

Lemmatisation, etymology and information overload on English and Swedish editions of Wiktionary

Allahverdi Verdizade

Uppsala University, P.O. Box 256, SE-751 05 Uppsala
E-mail: allahverdi.verdizade@lingfil.uu.se

Abstract

Wiktionary is a user-generated wiki-project with the goal of building a universal dictionary covering all words in all languages. Various language editions of Wiktionary have community-specific policies regulating concrete lexicographic questions. The distinct entry structures of English and Swedish Wiktionaries are examined in the context of the relation between headword and etymological information, under special consideration of the user-friendliness of the respective approach. The English Wiktionary applies the etymological approach in setting the headword, which splits identical forms into parts of speech, but also into headwords based on word origin. Additionally, the semantic information is separated from non-semantic more rigorously than is done in the Swedish Wiktionary, placing lists of related and derived terms below the headword rather than under each definition. The Swedish Wiktionary applies the formal-grammatical approach, where division into headwords is made strictly based on identical form and part of speech. In this approach, homonymy is disregarded. The etymological information is nested under each definition rather than having a separate section above the headword. The analysis of the two language editions suggests that the different approaches lead to different amounts of information overload in users, depending on the extent of non-semantic information. Equally extensive entries are handled better within the layout structure of the English Wiktionary.

Keywords: Wiktionary; information overload; etymology

1. Wiktionary, the universal dictionary

Wiktionary is a collaborative project aiming at creating a copyright-free, universal dictionary. The project declares as its goal nothing less than “describing all words in all languages”, including all living and extinct natural languages, as well as a selection of constructed languages. Wiktionary is currently available in 171 language editions. Each edition is characterised by information about the word, be it definitions, word etymologies, labels informing about the word’s register and usage, etc., provided in one meta-language¹. Each language edition housed under a domain prefix (en., sv., de. etc) thus has only one meta-language, but contains entries and definitions of words in (potentially) all languages.

Language editions vary strongly in coverage, quality and growth rate. It is hardly surprising that the large languages have the highest number of entries: the English Wiktionary, hereafter referred to as *en.wikt*, has as of now 3.6 million definitions distributed over 2.6 million entries in 4,500 target languages, out of which English is the largest, with 550,000 entries (21% of all entries). Three other languages – Chinese, Finnish and Italian – are also particularly well-represented on *en.wikt*, having over 100,000 entries each, whereas some 3000 other languages are represented by fewer than 10 definitions each. The Swedish edition, *sv.wikt*, is much smaller, at 356,000 entries, out of which 83,000 are entries on words in Swedish. The ratio between entries in the meta-language and other languages is approximately the same (23% of all entries in *sv.wikt* are entries of Swedish words).

Size and quality do not always go together, and one of the largest editions was until recently that in Malagasy. Wiktionary in Malagasy was able to keep up with *en.wikt* for a long time in terms of amount of entries, but the key to success was not the cumulative work

¹ This is referred to as "native language" in Meyer & Gurevych (2012), which provides an excellent and well-informed introduction to Wiktionary

of an active community, but machine-translation coupled with bot-assisted mass-creation of entries entirely without subsequent human involvement with a low accuracy of glosses and generally poor quality of entries as a result. Therefore, even if the size of the lexical stock covered and growth rate are not always associated with the size of the active community, the overall quality tends to be. As such, en.wikt has 6,000 active editors, sv.wikt has 170, whereas the Malagasy edition has 14. An “active editor” is defined broadly as a user with at least one edit in the past month. As has been noted in the literature, a collaborative project needs to reach a “critical mass” of active editors in order for the lexicographical work to take off in earnest (Törnqvist, 2015).

2. Target audience and functions

Svensén (2009: p. 482-3) lists criteria that can be used to assess a dictionary. Some of the aspects to take into account when critically reviewing a dictionary are: 1. the amount of information provided by a dictionary, 2. the quality of the provided information, and 3. the way it is presented. It is emphasised that every dictionary review must depart from the dictionary’s own idea of the target audience and functions it intends to fill. Neither the quality (1) nor the quantity (2) of the word-stock provided by any edition of Wiktionary is within the scope of this paper: only the various approaches chosen to present it in the relation to lemmatisation (3) are examined. Fuertes-Olivera (2009) evaluates and compares the quantity and the quality of the coverage of English and Spanish lemmas on en.wikt at the time, although findings of a qualitative analysis of Wiktionary like this quickly become outdated in view of the high growth ratio of the project.

Compared to printed dictionaries, the aspects listed above can be somewhat hard to apply when dealing with web-based collaborative projects. Wiktionary is, strictly speaking, not a dictionary, but a dictionary project, which unlike most products developed by private companies or other organizations (referred to as “institutional internet reference works” by Fuertes-Olivera (2009) is not intended to be complete within a certain time framework. This is partly due to the declared goal of “describing all words in all languages”, partly because human languages are in a constant state of change, with new words and senses emerging by the day, while others fall out of use or change their meaning. Seen from this perspective, all Wiktionary editions have the same, next to indefinite, potential to grow and to be reworked. This is only limited by the number of active editors and their interest in different aspects of lexicographic work.

