

# Information Needs and Contextualization in the Consultation Process of Dictionaries that are Linked to e-Texts

Theo J.D. Bothma, *Department of Information Science,  
University of Pretoria, Pretoria, South Africa (theo.bothma@up.ac.za)*  
and

Rufus H. Gouws, *Department of Afrikaans and Dutch,  
Stellenbosch University, Stellenbosch, South Africa (rhg@sun.ac.za)*

---

**Abstract:** This article focuses on various aspects regarding contextualization when e-texts are linked to integrated dictionaries. The article responds to a twofold problem statement: (1) Dictionaries linked to e-texts do not sufficiently take into account the contextualization and cotextualization of words when providing information to users. (2) The integrated dictionary may contain the items needed for contextualization and cotextualization, but the e-device cannot interpret the context of a word and link the word to the relevant item in the dictionary article. The aim of the article is to show the need of linking a word from a text on an e-device to the correct sense in the integrated dictionary. This presupposes dynamic dictionary articles and lexicographic structures in which a relation between words in an e-text and user-specified lexicographic sources is established. Some existing projects that perform such linking are discussed and evaluated. Based on these results this article makes some suggestions. It is foreseen that there will be a "black box" of software between the selected word and the dictionary that will determine the correct lemma and sense to be selected from the e-dictionary. Having discussed various alternatives, the article suggests parallel contextualization between the dictionary and the software of the e-device. Many aspects discussed in this article require further research. Relevant proposals are made with regard to this research.

**Keywords:** CONTEXT, CONTEXTUALIZATION, COTEXT, DICTIONARY CONSULTATION, E-DEVICE, E-READER, E-TEXT, INTEGRATED DICTIONARY, LEXICOGRAPHIC NEEDS, LINKING, PARALLEL CONTEXTUALIZATION, TEXT RECEPTION

**Opsomming:** Inligtingsbehoefte en kontekstualisering in die raadpleging van woordeboeke wat aan e-tekste gekoppel is. Hierdie artikel fokus op verskeie aspekte van kontekstualisering wanneer e-tekste gekoppel word aan geïntegreerde woordeboeke. Die artikel het 'n tweevoudige probleemstelling: (1) Woordeboeke wat aan e-tekste gekoppel is, verreken nie die kontekstualisering en kotekstualisering van woorde genoegsaam wanneer inligting aan gebruikers gebied word nie. (2) Die geïntegreerde woordeboek mag wel die aanduiders

wat nodig is vir kontekstualisering en kontekstualisering bevat, maar die e-apparaat kan nie die konteks van 'n woord interpreteer en dit aan die tersaaklike aanduiders in die woordeboek koppel nie. Die doel van hierdie artikel is om die behoefte aan te toon om 'n woord in 'n teks op 'n e-apparaat aan die regte betekenisonderskeiding in die geïntegreerde woordeboek te koppel. Dit voorveronderstel dinamiese woordeboekartikels en leksikografiese strukture waarin 'n verhouding tussen woorde in 'n e-tekste en gebruikerbepaalde leksikografiese bronne gevestig word. Enkele bestaande projekte waarin hierdie soort koppeling voorkom, word bespreek en geëvalueer. Na aanleiding van die resultate hiervan word bepaalde voorstelle gemaak. Dit word voorsien dat daar 'n "black box" met sagteware tussen die gekose woord en die woordeboek sal wees wat die korrekte lemma en betekenisonderskeiding in die e-woordeboek sal bepaal. Verskeie alternatiewe word bespreek waarna parallelle kontekstualisering tussen die woordeboek en die sagteware van die e-apparaat voorgestel word. Baie aspekte wat in hierdie artikel bespreek word, vereis verdere navorsing en relevante voorstelle word in hierdie verband gemaak.

**Sleutelwoorde:** E-APPARAAT, E-LESER, E-TEKSE, GEÏNTEGREERDE WOORDEBOEK, KONTEKS, KONTEKSTUALISERING, KOTEKSE, KOPPELING, LEKSIKOGRAFIESE BEHOEFES, PARALLELE KONTEKSTUALISERING, TEKSEBEGRIPE, WOORDEBOEKRAADPLEGING

## 1. Introduction

Lexicographic needs arise in extra-lexicographical situations and such needs initiate the execution of dictionary consultation procedures. A dictionary user finds a word in a text that causes text reception problems, and he/she consults a dictionary to find that word and the appropriate guidance to solve the problem.

Dictionary users typically do not read a dictionary, but they consult a dictionary for immediate needs, for example, to retrieve a limited amount of information to solve a specific problem (Tarp 2012). Successful dictionary consultation is achieved when these punctual needs can be satisfied because the information that had to be retrieved falls within the scope of the genuine purpose of the specific dictionary. The genuine purpose of a dictionary, according to Wiegand (1998: 299), lies therein that it can be used to retrieve specific information from the lexicographic data accommodated in the partial texts with outer access structure, regarding certain features of those linguistic expressions that belong to the subject matter of the dictionary. Achieving the genuine purpose of a dictionary and satisfying a punctual need are often impeded not by the data available in the dictionary articles but by the lack of supporting items to ensure an optimal retrieval of information. This is because dictionary articles contain a sufficient variety of items that convey the relevant lexicographic data but an insufficient number of contextual and cotextual items to supplement the other items.

A dictionary reflects the lexicon of the specific language by means of the lemma selection that enables a representative macrostructural coverage. However, in a dictionary a lemma sign as guiding element of an article is isolated from its occurrence in the real language. Sufficient addressing procedures are

required to counter this lexicographic isolation.

In the structuring of their dictionaries and the way in which data are presented, lexicographers need to take cognizance not only of the lexicographic needs of their intended target users but also of their reference skills. Although it is often required from the users to apply their mind when consulting a dictionary (i.e. carefully evaluate the search results to ensure that the suggestion by the system is correct and/or acceptable in the given context), a user-friendly approach is needed because the default presentation in many dictionaries does not guarantee consultation success. Retrieving information from the lexicographic data is often further impeded by the density of dictionary articles, unnatural syntax and data overload. In the online environment, a specific dictionary is often linked to a specific device, for example an e-reader. The user is guided from a word found in a text on the e-reader to the treatment of that word in the integrated dictionary. However, such a consultation procedure often fails because the user cannot retrieve the required information due to the linking not being directed at the appropriate item in the dictionary article, or even to an incorrect dictionary article. This could be because the dictionary does not offer enough contextual and cotextual assistance or because the software of the e-reader cannot identify the appropriate context in the text or link it to the appropriate item in the dictionary.

This leads to the following twofold problem statement to which this article will respond:

- Dictionaries linked to e-devices do not sufficiently take into account the contextualization and cotextualization of words when providing information to users.
- The integrated dictionary may contain the items needed for contextualization and cotextualization, but the e-reader cannot interpret the context of a word and link the word to the relevant item in the dictionary article.

Although some traditional procedures can be maintained to provide a certain degree of contextualization and cotextualization, lexicography in a new era is in need of new procedures. One such relatively new procedure is to integrate writing assistants in dictionaries with a text production function. This paper focuses primarily on text reception needs and the use of linking procedures between an e-device and the integrated dictionary to enhance contextualization and cotextualization. A point of departure is that lexicographers should be aware of the typical occurrence of words in real texts and the lexicographic process prevailing in integrated products should enable the recontextualization of these words.

## **2. A traditional approach to context and cotext in dictionaries and a wider use of the terms**

When discussing a topic like contextualization and dictionaries, it is important to

have a clear understanding of the use of the relevant terms in the field of metalexicography and the lexicographic practice. In metalexicography, cf. Wiegand (1988), Gouws (2002), Gouws and Prinsloo (2005), Lettner (2020), Domínguez and Gouws (in print), a distinction is made between items giving the context and those giving the cotext in dictionaries. Context is regarded as the pragmatic environment of an item and is indicated in dictionaries by, for example, labels, glosses and cultural notes. Cotext refers to the textual environment of an item and is typically indicated in a dictionary article by means of example sentences and collocations. The position allocated to items giving the cotext is determined by the type of microstructure of a specific dictionary. If the dictionary has an integrated microstructure, the cotextual items are given in the same subcomment on semantics where the relevant translation equivalent or paraphrase of meaning is given. This leads to a process of direct non-lemmatic addressing. If the dictionary has a non-integrated microstructure, the cotext items are presented in a separate text block but with a clear indication of which cotext item belongs to which translation equivalent or paraphrase of meaning. Remote addressing prevails in such a dictionary article.

In this contribution, the lexicographical use of the terms context and cotext will be maintained. However, because the dictionaries discussed in this paper are not used in isolation but always as part of an integrated product with an e-device as the other component, a slightly wider use of these terms will be proposed. They will have a more comprehensive scope than a mere use in dictionaries. The context of a word is therefore also regarded as its occurrence in a dictionary-external text, for example in a text downloaded onto an e-reader or viewed in a browser on any electronic device. Here the context includes the source and specific volume of the text, the chapter, section, and paragraph where the word occurs as well as extra-textual information regarding the author of the text and the period and geographical environment where the text is situated. The cotext of the word remains its syntactic environment, but, besides its occurrence in a sentence, also the paragraph in which it occurs.

