
Spelling Variations in *A Dictionary of Setswana Personal Names*

Thapelo J. Otlogetswe, *Faculty of Humanities,
University of Botswana, Gaborone, Botswana*
(otlogets@ub.ac.bw) (<https://orcid.org/0000-0003-3887-2982>)

Abstract: This paper investigates spelling variation in *A Dictionary of Setswana Personal Names* (Otlogetswe and Ramaeba 2024), addressing a gap in existing research which has previously focused on general-purpose monolingual Setswana dictionaries (Otlogetswe 2015) but not specialized dictionaries such as the one on personal names. The study draws on orthographic guidelines outlined in the *Standard Orthography of 1981* (Ministry of Education 1981) and *Mokwalo o o Lolameng wa Setswana* (Chebanne et al. 2008), while also tracing the historical development of Setswana orthography to early works by Lichtenstein (1928–1930; 1973) and Livingstone (1875). Using textual analysis, the paper critically examines dictionary headwords to identify and categorize spelling variants. The findings reveal numerous inconsistencies in the spelling of Setswana personal names, influenced by factors such as the coexistence of older and newer orthographic standards, dialectal variation across Setswana-speaking regions, and underlying morphological and phonological processes.

Keywords: SPELLING VARIATION, *A DICTIONARY OF SETSWANA PERSONAL NAMES*, SETSWANA, ORTHOGRAPHY

Opsomming: Spellingvariasie in *A Dictionary of Setswana Personal Names*.

In hierdie artikel word die spellingvariasie in *A Dictionary of Setswana Personal Names* (Otlogetswe en Ramaeba 2024) ondersoek, om sodoende 'n gaping te vul in die bestaande navorsing wat voorheen gefokus het op algemene Setswana-woordeboeke (Otlogetswe 2015), maar nie op gespesialiseerde woordeboeke soos die een oor persoonsname nie. Die studie steun op ortografiese riglyne, soos uiteengesit in die *Standard Orthography of 1981* (Ministerie van Onderwys 1981) en *Mokwalo o o Lolameng wa Setswana* (Chebanne et al. 2008), terwyl die historiese ontwikkeling van Setswana-ortografie na vroeë werke deur Lichtenstein (1928–1930; 1973) en Livingstone (1875) teruggespoor word. Met behulp van tekstuele analise ondersoek die artikel woordeboektrekwoorde krities om spellingvariante te identifiseer en te kategoriseer. Die bevindings toon talle inkonsekwenthede in die spelling van Setswana-persoonsname wat beïnvloed word deur faktore soos die gelyktydige voorkoms van ouer en nuwer ortografiese standaarde, dialektiese variasie in Setswana-sprekende streke, en onderliggende morfologiese en fonologiese prosesse.

Sleutelwoorde: SPELLINGVARIASIE, *A DICTIONARY OF SETSWANA PERSONAL NAMES*, SETSWANA, ORTOGRAFIE

1. Introduction and background

Spelling variations are ubiquitous in natural languages, and they have become

an essential element of any dictionary since they capture regional nuances, historical contexts, and the dynamic nature of language. In a globalised world, people encounter texts from various regions characterised by multiple variations. Capturing spelling variations can, therefore, go a long way to help readers understand and recognise different spellings, as such reducing confusion in text reception. For instance, English is spoken and used in many countries with different spelling conventions. Knowing that "color" and "colour" are the same word but used in different regions can prevent misunderstandings and minimize confusion. For language learners, encountering different spellings can be challenging. Dictionaries that include variations provide a comprehensive learning tool, helping learners become familiar with regional differences and enhancing their overall language proficiency.

Similarly, Africa is home to a multitude of languages and dialects, each with its own unique spelling conventions. The Centre for Advanced Studies of African Society (CASAS) leads the work to harmonize and standardize language orthographies in Africa to minimise the existence of competing orthographies of the same language (Chebanne et al. 2008, Prah 1998). A typical case study on how to handle spelling variations in an African language dictionary has been discussed by Otlogetswe (2015) looking at variation in the Setswana monolingual dictionary *Tlhalosi ya Medi ya Setswana* (Otlogetswe 2012). Setswana has borrowed many words from English and Afrikaans, and these words are often spelled in different ways because they are borrowed as they are heard. For example, the word "computer" has been borrowed as *khompiutara*, *khomphiutara*, *khomputara*, *komputara*, *khomputa*, or *khompiuta*. Since there is no language board that Setswana spellings of borrowed words in Botswana, variants remain in the language. Otlogetswe argues that spelling variants should be included in general monolingual dictionaries and that the way they are handled should be informed by frequency information from corpus data.

Dictionaries like the *African Pidgin Dictionary* (see www.pegeen.com) list spelling variations to ensure users from different regions can find the relevant spelling. For instance, a word might have different spellings in Nigerian, Ghanaian, and Cameroonian pidgin. By including these variations, dictionaries capture and document the rich linguistic diversity of a language. This is particularly important for preserving regional dialects and cultural identities. To help users understand the appropriate context for each spelling variant, African dictionaries can use usage labels and contextual notes (Otlogetswe 2014). These labels might indicate whether a spelling is specific to a particular country or region. For example, the South African National Lexicography Units (SANLU) provide detailed notes on the usage of words in different South African languages, helping users navigate the regional differences. This approach ensures that users can select the correct spelling based on their specific needs and contexts.

Additionally, language evolves over time, and spelling variations often reflect historical changes (Murray 1996). Including archaic or historical spellings in dictionaries helps users understand and trace the development of the language

and provides context for older texts. For example, knowing that "jail" was once spelled "gaol" can be useful for reading and understanding historical documents. Some African dictionaries include historical or archaic spellings to provide a comprehensive view of a word's evolution. This is particularly important for languages with rich oral traditions and historical texts. By including these spellings, dictionaries help users understand the historical context and development of the language. For instance, older spellings of Setswana words might be included to aid in the study and appreciation of historical documents.

Spelling variations also support accurate and contextual usage (Motzafi-Haller et al. 2006). Some words have different spellings depending on their context. For example, "dialog" is often used in computing contexts, while "dialogue" is preferred for conversations between people. Dictionaries that include these variations help users choose the correct spelling for their specific context. Lexicographers make editorial decisions about which variants to include in a dictionary based on the word's usage frequency and relevance. For example, if a particular spelling is commonly used in contemporary literature or media, it might be prioritised in the dictionary. This ensures that the dictionary remains relevant and useful for modern users. For instance, the inclusion of both "doughnut" and "donut" in dictionaries reflects their usage in different contexts. The "donut" spelling has been widely used in the USA from the mid-20th century, while the "doughnut" spelling has been used consistently elsewhere.

The inclusion of spelling variations in a dictionary can also go a long way in preserving the cultural heritage of a linguistic group (Scontras et al. 2015). Including spelling variations for indigenous and local terms helps preserve cultural heritage and ensures that these terms are recognised and respected. This is especially important in African dictionaries, where capturing the diversity of languages and dialects is crucial. Proper nouns and brand names often have unique spellings. Including these in dictionaries helps users understand that while these spellings are correct for specific entities, they are not generalisable to other contexts. This distinction is important for accurate communication and brand recognition.

The inclusion of spelling variation also has practical applications. In the digital age spellcheckers and autocorrect tools rely on comprehensive dictionaries (Mulingi 2023). Including spelling variations ensures that these tools can recognise and correct different spellings, enhancing their effectiveness and user experience. As English continues to be a global lingua franca, understanding and recognising spelling variations is essential for effective communication across different regions. Dictionaries that include these variations support global communication by providing a common reference point (<https://blog.spellquiz.com/spelling-variants/>).

The treatment of spelling variants in Setswana dictionaries provides a useful case study. The dictionaries *Thanodi ya Setswana* (Kgasa and Tsonope 1995), *Thanodi ya Setswana* (Mareme 2007), and *Tlhalosi ya Medi ya Setswana* (Otlogetswe 2012) have all dealt with spelling variation in different ways. These dictionaries reflect the influence of regional dialects and the evolution of the language over time.

As previously indicated, the *African Pidgin Dictionary* is another example of how spelling variations are handled. This dictionary includes words from various Pidgin English dialects spoken across Africa, such as Nigerian Pidgin and Ghanaian Pidgin (<https://peegeen.com/>). It lists multiple spellings for words and provides contextual notes to help users understand their regional differences. This approach ensures that the dictionary is a comprehensive resource for speakers of different Pidgin English dialects. The Shona monolingual dictionary developed by the African Languages Research Institute (ALRI) at the University of Zimbabwe grapples with dialectal diversity in Shona (Chabata 2003). It addresses tone marking, sense ordering, and synonym treatment to reflect dialectal nuances. However, representativeness remains a challenge, as some dialects are underrepresented.

In practical lexicography, cross-references are commonly used to guide users from one spelling variant to another (Prinsloo and Van Graan 2021). If a user looks up a word in one spelling, the dictionary might direct them to another variant with a note explaining the regional or contextual differences. This method ensures that users find the information they need, regardless of the spelling they initially searched for. For example, a dictionary might list "catalogue" and direct users to "catalog" with a note on regional preferences.

African dictionaries, however, face unique challenges due to the continent's linguistic diversity and the dynamic nature of its languages. One challenge is the standardisation of spelling in languages that have multiple dialects and regional variations (Baugh 2009). To address this, lexicographers often collaborate with local language experts and communities to ensure that the dictionary accurately reflects current usage and captures variations as synonyms in dictionaries (Chabata 2003). Innovations in technology have also played a significant role in developing African dictionaries. Digital platforms allow for more frequent updates and the inclusion of multimedia elements such as audio pronunciations and video clips. These features enhance the usability of the dictionary and provide a richer linguistic resource for users. African dictionaries handle spelling variations by including regional variants, using usage labels and contextual notes, providing cross-references, including historical spellings, making editorial decisions based on usage, and addressing proper nouns and indigenous terms. These strategies help users navigate the complexities of African languages, ensuring they can find and use the correct spelling for any given context. By reflecting the linguistic diversity and cultural richness of the continent, African dictionaries play a vital role in education and language preservation.

2. Methodology

This paper analyses spelling variations found in names that are headwords in *A Dictionary of Setswana Personal Names*. The names are analysed based on *Standard Orthography of 1981* (Ministry of Education 1981) and *Mokwalo o o lolameng wa*

Setswana (Chebanne et al. 2008). While the historical roots of Setswana orthography may be traced to Lichtenstein (1928) and Livingstone (1875), the Ministry of Education (1981) and Chebanne et al. (2008) present recent rules to Setswana spelling in Botswana. The study uses a textual analysis method to critically assess the dictionary headwords for spelling variants. The results show that there are numerous inconsistencies in the spelling of Setswana personal names. The spellings are motivated by how speakers pronounce the names and not necessarily how they ought to be spelt, according to the standard Setswana orthography. The variations may be clustered into four categories of: (a) vowel variations, (b) consonantal variations, and (c) dialect influenced variations. The final section of the paper deals with names that use discontinued consonants.

Results and discussion

Spelling variations in African personal names are a widespread and multifaceted issue, shaped by linguistic diversity, colonial legacies, orthographic inconsistencies, and sociocultural dynamics. They are a consequence of variation at consonantal and vowel levels as demonstrated in the following sections.

3. Vowel variations in Setswana names

Setswana has seven vowels. These are i, e, ε, u, ɔ, u, a (DALL 1999: 17). The seven vowels may be represented as in Table 1:

Height	Localisation	
	Front	Back
Close	i	U
Half-close	e	ɔ
Half-Open	ε	ɔ
Open	A	

Table 1: Setswana vowels

Orthographically, Setswana does not mark a distinction between /ε/ and /e/. They are both represented by [e] in normal written texts, except in the Botswana lower primary school syllabus where /ε/ is represented by [e] with a circumflex as [ê]. Equally, Setswana orthography does not mark a distinction between /u/ and /ɔ/ except in the same syllabus where the /ɔ/ is represented by [o] with a circumflex as [ô]. From the Botswana upper primary syllabus and beyond, written Setswana does not mark a distinction between /u/ and /ɔ/ and between

/ɛ/ and /e/ (Sebina and Otlogetswe 2023a and b). Since Setswana does not mark the /ɛ, e/ and /ʊ, ɔ/ distinction orthographically, there are no names that are spelt with [ê] or [ô]. Just as in common writing, Setswana names use the five orthographic vowels, a, e, i, o, u. This discussion of vowels is relevant to the vowel variations that follow below since several names do not follow standard orthography where vowels are concerned.

3.1 Names that use [i] instead of [e]

Name spelling variation in Setswana is found in names which in their articulation speakers use the half-close front vowel /e/, instead of the close front vowel /i/. This practice is then reflected in the spelling of the same name with one spelt with an /e/ while the same name is also spelt with /i/. Some speakers therefore spell their names as *Baaitsi* instead of *Baaitse*, *Matshidiso* instead of *Matshediso*, *Balibi* instead of *Balebi*, *Kelitile* instead of *Keletile*, *Baipidi* instead of *Baipedi*, *Baoliki* instead of *Baoleki*, *Bafiti* instead of *Bafeti* and *Goitsimodimo* instead of *Goitsemodimo*. This may be explained in terms of close vowel harmony where the preceding vowel harmonises in height with the following one. The /e/ is therefore pronounced as an /i/. The correct spelling of the name is traceable to the verb from which the name is derived. *Baaitse* is from *itse* "know", *Matshediso* from *tshela* "live", *Balebi* from *leba* "look", *Keletile* from *leta* "wait", *Baipedi* from *ipela* "celebrate", *Baoleki* from *leka* "try/test", *Bafeti* from *feta* "pass" and *Goitsemodimo* from *itse* "know".

The extracts from the dictionary in Table 2 illustrate how the dictionary treats both the wrong and the correct spellings.

