Against "lexicalization" (and what to replace it with)

Martin Haspelmath
Max Planck Institute for Evolutionary Anthropology

1. Introduction

"Lexicalization" =   becoming a lexical item

But there are four ways in which "lexical (item)" can be understood:

  (A) lexical item =   **word**
                        (a leaf of a syntactic tree; a form that is written between spaces)

  (B) lexical item =   **content word**
                        (as opposed to function word, or grammatical item)

  (C) lexical item =   **dictionary item** (*inventorial item*)
                        (something that must be learned by speakers and cannot be constructed)

  (D) lexical item =   **mentally stored item** (*mentalic item*)
                        (something that is stored and accessed holistically by a speaker)

These different concepts are often confused (or not clearly kept apart) by linguists, and the term "lexicalization" may be used in different and confusing ways.

I propose that we replace "lexicalization" by four different terms, broadly corresponding to the four senses of lexical item:

(α) *univerbation* =   the transition from a word combination to a word (§3)

(β) *lexemization* =   the transition from an unrestricted combination to a lexeme (§4)

(γ) *inventorization* = the passing of an unrestricted combination into the *inventorium* (the set of morphs, constructions and phrasemes of a language; §5)

(δ) *mentalicization* = the passing of an unrestricted combination into a speaker's *mentalicon* (= "mental lexicon"; §6)

NOTE:
I am generally talking about languages as sets of speaker conventions, not about languages as mental entities. A speaker must have knowledge of a language (i.e. it must be mentally represented), but linguists generally study **social languages**, not individuals' knowledge of languages.

2. The term "lexicalization" in practice

– packaging meaning components into lexical forms: ***lexification patterns***
    (e.g. Talmy 1985; Levin & Rappaport Hovav 2019)

    English: *I kicked the door open.*   (*kick* = manner)

    French: *J'ai ouvert la porte d'un coup de pied.*   (*ouvrir* = cause + open)

– "lexicalized compounds" = **not fully compositional** compounds
                                                    (e.g. Jackendoff 2016)

    e.g.   English   *honeymoon*
                         strawberry
                         eggplant
                         steamboat

    When the meaning is not fully predictable, the form must be lexically listed – this is often called **"lexicalized"**.

– lexicalized derivatives = not fully compositional derivatives

    e.g.   English   *curiosity*
                         oddity
                         personality   (Bauer et al. 2013: 258)

– "more or less lexicalized": **lexicalization is thought to be a matter of degree**

    cf. also expressions such as "highly lexicalized", "weakly lexicalized"

– **semantic** vs. **phonological** lexicalization   (Bauer 2001: 45)

    semantically lexicalized = idiomatic,   e.g.   *eggplant* (phonologically regular)
    phonologically lexicalized = irregular,   e.g.   *leng-th*   (semantically regular)

– **institutionalization** vs. **lexicalization**

    (Bauer 1993: 48): if a new word starts to be accepted as a new word, it is
                    "institutionalized", e.g. *telephone box*
                                        (not e.g. 'box shaped like a telephone')
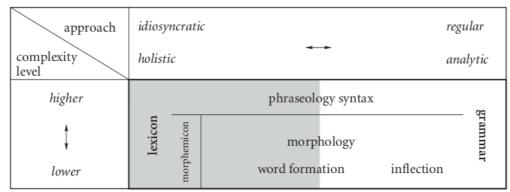
    But is this necessarily a *diachronic* process,
                    from <u>nonce word</u> to <u>institutionalized</u> to <u>lexicalized</u> (Bauer 1983: §3.2)?

        Some words start out as deliberate coinages, e.g. *mentalicization* (above).

– lexicalization cannot simply be "institutionalized **adoption into the lexicon**" (cf. Brinton & Traugott 2005: 89), because we do not want to include loanwords or neologisms such as *texting*.

– Maybe lexicalization can be understood as any change from a more regular to a more idiosyncratic pattern, as in Lehmann (2002: 3).

S1. Lexicon and grammar

approach	idiosyncratic		regular
complexity level	holistic	←→	analytic
higher		phraseology syntax	grammar
↕	lexicon / morphemicon	morphology	
lower		word formation inflection	

But: <u>dictionary item</u> is not the same as <u>mentally stored</u> item!