As indicated earlier in this article, the immediate need that leads to a dictionary consultation originates in an extra-lexicographic situation. In this paper the extra-lexicographic environment where the need originates, will be a specific text downloaded on an e-reader or viewed in a browser on any electronic device. The user is guided from a word in such a dictionary-external context to the word presented as lemma sign in a specific dictionary integrated with the e-device. The e-device may contain more than one dictionary. In the selected dictionary the items giving context and cotext should enable the user to link a specific treatment of the word in the dictionary to a specific occurrence of that word in the extra-lexicographic environment.

The focus of this article is to negotiate the contextualization and cotextualization of words as they occur in texts to improve the satisfaction of dictionary users with regard to especially their text reception information needs. This kind of contextualization implies that lexicographers should be acutely aware of the typical dictionary-internal contexts, and they should be able to relate dic-

tionary-external words to these items. When developing the software associated with an e-reader or other e-device, one has to be aware of the extent of contextualization in the linked dictionary. The software needs to be adapted to identify the context of a word and to use that to ensure a successful linking to an item in the dictionary.

### 3. Dictionary-external context

Depending on their functions, dictionaries should include ample items to supply the appropriate contextual and cotextual guidance to their users. In a dictionary article the word represented by the lemma sign is treated in isolation. The contextual and cotextual items provided as part of the lexicographic treatment should not be selected in a haphazard way but, utilising a balanced and representative corpus, it should reflect something of the typical dictionary-external occurrence of the word. This should enable the user to link the word as it was encountered in an extra-lexicographic environment (and context) to a specific search zone in the dictionary that contains the relevant treatment of that word. The context and cotext from dictionary-external occurrences of the word should be transferred to the dictionary article to enhance text reception and text production procedures.

Employing a more comprehensive use of the terms *context* and *cotext* (and also *contextualization* and *cotextualization*), contextualization should not only be seen as referring to a dictionary-internal procedure. Within a dictionary, lexicographers focus on giving the context of the treated word. However, another context and another procedure of contextualization should also be recognized by lexicographers. This is the context outside the dictionary, in this paper the texts found on an e-device. This will determine the dictionary or dictionaries to be integrated with the e-device. Contextualization then also implies an anchoring between this dictionary-external context and items presented in dictionary articles, and it determines the way in which the relevant data are negotiated in the dictionary-internal ordering and presentation procedures.

The online environment offers different possibilities of satisfying lexicographic information needs by means of contextualization. Dictionaries still function as stand-alone products or they can be part of a dictionary portal (Engelberg and Müller-Spitzer 2013: 1023). In both these instances the contextualization in the dictionary is not motivated by specific texts in the dictionary-external environment. The editorial system of the dictionary requires that certain items in the dictionary article, for example, paraphrases of meaning or translation equivalents, should be addressed by items giving context and cotext, and these supporting data are either made-up by the lexicographer or extracted from the specific corpus used by the lexicographer for the specific dictionary. Context can be obtained beyond the stand-alone dictionary of the dictionary portal. The online environment enables such possibilities.

Where dictionary users get access from within a dictionary, a search region, or a dictionary portal, a search domain, to the internet and other sources outside the dictionary portal, a search universe (cf. Gouws 2021: 7), a comprehensive but often unspecified and uncurated pool of supporting data is at the disposal of the user. This offers an opportunity to link an item in a dictionary to a dictionary-external source, but the users must apply their mind to make the appropriate pairing. The search moves from the dictionary to the external source to satisfy a specific lexicographic need of a user. For the current paper, the focus is on the reverse search direction — and this is also possible in an online environment. Users of an e-device should have the opportunity to link an item in a dictionary-external source to a lemma sign in a specific dictionary and the items in a specific subcomment on semantics in the article of that lemma sign. The ideal is that the pairing will link the word in the dictionary-external text to the appropriate items in the dictionary article so that a user can achieve an unambiguous retrieval of information.

To achieve the above-mentioned consultation, lexicographers need to take cognizance of another type of contextualization procedure. The e-device and its software should be able to identify the context of a word in the text and link this context with the appropriate context in the integrated dictionary. This type of contextualization by means of linking is discussed in the next section.

#### **4. Linking**

Linking forms the basis for establishing a mapping between a word in a text and an item in an e-dictionary article. Linking as a contextual procedure presupposes dynamic dictionary articles and lexicographic structures in which a relation between words in an e-text and user-specified lexicographic sources is established. A problematic word encountered in an extra-lexicographic environment needs to be linked to the treatment of that word and its specific sense in a dictionary. Successful linking would map the contextualization/cotextualization of a word in the source to contextualization/cotextualization of a lemma in the dictionary which would result in users being linked to exact and relevant items. It therefore needs both texts and dictionary articles with higher contextualization potential.

In the remainder of this section the focus will be limited to systems that make use of linking, and to context and cotext in dictionaries.

#### **5. Selected projects that link e-texts to e-dictionaries**

Linking words in an e-text to language tools, especially e-dictionaries, is not a new concept, and has been implemented in various projects. The following projects are discussed and briefly evaluated:

- The Perseus Project
- Amazon Kindle dictionary linking
- Browser-based linking
- Linking in an e-learning environment

In each case the project is briefly outlined and examples are discussed. Following these discussions, the principles involved in the projects are briefly evaluated.

### 5.1 Perseus Project

The Perseus Digital Library (<http://www.perseus.tufts.edu/hopper/>) is a project that explores the possibilities that online digital collections offer. The project "covers the history, literature and culture of the Greco-Roman world" (Perseus Digital Library, n.d.-b), and, since its inception in 1987, expanded to include "a massive library of art objects, sites, and buildings", Arabic, Germanic, 19th-Century American Materials etc. (Perseus Digital Library, n.d.-c). According to Crane (1998) the "long-term goal must be to make accessible, both physically and intellectually, to every human being on this planet the complete record of humanity".

The Greek and Roman collections currently contain 44,462,693 English words, 13,507,448 Greek words and 10,525,338 Latin words. The Greek and Latin texts are all encoded with TEI (Perseus Digital Library, n.d.-a, Rydberg-Cox et al. 2000) to provide easy access to properties of individual words so that they can be studied in depth. The encoding allows the user to search for a lemma, and obtain all inflected forms related to the lemma, either in all the texts, or in a specified subset. For example, in Figure 1, the word "*bellum*" is searched in the *De Bello Gallico* by Julius Caesar, which results in the highlighted words in the text; at the bottom right of the image, the three possible lemmas are given, viz. "*bellus*", "*bellum*" and "*bello*". By clicking on a specific occurrence, in this case "*bello*", all possible parts of speech are given, as illustrated in Figure 2. The Latin Word Study Tool (Figure 3) provides a statistical probability of all possible correct part of speech (PoS) analyses, and selects one of the options as the most likely one in context, but adds a caveat: "It may or may not be the correct form." It also provides a link to two online Latin dictionaries, viz. Lewis and Short (see Figure 3 for a short extract, which offers the meaning "war" as translation option) and Elementary Lewis and Short. This recommendation is correct, but this is unfortunately not always the case; if the PoS parsing statistical recommendation were to be incorrect, translations of "pretty, handsome" or "to wage war" would be possible, as is evident from the three possible lemmas listed in Figure 1, and with the PoS analyses in Figures 2, 3 and 4.

**Search Results**

Home Help Collections/Texts Perseus Catalog Research Grants Open Source About

Currently searching the following texts in Latin:

- C. Julius Caesar, *De bello Gallico* (ed. T. Rice Holmes)

Showing 1 - 1 of 1 document results in Latin.

**C. Julius Caesar, *De bello Gallico*** [Less](#)  
(Latin) (English)

**book 1, chapter 1:** ... , proximique sunt Germanis, qui trans Rhenum incolunt, quibuscum continenter **bellum** gerunt. Qua de causa Helvetii quoque reliquos Gallos virtute ... aut suis finibus eos prohibent aut ipsi in eorum finibus **bellum** gerunt. [Eorum una, pars, quam Gallos obtinere dictum

**book 1, chapter 2:** ... fiebat ut et minus late vagarentur et minus facile finitimis **bellum** inferre possent; qua ex parte homines **bellandi** cupidi magno dolore adfliciebantur. Pro multitudine autem hominum et pro gloria **belli** atque fortitudinis angustus se fines habere arbitrabantur, qui in

**book 1, chapter 13:** ... ad eum mittunt; cuius legationis Divico princeps fuit, qui **bello** Cassiano dux Helvetiorum fuerat. Is ita cum Caesare egit... Helvetios ubi eos Caesar constituisset atque esse voluisset; sin **bello** persequi perseveraret, reminisceretur et veteris incommodi populi Romani et

**book 1, chapter 16:** ... non sublevetur, praesertim cum magna ex parte eorum precibus adductus **bellum** susceperit; multo etiam gravius quod sit destitutus queritur].