The formal absence of a target audience must therefore be addressed for a meaningful analysis to be possible. I will therefore exclude from the following groups of users: 1. language learners, typically benefiting from information about a word’s formal, semantic and pragmatic aspects. The core vocabulary, i.e. 2,000 of the most frequent words or so, is of primary interest for this group. Examples of usage and collocations are also of uttermost importance. 2. Users looking up words in their native language, such as less frequent words, specialist vocabulary, neologisms, controversial terms or usage prescription. The needs of both above-mentioned groups may include both reception and production; semantic relations (synonyms, antonyms) are thus important. 3. Users interested in linguistic history: here, word etymologies are of primary interest. The potential of Wiktionary is perhaps greatest precisely in this area, and its importance (at least that of en.wikt) in academic contexts as a resource for both finding etymological information and data for novel etymological research becomes increasingly salient (see, for example, Meyer &

Gurevych (2012); Khoury & Sapsford (2016); Sagot (2017) to name a few). It has at times even been proposed that Wiktionary *is* above all an etymological dictionary²³, constituting a secondary source, which, unlike tertiary sources, not only accounts for and summarises published research, but also evaluates its adequacy, comments, and complements it⁴. In view of this, group 3 is perhaps as important as the first two, which usually are the main target audience of a dictionary.

Finally, a fourth group of users can be discerned, since Wiktionary is a project run by unpaid enthusiasts: the editors themselves. They may be representatives of groups 2 and 3, too, and in addition to that native speakers of a project's meta-language, and thus might not have the language learner's perspective in mind. Paradoxically, absence of formally stated target audience can make the editors a target audience in themselves: the unpaid community of hobby lexicographers compiles entries (first of all) for their own community, constituting the primary readership and critics.

This may also be the reason why en.wikt can be perceived as less helpful for learners of English: if the main bulk of the editors are native speakers of English, they might not be interested in contributing information that would help learners of their language, disambiguating definitions, adding synonyms and example usages etc. This is hardly unique for Wiktionary, as monolingual dictionaries are normally written by native speakers regardless of medium. In the case of Wiktionary, however, there is no commissioner to set "production goals" regarding content and time framework. One could argue that en.wikt is not intended for learners of English: however, making English entries more elaborate and user-friendly is of course a legitimate way of contributing, and it also makes it more useful for learners of English. Thus, en.wikt being less suitable for learners of English is not a result of a specific policy, but a consequence of most editors' backgrounds and fields of interests.

The functions filled by Wiktionary can be inferred from the target groups listed above. Another function, that can be hard to tie to any of the above, is that which can be inferred from the slogan "all words of all languages" – that of documentation. A potential target audience benefiting from this is possibly researchers, enthusiasts and activists of linguistic revitalisation and language technology developers.

If the assumption put forward by Gouws & Tarp (2017) regarding too much information being at odds with the needs of users to the same extent as too little is to be accepted, it is easy to see that there is a potential conflict between the will to document everything and degree of user-friendliness. As they note: "In many consultation procedures where problems are experienced there is little doubt that the provision of less lexicographic data would have raised the success rate" (ibid.: 896). Removing valid lexicographic data from

² User Widsith, 2018.11.14, in Beer Parlour, internal discussion page: "I think earliest senses should be first, including when they're obsolete, as in any historical dictionary (which Wiktionary is, like it or not)".

³ User KevinUp, 2019.05.10, BP, "Since Wiktionary is an etymological dictionary, I would prefer to see native Japanese words being lemmatized at their kana forms and Sino-Japanese terms lemmatized at their kanji forms".

⁴ User Rua, 2015.09.1, BP, "Hence, the question that still remains to be answered is whether Wiktionary is an etymological dictionary (secondary source with its own interpretations) or an encyclopedia/compendium of etymological research (tertiary source). Currently, Wiktionary is an etymological dictionary/secondary source as it contains its own interpretations of the data."

Wiktionary is, however, disallowed. One can only seek to relieve the information overload that occurs in the reader, i.e. by reorganising the content visually.

The user groups listed in this section may seem a case of unnecessary coinage of novel terminology, considering the well-established concept of consulting situations, such as reception, perception, translation, etc. However, these would be more relevant for an investigation of the *contents* of Wiktionary, rather than its layout structure. The relationship between entry e.g. entry structure and the etymology affects readers in all these consulting situations to the same degree. It does not mean, however, that we cannot draw between the user typology proposed here and a traditional typology of dictionaries, as suggested for example by Tarp (2017: p. 247):

Adapted from Tarp (2017)		Target groups proposed in current paper
communicative	assist users in solving problems related to written and oral communication, such as text reception, text production, translation and text revision	language learners; users looking up words in their native language
cognitive	transmit knowledge to their users	readers interested in language history; Wiktionary editors; researchers
operative	assist users in performing specific types of action	language learners
interpretive	assist users in interpreting non-linguistic signs	-

3. The overall structure

The starting point in the access structure at Wiktionary is spelling, which means that words in different languages are displayed alphabetically on the same page⁵. The entry layout is originally not developed for the purposes of a dictionary, but for encyclopaedic articles, whence it has been “inherited” and subsequently adjusted to a certain degree, making it radically different from a printed dictionary in several ways. The alphabetical order of entries within a language is not visible for the reader: although the sought entry can be reached by consulting the alphabetical index, the usual way is by using the search function. In order to compensate for the absence of a natural connection with other relevant entries (which can often be found on the same or adjacent pages in printed dictionaries), hyperlinks are used to refer to derived terms, compounds or otherwise related terms. Entries interconnected through semantic relationships (synonymy, antonymy etc.), that are normally not found next to each other in printed dictionaries, are also connected via hyperlinking.

Except for some very general principles applying over the edition boundaries (such as criteria for inclusion⁶), specific lexicographic policies are decided over by the local communities of each edition. One such policy is the question of lemmatisation, or “how lexical units with identical citation forms be presented” Svensén (2004). The differences in how this affects the entry layout in each edition can be exemplified with two constructed entries from the focal editions.

