

Three Frequency Effects in Syntax

Author(s): Joan Bybee and Sandra Thompson

Proceedings of the Twenty-Third Annual Meeting of the Berkeley Linguistics Society: General Session and Parasession on Pragmatics and Grammatical Structure (1997), pp. 378-388

Please see “How to cite” in the online sidebar for full citation information.

Please contact BLS regarding any further use of this work. BLS retains copyright for both print and screen forms of the publication. BLS may be contacted via <http://linguistics.berkeley.edu/bls/>.

The Annual Proceedings of the Berkeley Linguistics Society is published online via [eLanguage](#), the Linguistic Society of America's digital publishing platform.

Three Frequency Effects in Syntax

Joan Bybee and Sandra Thompson

University of New Mexico and University of California, Santa Barbara

The primary hypothesis of functionalist or usage-based linguistics is that language use shapes grammar. The recognized mechanism by which this language-shaping process occurs is frequent repetition (Givón 1979, DuBois 1985, Hopper 1987, Hopper and Thompson 1993). In this paper we will look at the three types of repetition or frequency effects that have been identified as important in phonology and morphology, and demonstrate that these same effects are operative in the shaping of the grammar of larger units, that is, in syntax.

First it is important to distinguish two kinds of frequency: token frequency and type frequency. Token frequency is the count of the occurrence in texts of particular words, such as *broken*, or *have*, or of specific phrases, such as *I don't think*. Type frequency, on the other hand, counts how many different lexical items a certain pattern or construction is applicable to. Using a morphological example, regular English Past Tense *-ed* has a very high type frequency because it applies to thousands of different verbs. The vowel-change pattern exemplified by *strung* and *stung* has a much lower type frequency since it applies to fewer than twenty verbs. Type frequency in syntactic constructions would count how many distinct items of a particular lexical or grammatical class (e.g. verbs) can be used in the construction.

In this paper we discuss two effects of token frequency, and show how these effects interact with a third effect, which is due to type frequency. The three effects we have chosen to discuss are not the only frequency effects important in shaping grammar, but we have chosen these three for a specific reason: the Reduction Effect and the Conserving Effect of high token frequency appear to condition opposite results, i.e. on the one hand, high token frequency promotes change and on the other hand it renders constructions resistant to change. We demonstrate that these two effects of high token frequency are applicable at different stages in the life of a construction, and the two types of change involved are very different in nature. The Reduction Effect plays a major role in grammaticization, where high token frequency promotes all types of reductive change, while the type of change that is resisted by old, highly entrenched constructions is change on the basis of productive patterns with high type frequency.

1. The Reduction Effect

First we consider the Reduction Effect of high token frequency, which has a phonetic, a syntactic, and a semantic dimension. Phonetic change often progresses more quickly in items with high token frequency. This effect is particularly noticeable in grammaticizing elements or phrases which undergo

drastic reduction as they increase in frequency. Thus *be going to*, which is becoming a future marker in English, is reduced to [ˈgʌnə] or even further reduced in phrases such as *I'm gonna* to [ˈaɪmənə]. Similarly the conventionalized contractions of English are reduced due to their high frequency: *I'm, I'll, I've, can't, don't, won't*, etc. But the effect occurs on a more subtle level as well: regular sound change in many cases progresses more quickly in items of high token frequency. For instance, there is a tendency in American English for the loss of syllabicity in post-stress schwa + resonant sequences, as in *every, camera, memory, family*. This reduction is more advanced in words of higher frequency (such as those just named) than in words of lower frequency, such as *mammary, artillery, homily* (Hooper 1976). The loss of final [t] or [d] after a consonant is also more common in words of higher frequency, such as *went, just, and*; in fact a general effect of token frequency on the rate of deletion has been found for 2000 tokens of final [t] or [d] (Bybee 1997).

If sound changes are the result of phonetic processes that apply in real time as words are used, those words that are used more often will have more opportunity to be affected by phonetic processes. If representations are changed gradually, with each token of use having a potential effect on representation, then words of high frequency will change at a faster rate than words of low frequency. The streamlining of high frequency words and phrases has the effect of automating production. Any motor activity that is repeated often will become more efficient (Bybee 1997).

