Comprehensive Journal of Humanities and Educational Studies
المجلة الشاملة للدراسات الإنسانية التربوية
Volume 2, Issue 2, 2026, Pages: 193-207
Journal homepage :http://cjhes.com.ly/index.php/cjhes/en/index

Variation in Grammar as a Theoretical Issue: A Usage-Based Approach to Explain Language Patterns

Zeena Al-Asi *
Department of English, Faculty of Education, Zuwara, University of Zawia, Libya

التباين في القواعد النحوية كقضية نظرية: مقاربة قائمة على الاستخدام لتفسير الأنماط اللغوية

زينة العاصي*
قسم اللغة الإنجليزية، كلية التربية - زوارة، جامعة الزاوية، ليبيا

*Corresponding author: zynhalasy7@gmail.com

Received: February 05, 2026	Accepted: April 04, 2026	Published: April 15, 2026

Abstract:
This paper argues that variation in grammar is not a side effect. It is part of grammar itself. A usage-based view explains this point with stored experience, repeated exposure, and graded choices. Grammar is shaped by what speakers hear, say, and remember. Patterns grow strong through repetition. Rare options stay weak, local, or genre-bound. This view helps explain why languages show both stable structure and flexible choice. The paper first reviews major work on usage-based grammar, variation, and constructional learning. It then presents a small open-data experiment based on five Universal Dependencies treebanks and one genre study from the English Web Treebank. The analysis tracks pronoun subjects, auxiliary density, subject-verb order, verb-object order, and adjective position. The results show clear cross-linguistic contrasts and clear genre effects. English shows frequent pronoun subjects and high auxiliary density. Spanish shows low overt subject use. Turkish strongly prefers object-before-verb order. French and Spanish favor many postnominal adjectives. These patterns fit a model in which grammar is a probabilistic network of constructions shaped by discourse, processing, and repeated use. The study supports a theoretical claim. Variation should not be placed outside grammar. It should be treated as evidence for how grammar is built, stored, and changed over time.

Keywords: grammar variation, usage-based linguistics, constructions, corpus linguistics, Universal Dependencies, probabilistic grammar, language patterns.

الملخص
تجادل هذه الورقة البحثية بأن التباين في القواعد النحوية ليس مجرد أثر جانبي، بل هو جزء أصيل من القواعد نفسها. ويوضح المنظور القائم على الاستخدام (Usage-based view) هذه النقطة من خلال الخبرات المخزنة، والتعرض المتكرر، والخيارات المتدرجة؛ حيث يتشكل النحو بناءً على ما يسمعه المتحدثون، وما يقولونه، وما يتذكرونه. تكتسب الأنماط قوتها من خلال التكرار، بينما تظل الخيارات النادرة ضعيفة، أو محلية، أو مرتبطة بنوع أدبي (Genre) معين. يساعد هذا المنظور في تفسير سبب

إظهار اللغات لبنية مستقرة وخيارات مرنة في آن واحد. تبدأ الورقة بمراجعة الأعمال الرئيسية حول النحو القائم على الاستخدام، والتباين، والتعلم البنائي .(Constructional learning) ثم تعرض تجربة صغيرة تعتمد على البيانات المفتوحة من خمسة بنوك شجرية (Treebanks) تابعة للمشروع العالمي "Universal Dependencies"، ودراسة نوعية واحدة من البنك الشجري للويب الإنجليزي English) Web Treebank). يتتبع التحليل الفاعل الضميري، وكثافة الأفعال المساعدة، وترتيب (الفاعل-الفعل)، وترتيب (الفعل-المفعول)، وموقع الصفة. وتظهر النتائج تباينات واضحة بين اللغات وتأثيرات جلية لنوع النص؛ حيث تظهر اللغة الإنجليزية استخداماً متكرراً للضمائر كفاعل وكثافة عالية للأفعال المساعدة، بينما تظهر الإسبانية استخداماً منخفضاً للفاعل الظاهر. كما تفضل اللغة التركية بقوة ترتيب (المفعول قبل الفعل)، في حين تميل الفرنسية والإسبانية إلى وضع العديد من الصفات بعد الاسم. تتفق هذه الأنماط مع نموذج يُنظر فيه إلى النحو كشبكة احتمالية من البنى التي يشكلها الخطاب، والمعالجة الذهنية، والاستخدام المتكرر. وتدعم الدراسة ادعاءً نظرياً مفاده: لا ينبغي وضع التباين خارج نطاق النحو، بل يجب التعامل معه كدليل على كيفية بناء القواعد وتخزينها وتغيرها بمرور الوقت.

الكلمات المفتاحية: التباين النحوي، اللسانيات القائمة على الاستخدام، البنى اللغوية، لسانيات المدونات الحاسوبية، التبعيات العالمية (Universal Dependencies)، النحو الاحتمالي، الأنماط اللغوية.

Introduction

Variation in grammar has often been treated as a side matter. It was pushed to the edge of theory. Many models gave center stage to fixed rules. Variable patterns were then placed in performance, style, or noise. That move hides a basic fact. Speakers do not use one form in all settings. They choose among options. Those options are not random. They follow strong tendencies. They react to genre, discourse, processing load, and repeated practice. A grammar model must explain that patterned choice. If it cannot, the model stays incomplete.

Table 1 Two ways to think about grammar and variation.

Issue	More categorical reading	Usage-based reading	Why it matters
Source of structure	Grammar is mainly a fixed rule set.	Grammar grows from stored events and larger schemas.	Repeated use can shape the system itself.
Status of variation	Variation sits at the edge of grammar.	Variation is part of grammatical knowledge.	Probabilities become theoretical evidence.
Role of frequency	Frequency is secondary or external.	Frequency changes strength, access, and generalization.	Common forms become entrenched routines.
Genre and discourse	Genre effects are outside core grammar.	Genre leaves traces in construction choice.	Context helps organize the grammar network.
Path of change	Change looks like rule replacement.	Change starts with skewed local use.	Variation can become the seed of change.

A usage-based approach offers a direct answer. It treats grammar as a record of language experience. Speakers store many concrete events. They also build larger schemas from them. Frequent strings become strong. Frequent mappings become easy to access. Competing forms stay in the system with different strengths. Choice is therefore graded, not empty. Variation becomes evidence about the shape of grammar, not evidence against grammar (Bybee, 2006; Langacker, 1987).