Wrong spelling entry	Correct spelling entry
Balibi (the correct spelling is <i>Balebi</i> from the verb <i>leba</i> "look") <i>M</i> or <i>F</i> those who look.	Balebi <i>M</i> onlookers; those who are looking.
Baliki (the correct spelling is <i>Baleki</i> from the verb <i>leka</i> "try; attempt") <i>M</i> those who try; those who test someone or something.	Baleki <i>F</i> the trying ones.
Balisi (the correct spelling is <i>Balesi</i> from the verb <i>lesa</i> "leave; let go") <i>M</i> those who leave something.	Balesi <i>M</i> those who leave something.
Goitsimodimo (the correct spelling is <i>Goitsemodimo</i>) <i>M</i> it is God who knows.	Goitsemodimo <i>M</i> it is God who knows.

Table 2: Names that incorrectly use [i] instead of [e]

The wrong spelling entry is cross-referenced to the correct spelling entry though the correct spelling entry is not cross-referenced to the wrong spelling. The cor-

rect spelling entry could be improved by cross-referencing the wrong spelling entry to make the user aware of the misspelt version in the following manner:

Balebi *M* on lookers; those who are looking
 (usually misspelt as *Balibi*).

3.2 Names that use [e] instead of [i]

Other spelling variations are introduced in a reversal of the situation discussed in 3.1, where name givers use the half-close front vowel and write /e/ where the close front vowel /i/ should be used, thereby violating the agentive rule that derives -i verbs. Cole (1955: 113) has demonstrated that one way of forming Setswana nouns is from verb stems, by prefixing the class prefixes, while the final vowel of the stem is changed to the agentive marker -i (e.g. mo- + dira + -i = modiri). They, therefore, spell their names as *Gabaiphewe* instead of *Gabaiphiwe*, *Molapise* instead of *Molapisi*, *Baitshenyetse* instead of *Baitshenyetsi*, *Molebatse* instead of *Molebatsi*, *Kebaitsele* instead of *Kebaitsile*, and *Modiege* instead of *Modiegi*. This is possibly because the name givers gave the name by ear as in hypothesised in 3.2 and in their hearing, the /i/ sounded like an /e/. Table 3 further illustrates this:

Wrong spelling entry	Correct spelling entry
Molapise (the correct spelling is <i>Molapisi</i>) <i>M</i> one that tires others.	Molapisi (see <i>Molapise</i>) <i>M</i> one that tires others.
Molebatse (the correct spelling is <i>Molebatsi</i>) <i>M</i> one that makes you forget.	Molebatsi <i>M</i> one that makes you forget.
Modiege (the correct spelling is <i>Modiegi</i>) <i>F</i> one that delays.	Modiegi <i>F</i> one that delays.
Baitshenyetse (the correct spelling is <i>Baitshenyetsi</i>) <i>M</i> those that spoil things for themselves.	Baitshenyetsi <i>F</i> those that spoil things for themselves.

Table 3: Names that incorrectly use [e] instead of [i]

The wrong spelling entry is cross-referenced to the correct spelling entry to direct the users to the correct-spelt entry. The correct spelling entry is not cross-referenced to the wrong spelling entry. As in 3.1 above, the correct spelling entry could be improved by cross-referencing the wrong spelling entry to make the user aware of the misspelt version in the dictionary like this:

Molapisi (see *Molapise*) *M* one that tires others
 (usually misspelt as *Molapise*).

3.3 Names that use double [o] instead of [eo]

There are names that use [oo] where the standard orthography [eo] should be used. Such names are prevalent in the central and southern regions of Botswana. This spelling variation is typical of the pronunciation of central and southern Setswana dialects amongst the Ngwaketse, Rolong and Hurutshe. This spelling change mirrors vowel assimilation in Setswana speech where one vowel influences another — typically across syllable boundaries — so that they become more similar in quality. Vowel assimilation maintains phonotactic harmony and ease of articulation in speech. While this is true in speech, spelling rules should not be violated as they have been in some of the names. Vowel assimilation has therefore influenced the spelling of certain names such that we have *Kooagile* instead of *Keoagile*, *Koobake* instead of *Keobake*, *Koobokile*, instead of *Keobokile*, *Koolebile* instead of *Keolebile*, *Koorapetse* instead of *Keorapetse* and *Koobame* instead of *Keobame*. The sound change from [eo] to [oo] is a consequence of total regressive assimilation where a preceding sound becomes the same as the following one. For example,

Ke + obame = Koobame

This process helps maintain phonotactic harmony and ease of articulation in speech. In this instance, the [e] of [Ke] becomes like [o] of [obame] to result in *Koobame*. Those who spell the names with [oo] instead of [eo], therefore, write as they speak, and do not adhere to the Setswana standard orthography. *A Dictionary of Setswana Personal Names* has captured both these spelling variations as demonstrated in Table 4:

Wrong spelling entry	Correct spelling entry
Kooagile (the correct spelling is <i>Keoagile</i>) <i>M</i> I have built it.	Keoagile <i>M</i> I have built it.
Koobake (the correct spelling is <i>Keobake</i>) <i>M</i> that I should praise Him (God).	Keobake 1. <i>M</i> I should praise Him (God) 2. should I praise Him (God)? 3. (short for <i>Gakeobake</i>) I am not praising Him (God).
Koobame (the correct spelling is <i>Keobame</i>) <i>M</i> I have worshipped; I am bent.	Keobame 1. <i>F</i> I am bending forward 2. I worshipped.
Koobokile (the correct spelling is <i>Keobokile</i>) <i>M</i> or <i>F</i> I have praised Him (God).	Keobokile <i>M</i> I praised Him (God).

Table 4: Names that incorrectly use [oo] instead of [eo]

The wrong spelling entry is cross-referenced to the correct spelling entry though the correct spelling is not cross-referenced to the wrong spelling entry. The correct

spelling entry could be improved by cross-referencing the wrong spelling entry to make the user aware of the misspelt version in the following manner:

Keobame 1. *F* I am bending forward **2.** I worshipped
 (usually misspelt as *Koobame*).

3.4 Names that delete the vowel [e] in the context of the lateral approximant /l/

There is evidence of names where the vowel [e] gets deleted when it occurs between two lateral approximants /l/ as in the word *molelo* "fire", which is then spelt as *mollo*, after deleting the vowel [e], and retaining the second vowel [o]. This then results in the first /l/ becoming a tone-bearing syllabic consonant. This gives rise to names such as *Mollowakgosi* instead of *Molelowakgosi* and *Mollowakgotla* instead of *Molelowakgotla*. While the [e] is elided in speech, such an elision violates standard Setswana orthography (Chebanne et al. 2008: 64) that states that "... tlogelo ya tumanosi fa gare ga bo[l] ga e letlelelwe ..." (an elision of a vowel in between a sequence of the lateral /l/ is not permitted.) Table 5 further illustrates this:

Wrong spelling entry	Correct spelling entry
Mollowakgosi (the correct spelling is <i>Molelowakgosi</i>) <i>M</i> the chief's fire.	Molelowakgosi <i>M</i> the chief's fire.
Mollowakgotla (the correct spelling is <i>Molelowakgotla</i>) <i>M</i> the fire at the traditional court.	Molelowakgotla <i>M</i> the fire at the traditional court

Table 5: [e] elision between laterals

The wrong spelling entry is cross-referenced to the correct spelling entry though the correct spelling is not cross-referenced to the wrong spelling entry. The correct spelling entry could be improved by cross-referencing the wrong spelling entry to make the user aware of the misspelt version in the following manner:

Molelowakgotla *M* the fire at the traditional court
 (usually misspelt as *Mollowakgotla*).

4. Consonant variations in Setswana personal names

Some name variations are a result of a difference in consonantal use for the same part of a name. Before we delve deeper into consonantal differences, first we present the Setswana consonantal inventory. Setswana has 29 phonemic consonants (Batibo 2013 and DALL 1999: 12) as illustrated in Table 6.

		Labial	Alveolar		Post alveolar	Palatal	Velar	Uvular	Glottal
			Central	Lateral					
NASAL		/m/	/n/			/ɲ/	/ŋ/		
PLOSIVE	Unaspirated	/p/ /b/	/t/ /d/				/k/		
	Aspirated	/pʰ/	/tʰ/				/kʰ/	/qʰ/	
AFFRICATE	Unaspirated		/ts/	/tʃ/	/ʃ/ /dʒ/				
	Aspirated		/tsʰ/	/tʃʰ/	/ʃʰ/				
FRICATIVE		/f/	/s/		/ʒ/			/χ/	/h/
Trill			/r/						
Approximant		/w/		/l/		/j/			

Table 6: Setswana consonants

Four of these (p^h, t^h, k^h and q^h) are aspirated voiceless plosives, three are non-aspirated voiceless plosives (p, t and k), one voiced plosive /b/, three aspirated voiceless affricates (tʃ^h, tʃ^h, ts^h), three non-aspirated voiceless affricates (ts, tʃ, and ʃ), one voiced affricate (dʒ), five voiceless fricatives (f, s, ʃ, χ, and h), four nasals (m, n, ɲ, and ŋ), and four sonorants (r, l, j and w).

The voiced plosive /d/ is an allophone of /l/ where /d/ is usually followed by close vowel /u/ and /i/ as in *dika* (surround) and *huduga* (relocate) while /l/ is followed by open vowels such as /a/ and /o/ as in *lala* (spend the night) and *lole* (fought) (see Cole 1955: 28). The exception to this rule is borrowed words such as *lithara* (liter) and *dolara* (dollar).

The Setswana consonantal phonemes are represented orthographically with their equivalent forms as demonstrated in Table 7.

Phonetic	Orthography	/ts/	Ts	/ɲ/	ny
/m/	m	/ts ^h /	Tsh	/j/	y
/p/	p	/s/	S	/ŋ/	ng
/p ^h /	ph	/r/	R	/k/	k
/b/	b	/tʃ/	Tl	/k ^h /	kh
/f/	f	/tʃ ^h /	Tlh	/q ^h /	kg
w/	w	/l/	L	/χ/	g
/n/	n	/ʃ/	Tš	/h/	h
/t/	t	/dʒ/	J	Total	29
/d/	d	/ʃ ^h /	Tšh		
/t ^h /	th	/ʃ/	Š		

Table 7: Phonetic consonants and their orthographic forms

A discussion of Setswana consonantal inventory will prove relevant to sections 4.1–4.3 where consonantal variation in names is discussed.

The use of /h/ instead of /f/

Some name spelling variations in Setswana are because the same name is spelt either with a voiceless, labio-dental fricative [f] in some cases, while in other cases it is spelt with the glottal fricative [h]. Using the sound [h] instead of [f] is a common feature of Setswana speech and there is much confusion in written Setswana about where [h] and [f] should be used (Cole 1955: 26). There is, however, a Setswana morphological rule which illustrates that [h] verbs result with [kh] deverbative nouns and [f] verbs result with [ph] deverbative nouns in strengthening (Cole 1955: 40). For instance, the verb *hupela* generates the noun *khupelo* while the verb *fenya* generates the noun *phenyo*. The [h] and [f] spelling confusion are pervasive in personal names. This results in names such as *Bahedile* instead of *Bafedile*, *Bahemedi* instead of *Bafemedi*, *Gaehole* instead of *Gaefole*, *Hetolang* instead of *Fetolang* and *Hetanang* instead of *Fetanang*. One way of determining the correct spelling for each name is therefore to trace it to its original verb. Table 8 further illustrates how these are captured in the dictionary:

Wrong spelling entry	Correct spelling entry
Bahedile (the correct spelling is <i>Bafedile</i> from the verb <i>fela</i> "end; finish") <i>M</i> they are finished.	Bafedile <i>M</i> they are finished.
Bahemedi (the correct spelling is <i>Bafemedi</i> from the verb <i>fema</i> "defend") <i>M</i> the defenders.	Bafemedi <i>M</i> the defenders.
Bahentse (the correct spelling is <i>Bafentse</i> from the verb <i>fenya</i> "win; conquer") <i>F</i> they have conquered.	Bafentse <i>M</i> they have conquered, won.
Bahenyi (the correct spelling is <i>Bafenyi</i> from the verb <i>fenya</i> "win; conquer") <i>M</i> winners, victors, conquerors.	Bafenyi <i>M</i> winners; victors.
Henyang (the correct spelling is <i>Fenyang</i> from the verb <i>fenya</i> "win; conquer") 1. <i>M</i> win 2. Defeat.	Fenyang 1. <i>M</i> win 2. Defeat.
Hetanang (the correct spelling is <i>Fetanang</i> from the verb <i>feta</i> "pass; exceed") <i>M</i> pass each other.	Fetanang <i>M</i> pass each other.
Hetang (the correct spelling is <i>Fetang</i> from the verb <i>feta</i> "pass") <i>M</i> move on, pass.	Fetang <i>M</i> move on, pass.
Hetolang (the correct spelling is <i>Fetolang</i> from the verb <i>fetola</i> "change") 1. <i>M</i> reply, respond 2. Change, transform.	Fetolang 1. <i>M</i> reply, respond 2. Change, transform.

Table 8: Names that incorrectly use [h] instead of [f]

The dictionary entry indicates that the wrong spelling entry is cross-referenced to the correct spelling entry though the correct spelling is not cross-referenced to the incorrect spelling entry. The correct spelling entry could be improved by cross-referencing the wrong spelling entry to make the user aware of the misspelt version in the following manner:

Fetolang 1. M reply, respond 2. Change, transform.
(usually misspelt as *Hetolang*).

In many old Setswana names, instead of choosing either [h] or [f], name givers have used both consonants spelt [fh]. There are therefore many names which use [fh] instead of [f], such as *Gadifhele* instead of *Gadifele*, *Kefhilwe* instead of *Kefilwe*, *Molefhe* instead of *Molefe*, *Molefhi* instead of *Molefi*, *Nonofho* instead of *Nonofo*, *Baikalafhi* instead of *Baikalafi*, and *Merafhe* instead of *Merafe*. Table 9 further illustrates this:

Wrong spelling entry	Correct spelling entry
Gadilefhe (the correct spelling is <i>Gadifele</i> from the verb <i>lefa</i> "pay a charge") M they do not pay a fine or a charge.	Gadilefe M they do not pay a fine or a charge.
Merafhe (the correct spelling is <i>Merafe</i>) 1. M tribes, ethnic groups 2. Communities 3. Nations.	Merafe 1. M tribes, ethnic groups 2. Communities 3. Nations.
Nonofho (the correct spelling is <i>Nonofo</i>) M strength.	Nonofo M strength.
Molefhi (the correct spelling is <i>Molefi</i>) M one that pays damages.	Molefi M one that pays damages.
Kehilwe (the correct spelling is <i>Kefilwe</i>) F I have been given a gift	Kefilwe F I have been given a gift.

