Title: What if it's lexicalization?

Author: Joanna Nykiel

Joanna Nykiel

What if it's lexicalization?[1]

1. Introduction

The term lexicalization is used to refer to the formation of identifiable units out of (more) complex constructions; more specifically, there is recognition that the process involves fusion (Traugott 1994, Wischer 2000, Blank 2001, Brinton and Traugott 2005). Brinton and Traugott argue that the resultant unit is not semantically transparent and gradually moves away from productive use, showing different degrees of lexicalization along the way. Hence, it comes as no surprise that there is a scalar quality about lexicalized units: they may range from free to inseparable sequences. The scope of lexicalization effects in English has been discussed in connection with the ordering of post-verbal constituents, where adjacency of a verb and preposition is a general preference if they are semantically dependent on each other (Wasow 2002, Hawkins 2004).

(1)   attribute to me a statement I didn't make
(2)   attribute a statement I didn't make to me[2]

Hawkins (2004:114) proposes a means of testing semantic dependence (Pro-verb Entailment Test) based on the possibility, or lack thereof, of using

---

a Pro-V *do something* in place of the original verb. For example, *work with someone* is replaceable by *do something with someone*, but *argue for something* is not replaceable by *do something for it*; hence, only the latter is a lexicalized unit. The generalization is, according to Hawkins (2004) and Wasow (2002), that all the environments in which a separation of a verb and preposition could occur are likely to show lexicalization effects. This raises the question of whether the same forces influence preposition doubling (P doubling) illustrated in (3–4).

(3)   That doesn't have anything to do with the fact that I did something I shouldn't have done, **of** which I am ashamed **of**, and which I apologize for.

(4)   Or Virginia, depending **on** which side of the line you're standing **on**, right? (*Public Papers of the Presidents of the United States*)

Such doubling is mostly associated with *wh*-phrases embedded in relative- (3) and interrogative-clause (4) environments. Both these uses tie in with the possibility of preposition stranding (*the world (which) we live in*) as an alternative to pied-piping (*the world in which we live*), and may stem from a variety of performance-based factors. Riley and Parker (1986) suggest a reference-tracking mechanism that inserts a preposition to mark case on a *wh*-phrase, while another preposition is stranded. This strategy could be part of the *Wh*-Processing Hypothesis due to Arnon et al. (2005)[3] in that the presence of a preposition increases the accessibility of an extracted *wh*-phrase. The strength of this argument derives from statistical preferences for P doubling in English relative clauses, whose *wh*-phrases are low in accessibility, because they may be bare but not complex (Nykiel, forthcoming). Compare bare *which* (5) and complex *which side of the line* (6).

(5)   ... the side of the line which you're standing on
(6)   Which side of the line are you standing on?

In the absence of other clues, a pied-piped preposition in (5) would limit the search for the gap (subcategorizer) evoked by the dislocated *wh*-

---

[3] On this theory, a *wh*-phrase is easier to process if it is more accessible, that is, more informative.

phrase.[4] Beyond these effects, I found that a preference for doubling corre-
lates with lexical units in Middle English (ME), when doubling first arises,
and in present-day English. For example, the most frequent ME construc-
tion *speak of* appears in 19% of the instances (Nykiel, forthcoming).

Reference tracking and lexicalization are not the only purposes for
P doubling, though. Staum and Sag's (2008) experimental data demon-
strate how locality can interact with P doubling: the longer the distance
between two prepositions, the more acceptable they become. Examples
(7–8) illustrate.

(7)   I asked **from** which teacher my son had gotten the bad grade **from** at
        the end of the quarter at the new school he attended.
(8)   I asked **from** which teacher at the new school he attended my son had
        gotten the bad grade **from** at the end of the quarter.

This result, if correct, predicts that the preference for lengthy interveners
should be reflected in corpus data as well, because experimental and corpus
evidence have been shown to parallel each other (Bresnan, in press). Anoth-
er prediction is that P doubling itself is a gradient phenomenon, with some
environments clearly improving its acceptability.

In this paper, I argue that among reference tracking, locality and lexi-
calization, it is lexicalization that provides the most adequate account of P
doubling which occurs in simple affirmative sentences, e.g. (9).

(9)   dominated by pro-government and pro-American officers and com-
        posed **of** mostl**y of** Christians and Sunnis (http://www.time.com/time/
        world/article/0,8599,1738748,00.html, 9/24 2008)

Strikingly, such constructions are not unique to preposition-strand-
ing or analytic languages like English. Section 2 examines data from Eng-
lish, German, Danish, and Polish, which all mark an intimate connection
between a verb and preposition. In section 3, I discuss the implications of
these results for theories of grammar. Section 4 concludes.