⁵ However, entries in meta-language are displayed at the top regardless of the language name’s initial letter. English always comes first on the page on en.wikt, Swedish on sv.wikt, etc.

⁶ Some differences regarding which words may be included do exist, too: i.e. given names as well as surnames may be included on en.wikt but are not permitted on sv.wikt.

Figure 1: A simplified basic entry on en.wikt

English

Etymology

From Old Swedish *asker*, from Old Norse *askr*, from Proto-Germanic **askaz*, ultimately from Proto-Indo-European **ōs-* (“ash”).

Noun

ask *ɑː*

1. the European ash (tree) *Fraxinus excelsior*
2. a little box.

Figure 2: A simplified basic entry on sv.wikt

Svenska

Substantiv

ask

1. ett trädslag (*Fraxinus excelsior*) i familjen syrenväxter; exemplar av detta träd

Etymologi: Av fornsvenska *asker* (endast belagt genom sammansättningar), av fornnordiska *askr*, av urgermanska **askaz*, slutligen av urindoeuropeiska **ōs-* (“ask”).

2. liten förslutningsbar *låda*

Etymologi: Samma som ovan; ursprungligen i åsyftande till lådor gjorda av askträ.

The main difference lies in how the entry is organised in relation to etymology: while en.wikt structures the content (primarily) around individual etymologies, it is organised (primarily) around the part of speech on sv.wikt. The contrast is most visible in the order of headers: the etymology section constitutes a higher-order section on en.wikt, and the etymological information is given above the definitions, at its own top-level on the page. On sv.wikt, the etymological information is provided inside the lexeme, under each definition. As can be seen, the division into parts of speech constitutes the higher-order hierarchy on sv.wikt, whereas it is subordinate to etymologies on en.wikt. It could be argued that sv.wikt has moved further away from the encyclopaedic entry layout inherited from Wikipedia and done away with the level in the page structure hierarchy, which on en.wikt is made up by the etymology section. The etymology has ceased to be central part of the macro-structure and is demoted to the micro-structure, under the individual definitions. The contrast can be presented schematically, and compared with printed dictionaries in Table (1).

The Swedish word *ask* featured in the constructed entries above presents a case of polysemy: the sense ‘a little box’ developed from the primary sense ‘ash (tree)’. This simplistic example does therefore not fully reflect the contrast in entry structure brought about by the different approaches to lemmatisation adopted by each edition, which is most evident with regard to homonyms. The constructed entries below exemplify each edition’s approach to the homonymous English word *bore* (figures 6 and 7), belonging to

Table 1: Comparison of layout structure in printed dictionaries, en.wikt and sv.wikt

Printed dictionaries	English Wiktionary	Swedish Wiktionary
Page (all lemmata sorted alphabetically, fitting in a single paper page)	Page (all lemmata with identical spelling)	Page
↓	↓	↓
↓	Etymologies (lemmata which can be derived from the same source)	↓
↓	↓	↓
Lemmata (independent entries with identical formal properties: spelling, part of speech, declension/conjugation)	Lemmata	Lemmata
↓	↓	↓
Definitions	Definitions	Definitions

several parts of speech⁷. Note the striking difference in the amount of screenspace used by the entries: the entry on en.wikt is visually much larger than the one on sv.wikt.

4. Lemmatisation

The principles that can be discerned behind the organisation of the entry structure can and should be contextualised within the ones traditionally applied in printed dictionaries. A central reason for the different appearance of the entries on en.wikt and sv.wikt is lemmatisation. Below follows a short review of how it is approached in paper dictionaries, and, by extension, how the question of polysemy vs. homonymy is resolved there. Svensén (2004) lists four approaches: the etymological, the semantic, the morpho-semantic and the formal-grammatical. These four approaches can also be seen as four ways of answering the question “what is a word, in the lexicographical sense?” (and, by extension, “what is another?”).

4.1 Approaches to lemmatisation in printed dictionaries

The etymological method⁸ in its strict application departs from wordhood based on forms of shared origin. Such lexical units are treated as polysemous and lemmatised under the same entry. The readers’ intuitions regarding which forms belong together



⁷ Certain departures were made from the actual entries in order to secure the same amount of information in both constructed entries. For example, the entry on en.wikt is in reality much larger and the one on sv.wikt is smaller. Some lemmata belonging to other parts of speech have been left out. The translation section in the Swedish entry is given merely for comparability, as translations to other target languages are only allowed from the entry in the meta-language. Figures (3) and (4) show parts of the actual entries

⁸ The English-language edition of Svensén (2009) does not include the etymological approach as a distinct way of organizing the entries and concludes further that “the place of etymology in the micro-structure is usually uncomplicated“. Since our analysis suggests that etymology is far from uncomplicated in the context of Wiktionaries, we will utilise the original analysis proposed in Svensén (2004).

Figure 3: Parts of actual entries for the word *bore*

English [edit]

Pronunciation [edit]

- (General American) IPA^(key): /bɔːɹ/
- (Received Pronunciation) IPA^(key): /bɔː/
- (rhotic, without the horse–hoarse merger) IPA^(key): /bɔː(ː)ɹ/
- (non-rhotic, without the horse–hoarse merger) IPA^(key): /bɔə/
- Audio (US)  0:00  MENU
- Rhymes: -ɔː(ː)ɹ
- Homophones: boar, Bohr, boor (accents with the pour–poor merger)

Etymology 1 [edit]

From Middle English *boren*, from Old English *borian* (“to pierce”), from Proto-Germanic **burōną*. Compare Danish *bore*, Norwegian Bokmål *bore*, Dutch *boren*, German *bohren*, Old Norse *bora*. Cognate with Latin *forō* (“to bore, to pierce”), Latin *feriō* (“strike, cut”) and Albanian *birë* (“hole”). Sense of wearying may come from a figurative use such as “to bore the ears”; compare German *drillen*.