Table 9: Names that use [fh] instead of [f]

Standard Setswana orthography does not have /fh/. The table indicates that the wrong spelling entry is cross-referenced to the correct spelling entry although the correct spelling is not cross referenced to the wrong spelling entry. The correct spelling entry could be improved by cross-referencing the wrong spelling entry to make the user aware of the misspelt version in the following manner:

Merafe 1. M tribes, ethnic groups 2. Communities 3. Nations.
(usually misspelt as *Merafhe*).

4.1 The use of the [ny] and [n] variants

There is confusion amongst Setswana speakers between words that end with the causative [ny-] and the reciprocal [an-]. For instance, in daily speech there is an enduring confusion between *puisano* and *puisanyo* where *puisano* means a discussion or dialogue and *puisanyo* strictly speaking is an act of making each other talk. This confusion is reflected in many names which should end in [-n-] but end in [-ny-]. There is therefore *Kedibonye* instead of *Kedibone*, *Golekanye* instead of *Golekane*, *Kebabonye* instead of *Kebabone*, *Kemmony*, instead of *Kemmone*, *Gareitsanye* instead of *Gareitsane*, *Kelekanye* instead of *Kelekane* and *Lesanyang* instead of *Lesanang*. The dictionary captures both spelling variants in its macro-structure as shown in the sampled entries in Table 10.

Wrong spelling entry	Correct spelling entry
Kedibonye (the correct spelling is <i>Kedibone</i>) <i>F</i> I saw them, or I have seen them; I have been a victim of circumstances.	Kedibone <i>F</i> I saw them, or I have seen them; I have been a victim of circumstances.
Kemmony (the correct spelling is <i>Kemmone</i>) 1. <i>F</i> I have seen him or her 2. I have found him or her.	Kemmone <i>F</i> 1. I have seen him or her 2. I have found him or her.
Golekanye (the correct spelling is <i>Golekane</i>) <i>F</i> it is enough; it is sufficient.	Golekane <i>F</i> it is enough; it is sufficient.
Gareitsanye (the correct spelling is <i>Gareitsane</i>) <i>M</i> we do not know each other.	Gareitsane <i>M</i> we do not know each other.

Table 10: Names that incorrectly use [ny-] instead of [n-]

The wrong spelling entry is cross-referenced to the correct spelling entry even though the correct spelling is not cross referenced to the wrong spelling entry. The correct spelling entry could be improved by cross-referencing the wrong spelling entry to make the user aware of the misspelt version in the following manner:

Kedibone *F* I saw them, or I have seen them; I have been a victim of circumstances (usually misspelt as *Kedibonye*).

4.2 Names which should have a double [n], or double [m] but lack this feature

In Setswana orthography, the long nasal sound is represented orthographically by doubling the concerned nasal consonant, whether it is /n/ or /m/, while a short nasal is written as a single nasal consonant. Doubling the long nasal in orthography is supported by the fact that one of the nasal consonants is a tone

bearing syllabic nasal. Spelling variation is, however, introduced when there is an inconsistency in the writing of a long nasal. Some Setswana names that are pronounced with a long nasal are spelt with a single letter as if they use a short nasal which is represented by a single character. This has caused variation in the spelling of names, with some spelt with a double nasal while others have a single nasal character. The dictionary therefore has *Gaofengwe* /χáúfɪŋwí/ as well as the correctly spelt *Gaofenngwe* /χáúfɪŋwí/, and *Mamosadinyana* /mámósádɪjáná/ as well as the correctly spelt *Mmamosadinyana* /mmámósádɪjáná/ as shown in Table 11.

Wrong spelling entry	Correct spelling entry
Gaseikangwe (the correct spelling is <i>Gaseikanngwe</i>) <i>F</i> it is not trusted or relied on.	Gaseikanngwe <i>F</i> it is not trusted or relied on.
Molotsi (the correct spelling is <i>Mmolotsi</i>) <i>M</i> one who takes animals from the kraal.	Mmolotsi <i>M</i> one who takes animals from the kraal.
Mamosadinyana <i>F</i> (the correct spelling is <i>Mmamosadinyana</i>) queen or Queen of England.	Mmamosadinyana <i>F</i> the queen of England.

Table 11: Names that incorrectly miss a pre-nasal nasal consonant

The wrong spelling entry is cross-referenced to the correct spelling entry though the correct spelling is not cross-referenced to the wrong spelling entry. The correct spelling entry could be improved by cross-referencing the wrong spelling entry to make the user aware of the misspelt version in the following manner:

Mmolotsi *M* one who takes animals from the kraal. (usually misspelt as *Molotsi*).

5. The impact of Northern Setswana on name spellings

Some spelling variations in the dictionary were brought about by the influence of Northern Setswana. Northern Setswana is a cluster of dialects spoken by Bangwato, Batawana, and Bakhurutshe, amongst other groups who are found in the central, northern and parts of north-eastern Botswana (Cole 1955). Northern Setswana is sometimes referred to as Sengwato in the literature (Tlale 2005, Boyer and Zsiga 2013).

5.1 The loss of /t/ and /tʰ/

Historically, Northern Setswana uses the voiceless alveolar plosive /t/ where the

standard language uses the voiceless unaspirated affricate /tʃ/ and usually replaces the voiceless aspirated affricate /tʃh/ consonantal sound (Batibo 1999) with the voiceless aspirated explosive /tʃh/. This is reflected in many words in the Setswana vocabulary as in Table 12 that illustrates the different orthographic representations of some words between Northern Setswana and other Setswana dialects.

Standard Setswana orthography	Northern Setswana	English
tlhaga	Thaga	Grass
otlhaya	othaya	Discipline
setlha	Setha	Bladder
tlhaba	Thaba	stab, pierce
tla	Ta	Come

Table 12: Standard Setswana vs Northern Setswana dialect

This evolution has had a lasting influence on how speakers of the Northern Setswana dialect spell personal names. The spelling of the personal names, therefore, follows their pronunciation and not the Setswana standard orthography. The dictionary has included spellings of standard Setswana and that of Northern Setswana leading to a variation of names in which the same name is spelt with /tʃh/ and /tʃ/. For instance, *Tlhalefang* and *Thalefang*, *Atlholang* and *Atholang*, *Baikatlhodi* and *Baikathodi*, *Baiphethli* and *Baiphethi*, *Batlang* and *Batang*, *Batlhaetse* and *Bathaetse*, and *Batlhalefi* and *Bathalefi*. The Northern Setswana name spellings are captured in the dictionary as they are so spelt by their owners. This is in line with the recommendations of Setswana Standard Orthography of 1981 which notes that, "As a general rule, proper names of people should be spelled as used by the persons themselves" (p.19). The dictionary, however, cross references the standard spelling entry at each Northern Setswana spelling entry, for the benefit of the user. This is demonstrated in Table 13.

Northern Setswana spelling entry	Standard Setswana spelling entry
Atholang (Northern Setswana version of <i>Atlholang</i>) <i>M</i> pass/give judgement (pl.).	Atlholang <i>M</i> pass/give judgement (pl.).
Baikathodi (Northern Setswana version of <i>Baikatlhodi</i>) <i>M</i> those who judge themselves; those who find themselves guilty.	Baikatlhodi <i>M</i> those who judge themselves; those who find themselves guilty.
Molothanyi (Northern Setswana version of <i>Molotlhanyi</i>) <i>M</i> one who causes conflicts between others	Molotlhanyi <i>M</i> one who causes conflicts between others.

Mothagodi (Northern Setswana version of <i>Motlhagodi</i>) <i>M</i> one who cultivates; one who clears grass.	Motlhagodi <i>M</i> one who cultivates; one who clears grass.
---	--

Table 13: Northern Setswana names

5.2 Conflation of prefix *le-* and *lo-*

Northern Setswana also does not make a distinction between the noun class prefix /*lo-*/ of class 11 and the noun class prefix /*le-*/ of class 5. The distinctions have been collapsed into singular prefix [*le-*] of class 5. This difference has been discussed by Cole (1955). All the words which in other Setswana dialects belong to class 11 noun class, in Northern Setswana they are found in class 5. According to general Setswana books (e.g. Mogapi 1984; Cole 1955) and Setswana Standard Orthography, the determination of whether a word takes a [*le-*] or [*lo-*] prefix is based on the plural prefix that the word takes (Chebanne et al. 2008). When class 11 nouns are pluralised, they take class 8 prefix [*di-*] and when class 5 nouns are pluralised, they take the class 6 prefix [*ma-*] (Cole 1955; Chebanne et al. 2008). This may be summarised as follows:

Singular noun class prefix	→	Plural noun class prefix
noun class 5 [<i>le-</i>]	→	noun class 6 [<i>ma-</i>]
noun class 11 [<i>lo-</i>]	→	noun class 8 [<i>di-</i>]

Other Setswana dialects	Northern Setswana	English
logong	Legong	Wood
logopo	Legopo	Rib
loleme	Leleme	tongue
lobadi	Lebadi	scar
lofuka	Lehuka	feather

Table 14: [*lo-*] and [*le-*] noun variation

These /*le-*/ and /*lo-*/ variations have been extended to personal names, and as result we have the same name with different orthographic representations, for example, *Leungo/Loungo*, *Lerato/Lorato*, *Losika/Lesika*, *Lonyatso/Lenyatso*, *Logong/Legong* and *Loapi/Lewapi*. If we are to apply the pluralisation rule to determine the correct spellings of the names, then the variations of *Lotshwao*, *Loeto*, *Loungo* and *Legong* will be incorrect. This is because their pluralisation would flout the standard pluralisation rule. The first three names would take the noun class 8 prefix [*di-*] to become *ditshwao*, *dieto* and *diungo*, while the last word will take

the noun class 6 prefix [ma-] and become *magong*, all of which are linguistically non-existent in Setswana vocabulary. Table 15 further illustrates this as captured in the dictionary:

Correct spelling entry	Wrong spelling entry
Lebopo (also <i>Lobopo</i>) <i>M</i> the universe; creation.	Lobopo (also <i>Lebopo</i>) <i>M</i> creation; the universe.
Letshwao 1. <i>M</i> a sign 2. a symbol 3. a brand.	Lotshwao (the correct spelling is <i>Letshwao</i>) <i>M</i> a mark or sign.
Leungo <i>F</i> a fruit	Loungo (the correct spelling is <i>Leungo</i>) <i>F</i> a fruit
Leeto 1. <i>M</i> a visit 2. a journey.	Loeto (the correct spelling is <i>Leeto</i>) <i>M</i> a visit; a journey.
Logong <i>M</i> a piece of wood.	Legong (the correct spelling is <i>Logong</i>) <i>M</i> a piece of wood.

Table 15: [le] and [lo-] Setswana names

As previously stated, the Northern Setswana name spellings are captured in the dictionary as they are spelt by their owners, as recommended by the *Standard Setswana Orthography of 1981*, however, both names would be captured twice, one being the correct version.

The wrong spelling entry is cross-referenced to the correct spelling entry though the correct spelling is not cross-referenced to the wrong spelling entry. The correct spelling entry could be improved by cross-referencing the wrong spelling entry to make the user aware of the misspelt version in the following manner:

Leeto 1. *M* a visit 2. a journey. (usually misspelt as *Loeto*).

6. Discontinued consonants in Setswana names

An orthographic decision was taken to discontinue the use of certain consonants in the orthography of Setswana. A conference held in Johannesburg, South Africa on April 25, 1937 attended by representatives of the Education Departments of Bechuanaland Protectorate, the Cape, the Orange Free State, and the Transvaal, together with the Inter-University Committee for African Studies, resolved to change how it spelt two consonants: the voiceless postalveolar affricate [tʃʰ] and the voiceless postalveolar fricative /ʃ/. The conference resolved that [tʃʰ] which was then spelt as [ch], should henceforth be spelt as [tʃh], while /ʃ/ which was then spelt as [sh] was to be spelt as [ʃ]. It was noted that:

The symbol š, with its movable diacritic, was chosen by the Committee to represent the voiceless hushing fricative in preference of the I.P.A. symbol on the principle of avoiding new types, and to the symbol *sh* on account of the latter's ambiguousness and the difficulties it creates in connection with the representation of the hissing and hushing fricative combinations. It was felt that the symbol š possessed all the advantages of the I.P.A. symbol, without its disadvantages as a new type, and with the added advantage that its movable diacritic indicates, without exaggerating, the dialectal differences between various Tswana dialects among themselves and between Tswana and Transvaal Sotho and Basutoland Sotho in general. (Lestrade 1937: 8-9)

While this decision was taken over 80 years ago and has been adopted in much of Setswana writing patterns, it has had no impact on the spelling of Setswana personal names. Batswana still use [ch] and [sh] in the spelling of their names. No name in the Setswana database that informed the writing of *A Dictionary of Setswana Personal Names* used [tšh] or [š]. This may be because most Batswana are bilingual in Setswana and English which uses the [ch] and [sh] characters and therefore associate the [tʃʰ] and /ʃ/ sounds with the English [ch] and [sh] representations. The other reason may be related to advances in technology — that [š] is a more challenging character to write on a typewriter or computer. Consequently, [ch] and [sh] have endured in Setswana names. Names that use [ch] or [sh] are therefore entered in the dictionary as they are spelt by their owners. They are not spelt using the agreed post 1937 orthography. However, there is a note in the microstructure of the dictionary on the correct spelling as demonstrated in the examples below:

- Chaba** *M* (an old spelling of *Tšhaba*); nation.
- Chabadiile** *M* (an old spelling of *Tšhabadiile*) nations have gone.
- Chabaesele** *M* (an old spelling of *Tšhabaesele*) a different nation.
- Chabaditsile** *M* (an old spelling of *Tšhabaditsile*) nations have arrived.
- Chabaele** *M* (an old spelling of *Tšhabaele*) which nation?
- Cheleketo** *M* (old spelling of *Tšheleketlo*); a struggle; suffering.