**Refine This Search** [hide](#)

Language: Latin

Required words:   Expand

Required phrase:

Allowed words:   Expand

Excluded words:   Expand

(This searches within the currently selected documents. To search within all documents, use the form below.)

**Relevant Works (8)** [show](#)

**All Matching Documents (1)** [show](#)

**Matching Lemmas (3)** [hide](#)

- bellus: "pretty, handsome, neat, pleasant, fine, agreeable" (entry in [Lewis & Short Elem. Lewis](#))
- bellum: "war" (entry in [Lewis & Short Elem. Lewis](#))
- bello: "to wage war, carry on war, war" (entry in [Lewis & Short Elem. Lewis](#))

Figure 1: The search word is "bellum", and all potential derivatives in the specific text are found; the selection lists "bellum", "bellandi", "belli" and "bello"

**Latin Word Study Tool**

Home Collections/Texts Perseus Catalog Research Grants Open Source About Help

**Search** [hide](#)

Get Info for  in Latin

**Display Preferences** [hide](#)

Greek Display: Unicode (precombined)

Arabic Display: Unicode

View by Default: Translation

Browse Bar: Show by default

**bello** to wage war, carry on war, war  
(Show lexicon entry in [Lewis & Short Elem. Lewis](#)) (search)

bello verb 1st sg pres ind act

[Word frequency statistics](#)

**bellum** war  
(Show lexicon entry in [Lewis & Short Elem. Lewis](#)) (search)

bello noun sg neut dat  
bello noun sg neut abl

[Word frequency statistics](#)

**bellus** pretty, handsome, neat, pleasant, fine, agreeable  
(Show lexicon entry in [Lewis & Short Elem. Lewis](#)) (search)

bello adj sg neut dat  
bello adj sg neut abl  
bello adj sg masc dat  
bello adj sg masc abl

[Word frequency statistics](#)

Figure 2: Potential part-of-speech analyses of the selected item, "bello"



**Latin Word Study Tool**

Search:  Search  
 ("Agamemnon", "Hom. Od. 9.1", "denarius")  
 All Search Options [\[view abbreviations\]](#)

Home Collections/Texts Perseus Catalog Research Grants Open Source About Help

**bello** to wage war, carry on war, war  
 (Show lexicon entry in [Lewis & Short Elem. Lewis](#)) (search)

bello verb 1st sg pres ind act *no user votes* 2.2%

Word Frequency Statistics ([more statistics](#))

Words in Corpus	Max	Max/10k	Min	Min/10k	Corpus Name
51,295	167	32.557	5	0.975	C. Julius Caesar, De bello Gallico

**bellum** war  
 (Show lexicon entry in [Lewis & Short Elem. Lewis](#)) (search)

bello noun sg neut dat *no user votes* 2.3%  
**bello †** noun sg neut abl *no user votes* 88.1%

† This form has been selected using statistical methods as the most likely one in this context. It may or may not be the correct form. ([More info](#))

Word Frequency Statistics ([more statistics](#))

Words in Corpus	Max	Max/10k	Min	Min/10k	Corpus Name
51,295	516	100.595	0	0	C. Julius Caesar, De bello Gallico

**bellus** pretty, handsome, neat, pleasant, fine, agreeable  
 (Show lexicon entry in [Lewis & Short Elem. Lewis](#)) (search)

bello adj sg neut dat *no user votes* 1.9%  
 bello adj sg neut abl *no user votes* 1.9%  
 bello adj sg masc dat *no user votes* 1.9%  
 bello adj sg masc abl *no user votes* 1.9%

Word Frequency Statistics ([more statistics](#))

Words in Corpus	Max	Max/10k	Min	Min/10k	Corpus Name
51,295	516	100.595	0	0	C. Julius Caesar, De bello Gallico

Search [hide](#)  
 Get Info for bello in  
 Latin Go  
 Display Preferences [hide](#)  
 Greek Display: Unicode (precombined)  
 Arabic Display: Unicode  
 View by Default: Translation  
 Browse Bar: Show by default  
 Update Preferences

**Figure 3:** Statistical analysis suggests that "noun sg neut abl" is correct, and provides the meaning "war"

**A. [select]** *War, warfare* (abstr.), or *a war, the war* (concr.), i.e. *hostilities between two nations* (cf. *tumultus*).

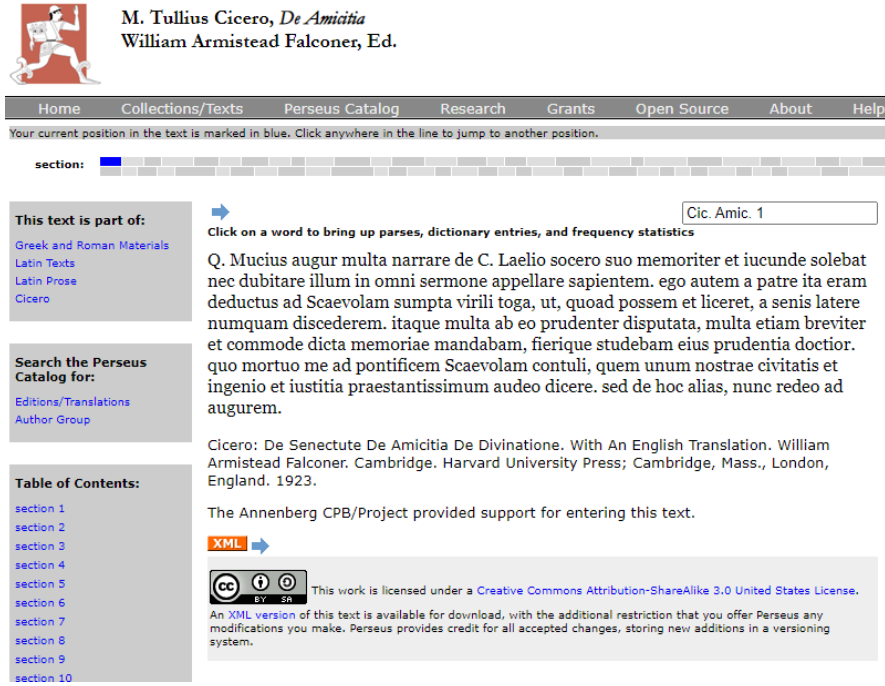
**1. [select]** Specifying the enemy.

- a. [select]** By *adj.* denoting the nation: “omnibus Punicis Siciliensibusque bellis,” Cic. Verr. 2, 5, 47, § 124: “aliquot annis ante secundum Punicum bellum,” *id. Ac.* 2, 5, 13: “Britannicum bellum,” *id. Att.* 4, 16, 13: “Gallicum,” *id. Prov. Cons.* 14, 35: “Germanicum,” Caes. B. G. 3, 28: “Sabinum,” Liv. 1, 26, 4: “Parthicum,” Vell. 2, 46, 2; “similarly: bellum piraticum,” *the war against the pirates*, Vell. 2, 33, 1.— Sometimes the *adj.* refers to the leader or king of the enemy: “Sertorianum bellum,” Cic. Phil. 11, 8, 18: “Mithridaticum,” *id. Imp. Pomp.* 3, 7: “Jugurthinum,” Hor. Epod. 9, 23; Vell. 2, 11, 1; “similarly: bellum regium,” *the war against kings*, Cic. Imp. Pomp. 17, 50. —Or it refers to the theatre of the war: “bellum Africanum, Transalpinum,” Cic. Imp. Pomp. 10, 28: “Asiaticum,” *id. ib.* 22, 64: “Africanum,” Caes. B. C. 2, 32 *fin.*: “Actiacum,” Vell. 2, 86, 3: “Hispaniense,” *id.* 2, 55, 2.—
- b. [select]** With *gen.* of the name of the nation or its leader: bellum Latinorum, *the Latin war*, i. e. *against the Latins*, Cic. N. D. 2, 2, 6: “Venetorum,” Caes. B. G. 3, 16: “Helvetiorum,” *id. ib.* 1, 40 *fin.*; “1, 30: Ambiorigis,” *id. ib.* 6, 29, 4: “Pyrrhi, Philippi,” Cic. Phil. 11, 7, 17: “Samnitium,” Liv. 7, 29, 2.—
- c. [select]** With *cum* and *abl.* of the name.

**(α). [select]** Attributively: “cum Jugurthā, cum Cimbris, cum

**Figure 4:** An extract from the linked version of the Latin–English dictionary by Lewis and Short in the Perseus Project, which provides more detailed information about “*bellum*” and links to various texts

The following figures provide examples of incorrect morphological parsing and/or the complexity of finding the correct translation equivalent. The text, from the *De Amicitia* by Cicero, is given in Figure 5.