---

[4] In English, since a relative *wh*-phrase can be subcategorized for by verbs or preposi-
tions, a preceding preposition indicates that the former is the case (Hawkins 2004: 204).

2. The data

The data were collected by means of a Google search; therefore, a cautionary note is in order. Given a possibly unclear status of their sources, Google hits may not always accurately reflect syntactic patterns. If the pattern in question is, however, one that cannot be found in the extant corpora, the benefit of a Google search follows from lack of editorial intrusion. The results may thus more readily indicate speakers' spontaneous use of a particular construction, not least the kind that could be considered incorrect by a careful editor. This approach permits ballpark figures for the three-word English string *of mostly of*, where P doubling occurs spontaneously in discourses attributed to native speakers of English. The reason for selecting this pattern is that there is a particularly heavy concentration of constructions containing preposition *of*.

For the first 81 hits, I have given a list of the constructions that contain the pattern and their token frequencies (Table 1).[5]

Table 1. Constructions by token frequency

Construction	Frequency
be made (up) of	27
consist of	20
be composed of	13
be comprised of	9
a gang of	2
a community of	1
be full of	1
spiral crime of	1
be built of	1
be of … composition	1
graphics of	1
a collection of pictures of	1
a study of	1
be constructed of	1
be of … interest	1

---

[5] The total numbers are 3,770 (*of mostly of*) and 1,826 (*consist/be composed/comprised/ made (up) of mostly of*), googled on October 4, 2008.

It emerges that the most frequent construction is *to be made (up) of* along with its three synonyms.

(10) Made up **of** mostly **of** native American people, a large proportion of them directly from the smaller towns and open country, or only one generation removed, the city problem in this area has not yet become so compelling or so complex as of the industrial cities of the east. (http://www.kansasheritage.org/um/worldserv.html)

Beyond English, P doubling in this construction appears in German (11), Polish (12) and Danish (13), as various Google hits attest. The constructions googled on October 4, 2008 were: *składać się z głównie z* (Polish), *bestehen aus meistens aus* (German)[6], and *bestå af mest af* (Danish), together with all the inflected verb forms.

(11) Die Gruppe besteht **aus** meistens **aus** 5–10 Personen.
the group   consists of mostly   of 5–10 people
(www.skate4you-bremen.de/index.php)

(12) Kolekcja Tadeusza Szumańskiego   składa się **z** głównie
collection Tadeusz Szumański.GEN consists   of mainly
z fotografii z   lat   1940–1947.
of pictures   from the years 1940–47
(http://www.nac.gov.pl/cms/tekst.php)

(13) Bornholms økonomi består   **af** mest   **af** landbrug,
Bornholm's economy   consists of mostly of farming,
turisme og   en lille smule fiskeri.
tourism and a   little bit of fishing
(www.gaarslev-skole.dk/klasser/6b/frederik.html)

Importantly, among these languages, Polish and German are case-marking and non-preposition-stranding, while Danish is not. This means that it should be possible to establish what purpose there is for doubling.

In all four languages, the verbs require prepositions that are dependent on them, that is, the prepositions are specified rather than free in the sense

---

[6] The German construction googled includes adverbs *meistens* and *größtenteils*, both of which mean *mostly*.

of Huddleston and Pullum (2002).[7] By this logic, there is a semantic link between the verb and preposition, which indicates a certain degree of lexicalization. Based on the location of the first preposition in each example, it is possible to generalize that the verbs attract the prepositions. Note that, without doubling, a preposition follows *mostly*. Clearly, there is no effect of locality along the lines of Staum and Sag (2008), since the distance between any doubled prepositions is exactly one word. The next section discusses the consequences of these findings.

3. Discussion

These cross-linguistic results are highly consistent with the view that the products of lexicalization are semantic units. As units, they favor adjacency to the extent that parsing should be easier if the elements that make up a unit can be accessed over a minimal distance. Since *mostly*, and its equivalents, normally comes between a verb and preposition, it delays the processor's recognition of the lexical unit unless an extra preposition serves as a signal of it. It is clear that the second preposition appears in its default position. For English, the *made up mostly of* sequence yields 295,000 Google hits; post-prepositional *mostly*, though of course attested, has a different scope and reading. Hawkins (2004:103) operationizes this general preference for adjacency as a performance principle called Minimize Domains (MiD) given in (14) below.

(14)   The human processor prefers to minimize the connected sequences of linguistic forms and their conventionally associated syntactic and semantic properties in which relations of combination and/or dependency are processed. The degree of this preference is proportional to the number of relations whose domains can be minimized in competing sequences or structures, and to the extent of the minimization difference in each domain.