Examples from the dictionary of names that use the discontinued [sh] instead of the standard orthography character [š] are shown as follows:

- Bashapi** (an old spelling of *Bašapi*) 1. *M* swimmers 2. those who beat.
- Bashi** (an old spelling of *Baši*) *M* short for *Basimane*; boys. A pet name for a boy from *Bashimane*.
- Ramasheleng** (the correct spelling is *Ramašeleng*) 1. *M* the father of shillings 2. the man of shillings
- Boshafatso** (an old spelling of *Bošafatso*) *M* renewal.
- Boshula** (an old spelling of *Bošula* — the standard spelling of this word is *Bosula*) 1. *M* bad taste 2. evil
- Shianyana** (the correct spelling is *Šianyana*) *M* small boy.

This note on the correct spelling must not be misconstrued as a cross-reference. It does not point to another entry elsewhere in the dictionary. It merely states how an entry would be written in Setswana standard orthography. Entries with discontinued characters [sh] and [ch] are therefore entered in the dictionary exactly the way the owners of the names spell them.

7. Conclusion

This paper has examined spelling variations in *A Dictionary of Setswana Personal Names*, highlighting the phonological and morphological patterns that underlie these differences. It has shown that many of these variations stem from the influence of distinct Setswana dialects — particularly the northern and southern varieties — on the orthographic representation of personal names. While general dictionaries tend to be normative, adhering to standardized orthographies, this study reveals that formal spelling rules are sometimes bypassed in favor of forms that more closely reflect speakers' individual pronunciations.

Despite the introduction of the phoneme /š/ into Setswana orthography over eighty years ago, this study found no personal names utilizing this symbol. Instead, speakers continue to use [ch] and [sh], which are considered outdated under current orthographic standards. *A Dictionary of Setswana Personal Names* documents these spelling variants as provided by name givers or bearers, aligning with the 1981 Standard Setswana Orthography's recommendation that personal names be recorded as they are used by individuals.

Ultimately, this paper underscores the ubiquity of variation in natural languages and emphasizes the need for lexicographic strategies that enhance user accessibility. One such strategy proposed here is the documentation of personal names in their authentic forms, supplemented by cross-referencing and informational annotations to guide users through variant spellings.

References

Dictionaries

- Kgasa, M. and J. Tsonope. 1995. *Thanodi ya Setswana*. Gaborone: Longman.
Otlogetswe, T.J. 2012. *Tlhalosi ya Medi ya Setswana*. Gaborone: Medi Publishing.
Otlogetswe, T.J. and G.N. Ramaeba. 2024. *A Dictionary of Setswana Personal Names*. Gaborone: Mokala Publishing.

Online Sources

- African Pidgin Dictionary*.
www.pegeen.com
Spelling Variants: Why Are They Important?
<https://blog.spellquiz.com/spelling-variants>

Academic papers and books

- Batibo, H.M.** 1999. The Grammaticalization Process in Setswana. *Afrikanistische Arbeitspapiere (AAP)* 59: 131-150.
- Batibo, H.M.** 2013. The Evolution and Adaptation of Swahili and Tswana Syllable Structures. Legère, K. (Ed.). 2013. *Bantu Languages and Linguistics: Papers in Memory of Dr Rugatiri D.K. Mekacha*. Bayreuth African Studies 91: 13-34. Bayreuth: Bayreuth University Press.
- Baugh, J.** 2009. Linguistic Diversity, Access, and Risk. *Review of Research in Education* 33(1): 272-282.
- Boyer, O. and E. Zsiga.** 2013. Phonological Devoicing and Phonetic Voicing in Setswana. Orie, Olanike Ola and Karen W. Sanders (Eds.). 2013. *Selected Proceedings of the 43rd Annual Conference on African Linguistics (ACAL): Linguistic Interfaces in African Languages, New Orleans, USA, 15-17 March 2012*: 43: 82-89. Somerville, MA: Cascadilla.
- Chabata, E.** 2003. Linguistic Variation in Shona with Special Reference to Monolingual Dictionaries. *Lexikos* 13: 81-91.
- Chebanne, A., T.J. Otlogetswe, P. Seloma, A. Molosiwa, B. Moatlhodi, N. Kgolo and M. David.** 2008. *Mokwalo o o lolameng wa Setswana*. Gaborone: Pentagon Publishers.
- Cole, D.T.** 1955. *An Introduction to Tswana Grammar*. Cape Town: Longman.
- DALL (Department of African Languages and Literature), University of Botswana.** 1999. *The Sound System of Setswana*. Gaborone: Lightbooks.
- Lestrade, G.P.** 1937. *The Practical Orthography of Tswana — Recommendations of the Tswana District Orthography Committee. Approved by the Union Advisory Committee on African Studies and Research*.
- Lichtenstein, H.** 1928. *Travels in Southern Africa in the Years 1803, 1804, 1805 and 1806*. (A reprint of the translation from the original German by Anne Plumptre). Translated by Anne Plumptre. Cape Town: The Van Riebeeck Society.
- Lichtenstein, H.** 1973. *Foundation of the Cape (1811) and About the Bechuanas (1807)*. Edited and translated by O.H Spohr. Cape Town: A.A. Balkema.
- Livingstone, D.** 1875-1885. *Missionary Travels and Researches in South Africa*. London: J. Murray.
- Mareme, G.B.** 2007. *Thanodi ya Setswana*. Pietermaritzburg: Nutrend Publishers.
- Ministry of Education.** 1981. *Setswana Standard Orthography of 1981*. Gaborone: Ministry of Education, Government of Botswana.
- Mogapi, K.** 1984. *Thutapuo ya Setswana*. Gaborone: Longman.
- Motzafi-Haller, P., K. Keuthmann and R. Vossen.** 2006. Setswana Dialects and Inter-dialectal Variation in the Republic of Botswana. I. Setswapong. *Zeitschrift der Deutschen Morgenländischen Gesellschaft* 156(1): 9-41.
- Mulingi, Y.** 2023. *Development of Models for Non-word Error Detection for and Correction System for Kiswahili Language*. Master's dissertation, Kampala International University, Uganda.
- Murray, R.W.** 1996. Historical Linguistics: The Study of Language Change. O'Grady, W., M. Dobrovolsky and F. Katamba. 1996. *Contemporary Linguistics. An Introduction*: 313-371. London/New York: Longman.
- Otlogetswe, T.J.** 2014. Extracting Business Terms for Dictionary Subject Label. Khumalo, L. (Ed.). 2014. *African Languages and Linguistic Theory: A Festschrift in Honour of Professor Herbert Chimhundu*: 268-286. Cape Town: CASAS.
- Otlogetswe, T.J.** 2015. Treatment of Spelling Variants in Setswana Monolingual Dictionaries. *Lexikos* 25: 262-284.

- Prah, K.** 1998. *Between Distinction and Extinction: The Harmonization and Standardization of African Languages*. Cape Town: CASAS.
- Prinsloo, D. and N.D. van Graan.** 2021. Principles and Practice of Cross-referencing in Paper and Electronic Dictionaries with Specific Reference to African Languages. *Lexicography, Journal of ASIALEX* 8(1): 32-58.
- Scontras, G., Z. Fuchs and M. Polinsky.** 2015. Heritage Language and Linguistic Theory. *Frontiers in Psychology* 6: 1545.
- Sebina, B. and T.J. Otlogetswe.** 2023a. Vowel Sex-based Symbolism in Setswana Personal Names. *Studies in African Linguistics* 52(13): 30-44.
- Sebina, B. and T.J. Otlogetswe.** 2023b. Syllable Structure in Setswana Personal Names. *Voprosy onomastiki* 20(2): 174-192.
- Tlale, O.** 2005. *The Phonetics and Phonology of Sengwato, a Dialect of Setswana*. Unpublished PhD thesis, Georgetown University, Washington, DC.

Lexicographic Post-processing and Refinement Issues in Wordnets: The Ukrainian Wordnet Experience

Maksym Vakulenko, *Darmstadt University of Applied Sciences,
Darmstadt, Germany*
(maxvakul@yahoo.com)
(<https://orcid.org/0000-0003-0772-7950>)

Abstract: This paper considers challenges arising within the development of the Ukrainian wordnet, specifically focusing on the integration and alignment of vocabularies originating from diverse sources, such as conventional dictionaries, the Open English Wordnet, and other wordnets. Various cases of synonymy and hyper-hyponymy presented in the Open English Wordnet and in conventional terminological and phonetic dictionaries were analyzed, as well as the definitions of terms in these sources were compared. It was found that traditional dictionaries do not contain all the necessary hyper- and hyponymic relations, and definitions in wordnets are often inaccurate or incomplete. It is shown that creating wordnets based on the "merge" and combined approaches, exemplified by Ukrajinet, necessitates specific lexicographic post-processing. This procedure involves several crucial steps, including incorporating additional lexical units for lacking concepts, refining definitions, and introducing novel concepts and semantic structures, particularly those associated with hyper- and hyponymy, aligned with revised definitions.

Keywords: WORDNETS, UKRAJINET, CONCEPT, DEFINITION, LEXICAL-SEMANTIC RELATIONS, SYNONYMY, HYPER-HYPONYMY, LEXICOGRAPHIC DATA

Opsomming: Leksikografiese postverwerking en verfyningskwessies in woordnette: Die Oekraïense woordnet-ervaring. Hierdie bydrae beskou uitdagings wat binne die ontwikkeling van die Oekraïense woordnet (Ukrajinet) ontstaan, en fokus spesifiek op die integrasie en belyning van woordeskat wat afkomstig is van diverse bronne, soos konvensionele woordeboeke, die Open English Wordnet en ander woordnette. Verskeie gevalle van sinonimie en hipero-hiponimie wat in die Open English Wordnet en in konvensionele terminologiese en fonetiese woordeboeke aangebied word, is geanaliseer. Die definisies van terme in hierdie bronne is ook met mekaar vergelyk. Daar is bevind dat tradisionele woordeboeke nie al die nodige hipero- en hiponimiese verwantskappe bevat nie, en definisies in woordnette is dikwels onakkuraat of onvolledig. Daar word aangetoon dat die skep van woordnette gebaseer op die "samesmeltings-" en gekombineerde benaderings, soos geïllustreer deur Ukrajinet, spesifieke leksikografiese postverwerking noodsaak. Hierdie prosedure behels verskeie belangrike stappe, insluitend die inkorporering van bykomende leksikale eenhede vir ontbrekende konsepte, die verfyning van definisies en die bywerking van nuwe konsepte en semantiese strukture, veral dié wat met hipero- en hiponimie geassosieer word sodat dit met hersiene definisies strook.

Slutelwoorde: WOORDNETTE, UKRAJINET, KONSEP, DEFINISIE, LEKSIKAAL-SEMAN-
TIESE VERHOUDINGS, SINONIMIE, HIPERO-HIPONIMIE, LEKSIKOGRAFIESE DATA

1. Introduction

The rapid development of linguistic resources such as wordnets (Fellbaum 1998; Bond et al. 2016; McCrae et al. 2020; Siegel and Bond 2021) caused a breakthrough in lexicography, providing multilingual lexical databases containing definitions, use examples, lexical-semantic relations (LSRs), etc. Wordnets are useful in various natural language processing tasks, such as machine translation and semantic text analysis. In general, there are two approaches to developing wordnets: "expansion" and "merge", where the first is based on translations of existing lemma forms and the second foresees building semantic structures of the language in question (synonymy, hyper-/hyponymy, meronymy, and antonymy) with subsequent linking of concepts through interlingual indices — ILIs (Bond et al. 2016: 50-51). Then the development of wordnets based on the "merge" and combined ("expansion" and "merge") approaches necessitates the joined use of lexicographic information available in existing wordnets, mainly Open English Wordnet (OEWN)¹, and traditional dictionaries containing lexical units and their definitions.

In particular, the first version of the Ukrainian wordnet, Ukrajinet 1.0², was created based on the merge approach, making use of the "Explanatory Dictionary on Physics" (TSF)³ and the "Dictionary of Physical Synonymous Terms" (SFS)⁴. As the follow-up of this project, the 750 concept definitions and 700 examples of use have been added so far, and this resource was supplemented with a translation of OEWN into Ukrainian. The definitions were taken from the "Dictionary of Ukrainian Language in 20 volumes" (SUM)⁵, TSF, and OEWN. The current version of Ukrajinet contains 153 683 lexical entries (single and multi-word lexemes) combined in 124 214 synsets.

The linguistic data coming from different sources pose "a particular challenge for annotators" (Rudnicka et al. 2019: 310), requiring tedious refinement and adjustment. Extremely difficult is to find corresponding concepts in different languages where some of them may be absent, and others may differ in their scope that requires thorough analyzing definitions and use examples. For example, the concept i29195 "put" — 'cause to be in a certain state; cause to be in a certain relation' — corresponds to several Ukrainian concepts, in particular: 1 *викликати [в когось настрій/стан]* 'put [someone in a mood/state]'; 2 *оформляти* 'put [something] in form'; 3 *упорядковувати* 'put [something] in order'.

The most problematic are here the lexical units that enter the corresponding LSRs, and their definitions. For linking wordnets, the most important relations are synonymy and hypernymy, where the greatest number of equivalents across synsets comes from absolute synonymy (Rudnicka et al. 2019). So, synonymy plays a fundamental role in creating wordnets. It is remarkable that if synonymy is defined on the basis of hypernymy, treating hypernymy as the fundamental relation, the resulting synsets would differ (Halas Popović 2023).