This paper argues that variation is a theoretical issue in its own right. It is not only a descriptive problem. It bears on the nature of grammatical knowledge. A model that admits only rigid categories misses the way speakers actually organize form and meaning. A usage-based model can explain why a pattern is stable, why a rival pattern survives, and why both may shift over time (Ellis, 2002; Goldberg, 2006).

The paper has two goals. The first goal is conceptual. It shows why variation belongs inside grammatical theory. The second goal is practical. It tests the claim with open corpora. The study uses public treebanks from the Universal Dependencies project and a genre analysis from the English Web Treebank. These resources make the argument testable and transparent (Nivre et al., 2020).

The study asks three linked questions. Do open corpora show stable grammatical contrasts across languages. Do genres inside one language shift grammatical choice in regular ways. Can these patterns be read as signs of a probabilistic grammar built through use. The answer developed here is yes on all three counts.

Variation in grammar as a theoretical issue

The core problem is simple. Grammar can be described as a set of forms. Yet speakers never use all forms alike. Some variants cluster in speech. Some cluster in writing. Some depend on information flow. Some depend on processing ease. A theory must explain why one structure rises in one setting and falls in another. It must also explain why the same speaker can use both forms without losing competence.

Variationist work made this point early and clearly. Labov showed that variable patterns are orderly. They are shaped by linguistic and social factors. The famous study of copula variation did not reveal chaos. It revealed strong constraints on choice (Labov, 1969). That result matters for grammar. If variable forms show stable conditioning, then the grammar cannot be purely categorical. It must include graded tendencies or linked probabilities.

Hopper pushed the debate in another direction. He argued that grammar is emergent. Structure grows out of discourse and repeated use. On this view, grammar is not a sealed code that sits apart from interaction. It is formed and re-formed through recurrent practice (Hopper, 1987). This claim does not deny pattern. It changes where pattern comes from. Pattern comes from use, sedimentation, and analogy. Variation is expected in such a system, because systems formed in use remain sensitive to context.

A strict rule model often treats competing forms as equal outputs or as separate modules. That move may describe the forms, but it does not explain their uneven distribution. Why does a language keep a low-frequency option alive. Why does one option become normal in one genre but marked in another. Why do shifts spread slowly through a speech community. These questions force theory to face gradient facts.

Construction-based work adds a further insight. Speakers do not store only words. They also store pairings of form and meaning at many sizes. Some pairings are narrow. Others are broad. Some remain tied to one lexical item. Others generalize. This layered storage predicts that variants can live side by side. A narrow pattern may stay strong in one discourse niche, while a wider schema governs the rest (Goldberg, 2006; Hilpert, 2014).

The theoretical issue, then, is not whether variation exists. It is how grammar contains it. One answer says that grammar gives a single ideal pattern, and variation sits outside. Another

answer says that grammar is a network with uneven strengths, local clusters, and competing routines. The second answer fits the evidence better. It links competence to experience. It also explains why change often starts as skewed preference before it becomes broad convention (Croft, 2000; Diessel, 2007).

Table 1 frames this contrast. The point is not to reject abstraction. Usage-based work also uses abstraction. The difference lies in the path to abstraction. Generalizations arise from repeated tokens and repeated local analogies. They are not free from use. That makes room for stable variation inside the theory itself.

A usage-based approach to grammatical patterns

Usage-based theory starts with memory. Speakers hear countless utterances. They do not erase these events after comprehension. Repeated exposure leaves traces. Similar traces cluster. Strong clusters yield entrenched patterns. Weak clusters remain rare or local. Grammar is therefore an organized memory system, not only a rule list (Bybee, 2006; Tomasello, 2003).

Frequency matters in two linked ways. Token frequency strengthens a known unit. Type frequency supports a wider schema. A very common phrase may become easy to access as a chunk. A large family of similar phrases may support a productive pattern. Both effects shape grammar. Both also shape variation. A rare variant stays fragile. A frequent variant becomes fast, expected, and resistant to loss (Ellis, 2002; Diessel, 2007).

Chunking is central here. Speakers often process recurring strings as single units. This lowers effort. It also changes representation. Once a string becomes familiar, it may support wider analogies. Analogies then connect new utterances to older ones. A network grows. The network does not hold only one abstract rule. It holds many paths with different strengths. This is why variable patterns can remain orderly without becoming fully categorical (Langacker, 1987; Perek, 2015).

The model also explains why discourse matters. Some constructions fit foregrounded agents. Some fit background information. Some suit direct involvement between speaker and hearer. Some match dense exposition. If speakers meet a pattern again and again in one discourse task, the pattern becomes tied to that task. Later choices will reflect this history. Genre effects are therefore not external decoration. They are part of grammatical organization.

Auxiliaries offer a good example. They are common in languages and genres that favor explicit tense, aspect, stance, and speaker control. Overt pronoun subjects show a related logic. They tend to rise where clause-by-clause anchoring is frequent. They may drop where person marking and discourse continuity reduce the need for overt subjects. Word order behaves in the same way. A language may permit several arrangements, but use history makes one arrangement dominant, easy, and expected.

Usage-based theory also gives a clear view of change. Change does not begin with a sudden rewrite of grammar. It often begins with small skews in use. A form becomes common in one context. The form is then stored more strongly in that context. New analogies spread from there. Over time, what looked like free variation becomes patterned preference, then stronger convention, and sometimes later a new norm (Bybee, 2006; Croft, 2000).

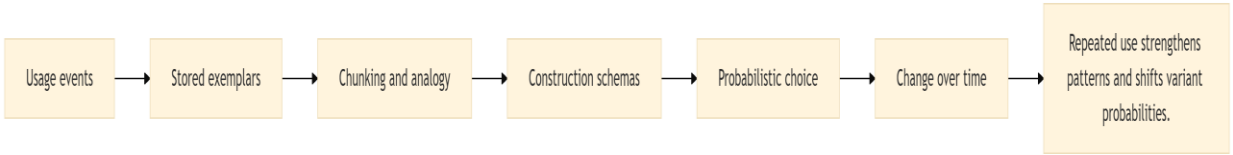

Figure 1 A simple usage-based cycle from repeated events to grammatical change.