The automation of processing has effects well beyond phonetics: the processing of frequently-used sequences as single chunks also leads to the loss of internal constituent structure. This effect is observable in grammaticization (Heine and Reh 1984, Heine, Claudi and Hünemeyer 1991). Consider, for example, the grammaticizing phrase (*be*) *supposed to*, which is reduced to [spostə]. It comes from a passive construction with an infinitive, as shown in (1), but now the *to* of the infinitive is phonologically fused with the verb, as in the case of other developing auxiliaries, such as *have to* [hæftə], *want to* [wʌnə], *going to* [gəʊnə] and so on, perhaps indicating a change in the affiliation of *to*. Additionally, in (*be*) *supposed to* the syntactic status as a passive has also been lost, as shown by the unacceptability of a passive agent with the reduced form:

- (1) He is supposed by most people to be very knowledgeable.
- (2) He's s'posed to be very knowledgeable *by most people.

Where formerly this phrase was an instantiation of a construction involving *be* + *Past Participle*, functioning as a main verb with an embedded clause complement expressed as an infinitive, it now has been fused into a single auxiliary, *s'posta*, which is structurally parallel to *gonna, wanna*, etc.

Accompanying phonological reduction and loss of internal structure of high frequency phrases is the well-known phenomenon of semantic bleaching. Bleaching is due to habituation whereby organisms cease to respond at the same level to repeated stimuli (Haiman 1994). Words or phrases that are much repeated lose their semantic force, which, in a spiraling effect, allows them to occur more often, which in turn conditions further semantic bleaching.

Some of the clearest examples of bleaching are cases in which formerly emphatic expressions lose their special status as emphatics and become the unmarked way to express certain concepts. For example, it is common for negative markers to be reinforced by the addition of emphatic elements, such as other negators combined with nouns, as in French *ne ... pas* 'not a step' or English *ne + a + wiht* 'not at all' which gave rise to *not*. Traugott (1972) reports that in Old English *noht* always had its full emphatic force, but in Middle English it gradually lost this force and came to be used as the normal, non-emphatic negation.

Phonetic reduction, loss of internal constituency, and bleaching all occur together in grammaticization because they are all promoted by frequent repetition, which is the mechanism that produces automation and habituation.

2. The Conserving Effect

The second effect of the repetition of tokens is due to the increased lexical strength (Bybee 1985) or entrenchment (Langacker 1987) of a particular word or phrase. Using a morphological example again, high frequency forms with alternations resist analogical leveling: while English *weep* / *wept*, *creep* / *crept*, and *leap* / *leapt* have a tendency to regularize to *weeped*, *creeped*, and *leaped* respectively, the high frequency verbs with the same pattern, *keep* / *kept*, *sleep* / *slept* show no such tendency (Hooper 1976, Bybee 1985). As a result, morphological irregularity is always centered in the high frequency items of a language.

In Bybee (1985), this conservative behavior of high frequency forms is related to the faster lexical access of high frequency forms: the more a form is used, the more its representation is strengthened, making it easier to access the next time. Words that are strong in memory and easy to access are not likely to be replaced by new forms created with the regular pattern. This effect of frequency on the strength of representation in memory we dub The Conserving Effect.

This conservatism of much-used expressions can also be found on the syntactic level: it has often been observed that pronouns show more conservative behavior than full noun phrases (Givón 1979). Pronouns and full NPs are related in the sense that pronouns diachronically derive from nouns, and synchronically, in that pronouns and NPs often occupy the same positions. However, a major difference between nouns and pronouns is that the latter are much more frequent than the former. This fact can be used as an explanation for why English pronouns

maintain distinct forms for nominative vs. dative/accusative case, while nouns have lost these case distinctions. It can also be used to explain why the position of pronouns sometimes reflects an earlier word order, as in Spanish, where object clitic pronouns occur before the verb (reflecting, it is argued, an older OV word order) while full NP objects occur after the verb (Givón 1979).

Similarly, verbal auxiliaries, which are highly frequent, often retain old conservative syntactic characteristics. The English auxiliaries, for instance, retain the ability to invert with the subject (which all verbs had previously, e.g. [3]) and they are followed, rather than preceded by, the negative, another property once shared by all verbs, as shown in (4) and (5) from Middle English (Mossé 1952).