Synonymy is widespread in common language and also in various terminologies (Vakulenko 2018; Bokshanj 2021). However, this concept is treated differently in different sources. For example, the Longman dictionary⁶ defines a synonym as "a word with the same meaning as another word in the same language" (Longman: 1685), where only lexical synonymy is considered and no partial synonymy and cross-language synonymy are recognized.

In contrast, Löbner understands synonymy in a way that:

"Two expressions are synonymous if they have the same meaning" (Löbner 2013: 203). This is an interesting approach where the synonymy is taken more generally, involving expressions instead of a single word and depending on their meanings.

The more general definition is based on the notion of a concept and suggests that: "Synonyms are the units of a language system (morphemes, lexemes, syntactic constructions) that within usually one linguistic level represent the same concept, and that can be substituted in the context without distorting the general sense" (Vakulenko 2018: 36). A similar approach is adopted for wordnets (Fellbaum 2006; Rudnicka et al. 2019: 299) where synonymy is considered in a narrower sense as restricted to the lexical level.

Another approach to synonymy that is aimed at meeting the lexicographer's requirements, derives synonymy from the so-called constitutive relations such as hypernymy/hyponymy, meronymy/holonymy, and antonymy (Maziarz et al. 2013; see also Rudnicka et al. 2019: 300).

The theory of synonymy in the context of terminological research was further developed in Vakulenko (2018: 40-43) where it was proposed to discern formal synonymy (covering lexical and grammatical synonymy) and contentual one. In particular, terminological dictionaries in the fields of physics (WPEDFR⁷; WPED⁸; TSF; SFS) and phonetics (Yurtbaşı 2017a; 2017b; Vakulenko 2018: 49-55) contain various kinds of full synonyms resulting from these types of synonymy. In what follows, these notions will be used.

A large problem in developing wordnets, particularly Ukrajinet, is the coordination and adjustment of lexicographic data coming from different sources. This problem will be addressed here. Synonymy and hyper-/hyponymy in wordnets will be considered and traditional dictionaries in connection with the existence of "hidden" senses not explicitly displayed in the last sources. A comparative analysis will be provided of definitions available in OEWN in comparison with those coming from academic dictionaries.

The article is organized as follows. In Section 2, we describe the method and material used for this research. Section 3 is devoted to the study of synonymy and hypernymy/hyponymy as they appear in traditional terminological dictionaries. In Section 4, we carry out a comparative analysis of term definitions available in OEWN and traditional dictionaries. Section 5 provides a discussion of related issues and future plans. The concluding remarks are presented in Section 6.

2. Methods and material

The statistical (in the symptomatic sense) and analytical methods of terminology science (see Vakulenko 2018) are applied to carry out the comparative analysis of terminological units and their definitions coming from conventional dictionaries in physics and phonetics (WPEDFR; WPED; TSF; Yurtbaşı 2017a; 2017b), with those of OEWN. The statistical method determines in this case the availability of terms and their definitions, as well as the use examples, while the analytical method reveals scientific quality of these. The study is restricted to English, German, and Ukrainian lexical units, and uses standard conventions for traditional dictionaries. Within this lexicographic treatment, the units appearing in the same terminological article, are considered full synonyms, and the nouns in the word combinations which are specified by preceding adjectives (modifiers), are considered hypernyms. For example, *shift* is a hypernym for *Doppler shift*. As for definitions, the term to which the described term is referred is regarded as its hypernym.

3. Synonymy and hypernymy

Approximately 10% of concepts available in the physical dictionaries (WPEDFR; WPED; TSF) have synonyms. The rise of lexical synonymy is conditioned by a certain degree of semantic similarity of concepts whose hyponyms have identical meanings:

- Doppler **shift** – Doppler **displacement** – Doppler **frequency shift** (WPEDFR: 375);
- Ger. Polschuh**abstand** m – Luftspalts**länge** f – Luftspalts**breite** f (WPEDFR: 570);
- **initiation** of the discharge – **ignition** of the discharge (WPEDFR: 726);
- level of significance – significance **level** – error **probability** (WPEDFR: 838);
- **level** of significance – confidence **level** – **degree** of confidence – confidence **probability** – confidence **coefficient** (WPEDFR: 838);
- Ger. statistische Sicherheit – Sicherheitss**chwelle** f – Sicherheits**grad** m – Vertrauens**koeffizient** m – Konfidenz**koeffizient** m – Konfidenz**niveau** n – Konfidenz**wahrscheinlichkeit** f – Bedeutsamkeit**stufe** f (WPEDFR: 838);
- contact **clearance** – contact **gap** – contact **spacing** (WPED: 594);
- **monitor** – monitoring instrument – survey **instrument** – survey **meter** (WPED: 594);
- Ger. spiegelnder **Körper** – weisses **Objekt** (WPED: 306);
- sighting **apparatus** – sighting **device** (WPED: 594);
- Ger. Visiervorrichtung f – Zielgerät n (WPED: 594);
- **density** – specific **weight**, optical **isomer** – optical **antipode** (TSF: 165);
- light-beam **oscillograph** – light-beam **instrument** (TSF: 165);
- **argument** – independent **variable**, **momentum** – **quantity** of motion (Vakulenko 2018: 41);

- *magnetometer* – *magnetic(-field) meter* – *magnetic detector* – *gaussmeter* (TSF: 296-297);
- *approximant* – *frictionless continuant* (Vakulenko 2018: 53);
- *diminuendo accent* – *falling accent* – *diminuendo stress* – *falling stress* (Vakulenko 2018: 53).

In these examples, there is a clear semantic correlation between the head terms, including hypernymy/hyponymy (*shift* – *frequency shift*, *instrument* – *monitor*, *instrument* – *meter*, *weight* – *density*, *instrument* – *oscillograph*, *variable* – *argument*, *quantity* – *momentum*, *meter* – *detector*, *continuant* – *approximant*) and synonymy (*shift* – *displacement*, *level* – *degree*, *accent* – *stress*). It can be seen from this that a necessary condition for the synonymy of two lexemes is the presence of hyponyms in the corresponding concepts that are part of the same synset. So, the equivalence of hyponyms is caused by synonymy of their hypernyms. As can be seen from the synonymous pair *Doppler shift* – *Doppler frequency shift*, the opposite is correct only when the hypernym (*shift*) has an additional meaning identical to the meaning of its hyponym (*frequency shift*).

Grammatical variability also often causes synonymy, primarily for compound terms. This is exemplified by the following series of synonyms: *unshared electron pair* – *electron lone pair* – *lone pair* – *lone-pair electrons* (TSF: 394). Here the element *pair* that is a hypernym to *electron lone pair*, can be considered as a hypernym to the unit *electrons* that is a hypernym to *lone-pair electrons*.

The situation is somewhat different for the **contentual synonymy** that arises when different terms reflect different aspects of the concept or different approaches to its study (Vakulenko 2018: 41). This kind of synonymy is also typical for composed terms, i.e. when some senses of their hypernyms are neutralized. In traditional dictionaries, the corresponding head terms for synset members often display either little semantic similarity or complicated and unstraightforward semantic relations:

piezo effect – *piezoelectric effect* – *polarization* of distortion (WPED: 594);
Ger. *schwarzer Körper* – *schwarzer Strahler* 'blackbody radiator' (WPED: 22);
density matrix – *statistical matrix* – *statistical operator* (TSF: 305);
Gibbs energy – *Gibbs thermodynamic potential* – *isobaric-isothermal potential* – *free enthalpy* (TSF: 166);
thermal electron emission – *Richardson effect* – *filament emission* (TSF: 165);
Planck constant – *quantum* of action (TSF: 608);
isotopic indicator – *tracer* – *tracer isotope* – *label* – *label(l)ed atom* (TSF: 225);
freezing point – *freezing temperature*; *dielectric loss coefficient* – *dielectric loss factor* – *dielectric loss index* – *dielectric power factor* – *dielectric dissipation factor* – *dielectric loss tangent* (Vakulenko 2018: 43).

Historically, these terms entered the conventional dictionaries due to their practical use that from the standpoint of terminological theory is the result of the statistical method in the symptomatic sense (Vakulenko 2018: 14-19). However, the

appearance of synonymy between composed terms assumes synonymy or hypernymy between head terms describing corresponding concepts. In other words, the conceptual approach underlying the development of wordnets necessitates the introduction of extra meanings of these head terms that have not been explicitly displayed in conventional dictionaries. Such augmentation of the human lexicographic practice manifests the application of the analytical method of terminology science (Vakulenko 2018: 14-19).

While the relative number of such correlated and under-refined concepts is an open question, their existence was already observed while working directly on Ukrajinet.

In particular, the synset **omw-1354-n** contains lexemes "НВЧ-генератор" 'microwave oscillator' and "лампа зворотної хвилі" 'backward-wave tube' which, according to the TSF, belong to the nests of different terms: "генератор" 'oscillator' (**omw-124123-n**) and "лампа" 'tube' (**omw-1617-n**), respectively. The synset **omw-689-n** (i61958) contains lexemes "гомоморфізм" 'homomorphism' and "відображення гомоморфне" 'homomorphic mapping' which, according to the TSF, are also included in different term nests: "подібність" 'similarity' (**omw-100-n**) and "відображення" 'mapping' (**omw-54804-n**). In such cases, lacking senses are added to the corresponding nesting terms.

4. Dictionary definitions vs OEWN definitions

The degree of semantic similarity between hypernyms to the abovementioned synsets cannot be easily calculated, as it strongly depends on the corresponding definitions (see Vakulenko 2019). For example, the Open English Wordnet treats energy as "a thermodynamic quantity equivalent to the capacity of a physical system to do work", connecting this concept with the hypernym "quantity" (<https://en-word.net/lemma/energy>). In turn, the TSF defines energy as "a general measure of various processes and types of interaction that allows measuring all kinds of physical forms of movement and interaction", relating this concept to the hypernym "measure". Using the latter is recommended, because some definitions of physical and technical terms available in the OEWN, fail to correctly convey significant features of the defined concepts such as those with the following ILIs.

Incorrect OEWN definitions

Determinant (i80590): 'a square matrix used to solve simultaneous equations'. This definition is completely incorrect. It is recommended to use the latter formulation from the TSF that is freely available and defines this concept as 'a polynomial, each term of which is a product of matrix elements, and only one coefficient is included in the product from each row and each column of the matrix' ['поліном, кожен член якого є добутком елементів матриці, причому з кожного рядка і кожного стовпця матриці в добуток входить лише один співмножник'].

Field, field of force, force field (i97792): 'the space around a radiating body within which its electromagnetic oscillations can exert force on another similar body not in contact with it'. This definition is not correct, because the force field is not about a body radiating electromagnetic oscillation. The Longman (which is freely available) defines this concept better as 'the area in which the natural force is felt or has an effect'. The TSF treats this concept more comprehensively, as 'a part of space (limited or unlimited), at each point of which a force with a certain numerical value and direction acts on a material particle located there, which depends only on the coordinates of this point' ['частина простору (обмежена чи необмежена), у кожній точці якої на розміщену там матеріальну частинку діє певна за числовим значенням і напрямком сила, що залежить тільки від координат цієї точки']. The last definition is recommended.

Luminescence (i97901): 'light not due to incandescence; occurs at low temperatures'. This description is not relevant to luminescence. However, the Longman also does not highlight essential properties of luminescence, defining this phenomenon as 'a soft shining light'. The TSF treats luminescence as 'radiation, which is an excess over the thermal radiation of the body and continues for a time that significantly exceeds the period of light oscillations' ['випромінювання, що являє собою надлишок над тепловим випромінюванням тіла і продовжується протягом часу, який значно перевищує період світлових коливань'] that should be used in wordnets.

Paramagnetism (i97922): 'materials like aluminum or platinum become magnetized in a magnetic field but it disappears when the field is removed'. This definition is incorrect, because paramagnetism is a property, not a material. The TSF contains a definition that should be used instead: 'the property of substances (paramagnets) to be magnetized in the direction of an external magnetic field' ['властивість речовин (парамагнетиків) намагнічуватися в напрямку зовнішнього магнітного поля'].

Vibration, oscillation (i75292): 'a regular periodic variation in value about a mean'. These terms are not synonyms, as can be seen from the Longman that proposes two senses for vibration(s): 1 'a continuous slight shaking movement'; 2 'vibes'; and four senses for oscillation(s): 1 'frequent changes between two extreme amounts of limits'; 2 'frequent changes between one feeling or attitude and another'; 3 'a regular movement of something from side to side'; 4 'a regular change in direction of an electrical current'. This treatment is much more relevant and should be applied in wordnets.

Inexact OEWN definitions

Elasticity (i63383): 'the tendency of a body to return to its original shape after it has been stretched or compressed'. This definition uses common language and lacks scientific exactness. The Longman suggests a somewhat better definition: 'the ability of something to stretch and go back to its usual length or size'. The

TSF contains an even more exact formulation: 'the ability of a material or product to undergo more or less significant springy reversible deformations without destruction' ['спроможність матеріалу або виробу зазнавати більш або менш значних пружних зворотливих деформацій без руйнування']. This last definition should be used in wordnets.

Flexure, flexion, flection (i110688): 'the state of being flexed (as of a joint)'. This definition refers the concept to the corresponding verb and is, therefore, not sufficient. The TSF defines this concept as 'a type of deformation characterized by a change in the curvature of the axis (bar, balk, rod) or the middle surface (plates, shells) under the influence of external force or temperature' ['вид деформації, що характеризується зміною кривизни осі (бруса, балки, стержня) або серединної поверхні (пластинки, оболонки) під дією зовнішньої сили або температури'] that is much more informative and can be recommended.

Fluorescence (i97800): 'light emitted during absorption of radiation of some other (invisible) wavelength'. This definition does not emphasize essential features of fluorescence and refers to light which is correct but not complete. The TSF gives a definition that refers fluorescence to luminescence and describes its essential features and, therefore, may be recommended: 'short-term luminescence with decay time of 10^{-8} – 10^{-9} sec' ['короткочасна люмінесценція з часом загасання 10^{-8} – 10^{-9} сек'].

