complexity of 1.7 bits and an integration cost of 3.41 bits in S_1—the sum of the integration complexity of *probably* and that of each word intervening between *probably* and *called*: 1.7 (*probably*) + 0.81 (*Sam*) + 0.71 (*has*) + 0.19 (*been*). The total integration cost of S_1 is 6.21 bits, the sum of the cost of integrating each dependent in the sentence.

The total integration cost of the disfavored S_4 is 9.6 bits, higher than the acceptable variants S_1 (6.21), S_2 (7.1), and S_3 (8.09). The unacceptability of S_4 may derive from its higher integration cost.

Integration cost as defined here rests on dependency distance minimization and a pattern of placing isomorphic cosisters with lower integration complexity closest to the head, both of which are evident as widespread structural regularities in corpora, and seems capable of addressing various ordering phenomena previously unexplored or explained by constituent-specific models.



S_1 *total integration cost* = 6.21 bits

S_2 *total integration cost* = 7.1 bits

S_3 *total integration cost* = 8.09 bits

S_4 *total integration cost* = 9.6 bits

Figure 7: Integration cost of adverb placement

7 Summary

This study addresses the order preference of isomorphic cosisters—pairs of sister constituents of the same syntactic form on the same side of their head—by building upon the insight that the cost of integrating dependents to their heads derives from the complexity of the integration and the distance between dependent and head (Gibson, 1998, 2000). Adopting methodology from Dependency Distance Minimization, which favors structures where the shorter of two cosisters appears closest to the head, this paper shows that as the integration complexity between two cosisters diverges, the tendency to place the constituent with the lower integration complexity closer to the head increases across most languages analyzed.

As such, this study contributes to the field by (1) providing a novel definition of integration complexity as the entropy of the probability distribution of the syntactic categories of a dependent word's heads; (2) demonstrating with a 70-language analysis that the order of isomorphic cosisters based on integration complexity describes a widespread cross-linguistic structural regularity; and (3) suggesting that many previously proposed constituent-specific ordering models can be subsumed by a more inclusive and externally motivated theory based on integration cost.

References

Steven P. Abney. 1987. *The English noun phrase in its sentential aspect*. Ph.D. thesis, Massachusetts Institute of Technology.

Otto Behaghel. 1930. Von deutscher Wortstellung [On German word order]. *Zeitschrift für Deutschkunde, Jargang 44 der Zeitschrift für deutschen Unterricht*, pages 81–9.

Otto Behaghel. 1932. *Deutsche Syntax eine geschichtliche Darstellung*. Carl Winters Universitätsbuchhandlung, Heidelberg.

Eva Berlage. 2014. *Noun Phrase Complexity in English*. Cambridge University Press, Cambridge.

Claude Boisson. 1981. Hiérarchie universelle des spécifications de temps, de lieu, et de manière. *Confluents*, 7:69–124.

Dwight Bolinger. 1967. Adjectives in English: Attribution and Predication. *Lingua*, 18:1–34.

Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, pages 31–40.

Ramon Ferrer-i Cancho. 2004. Euclidean distance between syntactically linked words. *Physical Review E*, 70(056135):1–5.

Ramon Ferrer-i Cancho. 2017. Towards a theory of word order. Comment on" Dependency distance: a new perspective on syntactic patterns in natural language" by Haitao Liu et al. *Physics of Life Reviews*.

Noam Chomsky. 1975. *The Logical Structure of Linguistic Theory*. University of Chicago Press, Chicago. 1955.

Kenneth Ward Church and Patrick Hanks. 1989. Word association norms, mutual information, and lexicography. In *Proceedings of the 27th annual meeting on Association for Computational Linguistics -*, pages 76–83, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Guglielmo Cinque. 2001. "Restructuring" and functional structure. *University of Venice Working Papers in Linguistics*, 11:45–127.

Guglielmo Cinque. 2010. *The Syntax of Adjectives: A Comparative Study*. The MIT Press, Cambridge, Massachusetts.

T. M. Cover and Joy A. Thomas. 1991. *Elements of information theory*. Wiley series in telecommunications. Wiley, New York.

Peter Culicover and Ray Jackendoff. 2005. *Simpler Syntax*. Oxford linguistics. Oxford University Press.

Joseph H. Danks and Sam Glucksberg. 1971. Psychological scaling of adjective orders. *Journal of Verbal Learning and Verbal Behavior*, 10(1):63–7.

Matthew S. Dryer. 2009. On the order of demonstrative, numeral, adjective, and noun: an alternative to Cinque. In *Conference on theoretical approaches to disharmonic word orders*.

James Murray Feist. 2012. *Premodifiers in English*. Cambridge University Press, Cambridge.

Richard Futrell, Roger Levy, and Edward Gibson. 2017. Generalizing dependency distance. *Physics of Life Reviews*, 21:197–9.

Richard Futrell, Kyle Mahowald, and Edward Gibson. 2015. Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, 112(33):10336–41.

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

Edward Gibson. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. *Image, language, brain*, pages 95–126.

Joseph Greenberg. 1963. Some universals of grammar with particular reference to the order of meaningful elements. In Joseph Greenberg, editor, *Universals of Grammar*, pages 73–113. MIT Press, Cambridge, Massachusetts.

Stefan T. Gries. 1999. Particle movement: A cognitive and functional approach. *Cognitive Linguistics*, 10(2).

Claude Hagège. 2010. *Adpositions*. Oxford University Press, Oxford.