(3) Gaf ye the chyld any thyng?
'Did you give the child anything?'

(4) My wyfe rose nott.

(5) cry not so

This, then, is the Conserving Effect: the idea that high frequency sequences take on a life of their own, and resist change on the basis of newer productive patterns for juxtaposing words and morphemes.

3. The Subjunctive in Canadian French

As an example of the conserving role of token frequency in syntax we take the example of the use of the Subjunctive mood in Canadian French, as studied in Poplack (1992, 1995). We choose this example because Poplack's study is based on a corpus of three and a half million words of naturally occurring spoken language from 120 adult native speakers, and because the use of the Subjunctive involves long-distance dependencies of main verb on complement verb, a canonical syntactic issue.

In general, subjunctive verb forms tend to be very old forms which have come to be used in subordinate clauses after a long history of main clause use as futures or other indicative forms (Bybee et al. 1994). So in Canadian French, the meaning and form of the Subjunctive have been highly eroded. Indeed, we argue, along with Poplack, that the Subjunctive has basically been lost, but residue remains in the most frequent contexts, with some indications of minor productivity.¹ The study we cite here shows that Subjunctive verb forms now occur only in the most frequent syntactic contexts and with the most frequent verbs.

Poplack's study focuses on noun clauses embedded as complements to certain matrix verbs. In her corpus, Poplack identified 6000 sentences with a matrix verb governing the Subjunctive at least once. It is important to note,

however, that one factor leading to the demise of the Subjunctive/Indicative distinction is the fact that for most verbs there is no phonological distinction between mood forms. So in this set of sentences, about half of the embedded verbs were ambiguous between Subjunctive and Indicative. This left 2694 instances in which Subjunctive and Indicative usage could be distinguished. It is particularly important for our point here to note that the verbs that do maintain a formal contrast between Indicative and Subjunctive are irregular and among the most frequent verbs of the language. That is, the maintenance of the mood distinction in a subset of verbs is at least in part due to the Conserving Effect of high token frequency.

In the sentences to be analyzed, then, the main verb is one which was used with the Subjunctive at least once and the embedded verb was one which distinguishes mood formally. In these sentences the Subjunctive was used 77% of the time. The goal of Poplack's study was to determine what factors prevent the Subjunctive from occurring in these positions 100% of the time.

A statistical analysis of a number of factors led Poplack to conclude that the Indicative/Subjunctive distinction is not performing any particular functional or semantic work. This conclusion is supported by examples such as those in (6) and (7) in which the same speaker repeats essentially the same message to the same interlocutor but alternates between Indicative and Subjunctive:

- (6) a. Faut que je lui *dis* (I) c'est vrai.
'I have to tell him it's true'.
- b. Faut je lui *dise* (S) c'est la vérité.
'I have to tell him it's the truth'.
- (7) Fallait qu'elle *répond* (I) 'oui, tu peux faire trois pas de géant.' Fallait qu'elle *réponde* (S) la phrase complète.
'She had to say "yes, you may take three giant steps." She had to say the whole sentence'.

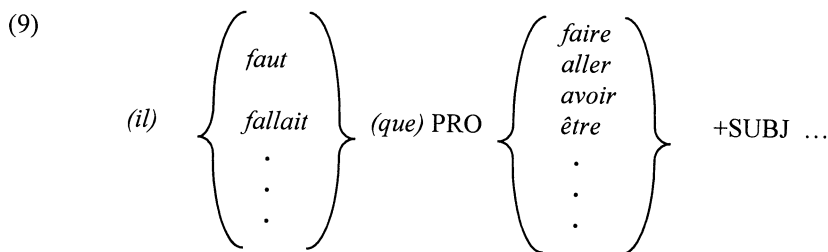
The critical question here is: if there is no functional difference in mood choice, why are French speakers still using Subjunctive verb forms? The evidence suggests that the answer lies in the fact that most of the Subjunctive forms occur in certain highly entrenched phrases with particular matrix verbs and particular embedded verbs. Thus, as we would predict from the Conserving Effect, these high-frequency expressions have maintained their traditional form despite general changes which allow the construction of sentences with Indicative verb forms in comparable, but less frequent contexts.