Verb [edit]

bore (third-person singular simple present **bores**, present participle **boring**, simple past and past participle **bored**)

1. (transitive) To inspire boredom in somebody. [quotations ▼]
Reading books really bores me, films are much more exciting.
bore someone to death
2. (transitive) To make a hole through something. [quotations ▼]
3. (intransitive) To make a hole with, or as if with, a boring instrument; to cut a circular hole by the rotary motion of a tool.
to bore for water or oil
An insect bores into a tree.

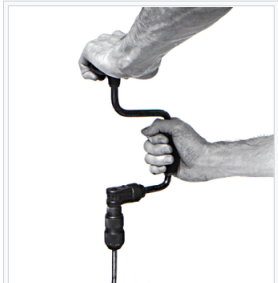



Figure 4: Parts of actual entries for the word *bore*

Engelska [redigera]

 Wiktionaryupplagan på engelska har ett uppslag för **bore**

Substantiv [redigera]

bore

- **uttal**: /bɔːɹ/ eller /bɔːɹ/
- 1. någonting tråkigt; en tråkig person, tråkmåns
- 2. en borrh

Böjningar av bore	Singular	Plural
Nominativ	bore	bores
Genitiv	bore's	bores'

Verb [redigera]

bore

- **uttal**: /bɔːɹ/ eller /bɔːɹ/
- 1. tråka ut
- 2. borra
- 3. böjningsform av bear

----- ⓘ

Besläktade ord: boring, boredom

----- ⓘ

Homofoner: boar

Böjningar av bore 1-2	Singular		Plural
	1-2:a pers.	3:e pers.	
Presens	bore	bores	bore
Preteritum	◀ bored ▶		
Perfektparticip	◀ bored ▶		
Presensparticip	◀ boring ▶		

are thereby of no importance. Words demonstrating identical formal properties (part of speech, inflection, pronunciation) but unrelated historically are seen as homonymous, unrelated forms merely coinciding on the surface and are treated under separate entries. As the name suggests, this approach is best suited for etymological dictionaries, but the principle has been adopted in general-purpose dictionaries too, such as the *Concise Oxford English Dictionary* (2011), which groups lemmata by word origin.

The semantic approach⁹, on the other hand, disregards etymology and groups words by (groups of) meaning. Words are treated as homonymous when their senses are deemed to be too divergent. Etymologically related words like the Swedish *ask* (1. ‘a kind of tree; 2. ‘a small box’) are divided between two entries, whereas i.e. English *crown* (‘a royal headdress; the top of a tree’) is viewed as polysemous and lemmatised under one entry. This approach is well suited for general-purpose dictionaries.

The morpho-semantic approach¹⁰ has the same view on the relationship between etymology and semantics as the previous approach but implies a more learner-friendly macro-structure since semantically related groups of lemmas are given under one “super-lemma” chosen to represent the word-family. This model deviates from formal properties (alphabetical sorting, part of speech and inflection) as a base for the access structure to a larger extent and lemmatises all members of the word-family under the main lemma (cf. Swedish *basal* ‘basal’ ADJ, *basning* ‘steaming’ VN, *basera* ‘to base’ V, under the superlemma *bas* ‘base’ N.). In addition to the learner-friendliness of this approach, it is also a natural choice for languages relying heavily on prefixation for word-formation, such as Indonesian (e.g. in Korigodskiy et al. (1990)), as it allows us to quickly find derived forms which otherwise would end up in another part of the volume. At the same time, Svensén (2004: p. 124) puts forward the argument that it can be harder, not easier, for the reader to arrive at the sought word if he isn’t able to identify the super-lemma¹¹. However, since Wiktionaries only have one form per language and page (entries with distinct spellings are not listed on under the same page and can be accessed via the search function), this weakness does not really apply.

The fourth logical way of handling homonymy and polysemy is the formal-grammatical approach¹², which bases lemmatisation entirely on formal properties of a word without any reference to either etymology or semantics. All forms with identical spelling, part of speech and inflection are treated under the same entry.

Although the formal-grammatical approach eliminates the need for making decisions on how to divide formally identical words into several entries based on extra-linguistic (historical) and semantic grounds, thus speeding up the compilation, it does have drawbacks, too. It is not well-suited for an etymological dictionary and, at the same time, can be somewhat counter-intuitive for readers looking up polysemous/homonymous words in their native language, where it can be assumed that semantic groupings and sub-groupings would facilitate successful look-up. The approach is fully implemented in the latest edition of the printed *Swedish Academic Word List* (SAWL, 2015), which focuses on listing the vocabulary of the Swedish language and attaches less importance to definitions.