**M. Tullius Cicero, *De Amicitia***  
William Armistead Falconer, Ed.

Home Collections/Texts Perseus Catalog Research Grants Open Source About Help

Your current position in the text is marked in blue. Click anywhere in the line to jump to another position.

section: █

**This text is part of:**  
[Greek and Roman Materials](#)  
[Latin Texts](#)  
[Latin Prose](#)  
[Cicero](#)

**Search the Perseus Catalog for:**  
[Editions/Translations](#)  
[Author Group](#)

**Table of Contents:**  
[section 1](#)  
[section 2](#)  
[section 3](#)  
[section 4](#)  
[section 5](#)  
[section 6](#)  
[section 7](#)  
[section 8](#)  
[section 9](#)  
[section 10](#)

Click on a word to bring up parses, dictionary entries, and frequency statistics


Cic. Amic. 1

Q. Mucius augur multa narrare de C. Laelio socero suo memoriter et iucunde solebat nec dubitare illum in omni sermone appellare sapientem. ego autem a patre ita eram deductus ad Scaevolam sumpta virili toga, ut, quoad possem et liceret, a senis latere numquam discederem. itaque multa ab eo prudenter disputata, multa etiam breviter et commode dicta memoriae mandabam, fierique studebam eius prudentia doctor. quo mortuo me ad pontificem Scaevolam contuli, quem unum nostrae civitatis et ingenio et iustitia praestantissimum audeo dicere. sed de hoc alias, nunc redeo ad augurem.

Cicero: De Senectute De Amicitia De Divinatione. With An English Translation. William Armistead Falconer. Cambridge. Harvard University Press; Cambridge, Mass., London, England. 1923.

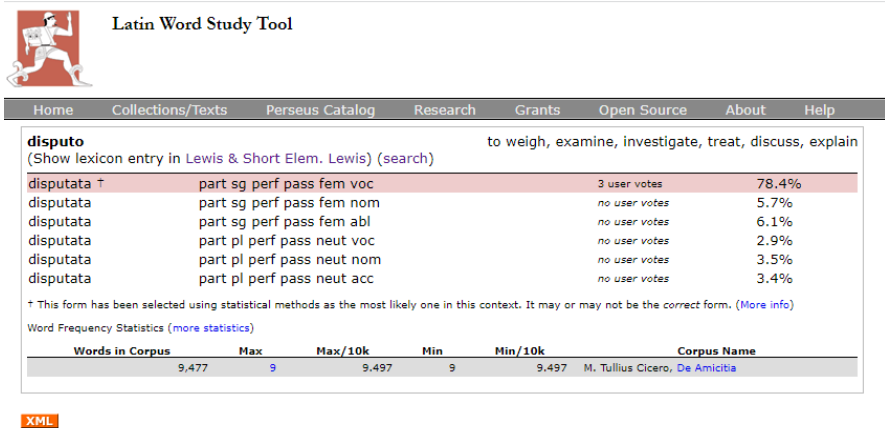
The Annenberg CPB/Project provided support for entering this text.

XML

 This work is licensed under a [Creative Commons Attribution-ShareAlike 3.0 United States License](#).

An XML version of this text is available for download, with the additional restriction that you offer Perseus any modifications you make. Perseus provides credit for all accepted changes, storing new additions in a versioning system.

**Figure 5:** Three words in line 4 are relevant — "ab", "eo" and "disputata" (discussed in inverse order)



**Latin Word Study Tool**

Home Collections/Texts Perseus Catalog Research Grants Open Source About Help

**disputo** to weigh, examine, investigate, treat, discuss, explain  
(Show lexicon entry in Lewis & Short Elem. Lewis) (search)

disputata †	part sg perf pass fem voc	3 user votes	78.4%
disputata	part sg perf pass fem nom	no user votes	5.7%
disputata	part sg perf pass fem abl	no user votes	6.1%
disputata	part pl perf pass neut voc	no user votes	2.9%
disputata	part pl perf pass neut nom	no user votes	3.5%
disputata	part pl perf pass neut acc	no user votes	3.4%

† This form has been selected using statistical methods as the most likely one in this context. It may or may not be the correct form. ([More info](#))

Word Frequency Statistics ([more statistics](#))

Words in Corpus	Max	Max/10k	Min	Min/10k	Corpus Name
9,477	9	9,497	9	9,497	M. Tullius Cicero, <i>De Amicitia</i>

XML

**Figure 6:** The lemma selection for "disputata" is correct. The PoS is partially correct, as it is not "fem voc", as suggested by the statistical analysis, but "neut acc"

**Latin Word Study Tool**

Home Collections/Texts Perseus Catalog Research Grants Open Source About Help

**eo** the dawn  
(Show lexicon entry in Lewis & Short Elem. Lewis) (search)

eo	noun sg fem dat	no user votes	3.5%
eo	noun sg fem abl	no user votes	6.2%

Word Frequency Statistics (more statistics)

Words in Corpus	Max	Max/10k	Min	Min/10k	Corpus Name
9,477	246	259.576	0	0	M. Tullius Cicero, De Amicitia

---

**eo** to go, walk, ride, sail, fly, move, pass  
(Show lexicon entry in Lewis & Short Elem. Lewis) (search)

eo	verb 1st sg pres ind act	no user votes	2.8%
----	--------------------------	---------------	------

Word Frequency Statistics (more statistics)

Words in Corpus	Max	Max/10k	Min	Min/10k	Corpus Name
9,477	136	143.505	0	0	M. Tullius Cicero, De Amicitia

---

**eo2** there, in that place  
(Show lexicon entry in Lewis & Short Elem. Lewis) (search)

eo	adv indeclform	no user votes	17.8%
----	----------------	---------------	-------

Word Frequency Statistics (more statistics)

Words in Corpus	Max	Max/10k	Min	Min/10k	Corpus Name
9,477	72	75.973	0	0	M. Tullius Cicero, De Amicitia

---

**is** he, she, it, the one mentioned  
(Show lexicon entry in Lewis & Short Elem. Lewis) (search)

eo †	pron sg neut abl indeclform	2 user votes	59.6%
eo	pron sg masc abl indeclform	no user votes	10.1%

† This form has been selected using statistical methods as the most likely one in this context. It may or may not be the correct form. (More info)

Word Frequency Statistics (more statistics)

Words in Corpus	Max	Max/10k	Min	Min/10k	Corpus Name
9,477	391	412.578	87	91.801	M. Tullius Cicero, De Amicitia

**Figure 7:** There are four possible lemmas for "eo"; the statistical analysis identifies the correct lemma, "eo", but the incorrect gender, as it is "masc", and not "neut"

**Latin Word Study Tool**

Home Collections/Texts Perseus Catalog Research Grants Open Source About Help

**ab** all the way from  
(Show lexicon entry in Lewis & Short) (search)

ab	prep indeclform		
----	-----------------	--	--

Word Frequency Statistics (more statistics)

Words in Corpus	Max	Max/10k	Min	Min/10k	Corpus Name
9,477	176	185.713	44	46.428	M. Tullius Cicero, De Amicitia

[XML](#)

**Figure 8:** The preposition "ab" is identified correctly; however, one translation equivalent is provided, "all the way from", which makes no sense in context

2. [select] In partic.

a. [select] To denote an agent from whom an action proceeds, or by whom a thing is done or takes place. *By*, and in archaic and solemn style, *of*. So most frequently with *pass.* or *intrans. verbs* with *pass.* signif., when the active object is or is considered as a living being: *Laudari me abs te, a laudato viro*, Naev. ap. *Cic. Tusc.* 4, 31, 67: *injuriā abs te afficior*, Enn. ap. *Auct. Her.* 2, 24, 38: "a patre deductus ad Scaevolam," *Cic. Lael.* 1, 1: "ut tamquam a praesentibus coram haberi sermo videretur," *id. ib.* 1, 3: "disputata ab eo," *id. ib.* 1, 4 al.: "illa (i. e. numerorum ac vocum vis) maxime a Graeciā vetere celebrata," *id. de Or.* 3, 51, 197: "ita generati a naturā sumus," *id. Off.* 1, 29, 103; cf.: "pars mundi damnata a rerum naturā," *Plin.* 4, 12, 26, § 88: "niagna adhibita cura est a providentiā deorum," *Cic. N. D.* 2, 51 al.—With *intrans. verbs*: "quae (i. e. anima) calescit ab eo spiritu," *is warmed by this breath*, *Cic. N. D.* 2, 55, 138; cf. *Ov. M.* 1, 417: (mare) quā a sole collucet, *Cic. Ac.* 2, 105: "salvebis a meo Cicerone," i. e. *young Cicero sends his compliments to you*, *id. Att.* 6, 2 *fin.*: "a quibus (Atheniensibus) erat profectus," i. e. *by whose command*, *Nep. Mil.* 2, 3: "ne vir

**Figure 9:** Clicking on "Show lexicon entry in Lewis and Short" provides the full entry for "ab" in the dictionary — a short selection of the correct sense is provided

The entry in Lewis and Short, however, provides a total overload of information: the article is more than 5,800 words long, and the correct translation equivalent occurs in a similar text (also by Cicero) in the hierarchy at II.B.2.a, at around word 3,400.