Loupe (i55838): 'small magnifying glass (usually set in an eyepiece) used by jewelers and horologists'. This definition is too narrow. The TSF gives a more general and comprehensive definition, which is recommended: 'an optical system consisting of a lens or several lenses, designed to observe small objects located at a finite distance, under an increased angle of view' ['оптична система, що складається з лінзи або кількох лінз, призначена для спостереження дрібних предметів, розташованих на скінченній відстані, під збільшеним кутом зору'].

Resistance, impedance, ohmic resistance, resistivity, electric resistance, electrical resistance (i98048): 'a material's opposition to the flow of electric current; measured in ohms'. This definition makes no distinction between direct and alternating currents, which is crucial here. In fact, impedance is analogical to resistance but represents a different concept. The Longman defines impedance as 'a measure of the power of a piece of electrical equipment to stop the flow of an alternating current'. The TSF gives a more precise definition of impedance: 'an analogue of electrical resistance for harmonic processes' ['аналог електричного опору для гармонічних процесів'], which is recommended.

Selectivity (i64341): 'the property of being selective'. The TSF gives another definition used in optics that should be also present in wordnets: 'the ability of a spectral device to distinguish narrow spectral intervals from a continuous spectrum' ['спроможність спектрального прилада виділяти вузькі спектральні інтервали з суцільного спектру'].

Sensing element, sensor, detector (i52820): 'any device that receives a signal or stimulus (as heat or pressure or light or motion etc.) and responds to it in a distinctive manner'. This formulation is too broad, and the term "detector" does not have this sense. The Longman defines sensor similarly but incompletely, as 'a piece of equipment used for discovering the presence of light, heat, movement etc.', and detector differently, as 'a machine or piece of equipment that finds or measures something'. The TSF defines this concept as 'a converter of a controlled or regulated value into an output signal, convenient for remote transmission and further processing' ['перетворювач контрольованої чи регульованої величини у вихідний сигнал, зручний для дистанційної передачі та подальшої обробки'], and detector as 1 'a device for detecting certain particles or radiations' ['прилад для виявлення тих чи інших частинок або випромінювань'] and 2 'a device for converting modulated high-frequency oscillations into oscillations with a modulation frequency' ['прилад для перетворення модульованих височастотних коливань у коливання з частотою модуляції']. These definitions are recommended for the use in wordnets.

Set, place, pose, position, put, lay (i29197): 'put into a certain place or abstract location'. This definition is not well-formed because it repeats the defined concept ("put"). The Longman avoids this inconsistency: 'to move something to a particular place or position, especially using your hands'. This definition should be recommended for wordnets.

Undulation, wave (i75294): '(physics) a movement up and down or back and forth'. This definition is too simplified to convey notable information. The Longman proposes a more relevant but too narrow definition: 'a line of raised water that moves across the surface of the sea'. The more strict definition is available in the "Explanatory dictionary of physics": 'a change in some set of physical quantities (fields), capable of moving (spreading), moving away from the place of its origin, or oscillating within limited regions of space' ['зміна деякої сукупності фізичних величин (полів), здатна переміщуватися (поширюватися), віддаляючись від місця її виникнення, або коливатися всередині обмежених областей простору'] which we recommend.

Incomplete OEWN definitions

Flint (i115216): 'a hard kind of stone; a form of silica more opaque than chalcedony'. The Longman distinguishes two meanings: 1 'a type of smooth hard stone that makes a small flame when you hit it with steel'; 2 'a piece of this stone or a small piece of metal that makes a small flame when you hit it with steel' and the TSF adds a sense 'optical glass that contains lead oxide and thus has a refractive index and its dispersion greater than, for example, crown' ['оптичне скло, яке містить окис свинцю і завдяки цьому має показник залому та його дисперсію більші, ніж, наприклад, крон']. The wordnet definitions should include all these meanings.

Graphite, black lead, plumbago (i114791): 'used as a lubricant and as a moderator in nuclear reactors'. This is not a complete definition, but rather a part of it. Another and essential part is available in the "Longman Dictionary of Contemporary English": 'a soft black substance that is a kind of carbon, used in pencils, paints, and electrical equipment'. The combined formulation for graphite is recommended: 'a soft black substance that is a kind of carbon, used in pencils, paints, and electrical equipment, and that serves as a lubricant and as a moderator in nuclear reactors'.

Median (i68434): 'the number midway between the two middle numbers in a series containing an even or odd number of items'. This definition is too complicated and encompasses only one sense of this polysemant. In comparison, the Longman distinguishes three senses of this noun: 1 'a narrow area of land that separates the two sides of a big road in order to keep traffic travelling in different directions apart'; 2 'the middle number of measurement in a set of numbers or measurements that have been arranged in order'; 3 'a line passing from one of the points of a triangle to the centre of the opposite side'. The TSF discerns also three senses of this term: 1 'the line that bisects the opposite side of the polygon' ['лінія, що ділить протилежну сторону багатокутника навпіл'], which is absent in OEWN and is more general than sense 3 of Longman; 2 'a point with the smallest sum of distances from given points' ['точка з найменшою сумою відстаней від заданих точок'], which is absent in OEWN and Longman; 3 'the value in the middle of a certain set of values' ['середнє значення певного набору величин'], which correlates with the OEWN and Longman definitions but is more comprehensive and clear. It is recommended to use the definitions of median given in the "Explanatory dictionary of physics".

Oscillator (i56791): 'generator that produces sonic oscillations or alternating current'. This treatment is too narrow. The Longman contains a more exact formulation: 'a machine that produces electrical oscillations', whereas the TSF adds two more definitions describing two other meanings of this term: 1 'a system (or a material point) carrying out oscillating periodic motion near the position of stable equilibrium' ['система (або матеріальна точка), що здійснює коливальний періодичний рух біля положення стійкої рівноваги']; 2 'exciter (source) of oscillations' ['збуджувач (джерело) коливань']. The last three definitions should be recognized also in wordnets.

Photometry (i40656): 'measurement of the properties of light (especially luminous intensity)'. There is also another meaning presented in the "Explanatory dictionary of physics": 'a section of physical optics dedicated to the measurement of electromagnetic radiation in the optical range' ['розділ фізичної оптики, присвячений вимірюванню електромагнітного випромінювання оптичного діапазону'] — that should be included in wordnets.

Radiometer (i57900): 'meter to detect and measure radiant energy (electromagnetic or acoustic)'. The TSF three more definitions that should be present in

wordnets: 1 'a device for measuring the energy of electromagnetic radiation, based on its thermal action' ['прилад для вимірювання енергії електромагнітного випромінювання, заснований на його тепловій дії']; 2 'a radio telescope receiver' ['приймальний пристрій радіотелескопа']; 3 'a device for measuring the activity of radioactive sources' ['прилад для вимірювання активності радіоактивних джерел'].

Spectrometry, spectroscopy (i38839): 'the use of spectrometers to analyze spectra'. This definition is incomplete and makes no difference between different notions of spectrometry and spectroscopy. The TSF defines spectrometry as 'a set of methods and theory of measuring the spectra of electromagnetic radiation and studying the spectral properties of substances and solids in the optical range of wavelengths' ['сукупність методів і теорія вимірювання спектрів електромагнітного випромінювання та вивчення спектральних властивостей речовин і тіл в оптичному діапазоні довжин хвиль'] and proposes two senses for spectroscopy: 1 'a field of physics devoted to the study of the distribution of the intensity of electromagnetic radiation by wavelengths or frequencies' ['область фізики, присвячена дослідженню розподілу інтенсивності електромагнітного випромінювання за довжинами хвиль або частотами']; 2 'study of decompositions into the spectrum' ['дослідження розкладів у спектр']. These two different notions should be treated in wordnets accordingly.

Transparency, transparence (i98144): 'permitting the free passage of electromagnetic radiation'. The Longman does not have an entry "transparence" and proposes three senses for "transparency": 1 'a sheet of plastic or a piece of photographic film through which light can be shone to show a picture on a large screen'; 2 'the quality of glass, plastic etc. that makes it possible for you to see through it'; 3 'the quality of being easy to understand or know about'. Here, the second sense gives qualitative description of this concept. In turn, the TSF presents more strict definition: 'a value that shows how much of the radiation flux (or for visible light – the luminous flux) falling on a certain surface passes through a layer of unit thickness without changing its direction' ['величина, що показує, яка частина потоку випромінювання (або для видимого світла – світлового потоку), що падає на деяку поверхню, проходить без зміни напрямку через шар одиничної товщини'], which is recommended.

The proposed definition amendments will, accordingly, cause the changes in hypo-hypernymy relations that should be correctly displayed in wordnets.

The relative number of such inconsistencies is small because the total number of analyzed synsets in OEWN is 3 220. However, each third definition therein requires minor refinement.

The described procedure may be executed both automatically — through hypothetical corresponding tool based on the findings of (Vakulenko 2019; 2021) — or manually. Human validation by an experienced lexicographer is required.

5. Discussion

The analyzed problems are typical for wordnets created on the basis of the "merge" and "combined" approaches. The "expansion" approach alone does not have these inconsistencies because the resulting data in this case come from only one source — pIWordNet (Rudnicka et al. 2019), which is linked to OEWN. However, its disadvantage is that in most cases, the resulting synsets and their semantic structures do not correspond to those inherent to the target language, not to mention individual lexical units, whose machine translation is often incorrect today, especially with regard to the Ukrainian language. This results in the need for tedious post-processing that cannot yet be performed by a machine.

The profound reason for these inconsistencies between concepts and their verbal designations lies in asymmetrical dualism of the linguistic sign that is the fundamental property of language. Asymmetrical dualism of the linguistic sign means that the word meaning (signifié) and its written form (signifiant) are set in an unstable balance and display dynamic behavior. The first is inclined to be expressed by other word forms, thus giving rise to synonymy, while the second tends to acquire additional meanings (see Karcevskij 1929). This continuous emergence of additional meanings cannot be properly tracked by conventional lexicography methods.

Fellbaum (2006: 665) makes a reserving remark about the interchangeability of synset members in different contexts, indicating that there exist synset members that cannot be substituted in all contexts. However, if synonymy is considered as a phenomenon formed at a higher, contextual level, it may be concluded that lexical units that cannot be substituted in all contexts, should belong to different synsets. From this standpoint, a revised definition of synonyms strictly requiring their interchangeability in all contexts should be adhered to.

Regarding different definitions of the same lexeme, a reasonable question may be asked: how crucial is the given form of a definition for a semantic load of the defined unit? In other words, how do the differences in definitions available in different sources affect the sense of the lexeme in question?

A metric method to measure semantic distances between lexemes was applied to build semantic fields using a recursive procedure that represents a lexeme's meaning by its semes taken from a dictionary definition (Vakulenko 2021). It was observed that two types of lexicographic effects appear in the Merriam-Webster dictionary⁹: the horizontal crossing of different semantic trees when some semes belonging to different trees, are the same at certain semantic depth; and the vertical looping (appearance of lexicographic hypercycles) when the same seme appears at some depth of the given semantic tree. A similar lexicographic effect was reported earlier in Shyrokov (2018: 12-18) on the material of the "Dictionary of Ukrainian Language in 20 Volumes" where the following semantic hypercycle was found: *prylad* 'instrument' – *prystrij* 'device' (hypernym) – *obladnannja* 'equipment' (hypernym) – *prylad* 'instrument' (meronym). Such hypercycles appear in sufficiently large dictionaries and manifest the rise

of semantic structures that are conditioned by vertical and horizontal semantic correlations between concepts. The existence of these structures indicates that the resulting semantic fields display a certain level of stability against different formulations in definitions.

The outcomes of the recent research on concept ontology (Bense 2021) suggest that complex concepts can finally be decomposed into a significantly smaller number of elementary concepts displaying more semantic clarity, which are the same across analyzed languages. This fact also supports the assumption that the semantic load of a given lexeme is less dependent on the exact formulation of the corresponding concept.

6. Prospects of future research

Some related questions that should be examined in more detail in future investigations are briefly discussed here.

1. The notion of "meaning dimensions" (Löbner 2013) has a major sense in semantics because it sets the weight of any individual meaning. The further development if this idea leads to the conclusion that any sense of a linguistic unit can be presented in the form of its semantic field represented by a sum of elementary meanings taken with corresponding weight coefficients (see Vakulenko 2021). In other words, a proper account of "meaning dimensions" requires basing on their weight coefficients which determine the importance of these dimensions.

The definition of synonyms by Löbner (op. cit.) uses a vague concept of "sameness" and tempts a logical question of what senses are considered the "same", or to what extent of precision the lexical units must have the "same" meaning, and what the tools to measure this "sameness" could be. This fuzzy definition makes the synsets themselves ambiguous and unclear. On the other hand, a mathematically strict approach to calculate semantic distances between lexical units based on their definitions (Vakulenko 2019) may resolve this issue. The possible use of underlying mathematics to determine and measure synonymy would be promising for further development of wordnets.

2. The semantic equivalence of the compound synonyms *electron lone pair* and *lone-pair electrons* (TSF: 394) is due to syntactic reasons. This kind of similarity may be displayed also by corresponding contextualized embeddings as described in (Liu et al. 2020; Lee 2021; Loureiro et al. 2021). However, the method described in Section 3 to refine relevant semantics seems to be much more straightforward and clearer.
3. The development of semantic structures obtained within the "merge" approach, requires assignment of semantic relations not displayed in the lexicographic data sources. For example, "A Multilingual Dictionary on Phonetics and Phonology" (Yurtbaşı 2017a; 2017b; Vakulenko 2018: 49-55) and "A Concise

Dictionary of Synonyms of the Ukrainian Language"¹⁰ contain synsets but have no information about hyper-/hyponymy, mero-/holonymy, antonymy, etc. In this case, tools for automatically assigning LSRs are useful, such as the tool for building hypernymy described in Romanyshyn et al. 2024.

However, such a semantic hierarchy is disrupted in multiword expressions where the overall meaning cannot be inferred from the meanings of individual components (Voula 2023).