⁹ Svensén (2009) provides a more clear-cut typology and calls this “macro-structure oriented homonymization of core senses“ (p.366)

¹⁰ “Homonymization of individual senses“ (Svensén, 2009: p. 365) and “non-strict-alphabetical macro-structure“ (pp. 374-276)

¹¹ As such, it can be challenging for the learner to recognise that the Indonesian *menyerahkan* ‘to hand over’ should be looked up under *serah* ‘to give up’ unless the former is referring to the latter in the overall alphabetic structure in addition to being placed under the base-form; providing such reference for all derived forms easily becomes exceedingly space-consuming in a printed dictionary, since all verbs have a derived form prefixed with *me-*

¹² “Strict alphabetical macrostructure” (Svensén, 2009: p. 371-374)

4.2 Approaches to lemmatisation on Wiktionary

The universality in respect to target groups and purpose reflects the relation to lemmatisation described above: elements pertaining to all three methods can be identified. En.wikt is organised almost entirely according to the etymological approach, but its lemmatisation strategy is in a sense even more radically etymological compared to printed dictionaries: all etymologically related lexemes with identical forms are treated under the same lemma. Several lexemes with distinct formal properties are organised under the same etymology section or divided between several sections if they have different etymologies. The English term *base* is divided between four etymologies: etymology 1 contains subsections both for the noun and the verb *base*, etymology 2 only the adjective *base* etc.

The fundamental structure of sv.wikt is, first and foremost, in line with the formal-grammatical approach, part of speech and inflection are central for lemmatisation. The etymological information is nested under one or several senses by means of so called templates, which automatise the formatting (the position, font size and colour) of different elements. Nesting of links to related terms, such as compounded forms, can be viewed as incorporation of elements of the morpho-semantic approach.

Figure 5: Elements of the morpho-semantic approach implemented on sv.wikt: compounded forms (*sammansättningar*), related terms (*besläktade ord*) and phrases (*fraser*) linked to from the relevant senses of the lemma *man* ‘1. male 2. husband 3. person’.

Substantiv [redigera]

man

- **uttal:** /man/ (betydelse 1-5), /ma:n/ (betydelse 6)

1. (ofta *vuxen*) **människa** av manligt kön

Riktiga män behåller sitt skägg och sin övriga kroppsbehåring.

Sammansättningar: **manfolk**, **manhaftig**, **mankön**, **mansdräkt**, **mansfigur**, **mansgris**, **manshat**, **manshora**, **manshuvud**, **manshög**, **manshöjd**, **manskör**, **manslem**, **manslukerska**, **mansnamn**, **mansperson**, **mansroll**, **mansyrke**

Besläktade ord: **mandom**, **manlig**

2. äkta **make**, man i äktenskap

Det är min mans kostym.

3. (mest i *sammansättningar*) **person** i allmänhet, oavsett kön

Fraser: **bli man för något**, **gemene man**, **god man**, **känna sin man**, **lite till mans**, **menige man**, **på tu man hand**, **tredje man**

Sammansättningar: **brandman**, **förman**, **förtroendeman**, **gärningsman**, **järnvägsman**, **landsman**, **ombudsman**, **riksdagsman**, **spelman**, **tjänsteman**, **enmans**, **tvåmanna**, **tremanna**, **fyrmanna**, **fåmans**, **mannaminne**, **mannamån**

Besläktade ord: **bemanna**, **manskap**

In sum, the community of sv.wikt decided to move away from the structure inherited from Wikipedia to a further extent in order to get closer to the formal-grammatical method. Remnants of the original layout can still be found in some entries: i.e. the entry *person* has etymology as a separate section under the noun rather than having a template inside the definitions. Sv.wikt’s layout policy page, *Stilguiden*, states that this way of including

Figure 6: A simplified homonymous entry on en.wikt

English

Etymology 1

From Middle English *boren*, from Old English *borian* (“to pierce”), from Proto-Germanic **burōnǵ*. Sense of wearying may come from a figurative use such as “to bore the ears”; confer German *drillen*.

Verb

bore (*third-person singular simple present bores, present participle boring, simple past and past participle bored*)

3. To inspire boredom in somebody.
4. To make a hole through something.

Translations

± to make a hole
± to inspire boredom

Related terms

- (to make a hole): borer
- (to inspire boredom): bored, boredom, boring

Noun

bore (*plural bores*)

1. A hole drilled or milled through something, or (by extension) its diameter
the bore of a cannon
2. The tunnel inside of a gun's barrel through which the bullet travels when fired
3. A tool, such as an auger, for making a hole by boring.
4. One who inspires boredom or lack of interest; an uninteresting person

Translations

± a hole drilled or milled through something
± the tunnel inside of a gun's barrel
± boring person

Etymology 2

From Middle English **bore*, *bare*, a borrowing from Old Norse *bára* (“billow, wave”).

Noun

bore (*plural bores*)

1. A sudden and rapid flow of tide occurring in certain rivers and estuaries which rolls up as a wave.

Translations

± sudden and rapid flow of tide

Etymology 3

Verb

bore

simple past tense of *bear*

etymological information is being phased out. En.wikt, on the other hand, has retained a more encyclopedic layout in order to structure entries around shared origin and, by keeping the screenspace intended for formal and semantic properties of the word visually apart from the screenspace intended for etymologies, created a solid groundwork for inclusion of elaborate etymological information. Indeed, insufficient space has historically limited proper etymologisation in printed dictionaries, e.g. when it comes to derived terms (Buchi, 2016: p. 345), and in order to fully utilise the advantages of the paperless format, access to enough (screen)space for the etymology section must be assured in one form or another.

Figure 7: A simplified homonymous entry on sv.wikt

Engelska

Verb

- borra
- tråka ut

----- ⓘ

Etymologi: Av medelengelska *boren*, av fornengelska *borian*, av proto-germanska **burōnaþ*
 Besläktade ord: borer, boredom, boring

- böjningsform av bear*

Översättningar

± borra
 ± tråka ut

Substantiv

- ett borrhål
- en borrhål
- lopp (insidan av röret på ett eldvapen som projektilen passerar igenom)
the bore of a cannon – kanonens lopp

----- ⓘ

Etymologi: Av medelengelska *boren*, av fornengelska *borian*, av proto-germanska **burōnaþ*

- en träkmåns
 Etymologi: samma som ovan; kan ha uppstått genom en metaforisk användning i konstruktioner som *to bore the ears* ("att borra hål i någons öron"), jfr. sv. *mala på* för en liknande utveckling i svenskan.
 Besläktade ord: boredom, boring
- en plötslig högtidvattenvåg
 Etymologi: Av medelengelska **bore, bare, ytterst av fornordiska bára*.