The most commonly occurring matrix verb is impersonal *falloir* 'have to', which accounts for 62% of the 2694 matrices, and is followed by a Subjunctive verb form in 89% of the cases. The main verb *falloir* is always impersonal,

meaning that if the subject is expressed, third person masculine singular *il* is always used, but different tense/aspect forms can also be expressed, such as Present *il faut* and Imperfect *il fallait*. The embedded verbs that occur most frequently in the Subjunctive are high frequency irregular verbs. In fact, only ten verbs account for two-thirds of the examples with Subjunctive, among these are *avoir* 'to have', *être* 'to be', *aller* 'to go', *faire* 'to make, do', etc. See the examples in (8).

- (8) a. Même pour un job aujourd'hui, faut tu *sois* (S) bilingue.
'Even for a job these days, you have to be bilingual'.
- b. Bien certain, faut qu'ils *aient* (S) une place eux-autres aussi pour vivre.
'Well, of course, they should have a place to live, too'.
- c. Faut j'*aille* (S) voir pour de l'ouvrage.
'I have to go look for a job'.
- d. Bien ça, fallait tu *fasses* (S) ton huit heures par jour.
'Well, there you had to do your eight hours a day'.

These facts suggest that these main verb - complement constructions are not generated from highly generalized syntactic schemas of the form [verb [S]], but rather that very specific constructions ('routines' in Poplack's terms), with some lexical items indicated, are stored and accessed in production, as shown in (9).



Further support for the position that constructions, complete with very specific lexical items, are accessed in these cases comes from the second factor that Poplack found to be significant, the distance factor. That is, if a word or some parenthetical material intervened between the main verb and the subordinate one, it was more likely that the Indicative form would be used. Since we are claiming entrenchment of the sequences in storage, much as though they were lexical items

rather than hierarchical syntactic structures, it is significant that intervening material favors the productive form, the Indicative. That is, if the speaker gets derailed from an automated sequence such as *Il faut que ...* then s/he is less likely to resume with the routinized form and more likely to access the more generally used Indicative form.

The importance of particular lexical items is also evident in the other 38% of the matrix verbs. Two verbs, *vouloir* 'to want' and *aimer* 'to like', make up 11% of the remaining cases and they show a high percentage of Subjunctive usage (91% and 67% respectively) (Poplack 1995). With these verbs, too, the irregular embedded verbs favor Subjunctive use.

Our point in this section is that distributions that appear to be very arbitrary should be checked for this conserving effect of token frequency. While grammatical analysis should proceed with the working hypothesis that formal distinctions represent functional distinctions, we also have to bear in mind that not all contrasts and distributions are meaningful or functional. Some patterns represent a lexically-arbitrary residue of formerly productive patterns.

4. Type Frequency

The type of change that is resisted by words or phrases of high token frequency is change on the basis of combinatorial patterns or constructions that are productive. To return to a morphological example, high frequency English Strong Verbs resist regularization by the productive suffixation pattern. But frequency also plays a role in the determination of productivity, where productivity is defined as the likelihood that a pattern will apply to new forms. However, in this case it is type frequency: the type frequency of a pattern determines its degree of productivity (MacWhinney 1978, Bybee 1985, 1995).

It is easy to see why type frequency determines productivity: type frequency refers to the number of distinct lexical items that can be substituted in a given slot in a construction, whether it is a word-level construction for inflection, or syntactic construction specifying the relation among words. The more lexical items that are heard in a certain position in a construction, the less likely it is that the construction will be associated with a particular lexical item and the more likely it is that a general category will be formed over the items that occur in that position. The more items the category must cover, the more general will be its criterial features and the more likely it will be to extend to new items. Furthermore, high type frequency ensures that a construction will be used frequently, which will strengthen its representational schema, making it more accessible for further use, possibly with new items.

As type frequency can range from one to a very large number, so there are varying degrees of productivity associated with ranges of type frequency. The relationship between type frequency and degree of productivity among English Strong Verbs has been shown by Bybee and Moder (1983) and Moder (1992).²

The most productive Strong Verb class, exemplified by *strung*, is the only class that has added new members since the Old English period, and also has the highest type frequency of any Strong Verb class. As a result, we expect to find a relation between the existing type frequency of a syntactic construction and its ability to apply to novel lexical items.