The preceding discussion is not intended to be a full evaluation of the Perseus project features and functionalities. It is, however, evident that this is an excellent tool to enable research in the Classics, as well as to support beginning and intermediary students of the Classics. It is also evident that the user has to apply their mind to make a selection from the list of morphological parsings, dictionary entries and word senses.

## 5.2 Linking in e-texts on a Kindle e-reader

The Kindle e-reader allows the reader to link to user-specified dictionaries. A large number of dictionaries is available free of charge. These dictionaries need to be downloaded by the user and the selected dictionary needs to be specified by the user when first being consulted. The specified dictionary can be changed at any time, for example, to link to either a UK or US English dictionary, or a translation dictionary, for example from English to German, or vice versa (if a German phrase occurs in the English text, or when a German text is being read). Linking occurs to the first occurrence of a lemma in the dictionary. This is usually fairly reliable, but, since there is no PoS/syntactic analysis, the landing place of the linking process is not always correct, both at the level of the morphological form (for example, verb instead of noun), the wrong lemma (in the case of homographs and homonyms), and no clarification about the correct sense of a polysemous word.

This process has been studied in a fair amount of detail in Bothma and Prinsloo 2013, where many examples are provided, as well as in Bothma and Gouws 2020, also with examples. Bothma and Prinsloo (2013: 169-184) provide a categorisation of problems that occur in the Kindle linking:

- Correct lemma but incorrect PoS (e.g. verb–noun–adjective–adverb confusion)
  - Incorrect linking of homographs, typically verb/noun confusion in dictionary linking
  - Inflected/conjugated form links to correct lemma but wrong PoS
  - Word is a homograph of an inflected/conjugated form of the verb/noun
- Incorrect lemma (and often incorrect PoS as well)
  - Word links to a homograph lemma which is etymologically not related
  - Word links to the incorrect lemma based on the inflection of either the word itself or a possible inflected/conjugated homograph of the lemma

Bothma and Prinsloo (2013) further discuss issues with compounds, phrases and phrasal verbs, the treatment of proverbs, idioms and similar fixed expressions, the treatment of apostrophes, hyphens and capitalization, the occurrence of wrong or inappropriate options, and cases where no option is given. In many of these cases, the linking is incorrect, or requires further evaluation of the results by the user to find the appropriate sense for the context.

One example of each of the level one bullets earlier in the discussion is provided in Table 1, for ease of reference.

Example	Links to	Correct linking in context
"watch" as in "My watch has stopped"	<b>watch</b> <i>v.</i> 1 [with obj.], "look at or observe attentively ..."	The noun is provided later in the same article, preceded by a small black square: ■ <i>n.</i>
"flags", as in "He washed the flags in the courtyard"	<b>flag</b> <sup>1</sup> <i>n.</i> , "a piece of cloth or similar material ..."	<b>flag</b> <sup>2</sup> <i>n.</i> , "a flat stone slab ..."

**Table 1:** Examples of incorrect linking in Kindle

The linking of e-texts to a user-specified dictionary in the Kindle usually works very well. However, due to the fact that there is no contextual parsing, errors sometimes occur. In addition, the system is not aware of the context or cotext, which results in further possible errors. It is therefore again up to the user to

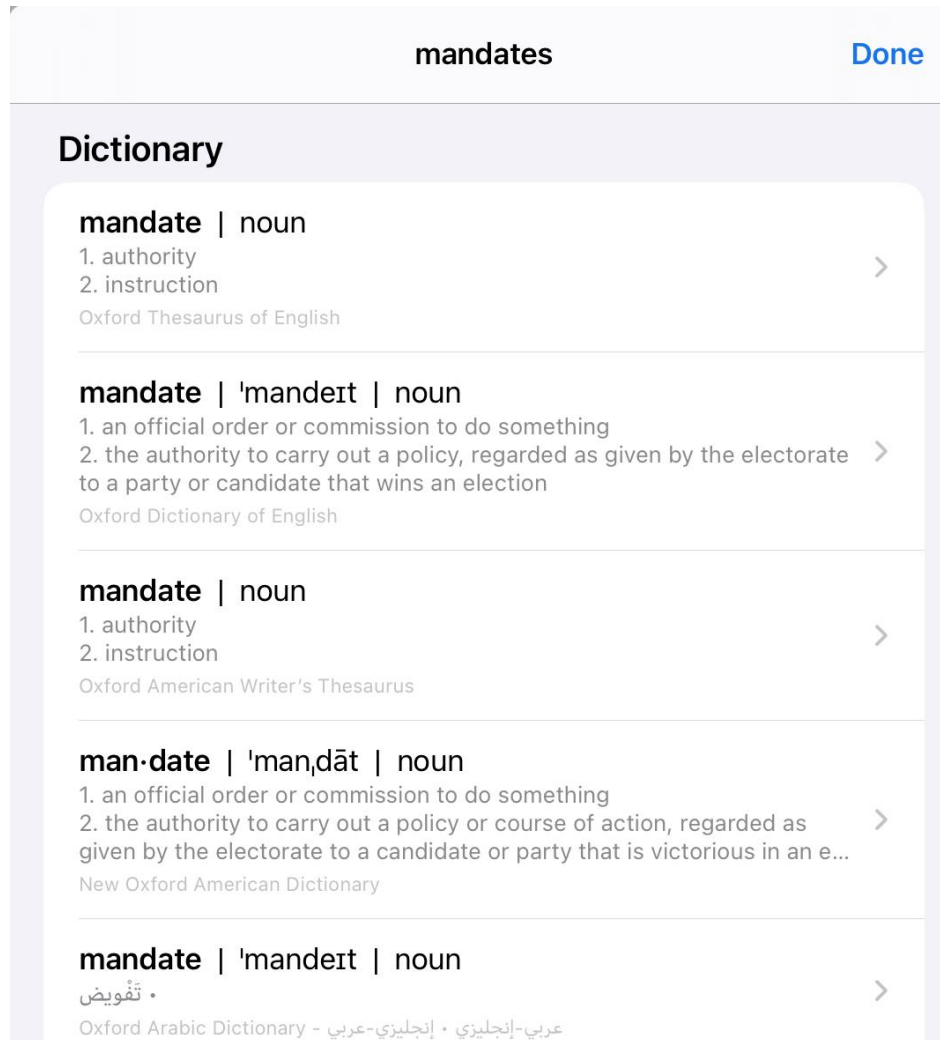
apply their mind to select the correct sense or meaning in each and every instance.

### 5.3 Browser-based linking

Browser-based linking from an e-text to an e-dictionary is an extension of the preceding process, in which linking is restricted to e-texts on a Kindle. In browser-based linking, any text in a browser can link to a user-selected dictionary, or set of dictionaries, which are fully customisable. This is discussed in Tarp and Gouws 2020 and in Bothma and Gouws 2020. From the examples and discussion in these two articles, it is evident that the system cannot be context-aware, and that the problems noted in the Kindle linking occur here as well. Different devices/operating systems have different interfaces, as is evident from the examples in Bothma and Gouws 2020, and Figures 10–15. Depending on the device/operating system, links to additional uncurated information sources can be accessed, which take the user outside the domain of the dictionary. One advantage of the browser-based linking on an iPad is that the user can select multiple dictionaries in the settings of their device. If a user then clicks on a word, a pop-up menu with three items is displayed (Figure 10); clicking on "Look Up" results in a customised dictionary portal with drill-down options to more information (Figure 11). This can obviously lead to information overload, but is, to a certain extent, under the control of the user, as they can select a larger or smaller number of dictionaries in the device settings. On an Android phone, however, a limited set of selection options appears (Figure 12); "Define" links to a dictionary article (Figure 13). Clicking on the three vertical dots produces a portal of unordered information sources, some of which are dictionaries, as illustrated in Figure 14. Clicking in this case on "Dictionary", a dictionary article based on the phone configuration is shown (Figure 15).