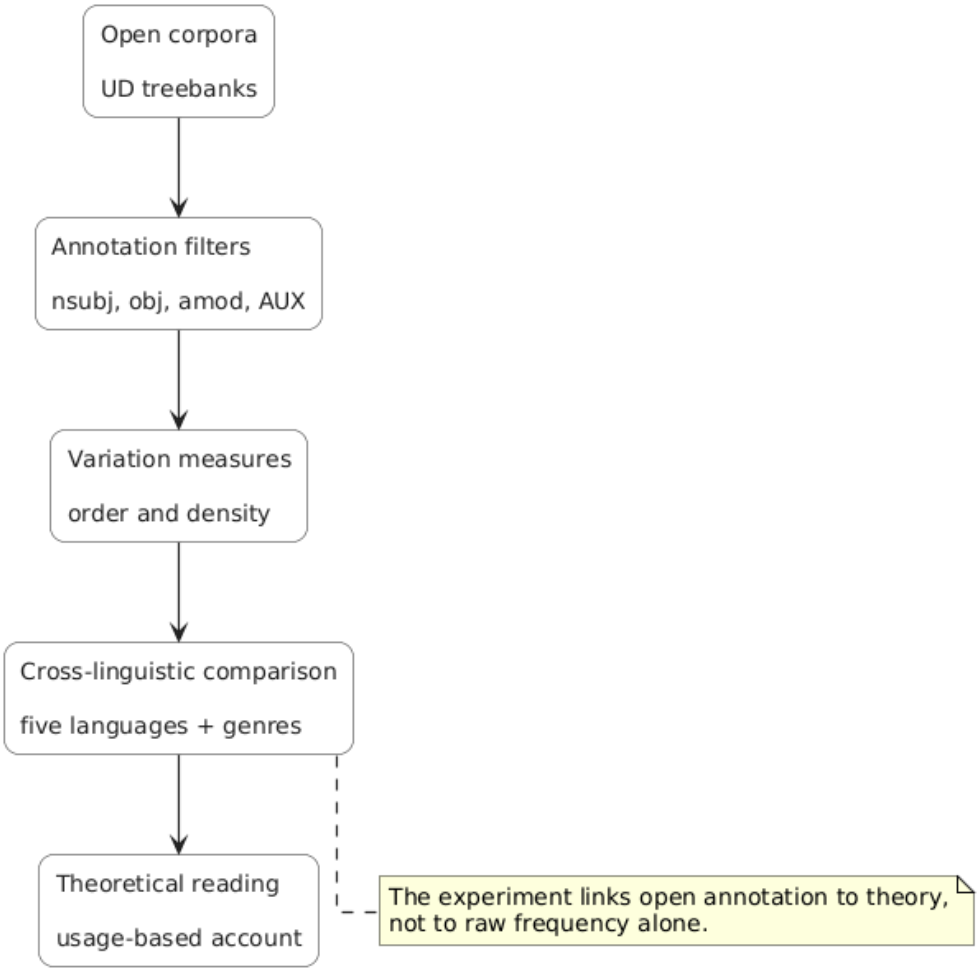

Figure 2 Workflow of the public corpus experiment used in this paper.

Figure 1 shows this logic in simple form. Usage events create stored traces. Traces support chunking and analogy. Larger constructional schemas then guide later choice. Repeated choice changes the system again. The cycle keeps grammar and use in contact. Figure 2 links this idea to the present experiment. Open corpora make the cycle visible through actual distributions.
This framework does not deny grammar. It gives grammar a richer shape. Grammar becomes a network of forms, meanings, cues, and probabilities. Variation is then expected wherever networks overlap. The task of theory is to explain those overlaps, their strength, and their limits.

Earlier research and the present gap

Many studies already connect usage and grammar. Tomasello showed that children build grammar from item-based learning and later generalization (Tomasello, 2003). Langacker described grammar as symbolic and usage-sensitive from the start (Langacker, 1987). Goldberg explained how constructions can range from fixed patterns to productive templates (Goldberg, 2006). Hilpert and Perek later showed how corpus evidence can reveal these gradients in fine detail (Hilpert, 2014; Perek, 2015).
Research on probabilistic syntax also matters here. Bresnan and Ford showed that speakers process syntactic alternatives in ways that reflect distributional expectations. Their work makes a strong point. Probabilistic structure is not a weak add-on after grammar. It is part of how grammar is learned and used in real time (Bresnan & Ford, 2010). This supports a usage-based view, because the same distributional patterns appear in both production and comprehension.

Even so, a gap remains. Some theoretical debates still treat variation as secondary. Others accept variation, but keep it separate from grammar proper. In both cases, the bridge from abstract theory to public data stays thin. That gap is important for students and new researchers. A strong theory should be testable with open materials. It should also explain simple contrasts that readers can verify for themselves.

The present study addresses that gap with a small but direct design. It does not claim to solve all questions of grammar. It asks whether a usage-based reading can explain visible patterns in public corpora. The answer matters because the data are open, replicable, and easy to inspect. A theory gains force when readers can retrace the evidence.

Data and method

The empirical section uses five public Universal Dependencies treebanks. The sample includes English EWT, German GSD, French GSD, Spanish GSD, and Turkish IMST. These corpora were chosen for two reasons. First, they are widely used open resources. Second, they display clear contrasts in subject expression, auxiliaries, and constituent order. The study also uses genre labels from the English EWT corpus for a smaller within-language test. Table 2 lists the resources and source links (Nivre et al., 2020).

Table 2 Public corpora used in the study and the main source links.

Language	Treebank	Sentences	Tokens	Official page	Download file
English	UD English EWT	12544	204577	Treebank page	Train file
German	UD German GSD	13813	263777	Treebank page	Train file
French	UD French GSD	14450	354648	Treebank page	Train file
Spanish	UD Spanish GSD	14187	382444	Treebank page	Train file
Turkish	UD Turkish IMST	3435	37522	Treebank page	Train file

The analysis focuses on simple, visible measures. Pronoun subject share is the percentage of subject tokens that are pronouns. Auxiliary density is the number of AUX tokens per 1,000 tokens. Subject-before-verb share measures how often a subject appears before its verbal head. Verb-object share measures how often an object follows its head verb. Adjective-after-noun share tracks the position of adjectival modifiers. These measures do not capture all grammar. They do capture regular differences in clause design and information packaging.

All counts were drawn from the training files of the public treebanks. Multiword tokens and empty nodes were excluded. Dependency labels and universal part-of-speech tags were read from the CoNLL-U files. The procedure stays close to the annotation scheme of Universal Dependencies, which is designed for cross-linguistic comparison (Nivre et al., 2020). The calculations are simple on purpose. A theory argument should not depend on hidden processing steps.