Some of the most difficult problems in syntax concern the competition among functionally similar constructions with varying degrees of productivity due to type frequency. For instance Goldberg (1995) argues that type frequency affects the productivity of the ditransitive construction (as in [10]). This construction is the continuation of an old construction that is gradually being replaced by the construction in which the dative is marked with the preposition *to* (as in [12]). In the ditransitive construction, which was common in Middle English, the indirect object noun phrase was formerly marked with the dative suffix, which explains why no preposition is present today. The ditransitive construction has some of the features of an archaic construction preserved through high frequency; in particular, it occurs only with specific lexical verbs, and most of these are of Germanic origin and of very high frequency. Thus, for instance, as has often been observed, this construction is not used with a verb such as *whisper*, as in (11). The prepositional construction in (12), on the other hand, is perfectly general, occurring with all verbs that can take a patient and recipient argument. Its high type frequency gives it full productivity, allowing it to be used productively, even with nonce forms (Gropen et al. 1989).

(10) He told the woman the news.

(11) *He whispered the woman the news.

(12) He whispered the news to the woman.

However, this case is made complex by the fact that the ditransitive construction does have some limited productivity, as evidenced by its occurrence with certain new verbs such as *to telephone*, *to e-mail* and *to fed-ex*, (see [15]). Goldberg (1995, Chapter 5) points out that this limited productivity corresponds to the type frequency of the semantic verb classes that occur in the ditransitive construction. Classes with large membership are able to attract new members to a limited extent while classes with smaller membership do not appear to attract new members. For example, a small class of verbs that can be found in the ditransitive construction is the class of verbs of permission. Only a very few verbs of this class can occur in the ditransitive, as illustrated in (13) and (14):

(13) Sally permitted/allowed Bob one kiss.

(14) *Sally let/enabled Bob one kiss.

Larger classes of verbs that can be used in the ditransitive construction are verbs of sending and communicating. This high type frequency is what allows the construction to be used with new verbs with related meanings, as in (15).³

(15) Sally telephoned / e-mailed / fed-exed Sam the offer.

The literature on the ditransitive construction shows that some arbitrariness must be recognized in the categorization of verbs according to whether or not they are permitted in this construction. On the other hand, certain generalizations have also been shown to be valid: that verbs indicating eventual possession of the object can be used in the ditransitive construction and that monosyllabic verbs are favored over polysyllabic verbs (Gropen et al. 1989). The existence of exceptions, arbitrariness, and non-categorical tendencies in productivity are properties of lexical classes. In fact, these are the properties found with morphological classes, such as the classes of English irregular verbs. It is not surprising, then, that type frequency in these classes affects productivity, just as it does in morphological classes.

5. Conclusion

Recognizing the two effects of token frequency on syntax and their interaction with type frequency over time allows us to explain many situations in language that otherwise appear very arbitrary. Not only can frequency be viewed as an integral part of an explanation for these situations, the effects of frequency have important implications for our notions of mental representation. There is not necessarily just one representation per construction; rather, a specific instance of a construction, with specific lexical items in it, can have its own representation in memory if it is of high frequency. As a result, it can undergo phonological reduction, and change in meaning or function independently of other examples of the same construction. At a later stage, highly entrenched examples of constructions with particular lexical items can continue to be used even though new productive patterns have become current in the language.

We are thus recognizing that some constructions are relics from earlier stages of a language, but they are not just arbitrary historical residue, rather such structures are still subject to categorization and generalization. There are not just two choices: fossilization or productivity, but many intermediate possibilities, depending both on categorization in terms of either meaning or form and in the type frequency of the construction.

On a more general level, we hope to have supported and further articulated John Haiman's (1994) claim that grammatical constructions, as conventionalized patterns of language use, are created and maintained through the same mechanisms associated with repetition as are other patterns of human culture, as well as ritualistic behavior found in animals other than humans.