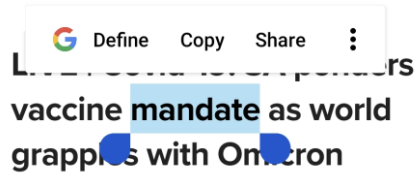


**Figure 10:** Options when clicking on a word in Google Chrome on an iPad

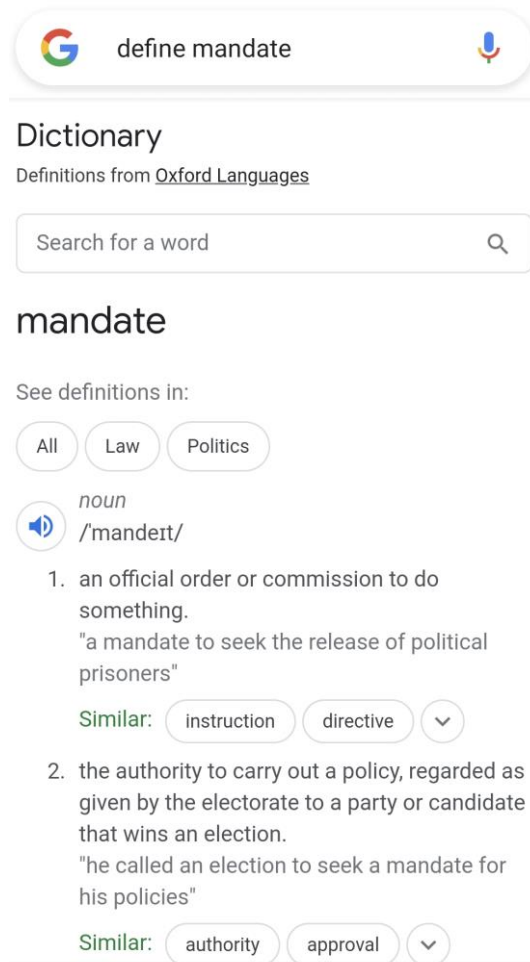


**Figure 11:** Dictionary portal available after clicking "Look Up" in Figure 10 in Google Chrome on an iPad





**Figure 12:** Options when clicking on a word in Google Chrome on an Android phone



**Figure 13:** Selecting "Define" in Figure 12 results in a dictionary article from *Oxford Languages*

Select all

Web search

Dictionary

Dictionary.com

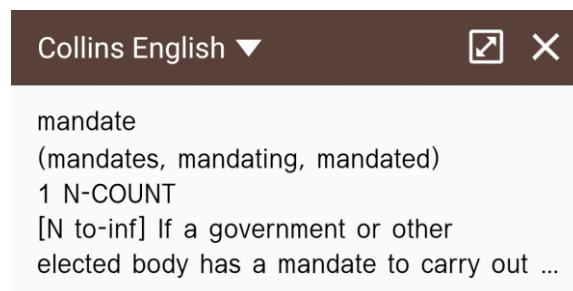
Translate

Search Wikipedia

Ox. Conc. En. Dict

Oxford Dictionary of English

**Figure 14:** Clicking on the three vertical dots in Figure 12 results in this unordered portal (scrolling required to see all options)



**Figure 15:** Clicking on "Dictionary" results, in the configuration on the specific device, on a full article from *Collins English Dictionary*

In addition to the issues of potential information overload (iPad) and unordered/illogical menu structures (Android phone), it is evident that the systems are not context sensitive and that the user has to apply their mind to obtain the correct sense or meaning in the context.

---

#### 5.4 Linking to dictionaries in a language learning environment

Huang and Tarp (2021) describe *Kaiyan OpenLanguage*, an English language learning application (app), for Chinese students. The app makes use of two dictionaries, one which the authors term "embedded", and one which they term "integrated" (Huang and Tarp 2021: 74). The embedded dictionary correlates with the dictionaries described in the previous sections of this paper, and the integrated dictionary is a dictionary that is linked to the course content provided in the app, and "only the words occurring in the text can be consulted"; "Embedded dictionaries cannot 'know' what information a user is looking for in a concrete consultation, whereas integrated dictionaries, if well designed, will be context-aware and, thus, 'know' the concrete sense of a word relevant to the user". According to the authors, "This context-awareness seems to be the most urgent lexicographical challenge to the *Kaiyan OpenLanguage* and other similar learning apps" (Huang and Tarp 2021: 74-75). From their discussions in the following sections of their article, it is evident that Huang and Tarp (2021: 75-85, section 4) are not impressed with the success of context-sensitive linking in this app, and they provide a number of examples. They summarise the problems as follows:

We have seen lexical units that are not treated in the pop-up window, and sometimes not even in the dictionaries when consulted from the front page. We have seen polysemous words where some senses are missing in the default dictionary or only available after accessing the whole article. We have seen words with inaccurate and even wrong definitions. We have seen examples of data overload with senses and equivalents that are irrelevant in the concrete context. We have seen users who have to click three or four times to get an answer or no answer at all. We have seen how the position of the pop-up window that is supposed to help users sometimes has the opposite effect and make it more difficult to pick up the meaning of a word (Huang and Tarp 2021: 86).

Contextualisation of lexicographic consultations is therefore not always very successful in this app, and Huang and Tarp (2021: 88) suggest that this should be fixed by a combination of programming and manual work, focusing on the course texts available in the app, and doing this for all texts that will be added to the app in future.

#### 5.5 Evaluation of linking systems

Linking systems definitely simplify the dictionary consultation process. In an e-reader it is a very easy way to do a quick search in the integrated dictionary, without the "hassle" of having to go to a dictionary app or a paper dictionary.

However, currently, the quality of the results is not guaranteed because in none of the cases the system is fully aware of either the pragmatic/functional environment (the context) or the grammatical environment (the cotext). Linking systems often are not directed at the needs of a specific user group or the interpretation of a specific context. Consequently, when linked to a dictionary, the user is offered a number of options from which to choose. This often leads to a haphazard selection and the users have to apply their mind to decide what the correct meaning/sense for the given context of the word in the dictionary-external text is.

Two of the systems discussed in the article provide links between e-texts and defined corpora, viz. in the Perseus project the corpus of Latin and Greek texts (and others), and in the *Kaiyan OpenLanguage* app the course texts used for language learning.

The Perseus project has extensive markup of the texts and makes a suggestion for the correct PoS analysis based on statistical analysis, linking all possible solutions to the e-dictionaries. The user therefore still has to apply their mind to decide on the correct PoS, as well as to select the correct sense, once the dictionary is accessed.

The linking in the *Kaiyan OpenLanguage* app currently requires that the user apply their mind to make the correct selection in the pop-up window of the dictionary, very similar to the Kindle linking. The suggestions for the improvement of this system (Huang and Tarp 2021) will require extensive input from experts to ensure context-sensitive linking for the course texts.

Based on the nature of the errors in the Kindle linking, there is no analysis of the selected word, and the word is linked to the first available lemma in the dictionary. The same applies to browser-based linking. There is, however, limited matching based on conjugated and inflected forms, but this is simply a matching of these items to the conjugated and inflected forms given in the dictionary, and there is no "intelligence" in the linking.

The amount of parsing in Perseus and the *Kaiyan OpenLanguage* app is insufficient to result in context-aware linking. The matching based on conjugated and inflected forms also does not lead to context-aware linking. The only currently available option is the laborious manual linking by experts. This can obviously be done only for a limited/small corpus, and other options should be investigated.

This article suggests the partial-automated and manual construction of a fully annotated (marked-up) corpus of limited size, to serve as the input for a machine-learning system with artificial intelligence. We therefore foresee that there will be a "black box" of software between the selected word and the dictionary that will determine the correct lemma and sense to be selected from the e-dictionary.

Based on the analyses of the four systems discussed thus far in this article, it is evident that, in all cases, the systems are not context aware, either in terms of the context/cotext of the item or in the broader context of the text:

- 
- Context/cotext issues relate to the system not being aware of, *inter alia*:
    - the PoS of an item;
    - the syntactic function of the item;
    - a distinction between homographs
    - the sense in context of a polysemous word
  - Broader context issues of the text relate to, *inter alia*:
    - the location in which the text is situated
    - the time period in which the text is situated
    - the style of the text, or the specific portion of the text, e.g. formal, informal, slang

In short, it is evident that linking systems need to be devised for a better pairing with specific dictionaries and dictionary entries, based on the context and cotext of the word in the text.

In the rest of this article, we differentiate between linking in a corpus with markup and linking in free text. Modular components and characteristics of linking software will also be discussed.

## 6. The construction of a corpus of limited size

When devising the linking system that helps a user to move from a word in a dictionary-external text to the appropriate item in the dictionary, it is necessary to start with the markup of a small corpus to enable the initial machine learning processes on the side of the e-device and its software. All texts to be downloaded onto the e-device cannot be marked-up and the ideal situation proposed in this paper is that the software of the e-reader or the browser will eventually be able to identify the context of a word in such a way that the linking to the dictionary can be done in an unambiguous way.

When one works with a small, defined corpus, it is possible to annotate the documents with metadata. Fine-grained metadata are required to describe the text sufficiently. This includes:

- PoS tagging, syntactic dependencies and certain semantic aspects;
- bibliographic detail, especially indicating the full volume as well as in-text occurrences
- functional/pragmatic data, including an indication of style, register and language varieties

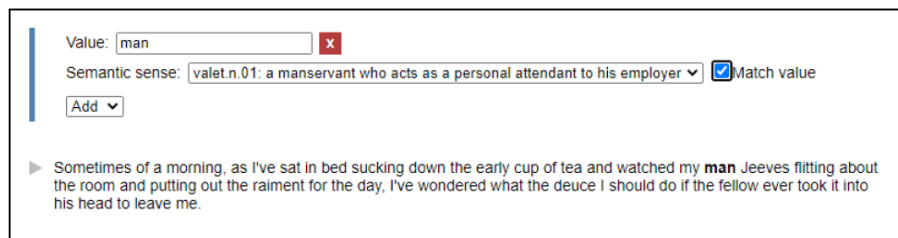
The linking process should then match the metadata of the word in a text with a dictionary article, and with a specific item or search zone in the dictionary article. The markup we propose to achieve this, is similar to markup for enhanced retrieval of specific words and phrases from a text, as described in

Ball (2020: 160-188). In this case, a specialised search engine retrieves a word or phrase that has specific attributes, by matching these attributes to the markup of the words in the database. Ball (2020) developed a prototype system to illustrate these functionalities; for details, see Ball (2020: 198-242).