---

[7] Huddleston and Pullum (2002) draw a distinction between Ps that are subcategorized for by the verb (specified) and those that are not (free), as in (1–2).
    (1)   I flew to Boston (free P)
    (2)   I referred to her book (specified P)

If, as indicated by the data, P doubling is part of MiD, we have a way of accounting for its occurrence in all environments, and even in languages in which a ban on preposition stranding prevents doubling in relative and interrogative clauses. Although MiD may be at work in these clauses, it is overridden by the formal pied-piping mechanism which requires that a preposition precede an extracted *wh*-phrase.[8] In non-extractions, however, where the constituent order is V-PP, MiD predicts doubling.

These facts provide further support for how performance preferences may generate constructions that would not be generated by competence grammar. Nor do they have to be. On a theory that accommodates an interaction between performance and competence, the observed linguistic behavior can be explained by appeal to ease of processing. Lexicalization effects are also predicted if one realizes that human discourse processing proceeds in a stepwise fashion, with speakers forming hypotheses about what comes next. Such hypotheses are based, among other things, on lexical valence information (Sag and Wasow, to appear).

4. Conclusion

In this paper I show the possibility of, and in fact a need for applying lexicalization criteria to P doubling. Whenever doubling appears in non-extractions and over short distances, it indicates a cross-linguistic preference for adjacency of a verb and preposition which form a semantically connected sequence. The same effect has been shown to be operative in relative and interrogative clauses (Nykiel, forthcoming); in such extractions, however, other factors are also involved, for example, locality and *wh*-processing (Hawkins 2004, Arnon et al 2005, Staum and Sag

---

[8] It seems that the greater the degree of dependency between a verb and preposition, the more likely their adjacency. For example, in Polish, which disallows preposition stranding, idioms can strand prepositions.

(1)   Byłem pierwszym człowiekiem, któremu wyszła **naprzeciw** i przyjąłem to jak
      I.was the first person who she.met halfway and I.considered it as
      niezasłużony, królewski podarunek.
      undeserved royal gift
      'I was the first person she tried to please, and I thought it was a gift I didn't deserve'
      (*Santa Maria*, 1971).

I found no examples of P doubling in such constructions, though.

2008). While lexicalization effects have been noticed before (Wasow 2002, Hawkins 2004), they have not yet been appealed to with respect to P doubling. The data presented here strongly suggest that theories of grammar can account for P doubling if it is interpreted as a performance strategy that interacts with competence.

References

Arnon, I., B. Estigarribia, Ph. Hofmeister, T. Florian Jaeger, J. Pettibone, I. A. Sag and N. Snider 2005. Long-distance dependencies without island constraints. Poster presented at HOWL 3: Hopkins Workshop on Language.

Blank, A. 2001. Pathways of lexicalization. In M. Haspelmath, E. Konig, W. Oesterreicher and W. Railble (eds.) *Language Typology and Language Universals.* Vol. II, 1596–1608. Berlin and New York: Walter de Gruyter.

Bresnan, J., in press. Is syntactic knowledge probabilistic? Experiments with the English dative alternation. In S. Featherston and W. Sternefeld (eds.) *Roots: Linguistics in Search of its Evidential Base.* Berlin: de Gruyter.

Brinton, L. J. and E. Traugott 2005. *Lexicalization and Language Change.* Cambridge: Cambridge University Press.

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

Huddleston, R. and G. K. Pullum 2002. *The Cambridge Grammar of the English Language.* Cambridge, UK: Cambridge University Press.

Nykiel, J., forthcoming. *Competence, performance and extra prepositions.* Journal of English Linguistics.

Riley, K. and F. Parker 1986. Anomalous prepositions in relative clauses. *American Speech* 61(4), 291–306.

Sag, I. A. and T. Wasow, to appear. Performance-Compatible Competence Grammar. In R. Borsley and K. Borjars (eds.) *Non-Transformational Syntax.* Cambridge: Blackwells.

Staum, L. and I. A. Sag 2008. *Antilocality in Ungrammaticality: Nonlocal grammaticality violations are easier to process.* Poster presented at CUNY 2008 in Chapel Hill, North Carolina.

Traugott, E. 1994. Grammaticalization and lexicalization. In R. E. Asher and J. M. Y. Simpson (eds.) *The Encyclopedia of Language and Linguistics.* Vol. III, 1481–1486. Oxford: Pergamon Press.

Wasow, T. 2002. *Post-verbal behavior.* Stanford, CA: CSLI Publications.

Wischer, I. 2000. Grammaticalization versus lexicalization – "methinks" there is some confusion. In O. Fischer, A. Rosenbach and D. Stein (eds.) *Pathways of Change*, 335–370. Amsterdam and Philadelphia: John Benjamins.