The genre analysis uses five English EWT genres that are easy to compare: answers, email, newsgroup, reviews, and weblog. For these texts, the study measures pronoun subject share and modal density. Modal density counts English modal lemmas such as can, may, must, will, and would per 1,000 tokens. The aim is modest. It asks whether genre changes grammatical choice in a regular way inside one language.

This is not a causal experiment in the strict laboratory sense. It is a corpus experiment with public data. That choice fits the paper's goal. The central claim concerns patterned distribution in actual use. Open treebanks offer a good test bed for that claim. They show what speakers and writers repeatedly do, not only what analysts imagine they could do.
The method has limits. Treebanks differ in genre balance and size. They also reflect editorial choices. Still, the design remains useful. The measures are broad. The languages contrast sharply on the target features. The goal is not fine statistical prediction. The goal is to show that variation is structured enough to require a grammatical account, and that a usage-based account can supply one.

Results

Table 3 reports the main cross-linguistic results. Figure 3 shows the sentence counts for the five treebanks. The corpora are not equal in size, but each is large enough for broad distributional patterns. The English, German, French, and Spanish files all contain over 200,000 tokens. The Turkish file is smaller, yet it still shows stable contrasts in word order and subject expression.

Table 3 Main cross-linguistic measures from the five public treebanks.

Language	Pronoun subjects (%)	AUX per 1,000	SV order (%)	VO order (%)	Adj after noun (%)
English	62.3	62.6	96.7	96.5	2.4
German	32.5	35.4	73.7	54.5	0.6
French	40.8	32.7	97.1	85.0	69.6
Spanish	9.8	24.8	81.2	90.7	71.9
Turkish	13.8	19.5	95.8	2.6	1.5

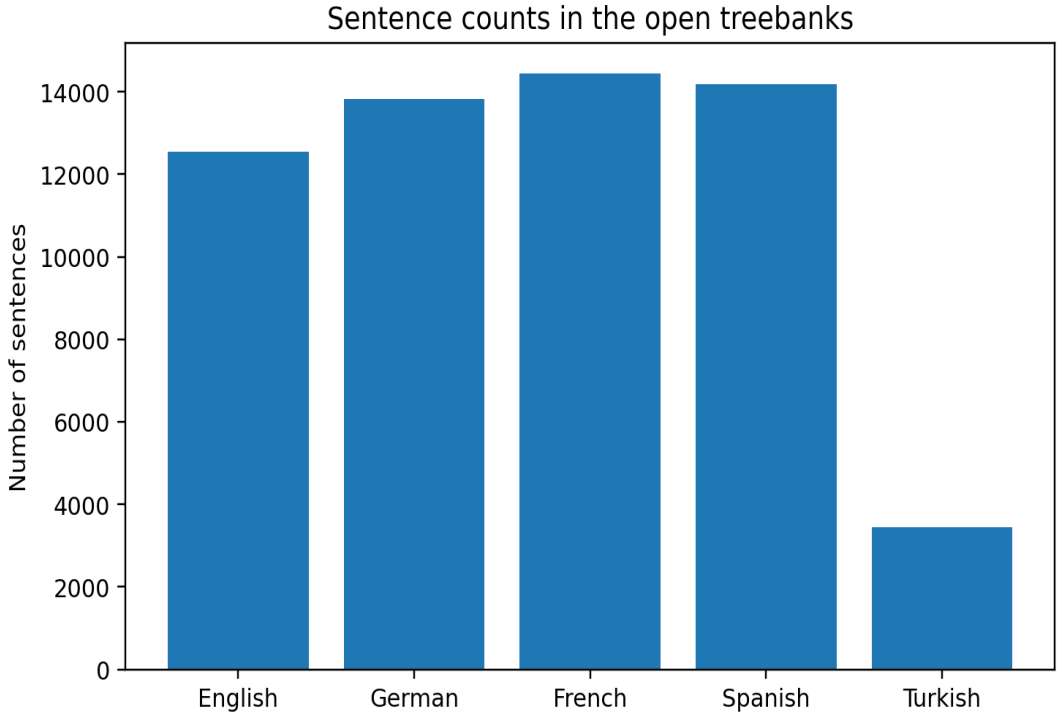

Figure 3 Sentence counts in the open treebanks used for comparison.

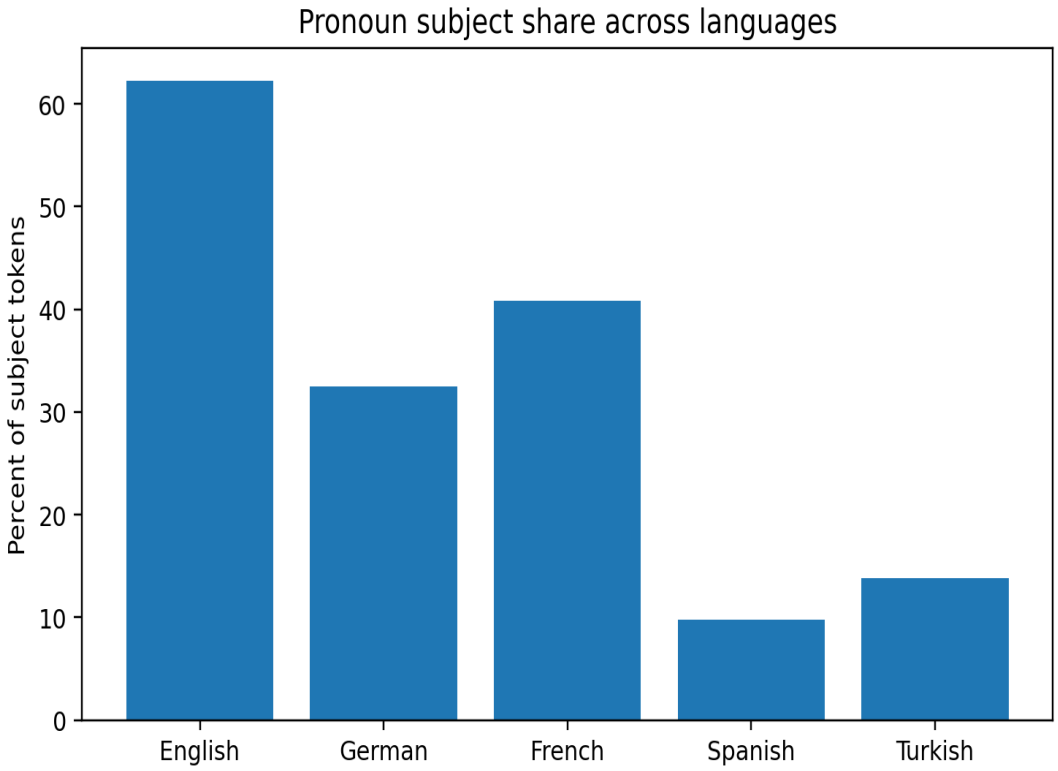

Figure 4 Pronoun subject share across the five languages.