   (C) lexical item =   **dictionary item** (*inventorial item*)
                            (something that must be learned by speakers and cannot be constructed)

   (D) lexical item =   **mentally stored item** (*mentalic item*)
                            (something that is stored and accessed holistically by a speaker)

Speakers **store** a large number of completely regular combinations, not only inflected words (Plag 2003: 48), but also many regular phrases (Arnon & Snider 2010).

3. Univerbation

(1) ***Univerbation*** is the diachronic change from a word combination to a word.
                                        (cf. Lehmann 2020)

e.g.

(2) a. Latin *cantare habeo* 'I have to sing' > Spanish *cantar-é* 'I will sing'
      b. Latin *quo modo* 'in what way' > *quomodo* > Spanish *cómo* 'how'
      c. Greek *θέλω ἵνα ὑπάγω* 'I want to go' > *θέ να πάω* > *θα-πάω* 'I'll go'

When one of the elements becomes an affix, the process is called ***agglutination***.

When there is no resulting affix, the process is called ***lexemization***.

But what is a "word"? In Haspelmath (2011), I noted that there is no generally applicable definition of "word" that gives the expected results.

Here is a new proposal: (see https://dlc.hypotheses.org/2621)

(3) A ***word*** is – a free morph,
                  – or a bound morph that is not an affix,
                  – or a combination of a root or a compound with its affixes.

"free morph":	e.g. *hello, water, now*
"bound morph (not affix)"	e.g. *to (London), my (garden)*
"root + affix(es)"	e.g. *cat-s, real-iz-ation-s*

(for definitions of "root" and "affix", see Haspelmath 2021)

(4) A ***compound*** is a combination of two roots which occur next to each other and which cannot be expanded by nominal or adjectival modifiers.

cf. Ralli (2013: 21) on Modern Greek:

(5) a. *áγria* *γáta*   'wild cat'
    wild cat

  b. *i* *áγria* *tis* *Marías i* *γáta* 'Maria's wild cat'
    the wild of Maria the cat

  c. *áγria* *ke* *meγáli* *γáta* 'wild and big cat'
    wild and big cat

(6) a. *aγrióγata*   'wild cat'

  b. **aγriomavrióγata*   'wild black cat'

  c. **poliaγrióγata*   'very wild cat'

Thus, there is a viable definition of "word", making it possible to define *univerbation* in terms of diachronic change to a word.

4. Lexemization

Stereotypically, "lexicalization" is a process in which a new lexeme arises. For this "core sense", we can create a new term that specifically refers to a lexeme:

(7) ***Lexemization*** is the diachronic change from an unrestricted combination to a lexeme (stem).

e.g.

(8) a. Latin *animum advertere > animadvert-ere*
  b. Latin *com-ed-ere* 'eat up' > Spanish *com-er* 'eat'
  c. Latin *terrae motum* 'earth movement' > Iralian *terremoto* 'earthquake'
  d. Ancient Greek *paid-íon* 'little child' > Modern Greek *peðí* 'child'

But what is a "lexeme"?

This term is used widely by morphologists, and there is even a distinct "theoretical approach" that is sometimes called "lexeme-based morphology" (as opposed to "morph(eme)-based morphology):

"Lexeme-based models of morphology reject the notion of morpheme and take the traditional dictionary entry as a point of departure. One set of morphological principles, inflectional morphology, defines the word forms of a lexeme and another set, derivational morphology, derives new lexemes from base lexemes." (Spencer 2019: 224)

But how is "lexeme" defined?

(9)   A *lexeme* is the set of forms that minimally contain the same *lexeme-stem*, or one of its suppletive counterparts, and that may only contain inflectional affixes in addition. (Haspelmath 2023: §8)

(10) | language | lexeme | lexeme-stem | some word-forms in the set |
|---|---|---|---|
| English | WALK | *walk-* | *walk-s, walk-ed, walk-ing* |
| Latin | LUPUS 'wolf' | *lup-* | *lup-us* 'NOM.SG', *lup-i* 'NOM.PL' |
| German | AUTOBAHN 'freeway' | *Auto-bahn-* | *Autobahn-en* 'freeways' |
| Spanish | JUGADOR 'player' | *juga-dor-* | *jugador-es* 'players' |

The definition of "lexeme" thus presupposes the notion of a lexeme-stem:

(11) A *lexeme-stem* is a form containing at least one *root* that can combine with *inflectional affixes* but does not contain any.

And what is an "inflectional affix"?