Översättningar

± borrhål
 ± borrhål
 ± träkmåns
 ± högtidvattenvåg

4.3 Implications for target groups

Taking apart definitions of a word and placing them in several sections based on origin (as done on en.wikt) can cause inconvenience for the casual reader uninterested in linguistic history and potentially impede a successful look-up. At the same time, it clears the micro-structure of all non-semantic information: no etymological information is given in the visual vicinity of definitions, being placed in a specially designated section. The part of speech section is reserved for definitions and language samples in the form of user-constructed example sentences, collocations and quotations. Lexical relations, such as synonyms and antonyms, which are deemed to be valuable for comprehension of the sense, are allowed too.

The etymology section is often made up of a short list of attested or reconstructed historical word-forms ancestral to the word in question and cognates in related languages, but there are also many instances of elaborate and sourced inquiries of a words history, including discussion of possible directions of borrowing, semantic shifts and typological parallels. Such inquiries often have a very high academic standard. In view of the very large number of contributors at en.wikt (as compared to sv.wikt), often with special interest in language history, it is not uncommon to see etymology sections of rather extensive size.

Having them nested among definitions would make the latter very hard to navigate, and likely reduce the editors' disposition to compile the often space-demanding review of the existing research, which should ideally be the basis of every etymology.

Derivations, otherwise related terms and translations, links to descendants in other languages all have their own sections visually separated from definitions. This results in an overall page structure with many sections and subsections. This might not be a problem for the seasoned readers of en.wikt, but it should be kept in mind that it was originally developed for encyclopaedic articles with relatively large amount of text in each section. Therefore, navigating a page with many sections, several of which only contain lists of links to other entries, can be challenging to first-time visitors, as it requires a lot of screenspace.

The question is, however, whether the overload of etymological information in the micro-structure (nested under the definitions) would not imply an even more severe impediment to successful look-up than a messy macro-structure. Compare the Swedish noun *bas*, mentioned by (Svensén, 2004: §52) as an example of a polysemous/homonymous word. The *Swedish Academic Dictionary* (SAD, the standard reference work for Swedish etymologies) lists five distinct homonyms belonging to the form. At present, the word encompasses 16 senses unsorted for etymology on sv.wikt. These senses could probably be derived from more than five etymologies provided by SAD at the time of the entry's compilation in the year 1900, as novel senses have emerged since. If fairly complete etymological information would be added under the definitions of the word on sv.wikt, the navigation and possibility for successful look-up would deteriorate for historically interested readers and learners alike.

However, this is in reality not much of a problem for sv.wikt in view of the fact that elaborate etymology sections are at present rare in homonymous words. It is not clear whether this depends on the entry layout reducing the willingness to compile elaborate etymologies, the small number of active editors, or a combination of both factors.

Considering the groups of users outlined at the beginning of this paper, it can safely be assumed that native speakers without interest in etymology and advanced learners benefit from this state of affairs at sv.wikt, as they are unlikely to look up highly frequent words. The latter are precisely the type of words that tend to be polysemous, homonymous and serve as bases of derivation for a great number of terms. Learners and readers who take interest in etymologies are more likely to look up frequent words with a potential for overloaded micro-structure. In particular, the decision to rely on templates nested under definitions for etymological information could discourage potential editors with interest in language history from making elaborate contributions.

5. Information overload

As indicated above, the extensive amount of etymological information on en.wikt results in slower look-up due to the definitions being split between several etymology sections, whereas sv.wikt is spared from this side-effect due to comparatively low amount of etymological information. The incorporation of elements of the morpho-semantic approach into sv.wikt, however, has a potential to slow-down the look-up, too. As such, compare the entry *stad* at sv.wikt (fig. 8) and en.wikt (fig. 9), where compounded terms are visually separated from the definitions to a greater extent. Both the messy macro-structure,

caused by splitting of the definitions between several etymology sections and the messy micro-structure caused by piling up of elements irrelevant for understanding the sense of the word in question are ultimately the results of the goal of including everything there is to say about a word (“all words of all languages”).

Figure 8: Excessive nesting of compounded terms into the micro-structure of the sv.wikt entry *stad*

Substantiv [redigera]

stad

• **uttal:** /sta:d/

1. större **tätort**, plats där många människor bor

*I **staden** finns det många bilister som inte stannar när man kommer till ett övergångsställe.*

*Jag är tillbaka i **stan** om en månad.*

Sammansättningar: **barndomsstad**, **födelsestad**, **förstad**, **hamnstad**, **handelsstad**, **hansestad**, **hemstad**, **huvudstad**, **innerstad**, **landsortsstad**, **miljonstad**, **provinnsstad**, **residensstad**, **småstad**, **sovstad**, **stadsbild**, **stadsdel**, **stadskarta**, **stadskontor**, **stadskultur**, **stadskärna**, **stadskörning**, **stadslag**, **stadslandskap**, **stadsliknande**, **stadsliv**, **stadsläkare**, **stadsmiljö**, **stadsmission**, **stadsmur**, **stadsmuseum**, **stadsmänniska**, **stadsmässig**, **stadsnotarie**, **stadsombudsman**, **stadspark**, **stadsplan**, **stadsplanearkitekt**, **stadsplaneförslag**, **stadsplanerad**, **stadsplanerare**, **stadsplanering**, **stadsplaneändring**, **stadssport**, **stadssprivilegier**, **stadsregister**, **stadsresa**, **stadsrevisor**, **stadsrum**, **stadsrättigheter**, **stadssekreterare**, **stadsstat**, **stadsstyrelse**, **stadsteater**, **stadstrafik**, **stadstrådgårdsmästare**, **stadstull**, **stadsvandring**, **stadsvapen**, **stadsveterinär**, **stadsvimmel**, **stadsvy**, **stadsöverhuvud**, **storstad**, **universitetsstad**, **villastad**, **vårdstad**