Figure 4 shows pronoun subject share. English has the highest value in this sample, at 62.3 percent. French also uses many overt pronoun subjects, though at a lower rate, with 40.8 percent. German stands lower still, at 32.5 percent. Spanish and Turkish are much lower, with 9.8 percent and 13.8 percent. This contrast matches a usage-based claim. Grammatical routines reflect what speakers repeatedly need to mark in discourse. Languages with rich agreement and frequent null subjects do not rely on overt pronouns in the same way.

Figure 5 shows auxiliary density. English stands highest with 62.6 auxiliary tokens per 1,000 tokens. German follows with 35.4. French shows 32.7, Spanish 24.8, and Turkish 19.5. The order fits everyday language use. English clauses often make tense, aspect, stance, and support verbs overt. Turkish expresses much of this material morphologically, so fewer separate auxiliaries are needed. The point is not that one grammar is richer. The point is that recurrent packaging choices leave measurable traces in the grammar network.

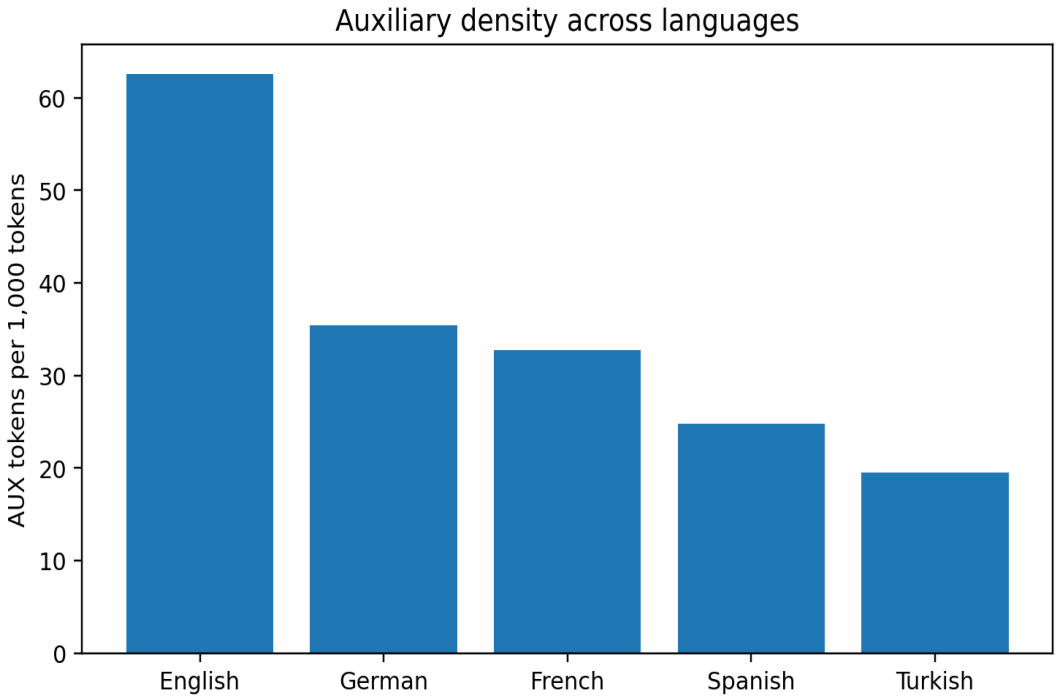

Figure 5 Auxiliary density across the five languages.

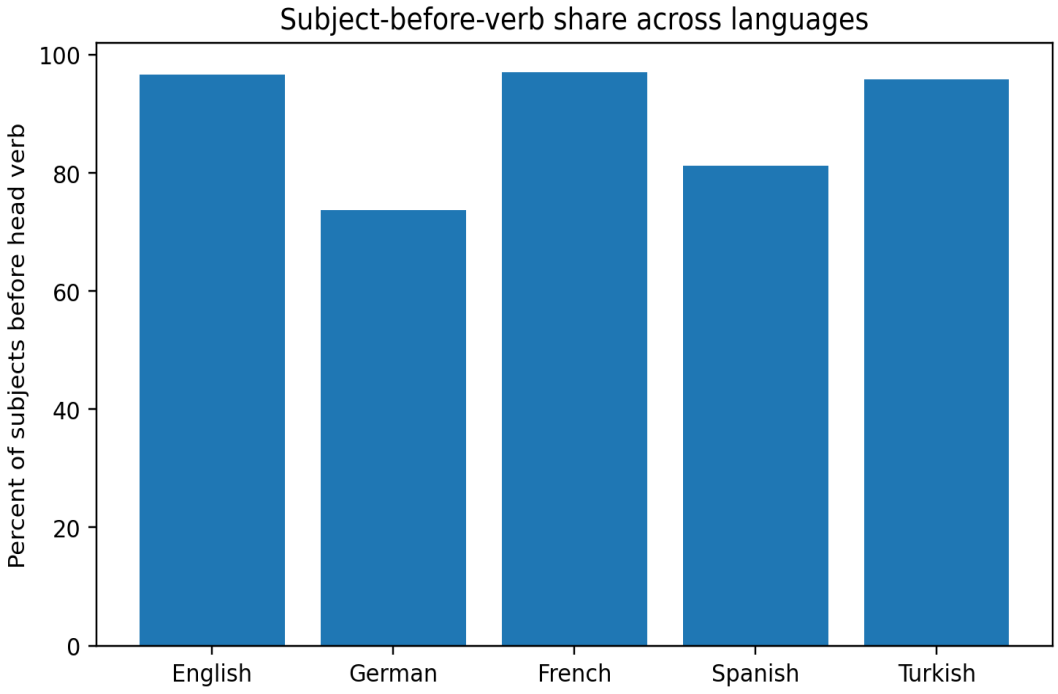

Figure 6 Subject-before-verb share across the five languages.

Figure 6 turns to subject position. English and French both place subjects before the verb in over 96 percent of relevant cases. Turkish also shows a high subject-before-verb value in this treebank sample, but this does not mean that Turkish has lost its broader flexibility. It means that the texts in this file still favor a common discourse routine. German and Spanish show lower values, at 73.7 and 81.2 percent. These differences remind us that grammar is not only about what is possible. It is also about what is usual.