(12) An *inflectional construction* is a construction that expresses an inflectional meaning (role, person, number, gender, tense, mood, evidentiality, polarity) by an *affix* or nonconcatenatively.

These definitions may appear unintutive – especially the extensional definition of "inflectional construction".

But they have the great virtue that they do not make any reference to "lexicon" or "lexical item" – and this is important, because we have seen that these terms are used in four different ways.

Is "lexicalization" restricted to the creation of lexemes? Hilpert (2019) says so:

"Lexicalization is the process of adding new open-class elements to a repository of holistically processed linguistic units."

But there is no reason to think that verb-creating univerbations and adverb-creating univerbations are crucially different:

animum advertere >	*animadevertere*	(VERB)
magno opere >	*magnopere*	(ADVERB)

Thus, *lexemization* (as a subcase of univerbation) is not a particularly interesting concept.

5. Inventorization

Most commonly, "lexicalization" is understood in the sense of "inventorization":

(13) *Inventorization* is a diachronic change by which an unrestricted combination acquires unpredictable properties so that it must be part of the *inventorium.*

The INVENTORIUM (a term coined on April 1st, 2022 in London) is the "dictionary" of a language as a set of social conventions. It is defined as in (8).

(14) The *inventorium* of a language is the set of its morphs, its constructions and its phrasemes (= phraseological patterns).

This is different from
– the set of words (because many words are constructed on the fly)
– the set of lexeme stems (because many lexemes are not inventorized)
– the set of idiosyncratic words (because the inventorium also contains phrasemes)
– the mental lexicon (because our mental storage space is so vast that we store a lot of completely regular combinations)

The term *lexicon* was originally introduced for the inventorium (Bloomfield 1933: 274), defined as "the list of basic irregularities";

But linguists have always had the intuition that **"the lexicon" is a list of words**, so word-formation rules were thought to be "in the lexicon", too ("lexical rules", e.g. Pollard & Sag 1994: 37).

The inventorium **is not a set of words** or word-like elements – in fact, most of its elements are phrasemes, e.g.

bone of contention, cheek by jowl, in other words, to make a long story short, take a shower, come to one's senses, know something by heart, start a family, heavy accent, soundly asleep, fasten the seatbelt, wet paint, be in despair, you have my support, happy birthday to you, will you marry me? (Mel'čuk 2012)

In addition, the inventorium includes the set of all constructions, i.e. it also contains the grammatical rules.

Inventorization is often mixed up with *mentalicization*, so we need to take a close look at this concept.

6. Mentalicization

The inventorium is the repository of elements that must be stored as they cannot be created on the fly, at the level of **social linguistic conventions**.

But linguists are often more intersted in **mental languages**, i.e. speakers' mental representations of their knowledge of the social conventions. (A mental language is often called "I-language" or "competence", especially in generative grammar contexts.)

The counterpart of lexicalization at the level of mental systems is "mentalicization":

(15) **_mentalicization_** is a psychological change by which an unrestricted combination comes to be part of a speaker's **mentalicon** (or "mental lexicon").

The term *mentalicon* refers to a speaker's "mental lexicon", corresponding to the inventorium of a language.

(16) A speaker's **_mentalicon_** is the set of morphs, constructions, phrasemes and predictable combinations that are stored in their brain.

Crucially, a speaker's mentalicon contains not only the unpredictable elements, but also many predictable elements.

7. Does all this help us understand inventorization ("lexicalization")?

A common type of change in languages can be described as inventorization (§6) –
the creation of new forms that have some unpredictable properties.

What might explain such changes?

– Mentalicization in some speakers seems to be a necessary prerequisite.
– But some high-frequency combinations (e.g. *I like it, that's a problem*, ...) are
    presumably mentalicized in all speakers, so it's not sufficient.
– It seems that inventorization can be seen as a type of random event
    that does not need explanation.

Grammaticalization is different because the extreme rarity of **_antigrammaticalization_**
is in need of explanation (Haspelmath 1999; 2004).

Antilexicalization   (= "folk etymology", Lehmann 1989; Lehmann 2002; Brinton &
             Traugott 2005: 102)

> *alcoholic*   >   *alc-oholic*
> *work-oholic*
> *choc-oholic*
>
> i.e. a former single lexeme becomes a kind of compound

Antilexicalization is very rare, but no special explanation is needed, because the conditions for it are rarely present.

8. Conclusion

"Lexicalization" is often explained with respect to speakers' "mental lexicon", but psychological approaches do not explain language change.