2. (något vardagligt, särskilt i singular bestämd form) centrum av stad (1.); **stadskärna**

*Jag ska in till **stan** och handla – är du på?*

3. (oftast i sammansättningar) ställe, plats; tillflyktsort

*Har du skaffat **bastad** än?*

*Hon gick ner till **verkstaden**.*

*Det brinner bra i **eldstaden**.*

Böjningar av stad 1-3.	Singular		Plural	
	Obestämd	Bestämd	Obestämd	Bestämd
utrum				
Nominativ	stad	staden, stan ¹	städer	städerna
Genitiv	stads	stadens, stans ¹	städers	städernas
Not:	1. något vardagligare, se under användning			

The information overload on Wiktionary is, in the typology of Gouws & Tarp (2017) a form of *concrete data overload*, where the formal properties of a word, formal lexical relations (derivations, compounded terms) and etymologies are incorporated into the micro-structure although not necessary demanded by the reader. The main bulk of readers are here assumed to be primarily interested in semantic and pragmatic information rather than etymology or formal lexical relations. Reducing the amount of information to remedy this kind of overload cannot be done, since Wiktionary strives to be as complete as possible.

The *perceptive data overload* (not presenting information optimally), however, can be dealt with. A perceptive data overload emerges when screen space is not used optimally. This is the case, for instance, with the pile up of compounded terms in the entry *stad*. Another example of this are translation sections at en.wikt: some translation sections of frequent terms grow so large that in order to navigate them meaningfully they must be moved to a separate page.¹³ The potential for (almost) infinite growth of entry contents, only limited by the number of active editors, makes this type of overload ever more pressing. The way it is dealt with (moving contents to separate pages or hiding them under “spoilers”) relieves some of the problems, but creates new ones, such as the need for more clicks to arrive at the sought content.

¹³ This is the case for example with the entry *hand*, for which there are 340 translations just for the primary, literal sense. Considering the fact that translations to any language (for which there are many more than 340) are allowed and welcome to be added, this constitutes a clear conflict between the ambition to include everything and reader-friendliness.

Figure 9: Compounded terms under a separate section, partially under a “spoiler”, at the en.wikt entry *stad*

Swedish [\[edit\]](#)

Etymology [\[edit\]](#)

From Old Swedish *staber*, from Old Norse *staðr*, from Proto-Germanic **stadiz*, from Proto-Indo-European **stéh₂tis*.

Pronunciation [\[edit\]](#)

- audio  0:00  MENU
- IPA^(key): /sta:d/

Note that when used in compound words (e.g. *stadsdel*), *stads-* is pronounced IPA^(key): /stats/.

Noun [\[edit\]](#)

stad ˈs̺tɑːd

1. city
2. town
3. (*obsolete, still in some compounds*) *stead*, place

Declension [\[edit\]](#)

Declension of *stad* [\[more\]](#) ▼

Derived terms [\[edit\]](#)

• annorstädes (“elsewhere”)	• stadsbyggnadskonst	• stadsmuseum
• huvudstad (“capital city”)	• stadsbyggnadskontor	• stadsmänniska
• stadsbibliotek (“city library”)	• stadsbyggnadsnämnd	• stadsmässig

[\[show more\]](#) ▼

To sum up, the perceptive data overload on en.wikt arises from the large number of sections and from fragmentation of definitions in homonymous words over multiple etymologies. The perceptive data overload on sv.wikt varies greatly with the amount of content at each individual entry, and arises from the too tight integration of the semantic and non-semantic information. The micro-structure becomes overloaded, since many different types of non-semantic information are placed under the definitions, impeding the chances of successful look-up for casual readers. While it is true that *some* related terms belonging to a definition would be beneficial for quick comprehension, a pile-up of the kind seen in the entry *stad* hardly serves the reader well.

These two degrees of integration could be contrasted with a third solution, presented by the German edition of Wiktionary. It is quite extreme and obviously suffers from too large *dis*integration of different types of information instead: here, every type of information is given under a separate section (see figure 10 for an example of this).

6. Concluding discussion

Every decision on entry layout, lemmatisation and visual integration of different types of information has its own (dis)advantageous effects. As such, the decisions made by the community of sv.wikt to move further away from the encyclopaedic layout of Wikipedia, abolishing separate sections for, for example etymologies, and adherence to the formal-grammatical approach made the screenspace of an entry much smaller (see figures 6 and 7), which is undoubtedly beneficial for the visual grasp of the contents. But this advantage lasts only as long as non-semantic information is held to a minimum. This is also the case for the majority of entries at sv.wikt¹⁴, which is why the decision can

¹⁴ As such, out of 303,373 pages on sv.wikt containing lexical entries, only 18,142 pages contain entries with compounded terms, 43,497 with otherwise related terms and 14,007 with etymologies. A page may include several entries in more than one language. For comparison: there are 2,594,263 lexical

Figure 10: Example sentences at the German Wiktionary entry *stad* separated from the definitions by a list of compounded terms.

stad (Schwedisch) [Bearbeiten]

Substantiv, u [Bearbeiten]

Anmerkung zur Flexion:
Die bestimmte Form Singular *stan* wird öfter benutzt für das Centrum (die Innenstadt) einer Stadt (umgangssprachlich).