4. AI technology has been successfully applied to the task of writing dictionary definitions (De Schryver 2023: 381) and text generation, including Ukrainian (*Ukrainian Text Generator*). This makes it possible to expect that this technology will be able to solve also the tasks described above.

7. Conclusion

This study has demonstrated that the development of the Ukrainian Wordnet (Ukrajinet) under the "merge" and "combined" approaches inevitably exposes inconsistencies between lexicographic sources and wordnet structures, especially regarding synonymy and hyper-/hyponymy. What is new and significant here is the identification of systematic mismatches between traditional dictionary definitions and their Open English Wordnet counterparts, and the formulation of a reproducible procedure for their correction. By applying methods of terminology science and comparative lexicography, this work clarified how missing or imprecise definitions distort conceptual hierarchies and proposed a structured post-processing framework that restores their integrity.

These findings contribute to the broader field of computational lexicography by highlighting that semantic refinement is not merely an editorial task but an essential step toward ensuring conceptual accuracy and interoperability of multilingual lexical databases. The recognition that synonymy and hyper-/hyponymy relations often depend on subtle shifts in definitions provides a pathway for developing more semantically consistent ontologies. In practical terms, the proposed procedures can inform the semi-automatic refinement of lexical databases, combining algorithmic processing with expert validation.

However, several limitations remain. The current refinement was performed on a subset of data and focused primarily on the domains of physics and phonetics. Extending this methodology to other domains will require additional testing and refinement of the semantic alignment rules. Moreover, while partial automation is achievable, full automation of lexicographic post-processing still depends on advances in AI-based semantic modeling.

The practical implications of this research lie in improving the quality of multilingual wordnets, supporting tasks such as machine translation, semantic search, and ontology construction. Future research should focus on implementing computational tools capable of identifying definitional inconsistencies automatically, experimenting with weighted semantic fields, and integrating contextual embeddings to enhance synonymy detection. Further theoretical work

on semantic metrics and concept ontologies may also provide mathematical foundations for refining lexical relations.

Ultimately, this study underscores that the future of lexicography lies in the synergy of human expertise and computational precision. Wordnets are not static repositories but evolving models of linguistic and conceptual knowledge. The refinement procedures outlined here contribute to moving closer to a truly interoperable and semantically transparent lexicon — a cornerstone for intelligent language technologies of the future.

8. Acknowledgments

The author is grateful to Prof. Melanie Siegel for useful comments and suggestions that helped to improve the content of this article.

9. Endnotes

1. <https://en-word.net/>
2. <https://ukrajinet.yoda.media.h-da.de/>
<https://github.com/hdaSprachtechnologie/ukrajinet>
https://code.fbi.h-da.de/ukrajinet/ukrajinet-dev/-/blob/main/20230728_merged_ukrajinet.xml
3. Vakulenko, M.O. and O.V. Vakulenko. 2008. *Tlumachnyj slovnyk iz fizyky* [Explanatory Dictionary on Physics]. Kyjiv: VPC "Kyjivskyj universytet". 767 p.
4. Vakulenko, M.O. and O.V. Vakulenko. 2017. *Slovnyk fizychnykh terminiv-synonimiv* [Dictionary of Physical Synonymous Terms]. Kyjiv: VPC "Kyjivskyj universytet". 191 p.
5. Rusanivskyj, Vitalij (Ed.). 2010–. *Slovnyk ukrajinsjkoji movy v 20 tomakh* [Dictionary of Ukrainian Language in 20 Volumes]. Kyjiv: Naukova dumka.
6. *Longman Dictionary of Contemporary English*. 2003. [Della Summers (ED.).] Harlow, Essex: Pearson Education.
7. Sube, Ralf and Günther Eisenreich. 1973. *Wörterbuch Physik. A–M. Englisch–Deutsch–Französisch–Russisch*. Zürich/Frankfurt am Main: Verlag Harri Deutsch.
8. Greulich, Walter and Dirk Meenenga. 1997. *Wörterbuch der Physik: Englisch–Deutsch/Deutsch–Englisch*. Heidelberg/Berlin: Spektrum, Akademischer Verlag.
9. *Dictionary by Merriam-Webster*. 2023.
<https://www.merriam-webster.com/>
10. Derkach, Pylyp. 1960. *Korotkyj slovnyk synonimiv ukrajinsjkoji movy* [A Concise Dictionary of Synonyms of the Ukrainian Language]. Kyjiv: Radjansjka shkola.
https://chtyvo.org.ua/authors/Derkach_Pylyp_Maksymovych/Korotkyi_slovnyk_synonimi_v_ukrainskoi_movy/

References

- Bense, Hermann. 2021. A Formal Ontology of Concepts for Deep Semantic Searching. *KONT-21 Knowledge — Ontology — Theory, 8th International Conference, 8–12 November 2021, Novosibirsk*.
https://www.ontology4.us/download/dot/ontopage/sub_Bens2021d.pdf

- Bokshanj, Ghalyna.** 2021. Synonimija v terminosystemi anghlomovnykh naukovykh tekstiv ekonomichnogho sprjamuvannja [Synonymy in Terminology System of English-language Scientific Texts in the Field of Economics]. *Pivdennyj arkhiv (filologichni nauky)* 86: 54-57 (in Ukrainian). DOI: 10.32999/ksu2663-2691/2021-86-8
- Bond, Francis, Piek Vossen, John P. McCrae and Christiane Fellbaum.** 2016. CILI: the Collaborative Interlingual Index. Fellbaum, Christiane, Piek Vossen, Verginica Barbu Mititelu and Corina Forascu. 2016. *Proceedings of the 8th Global WordNet Conference (GWC)*, Bucharest, Romania, 27–30 January 2016: 50-57. Global Wordnet Association.
<https://aclanthology.org/2016.gwc-1.9>
- De Schryver, Giles-Maurice.** 2023. Generative AI and Lexicography: The Current State of the Art Using ChatGPT. *International Journal of Lexicography* 36(4): 355-387.
DOI: <https://doi.org/10.1093/ijl/ecad021>
- Fellbaum, C.** 1998. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.
- Fellbaum, C.** 2006. WordNet(s). Keith Brown (Editor-in-Chief). 2006. *Encyclopedia of Language & Linguistics*. Second Edition. Volume 13: 665-670. Oxford: Elsevier.
- Halas Popović, Ana.** 2023. Synonymy from a Prototype Theory Perspective and its Symbiosis with Polysemy: Towards a New Dictionary of Synonyms. *Lexikos* 33: 1-23.
DOI: <https://doi.org/10.5788/33-1-1773>
- Karcevskij, Sergej.** 1929. Du dualisme asymétrique du sign linguistique [On the Asymmetric Dualism of the Linguistic Sign]. *Travaux du Cercle linguistique de Prague* 1: 88-92.
<https://www.degruyter.com/document/doi/10.7560/780439-004/html>
- Lee, Younghoon.** 2021. Systematic Homonym Detection and Replacement Based on Contextual Word Embedding. *Neural Processing Letters* 53: 17-36.
- Liu, Qi, Matt J. Kushner and Phil Blunsom.** 2020. A Survey on Contextual Embeddings. ArXiv.2003.07278v2 [cs.CL].
DOI: <https://doi.org/10.48550/arXiv.2003.07278>
- Loureiro, Daniel, Kiamehr Rezaee, Mohammad Taher Pilehvar and Jose Camacho-Collados.** 2021. Analysis and Evaluation of Language Models for Word Sense Disambiguation. *Computational Linguistics* 47(2): 387-443.
DOI: https://doi.org/10.1162/coli_a_00405
- Löbner, Sebastian.** 2013. *Understanding Semantics*. Second Edition. New York: Routledge.
- Maziarz, M., M. Piasecki and S. Szpakowicz.** 2013. The Chicken-and-Egg Problem in Wordnet Design: Synonymy, Synsets and Constitutive Relations. *Language Resources and Evaluation* 47(3): 769-796.
- McCrae, John Philip, Alexandre Rademaker, Ewa Rudnicka and Francis Bond.** 2020. English Wordnet 2020: Improving and Extending a Wordnet for English Using an Open-source Methodology. Declerk, Thierry, Itziar Gonzalez-Dios, German Rigau. 2020. *Proceedings of the LREC 2020 Workshop on Multimodal WordNets (MMW2020)*, Marseille, France, May 2020: 14-19. The European Language Resources Association (ELRA).
<https://aclanthology.org/2020.mmw-1.3/>
- Romanyshyn, Nataliia, Dmytro Chaplynskyi and Mariana Romanyshyn.** 2024. Automated Extraction of Hypo-Hypernym Relations for the Ukrainian WordNet. *The Third Ukrainian Natural Language Processing Workshop (UNLP) @LREC-COLING-2024, 25 May 2024*: 51-60. ELRA Language Resource Association.

- Rudnicka, Ewa, Maciej Piasecki, Francis Bond, Łukasz Grabowski and Tadeusz Piotrowski.** 2019. Sense Equivalence in plWordNet to Princeton WordNet Mapping. *International Journal of Lexicography* 32(3): 296-325.
DOI: <https://doi.org/10.1093/ijl/ecz004>
- Shyrokov, V.A. (Ed.).** 2018. *Linghivistychno-informacijni studiji. Praci Ukrajinsjkogho movno-informacijnogho fondu NAN Ukrainy: u 5 tomakh, tom 3. Tlumachna leksykoghrafija, knygha 2. Systemna semantyka tlumachnykh slovnykiv* [Linguistic Information Studies. Works of Ukrainian Lingua-Information Fund of National Academy of Sciences of Ukraine: in 5 Volumes, Vol. 3. Explanatory Lexicography, Book 2. Systemic Semantics of Explanatory Dictionaries] (in Ukrainian). Kyjiv: Ukrajinsjkyj movno-informacijnyj fond.
- Siegel, Melanie and Francis Bond.** 2021. OdeNet: Compiling a German Wordnet from Other Resources. Vossen, Piek and Christiane Fellbaum (Eds.). 2021. *Proceedings of the 11th Global Wordnet Conference (GWC 2021), Potchefstroom, South Africa, 18–21 January 2021*: 192-198.
<https://www.aclweb.org/anthology/2021.gwc-1.22>
- Ukrainian Text Generator.** Linguanex.
<https://lingvanex.com/services/ukrainian-text-generator/>
- Vakulenko, Maksym.** 2018. *From Terminology-Vocabulary to Terminology-Science: A Ukrainian Trend*. Beau Bassin: LAP.
- Vakulenko, Maksym.** 2019. Calculation of Semantic Distances between Words: From Synonymy to Antonymy. *Journal of Quantitative Linguistics* 26(2): 116-128. (Published online: 28 May 2018.)
<https://doi.org/10.1080/09296174.2018.1452524>
- Vakulenko, Maksym.** 2021. From Semantic Metrics to Semantic Fields. *Proceedings of the 2021 IEEE 16th International Conference on Computer Science and Information Technologies (CSIT), 22–25 September 2021, Lviv, Ukraine*: 44-47.
DOI: 10.1109/CSIT52700.2021.9648675
- Voula, Giouli.** 2023. A Model for Representing the Semantics of MWEs: From Lexical Semantics to the Semantic Annotation of Complex Predicates. *Frontiers in Artificial Intelligence* 6.
DOI: <https://doi.org/10.3389/frai.2023.802218>
- Yurtbaşı, Metin.** 2017a. The French Phonological Corpus with Equivalents in English, German and Turkish, as Part of a New Multilingual Dictionary Project. *International Journal of Arts, Humanities, and Social Sciences* 2(11): 27-77.
<http://www.ijahss.com/vol2-issue11.html>
- Yurtbaşı, Metin.** 2017b. A Multilingual Dictionary also Containing Qazaq Terminology in Phonetics and Phonology. *International Journal of Arts, Humanities, and Social Sciences* 2(11): 78-103.
<http://www.ijahss.com/vol2-issue11.html>

Out-of-the-Box GPTs for Lexicography

Gilles-Maurice de Schryver, *BantUGent & LT³, Ghent University,
Ghent, Belgium; and Department of African Languages,
University of Pretoria, Pretoria, South Africa*
(gillesmaurice.deschryver@UGent.be)
(<https://orcid.org/0000-0001-7272-9878>)

Abstract: A detailed analysis is offered of how ChatGPT-3.5 has treated the world's languages for lexicographic purposes, starting with hyper-visible English, moving to semi-visible Portuguese, and ending with low-visible Bantu languages. Case studies for the latter group are region-visible Swahili, nation-visible Xhosa and local-visible Lusoga. This impact-based scale is reflected in the uptake of GenAI technology for lexicography. The results of using out-of-the-box ChatGPT-3.5 are fantastic for English, fake for Portuguese, and simply fail for the Bantu languages. But there is a way out, and the seeds for that are sown in this article.