John A. Hawkins. 2000. The relative order of prepositional phrases in English: Going beyond Manner–Place–Time. *Language variation and change*, 11(03):231–66.

John A. Hawkins. 2004. *Efficiency and Complexity in Grammars*. Oxford University Press, Oxford.

John A. Hawkins. 2014. *Cross-linguistic variation and efficiency*. Oxford University Press, New York.

Richard Hudson. 1995. Measuring syntactic difficulty.

Richard Hudson. 2004. Are determiners heads? *Functions of Language*, 11(1):7–42.

Ray Jackendoff. 1980. *Semantic Interpretation in Generative Grammar*. Studies in linguistics series. MIT Press.

E. T. Jaynes. 1957. Information Theory and Statistical Mechanics. *The Physical Review*, 106(4):620–30.

Hans Kamp and Barbara Partee. 1995. Prototype theory and compositionality. *Cognition*, 57:129–91.

Sven Kotowski. 2016. *Adjectival Modification and Order Restrictions*. De Gruyter, Berlin.

Richard Larson. 1998. Events and modification in nominals. In *Proceedings from Semantics and Linguistic Theory (SALT)*, volume 8, pages 145–68.

Richard Larson. 2000. Temporal modification in nominals. *Handout of paper presented at the International Round Table "The Syntax of Tense and Aspect" Paris, France*.

Haitao Liu, Chunshan Xu, and Junying Liang. 2017. Dependency distance: A new perspective on syntactic patterns in natural languages. *Physics of Life Reviews*, 21:171–93.

Peter Matthews. 2014. *The Positions of Adjectives in English*. Oxford University Press, New York.

Igor Mel'čuk. 2000. Dependency in Linguistic Description.

Frederick Newmeyer and Laurel Preston. 2014. *Measuring Grammatical Complexity*. Oxford University Press, Oxford.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, and others. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1659–1666.

Timothy Osborne. 2007. The Weight of Predicates: A Dependency Grammar Analysis of Predicate Weight in German. *Journal of Germanic Linguistics*, 19(01):23–72.

Jinghui Ouyang and Jingyang Jiang. 2017. Can the Probability Distribution of Dependency Distance Measure Language Proficiency of Second Language Learners? *Journal of Quantitative Linguistics*, pages 1–19.

Barbara Partee. 2007. Compositionality and coercion in semantics: The dynamics of adjective meaning. *Cognitive foundations of interpretation*, pages 145–61.

Eric Potsdam. 1998. A Syntax for Adverbs. *Proceedings of the Twenty-seventh Western Conference on Linguistics*, 10:397–411.

Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of Contemporary English*. Longman, London.

Jan Rijkhoff. 1986. Word Order Universals Revisited: The Principle of Head Proximity. *Belgian Journal of Linguistics*, 1:95–125.

Jan Rijkhoff. 2000. When can a language have adjectives? An implicational universal. In Petra Vogel and Bernard Comrie, editors, *Approaches to the Typology of Word Classes*, pages 217–58. Mouton de Gruyter, New York.

Mats Rooth. 1992. A theory of focus interpretation. *Natural language semantics*, 1(1):75–116.

Walter Schweikert. 2004. The order of prepositional phrases. *Working Papers in Linguistics*, 14:195–216.

Gregory Scontras, Judith Degen, and Noah D. Goodman. 2017. Subjectivity Predicts Adjective Ordering Preferences. *Open Mind*, pages 1–14.

Gary-John Scott. 2002. Stacked adjectival modification and the structure of nominal phrases. In *Functional Structure in DP and IP: The Cartography of Syntactic Structures*, volume 1, pages 91–210. Oxford University Press, New York.

Claude Elwood Shannon. 1948. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55.

Jae Jung Song. 2012. *Word Order*. Cambridge University Press, New York.

Richard Sproat and Chilin Shih. 1991. The Cross-Linguistic Distribution of Adjective Ordering Restrictions. In Carol Georgopoulos and Roberta Ishihara, editors, *Interdisciplinary Approaches to Language*, pages 565 – 93. Kluwer Academic Publishers, Boston.

Robert Stockwell. 1977. *Foundations of syntactic theory*. Prentice-Hall foundations of modern linguistics series. Prentice-Hall.

Anna Szabolsci. 1983. The possessor that ran away from home. *The Linguistic Review*, 3(1):89–102.

David Temperley. 2008. Dependency-length minimization in natural and artificial languages. *Journal of Quantitative Linguistics*, 15(3):256–82.

Alexandra Teodorescu. 2006. Adjective ordering restrictions revisited. In *Proceedings of the 25th west coast conference on formal linguistics*, pages 399–407. Citeseer.

Lucien Tesnière. 1959. *Éléments de syntaxe structural*. Klincksieck, Paris.

Robert Truswell. 2009. Attributive adjectives and nominal templates. *Linguistic Inquiry*, 40(3):525–33.

Thomas Wasow. 1997. Remarks on grammatical weight. *Language Variation and Change*, 9(01):81–105.

Thomas Wasow and Jennifer Arnold. 2003. Postverbal constituent ordering in English. *Topics in English Linguistics*, 43:119–54.

Alfred Wehrl. 1978. General properties of entropy. *Reviews of Modern Physics*, 50(2):221–60.

Benjamin Lee Whorf. 1945. Grammatical Categories. *Language*, 21(1):1–11.

Paul Ziff. 1960. *Semantic Analysis*. Cornell University Press, Cornell, NY.