Notes

- ¹ Poplack's interpretations of descriptions of other dialects of French (including the standard) suggest that the same situation holds elsewhere.
- ² In these studies, other factors are also shown to interact with type frequency in determining productivity, i.e. phonological properties and high token frequency of individual types, which can detract from productivity. See also Bybee (1995).
- ³ Other properties of the two constructions, such as restrictions on the use of pronouns, appear to be less related to their relative age and frequency than to notions of topicality (Thompson 1990). In fact, this example shows that any two competing constructions can be affected by issues of semantics and pragmatics in addition to the complex interaction between token and type frequency.

References

- Bybee, Joan. 1985. *Morphology: a study of the relation between meaning and form*. Amsterdam: John Benjamins.
- Bybee, Joan. 1995. Regular morphology and the lexicon. *Language and Cognitive Processes* 10 (5).425-455.
- Bybee, Joan. 1997. The phonology of the lexicon: evidence from lexical diffusion. *Usage-based models of language*, ed. by M. Barlow and S. Kemmer, Stanford: CSLI.
- Bybee, Joan L., and Carol Lynn Moder. 1983. Morphological classes as natural categories. *Language* 59.251-270.
- Bybee, Joan, Revere Perkins, and William Pagliuca. 1994. *The evolution of grammar: tense, aspect and modality in the languages of the world*. Chicago: University of Chicago Press.
- DuBois, John. 1985. Competing motivations. *Iconicity in syntax*, ed. by John Haiman, 343-365. Amsterdam: John Benjamins.
- Givón, Talmy. 1979. *On understanding grammar*. San Diego: Academic Press.
- Goldberg, Adele. 1995. *Constructions: a construction grammar approach to argument structure*. Chicago: University of Chicago Press.
- Gropen, Jess, Steven Pinker, Michelle Hollander, Richard Goldberg, and Ronald Wilson. 1989. The learnability and acquisition of the dative alternation in English. *Language* 65(2).203-257.
- Haiman, John. 1994. Ritualization and the development of language. *Perspectives on grammaticalization*, ed. by William Pagliuca, 3-28. Amsterdam: John Benjamins.
- Heine, Bernd, Ulrike Claudi, and Friederike Hünemeyer. 1991. *Grammaticalization: a conceptual framework*. Chicago: University of Chicago Press.
- Heine, Bernd, and Mechtild Reh. 1984. *Grammaticalization and reanalysis in African languages*. Hamburg: Helmut Buske Verlag.

- Hooper, Joan B. 1976. Word frequency in lexical diffusion and the source of morphophonological change. *Current progress in historical linguistics*, ed. by William Christie, 96-105. Amsterdam: North Holland.
- Hopper, Paul J. 1987. Emergent grammar. *Berkeley Linguistics Society* 13.139-157.
- Hopper, Paul J., and Sandra A. Thompson. 1993. Language universals, discourse, pragmatics, and semantics. *Language Sciences* 15(4).357-376.
- Langacker, Ronald. 1987. *Foundations of cognitive grammar, Vol. 1. Theoretical prerequisites*. Stanford: Stanford University Press.
- MacWhinney, Brian. 1978. *The acquisition of morphophonology*. Monographs of the Society for Research in Child Development, 43 (1).
- Moder, Carol Lynn. 1992. Productivity and categorization in morphological classes. Buffalo: SUNY at Buffalo Dissertation.
- Mossé, Fernand. 1952. *A handbook of Middle English*. Baltimore: The Johns Hopkins Press.
- Poplack, Shana. 1992. The inherent variability of the French subjunctive. *Theoretical analyses in Romance linguistics*, ed. by Christiane Laeufer and Terrell A. Morgan, 235-263. Amsterdam: John Benjamins.
- Poplack, Shana. 1995. The sociolinguistic dynamics of apparent convergence. *Towards a social science of language: papers in honor of William Labov*, ed. by G. Guy, J. Baugh and D. Schiffrin, 285-309. Amsterdam: John Benjamins.
- Thompson, Sandra. 1990. Information flow and dative shift in English discourse. *Development and diversity: variation across time and space*, ed. by Jerold Edmonson, Crawford Feagin and Peter Mühlhäusler, 239-253. Dallas, Texas: SIL.
- Traugott, Elizabeth C. 1972. *A history of English syntax*. New York: Holt Rinehart and Winston.