Figure 7 gives the strongest typological contrast in the paper. English, French, and Spanish all show high verb-object rates. English reaches 96.5 percent. French reaches 85.0 percent. Spanish reaches 90.7 percent. German is lower, at 54.5 percent, which reflects its mixed clause patterns. Turkish is sharply different. Only 2.6 percent of objects follow the verb in this sample. The dominant routine is object-before-verb. A usage-based model explains this well. High-frequency order patterns become default expectations in production and comprehension.

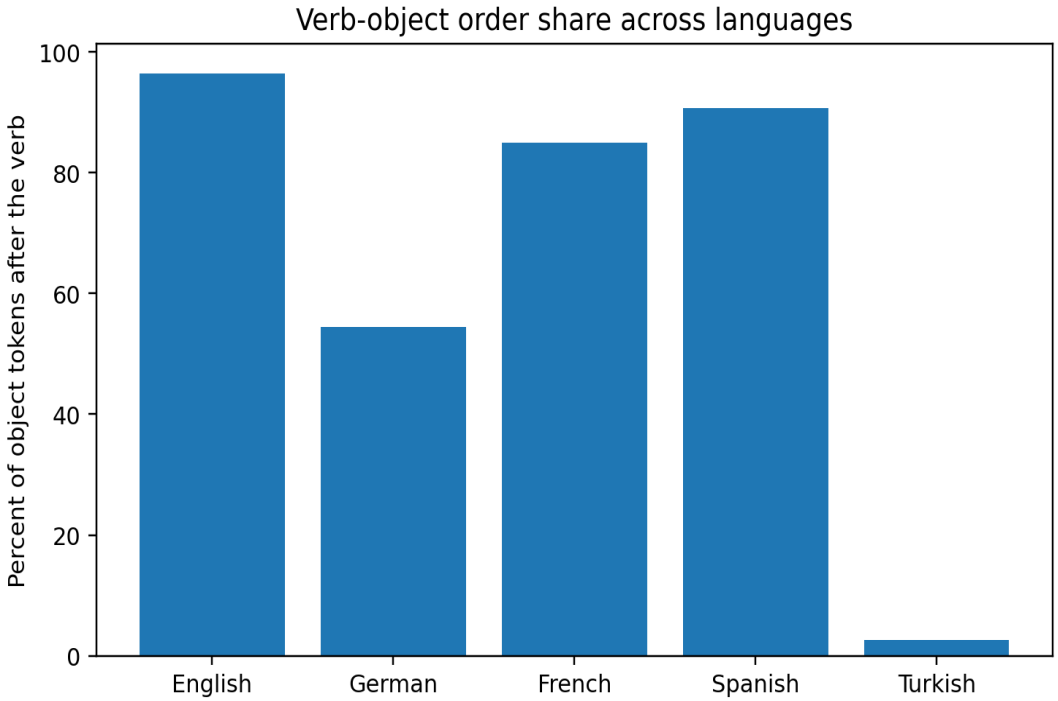

Figure 7 Verb-object order share across the five languages.

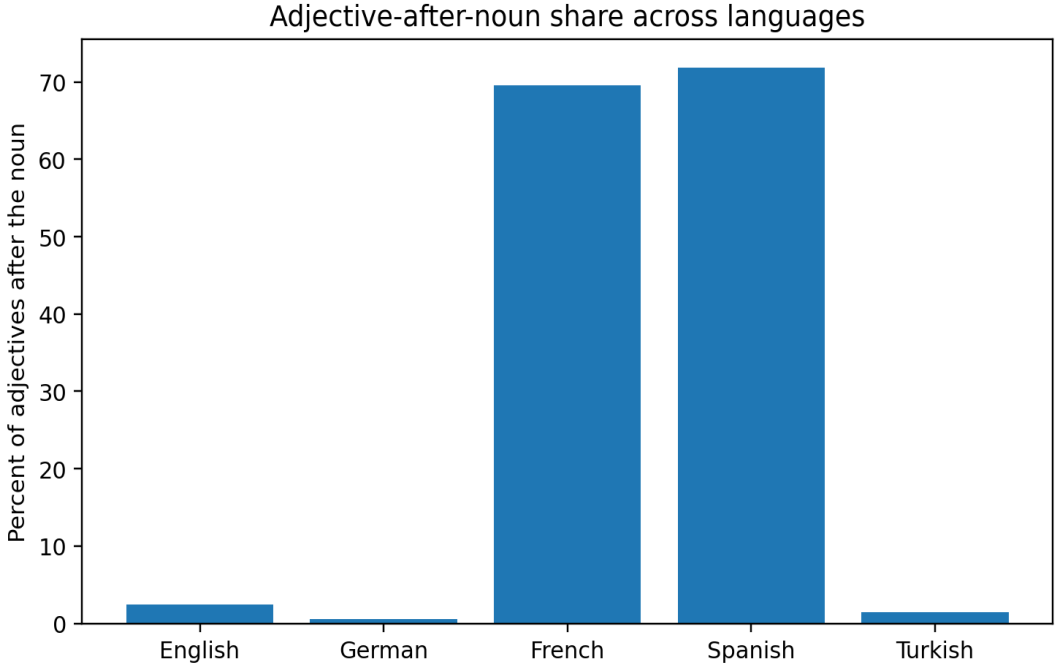

Figure 8 Adjective-after-noun share across the five languages.

Figure 8 shows adjective position. French and Spanish strongly favor adjectives after the noun in this sample, with 69.6 and 71.9 percent. English, German, and Turkish keep postnominal adjectives rare. The contrast is clear and stable. It is hard to treat such facts as marginal variation. They are part of how noun phrases are built in use. They also show how constructional preferences can remain strong across large amounts of data.

The within-language genre test adds a second kind of evidence. Figure 9 plots pronoun subject share against modal density in the English Web Treebank. Answers and reviews show very high pronoun subject rates, both above 71 percent. Email is close behind at 67.0 percent. Newsgroup texts sit lower at 47.6 percent. Weblogs show the lowest value, at 40.2 percent. Modal density follows a similar path. Answers and email have the highest rates. Weblogs again sit lowest. Table 4 gives the exact counts.

Table 4 Genre measures from the English Web Treebank sample.