Worttrennung:
stad, Plural: stä·der

Aussprache:
IPA: [...]
Hörbeispiele:  stad [\(Info\)](#)

Bedeutungen:
[1] Bebauung mit Gebäuden und Verkehrswegen, die als Zentrum für Verwaltung, Handel und Kultur fungiert; **Stadt**

Unterbegriffe:
[1] [förstad](#), [hamnstad](#), [hemstad](#), [huvudstad](#), [innerstad](#), [småstad](#), [storstad](#), [universitetsstad](#), [villastad](#)

Beispiele:
[1] Edinburgh tillhör Storbritanniens dyraste *städer*.
Edinburgh gehört zu den teuersten *Städten* Großbritanniens.
[1] Jag måste till *stan* i eftermiddags. Vill du följa med?
Ich muss nachmittags in die *Stadt*. Willst du mit?

Wortbildungen:
[stadsplanering](#), [stadsteater](#), [stadsdel](#), [stadskärna](#), [stadsliv](#), [stadsmänniska](#), [stadspark](#), [stadsstat](#), [stadsvapen](#)

be seen as justified. Since completeness of included information is the absolute ideal for Wiktionary, it would be beneficial for sv.wikt to find a way to sustain increasing depth of its content without increasing the concrete overload and exacerbating the user experience. One such way could be relying less on the use of micro-structure templates and establishing separate sections at least for some types of information. An alternative solution would be to introduce a so called "spoiler", or "fold/unfold" function, where the non-semantic information remains structurally subordinate to the definitions, but is hidden under a spoiler by default. This way, etymologies and lists of related terms would still be one click away without impeding the look-up for users who don't need them.

The comparatively large size of the editor community on en.wikt makes the vision of completeness, especially with regards to etymological information, much closer to the reality. As a result, the entry layout had to undergo a larger separation between semantic and non-semantic information, including fragmentation of definitions in homonymous entries between several etymologies. This has increased the concrete information overload in such entries for readers uninterested in language history, but enabled continued growth of high-quality content, such as elaborate etymology sections.

entries (distributed over a smaller number of pages) on en.wikt, 1,410,582 pages contain entries with etymologies, 69,194 pages contain homonymous entries with at least two etymologies. A total of 254,547 pages contain entries with derived terms and 267,975 pages contain entries with related terms.

Since Wiktionary is a project with enormous potential and increasing relevance to lexicography, it would be desirable to address some issues outlined but not examined in this paper. An in-depth study of the contents (in addition to the structure), its quality and adequacy in meeting the needs of target groups (both suggested here and derived from traditional consulting situations) are some of the topics for future research. Empirical verification of the findings of current paper using online user surveys or eye-trackers would also shed more light on the relation between the entry layout and various types of information overload.

7. References

- Buchi, É. (2016). *The Oxford Handbook of Lexicography*, chapter Etymological dictionaries. Oxford University Press.
- Fuertes-Olivera, P.A. (2009). The Function Theory of Lexicography and Electronic Dictionaries: Wiktionary as Prototype of Collective Multiple-Language Internet Dictionary. *Lexicography at a Crossroads. Dictionaries and Encyclopedias Today, Lexicographic Tools Tomorrow*, pp. 99–134.
- Gouws, R.H. & Tarp, S. (2017). Information overload and data overload in lexicography. *International Journal of Lexicography*, 30(4), pp. 389–415.
- Khoury, R. & Sapsford, F. (2016). Latin word stemming using Wiktionary. *Digital Scholarship in the Humanities*, 31(2), pp. 368–373.
- Korigodskiy, R., Kondrasykin, O.N., Zinowyeu, B.I. & Losyagin, W. (1990). *Kamus Besar Bahasa Indonesia-Rusia*. Moscow: Russkiy Yazik.
- Meyer, C.M. & Gurevych, I. (2012). *Wiktionary: A new rival for expert-built lexicons? Exploring the possibilities of collaborative lexicography*. na.
- Oxford Dictionaries (2011). *Concise Oxford English Dictionary: Main edition*. OUP Oxford. URL <https://books.google.se/books?id=DneZcQAACAAJ>.
- Sagot, B. (2017). Extracting an etymological database from wiktionary. In *Electronic Lexicography in the 21st century (eLex 2017)*. pp. 716–728.
- Sköldböck, E. & Wenner, L. (2020). Folkmun. se: A Study of a User-Generated Dictionary of Swedish. *International Journal of Lexicography*, 33(1), pp. 1–16.
- Svensén, B. (2004). *Handbok i lexikografi : Ordböcker i teori och praktik*. Norstedts.
- Svensén, B. (2009). *A handbook of lexicography: The theory and practice of dictionary-making*. Cambridge University Press Cambridge.
- Tarp, S. (2017). *The Routledge handbook of lexicography*, chapter The concept of dictionary. Routledge.
- Törnqvist, L. (2015). Nordiska dialekt-och slangordböcker på Internet. *LexicoNordica*, 22, pp. 57–75.
- Wolfer, S. & Müller-Spitzer, C. (2016). How many people constitute a crowd and what do they do? Quantitative analyses of revisions in the English and German Wiktionary editions. *Lexikos*, 26, pp. 347–371.

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>