Every speaker has a very rich mentalicon, but this does not explain **_inventorization_** –
the acquisition of idiosyncratic properties by an expression. Maybe we do not need an explanation, because random change happens all the time.

Univerbation and *lexemization* are often involved in grammaticalization, but they are not constitutive for it. In fact, "word" and "lexeme" are primarily traditional notions that are probably due to descriptive artifacts (spelling conventions, conventions of writing traditional dictionaries).

Institutionalization is probably a superfluous concept, because the inventorium defines the conventions of a language – there is no reason to assume that there is a difference between "the language system" and "linguistic norms" (Coseriu), or between "potential words" and "actual words".

Bauer (2001) mentions *incentivize*, which was first attested in 1989, and which he says was a potential word but not an actual word at the time.

But: – to the extent that *incentivize* has an idiosyncratic meaning, it became part of the inventorium in 1989

      – to the extent that *incentivize* does not have an idiosyncratic meaning, it existed before 1989, just as all novel sentences "exist" even if they have not been attested

References

Bauer, Laurie. 1983. *English word-formation*. Cambridge: Cambridge University Press.

Bauer, Laurie. 2001. *Morphological productivity*. Cambridge: Cambridge University Press.

Bauer, Laurie & Lieber, Rochelle & Plag, Ingo. 2013. *The Oxford reference guide to English morphology*. Oxford: Oxford University Press.

Bloomfield, Leonard. 1933. *Language*. New York: H. Holt and Company.

Brinton, Laurel J. & Traugott, Elizabeth Closs. 2005. *Lexicalization and language change*. Cambridge: Cambridge University Press.

Haspelmath, Martin. 1999. Why is grammaticalization irreversible? *Linguistics* 37(6). 1043–1068.

Haspelmath, Martin. 2004. On directionality in language change with particular reference to grammaticalization. In Fischer, Olga & Norde, Muriel & Perridon, Harry (eds.), *Up and down the cline: The nature of grammaticalization*, 17–44. Amsterdam: Benjamins.

Haspelmath, Martin. 2011. The indeterminacy of word segmentation and the nature of morphology and syntax. *Folia Linguistica* 45(1). 31–80. (doi:10.1515/flin-2017-1005)

Haspelmath, Martin. 2021. Bound forms, welded forms, and affixes: Basic concepts for morphological comparison. *Voprosy Jazykoznanija* 2021(1). 7–28. (doi:10.31857/0373-658X.2021.1.7-28)

Haspelmath, Martin. 2023. Inflection and derivation are merely ways of speaking. *(to appear)*.

Hilpert, Martin. 2019. Lexicalization in morphology. *Oxford Research Encyclopedia of Linguistics* 2019. (doi:10.1093/acrefore/9780199384655.013.622) Jackendoff, Ray. 2016. English noun-noun compounds in Conceptual Semantics. In ten Hacken, Pius (ed.), *The semantics of compounding*, 15–37. Cambridge: Cambridge University Press.

Lehmann, Christian. 1989. Grammatikalisierung und Lexikalisierung. *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung* 42(1). 11.

Lehmann, Christian. 2002. New reflections on grammaticalization and lexicalization. In Wischer, Ilse & Diewald, Gabriele (eds.), *New reflections on grammaticalization*, 1–18. Amsterdam: Benjamins.

Lehmann, Christian. 2020. Univerbation. *Folia Linguistica*. De Gruyter Mouton 54(s41–s1). 205–252. (doi:10.1515/flih-2020-0007)

Levin, Beth & Rappaport Hovav, Malka. 2019. Lexicalization patterns. In Truswell, Robert (ed.), *Oxford handbook of event structure*, 395–425. Oxford: Oxford University Press.

Mel'čuk, Igor. 2012. Phraseology in the language, in the dictionary, and in the computer. *Yearbook of Phraseology* 3(1). 31–56. (doi:10.1515/phras-2012-0003)

Plag, Ingo. 2003. *Word-formation in English*. Cambridge: Cambridge University Press.

Pollard, Carl J. & Sag, Ivan A. 1994. *Head-Driven Phrase Structure Grammar* (Studies in Contemporary Linguistics). Chicago: The University of Chicago Press.

Talmy, Leonard. 1985. Lexicalization patterns. In Shopen, Timothy (ed.), *Language typology and syntactic description (Volume III)*, 57–149. Cambridge: Cambridge University Press.