The matching of a search string with specialised metadata markup to a word or phrase with identical markup in the text database (marked-up corpus) is similar to matching a word with specialised metadata markup in an e-text to items with identical markup in the dictionary database. The database structures will obviously have to be adapted, but the underlying principles are similar. The examples based on retrieving items from a corpus are therefore relevant to illustrate the principle of matching items in an e-text to items based on identical metadata in a dictionary database.

Two examples are selected from Ball (2020) to illustrate such matching, in these cases searching according to a specified semantic sense, and searching according to functional properties, viz. the language of the word.

Figure 16 illustrates a search for the value "man" with the sense of "a man servant who acts as a personal assistant to his employer" (Ball 2020: 219). All other cases of "man" in the texts were not retrieved, and only the one with the required sense was retrieved. This sense was manually marked up in the database.



**Figure 16:** Searching for the value "man" with the sense of "a man servant who acts as a personal assistant to his employer" (Ball 2020: 219)

In the next three figures, the search was for the truncated form "\*men". In Figure 17, the language of the volume was specified as English, and three cases were retrieved, two cases of English words, and one of a Latin word; in Figure 18, English was specified as the "in-text" language, and only the English examples were retrieved; in Figure 19, Latin was specified as the in-text language, and only the Latin example was retrieved (Ball 2020: 230).

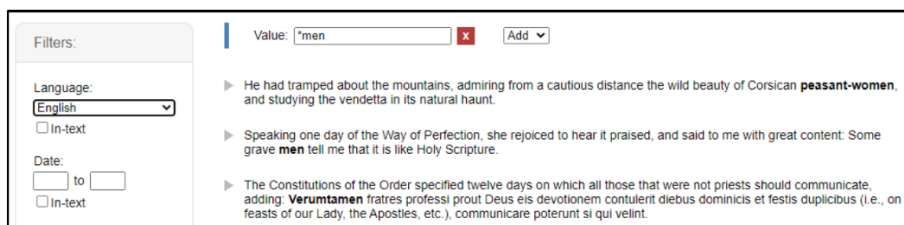


Figure 17: Searching for the truncated form "\*men", with English as the language of the volume (Ball 2020: 230)

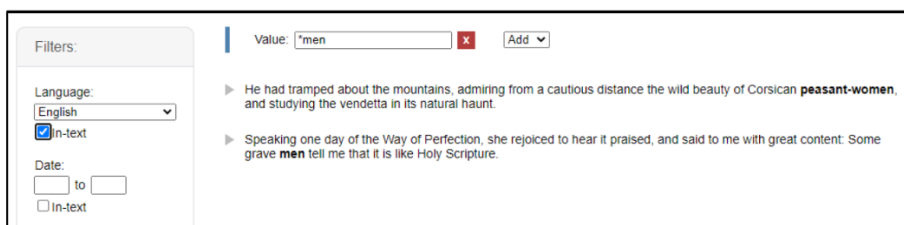


Figure 18: Searching for the truncated form "\*men", with English specified as the in-text language (Ball 2020: 230)

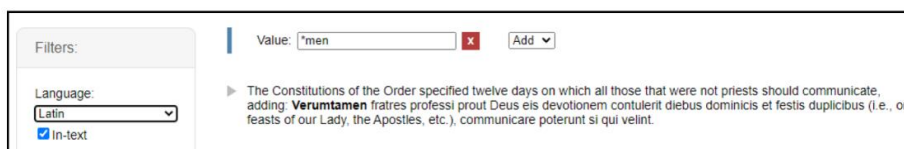


Figure 19: Searching for the truncated form "\*men", with Latin specified as the in-text language (Ball 2020: 230)

It is important to assess the role of a granular metadata markup of textual data in establishing context, as illustrated in the previous examples. Two significant aspects in this regard are the volume context and the section context. Volume context refers to bibliographic metadata, including a reference to the author and other relevant data. This type of context also includes functional/pragmatic metadata, for example style, location, dialectal, cultural data. On the other hand, section context refers to grammatical metadata, in-text biblio-

graphic metadata and functional/pragmatic metadata, for example, direct speech, style, location, dialectal, cultural data, and could also include functional/pragmatic metadata applicable to the paragraph or sentence.

This process can be partially automated, specifically for PoS tagging, and to a lesser extent for syntactic tagging, but all markup needs to be checked manually to ensure that the markup data are correct. Tagging at the semantic, functional and bibliographic levels cannot yet be semi-automated (Ball 2020: 194).

A relevant question is to what extent can such and more complex mappings be automated, and what technologies are required to do so? According to Tarp and Gouws (2019: 264) "... the long road to the required data means that full contextualization is still a challenge to modern lexicography ...". They state that comprehensive research is done "in order to develop a tool that can deduce the specific meaning of a word from the context."

Successful mapping will require a variety of parsing technologies and some of these applications will depend on the progress already made with regard to the specific language. For English a lot of success has been achieved but there are some unsolved challenges. The situation regarding English is currently as follows:

- Automated PoS tagging is good
- Automated syntactic dependencies are in general fair
- Automated semantic tagging is poor
- Automated bibliographic tagging is not yet possible
- Automated functional, pragmatic tagging is not yet possible.

It is important to note that tagging is not an isolated process, but certain inter-relationships are important. Semantic tagging depends on accurate context, syntactic, functional tagging and selection restrictions, whereas syntactic tagging depends on accurate PoS and valency tagging. In addition, accurate taggers at all levels are essential if one wants to automate tagging.

Full automation is therefore currently not yet possible, as described in Ball (2020: 243-279). Semi-automated and manual coding are very laborious and time-consuming tasks, and with current technologies it is not possible to encode very large data sets accurately. We therefore suggest that a fairly small corpus be annotated in detail, using currently available tools, and checking them very carefully manually, to ensure data quality and data integrity.

Based on the preceding descriptions, a limited corpus of annotated text can be created. On-the-fly tagging and analyses cannot simulate the functionalities of such a corpus, but the detail markup in the corpus is required for contextualization. Based on such a corpus, machine learning and artificial intelligence software can be trained to deduce many of these features for proper contextualization automatically. Such software, in a "black box" between the e-text and the e-dictionary can therefore facilitate proper contextualization. A number of further issues are addressed in the following sections of this paper.



## 7. Does a text need to have formal indications about context?

The success of mapping procedures between an e-reader (or similar device) and an e-dictionary demands innovative adaptations in both the e-reader and the e-dictionary that will result in a new perspective on the procedure of contextualization.

In the lexicographic treatment of words taken from texts, the items in the dictionary articles are complemented by items giving contextual and cotextual guidance. These complementing items may not be chosen at random but should be taken from corpus-based data that reflect the actual use of the linguistic expressions. For successful mapping between a dictionary-external text and the appropriate subcomment on semantics in a dictionary article, the notion of parallel contexts and parallel contextualization needs to be negotiated. This does not imply a marking-up of every text downloaded onto the e-device, although when only dealing with a small corpus formal indications of contexts are feasible and would be required to ensure accurate mappings of meanings and senses. Where such a corpus is compiled for language for general purposes, context will be difficult to deduce, as there are no formal markers. In the case of languages for special purposes it might be easier because the topic or discipline might offer limited context.

However, this cannot be done for each text read on an e-device. Parallel contextualization does not in the first instance prevail between the integrated dictionary and each individual text on the e-device, but between the dictionary and the software of the e-reader. Consequently, successful mapping does demand much stronger contextual considerations in the software of the e-reader or other device. This enhanced context awareness of the e-reader or other device could enable a better linking between a word in a text on that device and not only a lemma sign in a dictionary article that represents that specific word, but the specific item in such a dictionary article that presents the applicable treatment of the word in the dictionary-external text.

Parallel contextualization should be preceded by another phase in the integration of a dictionary into a device like an e-reader, namely a determining of the items giving context and cotext in the dictionary. The macrostructural coverage of the dictionary should be corpus-based and must reflect the active lexicon of the treated language as well as some items with a lesser usage frequency. An important early phase in the movement towards parallel contextualization is to ensure that the treatment in the dictionary displays a thorough account of the typical cotexts and contexts in which the treated words occur. This emphasises the significance of the correct choice of dictionaries to be integrated. The user-perspective, a dominant criterion in modern-day lexicography, should also be transferred to the e-device. The users will be primary users of the e-device and only secondary users of the integrated dictionary. Therefore, the candidate dictionary should not be randomly chosen only on account

of its lexicographic quality, important as it will always be, but the selection should be motivated by the intended use and users of the e-device.