Genre	Sentences	Tokens	Pronoun subjects (%)	Modals per 1,000
answers	2631	43480	71.4	23.9
email	3770	46255	67.0	22.4
newsgroup	1833	34978	47.6	13.9
reviews	2725	44811	71.3	14.3
weblog	1585	35053	40.2	9.9

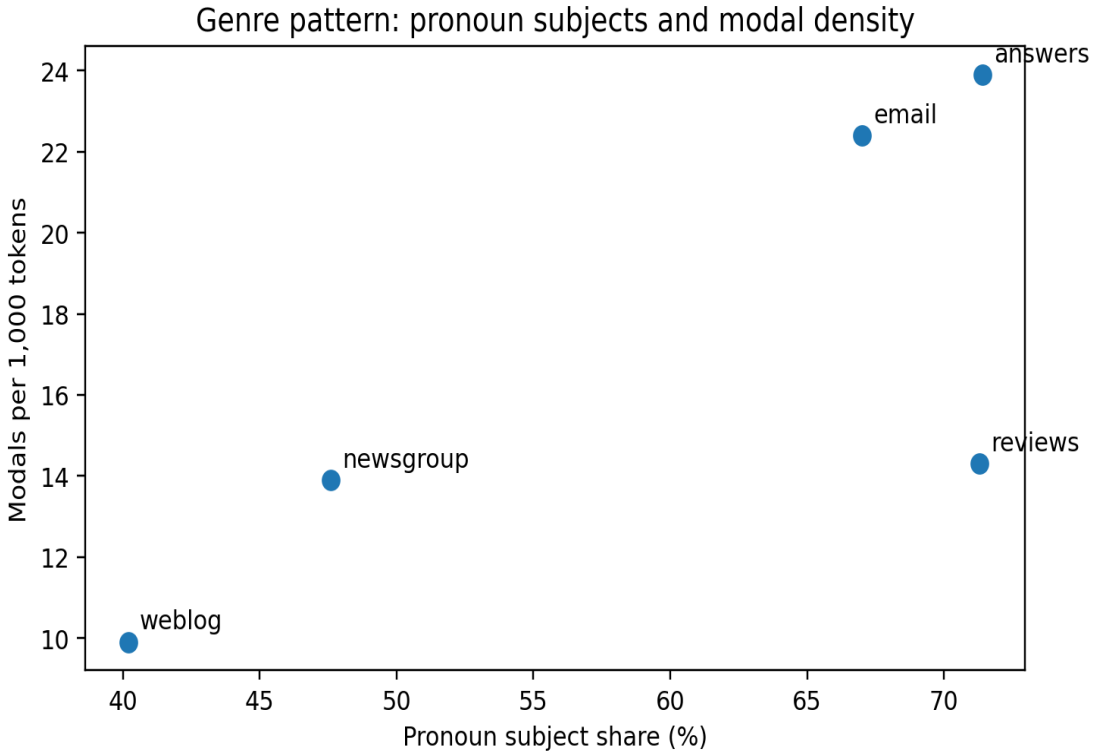

Figure 9 Genre pattern in English EWT: pronoun subject share and modal density.

This genre pattern matters for theory. The language is the same, but the grammar profile shifts with communicative task. Interactive writing invites direct reference to speaker and addressee. It also invites stance marking, advice, and prediction. Those functions raise pronouns and modals. Public commentary in weblogs relies less on that mix. The result is a different

grammatical texture. A rigid rule model can list all these forms. A usage-based model can explain why they cluster as they do.

The results therefore support two claims at once. Cross-linguistic contrasts are strong and orderly. Genre contrasts inside one language are also strong and orderly. Both kinds of evidence fit the same idea. Grammar is a network of learned routines with graded strengths. Variation is what that network looks like when it meets different histories of use.

Discussion

The findings support the paper's main claim. Variation is not outside grammar. It is one of the clearest windows into grammar. The cross-linguistic contrasts are too regular to be noise. The genre contrasts are too systematic to be free style. They show that grammar stores preferred solutions to recurrent communicative problems.

A usage-based account explains these results without forcing them into a rigid mold. Speakers store many exemplars. They also generalize over them. In that process, some options become dominant. Others remain possible but weaker. This is why a grammar can be stable and variable at the same time. Stability comes from repeated success. Variation comes from overlapping routines, discourse needs, and competing analogies (Bybee, 2006; Goldberg, 2006).

The English and Spanish contrast in overt subjects is especially revealing. A categorical description can say that Spanish allows null subjects. That statement is true, but it is thin. It does not explain when overt pronouns still appear, why they rise in some contexts, or how speakers learn those distributions. A usage-based approach can say more. It can tie overt subject choice to discourse prominence, contrast, recovery, and repeated local routines. Grammar then becomes a map of experienced probabilities, not a bare permission list.

The same logic applies to auxiliaries. English relies on separate auxiliaries for many clause functions. Turkish often packs such information into morphology. The difference becomes a difference in routine packaging. A usage-based model expects this. It treats grammar as a system of form-meaning pairings that have grown strong through repeated use. Different languages solve similar tasks with different recurrent packages. Those packages then shape expectation, speed, and acceptability.

Word order patterns also fit the theory well. High-frequency orders become easy to process and easy to predict. This does not erase alternatives. German shows that clearly. Yet the alternatives do not float without structure. They are tied to clause type, discourse design, and learned distribution. Usage-based theory can model this layered pattern better than a flat rule plus exception list (Hilpert, 2014; Perek, 2015).

The genre findings are equally important. They show that variation is not only a typological issue. It is also a local issue inside one language. The same grammatical resources are weighted differently in email, answers, reviews, newsgroups, and weblogs. This is what we should expect if grammar is shaped by recurrent tasks. Some tasks call for direct stance. Some call for compressed exposition. The grammar profile shifts because repeated use shifts what is easy, likely, and conventional.

A further point concerns theory building. Open data matter. When theory is tied to public corpora, readers can test the argument. They can inspect the files, recompute the counts, and extend the design. This is healthy for linguistics. It reduces the gap between abstract claims and visible evidence. It also makes classroom and early-stage research more robust.

The findings also connect to change. If grammar is built from repeated local choices, then small frequency skews can matter greatly over time. A genre may favor a variant first. The variant may then spread into nearby settings. Such paths are well known in grammaticalization and constructional change (Croft, 2000; Diessel, 2007). Variation is therefore not only a snapshot of grammar. It is also a clue to future change.

One strength of the usage-based view is that it links micro choice to macro pattern. A single utterance is small. Yet a long series of similar utterances can bend the system. Speakers then meet the bent system as ordinary grammar. This bridge from event to structure is hard to capture in approaches that separate grammar from use too sharply. It is easier to capture in a model where storage and abstraction work together. The corpus patterns in this paper are useful because they show both scales at once. They show broad language differences and small genre shifts inside one language.

The findings also speak to the status of optionality. Linguists often describe two forms as optional when both are allowed. That label can be helpful, but it can also hide the core issue. Very few options are truly free. One form is often easier, more frequent, or better matched to the discourse task. A usage-based approach asks what gives one form that edge. It looks for repeated alignment between form, meaning, and context. In that sense, optionality is often a surface label for deeper probabilistic organization. The present data support that reading.

The paper also has value for comparative grammar. Cross-linguistic comparison is often built from categorical claims such as null-subject language, verb-object language, or adjective-noun language. Those labels are useful first steps. Yet the corpus results show that each label covers a range of strengths. German does not behave like English on object position, but it also does not behave like Turkish. French patterns with Spanish on adjective position, but not on overt subject use. These graded relations are exactly what a usage-based model predicts. Languages share abstract pressures, but they resolve them through different use histories.

Another implication concerns pedagogy and writing support. Grammar teaching often presents one default form and treats alternatives as exceptions. That practice may help beginners, but it can also give a false picture. Real language use involves weighted choices. Learners need to know not only what is possible, but also what is common in a genre, what sounds direct, and what sounds dense or distant. A usage-based account can support that aim because it links grammatical choice to actual distribution. Public corpora make this especially useful in teaching, editing, and genre awareness.

The study supports a methodological shift in theoretical work. Strong theory should not fear simple counts. Basic distributional measures can reveal deep structural facts. They do not replace close analysis, but they sharpen it. When a theory claims that grammar is shaped by repeated use, corpus patterns become part of the proof. When a theory claims that variation is external, those same patterns become a problem. The present paper therefore argues for a balanced practice. Linguistic theory should keep conceptual depth, but it should also stay close to open and reproducible evidence.

None of this means that all grammar can be reduced to raw counting. Usage-based theory is not a simple frequency doctrine. Frequency matters because it affects memory, expectation, and analogy. Meaning, discourse role, morphology, and social indexical value matter too. The value of the present study lies in showing that these factors produce stable distributions. Grammar must be broad enough to contain them.

Limitations and future directions

The study has clear limits. It uses five treebanks and a small set of measures. A larger sample would add more languages, more genres, and more fine-grained predictors. The measures also stay at a broad level. They do not separate all clause types or all discourse functions. Some results, especially for Turkish subject position, likely reflect corpus composition as well as grammar.

Future work should extend the model in three ways. It should add more languages. It should test richer genre contrasts. It should also use multivariate models that combine morphology, information structure, and lexical preference. Such work would deepen the theoretical claim.

It would show more exactly how local use histories feed into constructional strength and variable choice (Bresnan & Ford, 2010; Zeldes, 2017).
Even with these limits, the present findings are enough for the paper's argument. The distributions are orderly. They reflect repeated language use. They fit a view of grammar as a learned probabilistic network. That is the main point.

Conclusion

Variation in grammar should be treated as evidence, not residue. It shows how grammar is organized, learned, and changed. A usage-based approach explains this with exemplar storage, chunking, analogy, and gradient constructional strength. The open corpus experiment supports that view. Pronoun subjects, auxiliaries, word order, and adjective position all vary in patterned ways. Genre inside English also reshapes these choices in regular ways.
Theoretical linguistics gains from taking this evidence seriously. Grammar is not weaker when it contains variation. It is more realistic. A strong theory should explain both what speakers can say and what they usually say. It should also explain why those tendencies shift across settings and across time. A usage-based model can do this. For that reason, variation is not a marginal issue. It is one of the main places where grammatical theory proves its worth.

Compliance with ethical standards

Disclosure of conflict of interest

The authors declare that they have no conflict of interest.

References

[1] Bresnan, J., & Ford, M. (2010). Predicting syntax: Processing dative constructions in American and Australian varieties of English. Language, 86(1), 168–213.
[2] Bybee, J. (2006). From usage to grammar: The mind's response to repetition. Language, 82(4), 711–733. https://doi.org/10.1353/lan.2006.0186
[3] Croft, W. (2000). Explaining language change: An evolutionary approach. Longman.
[4] Diessel, H. (2007). Frequency effects in language acquisition, language use, and diachronic change. New Ideas in Psychology, 25(2), 108–127. https://doi.org/10.1016/j.newideapsych.2007.02.002
[5] Ellis, N. C. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. Studies in Second Language Acquisition, 24(2), 143–188. https://doi.org/10.1017/S0272263102002024
[6] Goldberg, A. E. (2006). Constructions at work: The nature of generalization in language. Oxford University Press.
[7] Hilpert, M. (2014). Construction grammar and its application to English. Edinburgh University Press.
[8] Hopper, P. J. (1987). Emergent grammar. Proceedings of the Thirteenth Annual Meeting of the Berkeley Linguistics Society, 13, 139–157. https://doi.org/10.3765/bls.v13i0.1834
[9] Labov, W. (1969). Contraction, deletion, and inherent variability of the English copula. Language, 45(4), 715–762. https://doi.org/10.2307/412333
[10] Langacker, R. W. (1987). Foundations of cognitive grammar: Volume 1. Stanford University Press.
[11] Nivre, J., de Marneffe, M.-C., Ginter, F., Hajič, J., Manning, C. D., Pyysalo, S., Schuster, S., Tyers, F., & Zeman, D. (2020). Universal Dependencies v2: An evergrowing

multilingual treebank collection. In Proceedings of the Twelfth Language Resources and Evaluation Conference (pp. 4034–4043). European Language Resources Association.

[12] Perek, F. (2015). Argument structure in usage-based construction grammar: Experimental and corpus-based perspectives. John Benjamins Publishing Company.

[13] Sulubacak, U., Gokirmak, M., Tyers, F., Çöltekin, Ç., Nivre, J., & Eryiğit, G. (2016). Universal Dependencies for Turkish. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers (pp. 3444–3454). The COLING 2016 Organizing Committee.

[14] Tomasello, M. (2003). Constructing a language: A usage-based theory of language acquisition. Harvard University Press.

[15] Zeldes, A. (2017). The GUM corpus: Creating multilayer resources in the classroom. Language Resources and Evaluation, 51(3), 581–612. https://doi.org/10.1007/s10579-016-9343-x