Typical texts to be read on the e-reader (or similar device) will play a determining role in the selection of the dictionary that has to be integrated. Once this has been determined a context parallel to that of the dictionary needs to be accounted for in the software of the e-reader. The e-reader should contain context guidance that parallels that of the dictionary. When a reader clicks on a given word in a text on the e-reader the software interprets the context of that word and matches it with the contexts found in the dictionary article that has that word as lemma sign. This could then lead to successful mapping.

Each text encountered on the e-device does not need to have formal indications about context. Such indications should be found in the e-device. This mapping would still be difficult when dealing with free texts. The analysis of non-stop words in the text could probably give an indication, but, ever so important, the possible role of AI and machine learning in ensuring successful mapping should be investigated.

#### **8. Should the cotext be negotiated?**

Comparable arguments given for context could also be given for cotext. The software of the e-device should be able to analyse a limited section to be able to identify the relevant cotext of the section. In a corpus it has to be based on the markup of paragraphs or sentences, for example, in-text citation, direct speech, collocations and other metadata. In free text, probably the same features will apply, if sufficient metadata can be deduced by the software. The cotext in which a word occurs is important to ensure the correct mapping because the occurrence of a word in a text participates in activating a specific homonym/homograph or a specific sense of a polysemous word. The correct analysis of the cotext in the text could ensure the correct mapping with a specific item in the dictionary. The accurate linking of the cotext can enhance the speedy retrieval of information from the data presented in the dictionary article.

#### **9. The structure of the dictionary database**

Integrating a dictionary and an e-device implies that the dictionary becomes an instrument that must satisfy the lexicographic needs of the users of the e-device. The software linking the e-device to the e-dictionary will need to be adapted to enable the coordination of cotexts and contexts. Possible changes, including changes to the database of the dictionary, also need to be negotiated at an early phase of the marriage between dictionary and e-device.

Once integrated into the e-device the dictionary loses its independent status and use. Consultation is no longer open because the integrated dictionary has restricted access possibilities and can only be accessed via a click on a

---

word in a text on the e-device. If an existing dictionary is linked to the e-device the complexity of its former database will determine whether the database has to be changed. If the database contains a sufficient number of unique record types and a sufficient number of properties/attributes, it would most probably not be necessary to change the database structure. However, if there are insufficient record types and attributes, the structure should be adapted to make provision for these items. Nevertheless, one should be careful of over-complicating the database.

#### **10. What are modular components and characteristics of such software?**

Negotiating improved contextualization possibilities compels an investigation of the modular components and characteristics of the envisaged software. The dictionary and the e-device function as a unit and we foresee that a black box of software will be running in the background, to do the required analyses and mappings. The software in the black box will include PoS and other taggers, NLP technologies as well as AI and machine learning.

The software will interpret the text and interact with the dictionary database by mapping the attributes of the word in the text with the attributes of different fields in the database. If the mapping is successful, the user will be provided with a contextualized result.

However, it is important to note that the software should require no or limited manipulation by the user.

#### **11. Scaling the data set**

When embarking on a free text implementation it is a prerequisite to scale the data set and to have a gradual increase in the application possibilities. Starting with a small data set for markup, the markup can be partially automated. Fully automating the markup is problematical, because of inaccuracies at various levels, as discussed in previous sections. It will be essential to check the accuracy of all markup procedures manually to ensure quality. Working with only a small data set is very restrictive and any serious research in this environment should move from a defined corpus to general e-texts, to improve the quality and efficiency of the machine learning algorithms.

#### **12. Future research**

Most of what we have stated in this article requires further research. Some questions that could guide such research include whether contextual analysis should be provided for every word in a text? Unless the software of the e-device can link the context of the text to the appropriate search zone in a specific dictionary article, the answer would probably be yes. A follow-up ques-

tion is whether this would be based on a full set of parsing? This will most probably be required for an efficient disambiguation of homographs, etc. Further research could also result in presenting drill down options that could allow the user to retrieve other information types, like items giving morphology, inflection, collocations, etc.

Such an amount of work raises the question of processing overload. To a certain extent this depends on implementation efficiency, but also on device capabilities. Processing should be fast enough that the user is not frustrated by the lack of speed of the system. In terms of what current technologies offer, better quality parsers, specifically for syntactic, semantic and functional issues, are required. In addition, the technical specifications of artificial intelligence and machine learning algorithms need to be addressed, taking cognizance of the technological challenges and also the financial implications. Attempts should also be made to distinguish between what is traditionally known as a dictionary and what could in future perhaps be referred to as a lexicographical database.

## 11. In conclusion

Contextualization and cotextualization are essential goals that need to be addressed in future research. Extensive multi-disciplinary research and collaboration are required. Lexicographers, computer and information scientists, NLP and UX specialists and others will have to collaborate. It is going to be complex, and an easy solution is not envisaged. When attempting to develop genuine smart e-dictionaries, we concur with the final comment of the article by Tarp and Gouws (2019: 266): "We have work to do."

## Bibliography

- Ball, L.H.** 2020. *Enhancing Digital Text Collections with Detailed Metadata to Improve Retrieval*. PhD dissertation, University of Pretoria. Available: <https://repository.up.ac.za/handle/2263/79015> (accessed 18 June 2021).
- Bothma, T.J.D. and R.H. Gouws.** 2020. e-Dictionaries in a Network of Information Tools in the e-Environment. *Lexikos* 30: 29-56.
- Bothma, T.J.D. and D.J. Prinsloo.** 2013. Automated Dictionary Consultation for Text Reception: A Critical Evaluation of Lexicographic Guidance in Linked Kindle e-Dictionaries. *Lexicographica* 29(1): 165-198.
- Crane, G.** 1998. The Perseus Project and Beyond: How Building a Digital Library Challenges the Humanities and Technology. *D-Lib Magazine* 4(1). Available: <https://www.dlib.org/dlib/january98/01crane.html> (accessed 18 June 2021).
- Domínguez, M.J. and R.H. Gouws.** (In print). Regarding the Definition, Presentation and Automatic Generation of Contextual Data in Lexicography.

- Engelberg, S. and C. Müller-Spitzer.** 2013. Dictionary Portals. Gouws, R.H. et al. (Eds.). 2013. *Dictionaries. An International Encyclopedia of Lexicography. Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography*. Berlin: De Gruyter: 1023-1035.
- Gouws, R.H.** 2002. Equivalent Relations, Context and Cotext in Bilingual Dictionaries. *Hermes — Journal of Language and Communication in Business* 15(28): 195-209.
- Gouws, R.H.** 2021. Expanding the Use of Corpora in the Lexicographic Process of Online Dictionaries. Piosik, M. et al. (Eds.). 2021. *Korpora in der Lexikographie und Phraseologie*: 1-20. Berlin: De Gruyter.
- Gouws, R.H. and Prinsloo, D.J.** 2005. *Principles and Practice of South African Lexicography*. Stellenbosch: African SUNMedia.
- Huang, F. and S. Tarp.** 2021. Dictionaries Integrated into English Learning Apps: Critical Comments and Suggestions for Improvement. *Lexikos* 31: 68-92.
- Lettner, K.** 2020. *Zur Theorie des lexikographischen Beispiels*. Berlin/Boston: De Gruyter.
- Perseus Digital Library.** n.d.-a. *Greek and Roman Documents*. Available: <http://www.perseus.tufts.edu/hopper/collection?collection=Perseus:collection:Greco-Roman> (accessed 12 October 2021).
- Perseus Digital Library.** n.d.-b. *Perseus Digital Library — About*. Available: <http://www.perseus.tufts.edu/hopper/about> (accessed 12 October 2021).
- Perseus Digital Library.** n.d.-c. *Perseus Digital Library — Browse the Collections*. Available: <http://www.perseus.tufts.edu/hopper/collections> (accessed 12 October 2021).
- Rydberg-Cox, J.A., R.F. Chavez, D.A. Smith, A. Mahoney and G.R. Crane.** 2000. Knowledge Management in the Perseus Digital Library. *Ariadne* 25. Available: <http://www.ariadne.ac.uk/issue/25/rydberg-cox/> (accessed 18 June 2021).
- Tarp, S.** 2012. Theoretical Challenges in the Transition from Lexicographical p-works to e-tools. Granger, S. and M. Paquot (Eds.). 2012. *Electronic Lexicography*: 107-118. Oxford: Oxford University Press.
- Tarp, S. and R.H. Gouws.** 2019. Lexicographical Contextualization and Personalization: A New Perspective. *Lexikos* 29: 250-268.
- Tarp, S. and R.H. Gouws.** 2020. Reference Skills or Human-Centered Design: Towards a New Lexicographical Culture. *Lexikos* 30: 470-498.
- Wiegand, H.E.** 1988. Wörterbuchartikel als Text. Das Wörterbuch. Artikel und Verweisstrukturen. G. Harras (Ed.). 1988. *Jahrbuch des Instituts für deutsche Sprache* 1987: 30-120. Düsseldorf: Schwann.
- Wiegand, H.E.** 1998. *Wörterbuchforschung*. Berlin: De Gruyter.