Keywords: GENERATIVE ARTIFICIAL INTELLIGENCE (GENAI), LARGE LANGUAGE MODELS (LLMs), GENERATIVE PRE-TRAINED TRANSFORMER (GPT), OUT-OF-THE-BOX GPTs, LEXICOGRAPHY, DICTIONARIES, LEXICOGRAPHERS, ENGLISH, PORTUGUESE, BANTU, SWAHILI, XHOSA, LUSOGA

Obufunze: GPT ey'omutindo eya GenAI mw'isomo ly'okuwandiika amawanika. Okutangaazibwa ku ngeli ChatGPT-3.5 bw'ekozeseibwa mu nnimi dh'ensi yoonayoona mu kugezaaku okutuukiliza ebigendelelwa by'okuwandiika amawanika, kuweebwa mu lupapula luno. Okutangaaza kuno kutandiikila ku nkozesa eyolekebwa ku lulimi oluganzi mu nsi yoonayoona olw'Olungeleza, kw'amala kwaizibwa ku nnimi engazimuuku akatono ng'Olupootigo, kw'asembyayo enkozesa eyeeyolekela mu nnimi edhikozesebwa abantu abatono einho ng'edho edhigwa mu luse lwa Bantu. Ennimi dha Bantu edhiliku enkozesa entono einho edhigelaagela-nhizibwa ku y'edho emba oba ejjanziku n'omuli, olulimi lw'Oluswayili, Oluxhosa n'Olusoga. Ennimi dhino dhonsatule n'edhisinziilwaku okulaga tekinologiya wa GenAI bwe yeeyolekelamu ng'akozesebwa mu milimu gy'okuwandiika amawanika. Kizuuse nti enkozesa ya tekinologiya eno nnambulukufu inho mu milimu egikola ku mawanika g'Olungeleza, aye, ya kibogwe ku nnimi ejjanziku akatono ng'Olupootigo. Bwe kituuka ku nnimi dh'omu luse lwa Bantu kabili enkola eno eililila inho y'aba oti kumpi tekelela ilala. Aye, wakaaliwooku eisuubi eliva mu byandhulwa mu kitundu ky'olupapula luno eky'okubili. Mu lupapula olw'okubili n'omuli ebyava mu kwongela okunoonheleza ku ngeli eyinza okukozesebwa okuziba olutanda oluli wagati w'ennimi dh'ensi yoonayoona ejjanzu, n'edhitali, aye nga dhoonadhoona dhiluubilila okukoseza GenAI okuwandiika amawanika.

Ebigambo ebikulu: KALIMAGEZI-KISIBUKILA, OKUTETENKELEZA KW'ENNIMI ENNHINGI, KISIBUKILA AKAALI KUKENGULWA, GPT EY'OMUTINDO, EISOMO LY'OKUWA-

NDIIKA AMAWANIKA, AMAWANIKA, ABAWANDIISI B'AMAWANIKA, OLUNGELEZA,
OLUPOOTIGO, BANTU, OLUSWAYILI, OLUXHOSA, OLUSOGA

1. Prologue

In De Schryver (forthcoming-a, forthcoming-b) a sample of 100 contributions on Generative Artificial Intelligence (GenAI) in lexicography was studied, covering the period from the release of ChatGPT on 30 November 2022 up to 31 May 2025. Over those two years and a half, all but one of the contributions simply used *out-of-the-box* Generative Pre-trained Transformer (GPT) models. Those types of Large Language Models (LLMs) are known for their ability to generate sophisticated human-like texts. Just a single one of the 100 contributions moved to the next level for lexicography, namely to a *customised* GPT model (De Schryver 2024a). However, and despite the expressed wishes about it at gatherings on digital lexicography, starting with eLex 2023, not a single attempt has been made to date to move to the next-next level, namely to a *fine-tuned* GPT model. In the present article the first step is revisited for ChatGPT, and in a next article (De Schryver 2025) the second step is revisited with attention to both OpenAI's ChatGPT and Google's Gemini, expanding the four-page abstract on it prepared (but not delivered) for a workshop on LLMs and lexicography at Euralex 2024 (De Schryver 2024c).

2. Out-of-the-box GPTs

2.1 Hyper-visible English

By and large, the most successful experiments with LLMs for lexicography have been reported on for applications using English (1.45 billion speakers; spoken in 59 sovereign countries and 27 non-sovereign entities). See especially De Schryver (2023b), Lew (2023), Rees and Lew (2024), Lew (2024), Ptasznik et al. (2024) and Lew et al. (2024). At the same time, some colleagues correctly pointed out a few caveats, see especially McKean and Fitzgerald (2023), Jakubíček and Rundell (2023) and Nichols (2023). That said, virtually all scholars agree that it is not a good idea not to bring in today's GenAI for English lexicography.

2.2 Semi-visible Portuguese

The results of the first in-depth and detailed attempt to study the lexicographic potential of using ChatGPT for a language other than English, were presented at the inaugural Americalex-S conference in São Paulo, Brazil (De Schryver 2023a). The assumption was that ChatGPT would do an excellent job for the rather large language Portuguese (260 million speakers; spoken in 10 coun-

tries and territories). Following the success of using ChatGPT-3.5 for compiling a monolingual English COBUILD-style learners' dictionary — demonstrating its incredible performance on lemmas from a variety of word classes: verbs, nouns, conjunctions, adjectives and interjections, as well as metaphorical meanings and idiomatic expressions (see De Schryver 2023b: supplementary material) — the first goal was to test ChatGPT-3.5 on a monolingual Portuguese COBUILD-style learners' dictionary. There was an ongoing student strike on campus during the conference, so the polysemous lemma *strike* was chosen, and put into a Portuguese translation of the same prompt which had been used for English in De Schryver (2023b: supplementary material), as shown in Figure 1. Here and elsewhere, English translations of the Portuguese are to be found in Addendum A. Note that in De Schryver (2023b) the goal was to arrive at a single prompt with which each type of lemma, no matter its part of speech or any other aspects, may be generated. The 'prompt engineering' thus preceded the prompt eventually proposed, with which an entire COBUILD dictionary could be compiled. Copies of De Schryver's (2023b) prompt have subsequently been used by colleagues, such as Zhong (2025) using DeepSeek, and also in translated form, such as in Spanish by Fuertes-Olivera (2024), or in Lithuanian by Pavydytė (2024).

Default (GPT-3.5)



Por favor, forneça um verbete de dicionário no estilo COBUILD para "strike". Comece com as informações de frequência na parte superior. Cada sentido deve estar em um bloco numerado. Cada bloco começa com a classe gramatical e as formas morfológicas do respectivo sentido. Em seguida, há uma definição de sentido e exemplos de sentido que ilustram o uso e o significado de cada sentido específico. Para as frases de exemplo, certifique-se de usar diferentes estruturas de frases, referindo-se a diferentes pessoas; refira-se a situações passadas, presentes e futuras; varie frases de exemplo longas e curtas; e inclua outras elaborações para que você, caro ChatGPT, produza material de dicionário de alta qualidade para mim.

Figure 1: ChatGPT-3.5 prompt to obtain a COBUILD-style entry for a monolingual Portuguese learners' dictionary, here for the lemma *strike*

Even though the prompt was written in Portuguese, the reply from ChatGPT was entirely in English, as seen in Figure 2 (only showing the first part, the start of the verbal use).

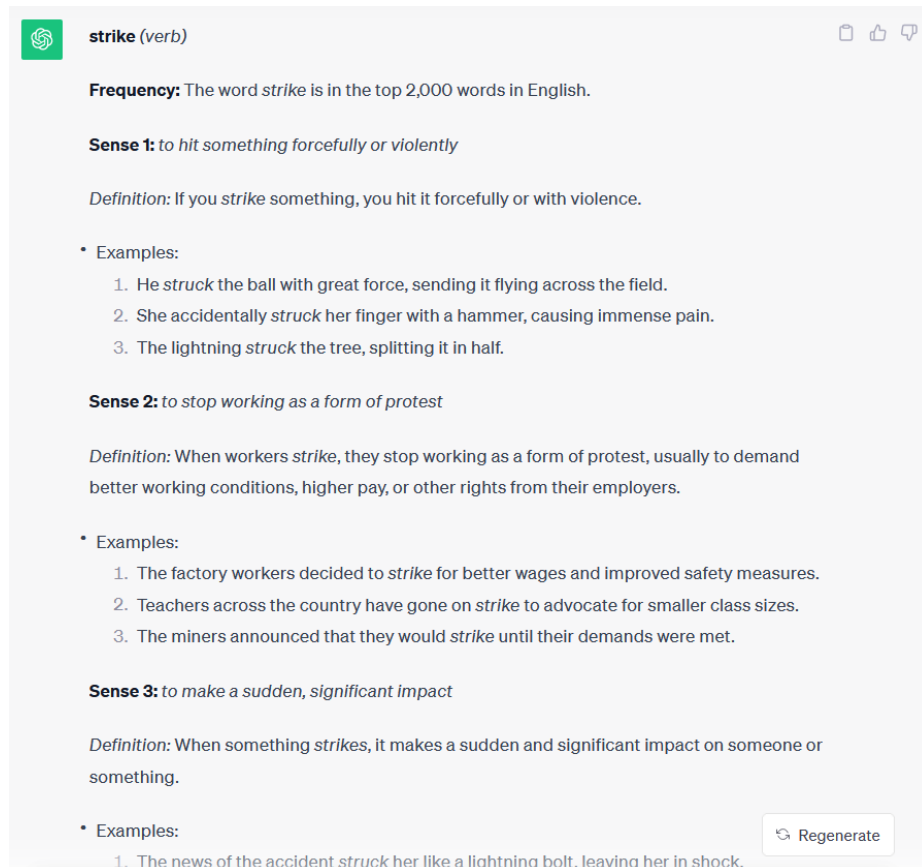


Figure 2: Start of the ChatGPT-3.5 output for the prompt seen in Figure 1

This was followed up with: 'Thanks, but actually I wanted to see a monolingual dictionary entry for the Portuguese entry '*greve*' (which is 'strike' in English)', after which ChatGPT replied with what is seen in Figure 3.

This time, ChatGPT presented a noun (correct), with an indication that it is feminine (correct), in a wider frequency band compared to English (ambiguous), a good definition (but not a COBUILD-style one), acceptable examples and an interesting encyclopaedic note. Observe that examples 1 and 2 in Figure 3, are very similar to examples 1 and 2 under the verbal sense 2 seen in English in Figure 2, except that the second one there is not even a verb, but part of a verb phrase where *strike* is used as a noun (so an error). From the very first reply by ChatGPT onwards, and now reinforced by the similar examples in the second reply, one starts wondering if one is not dealing with translation issues (in the sense that the model generates everything in and for English first, and only then maps the English onto another language as if it were English).



Obrigado, mas na verdade eu queria ver um verbete de dicionário monolíngue para o verbete português "greve" (que é "strike" em inglês).



Peço desculpas pela confusão anterior. Aqui está um verbete de dicionário monolíngue para a palavra "greve" em português:

greve (*substantivo feminino*)

Frequência: A palavra *greve* está entre as 10.000 palavras mais frequentes em português.

Definição: A *greve* é uma paralisação organizada por trabalhadores, na qual eles deixam de trabalhar como forma de protesto, geralmente para reivindicar melhores condições de trabalho, salários mais altos, ou outros direitos dos empregadores.

Exemplos:

1. Os funcionários da fábrica entraram em *greve* para exigir um aumento salarial.
2. Os professores decidiram iniciar uma *greve* devido à falta de investimentos na educação.
3. A *greve* dos caminhoneiros causou escassez de suprimentos em todo o país.

Nota: A palavra "greve" é frequentemente usada no contexto de protestos trabalhistas e é um direito garantido por leis trabalhistas em muitos países. Ela pode ser realizada por diversos setores profissionais, incluindo professores, médicos, motoristas, e outros, como um meio de pressionar empregadores ou governos a atenderem às demandas dos trabalhadores.

Regenerate

Figure 3: ChatGPT-3.5 output for the noun *greve* 'strike', for a monolingual Portuguese learners' dictionary

Next, the verb *recomendar* 'recommend', being one of the 15 verbs of communication used by Lew (2023), was submitted. The prompt (with an extra line) is shown in Figure 4. The reply by ChatGPT looked rather satisfactory, as may be seen from Figures 5a, b and c.



Por favor, forneça um verbete de dicionário monolíngue no estilo COBUILD para "recomendar". Comece com as informações de frequência na parte superior. Cada sentido deve estar em um bloco numerado. Cada bloco começa com a classe gramatical e as formas morfológicas do respectivo sentido. Em seguida, há uma definição de sentido e exemplos de sentido que ilustram o uso e o significado de cada sentido específico. Para as frases de exemplo, certifique-se de usar diferentes estruturas de frases, referindo-se a diferentes pessoas; refira-se a situações passadas, presentes e futuras; varie frases de exemplo longas e curtas; e inclua outras elaborações para que você, caro ChatGPT, produza material de dicionário de alta qualidade para mim. Inclua também uma pronúncia do verbete usando o IPA e inclua uma lista de sinônimos e antônimos nos sentidos em que essas listas forem relevantes.



Figure 4: ChatGPT-3.5 prompt to obtain a COBUILD-style entry for a monolingual Portuguese learners' dictionary, here for the verb of communication *recomendar* 'recommend'

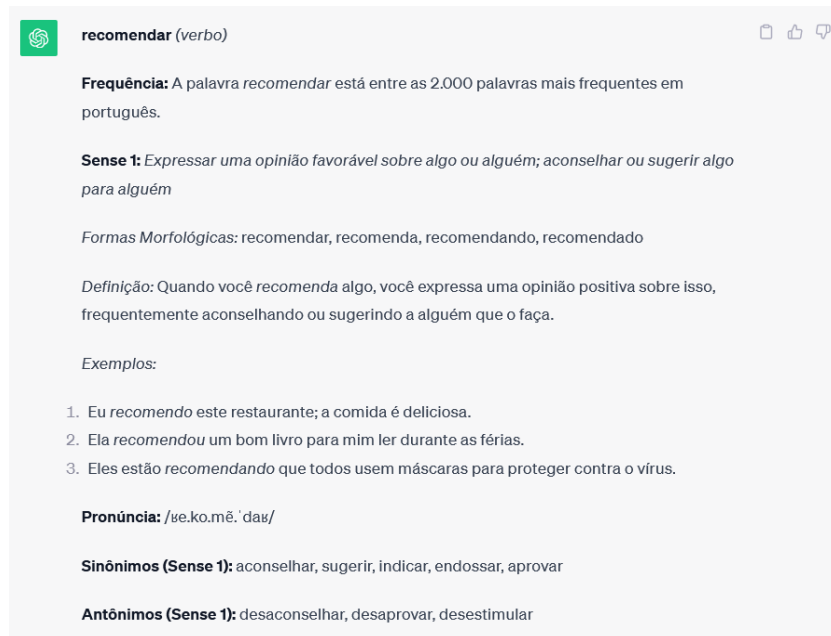


Figure 5a: ChatGPT-3.5 output for the Port. verb *recomendar* 'recommend', # 1, for the prompt seen in Figure 4

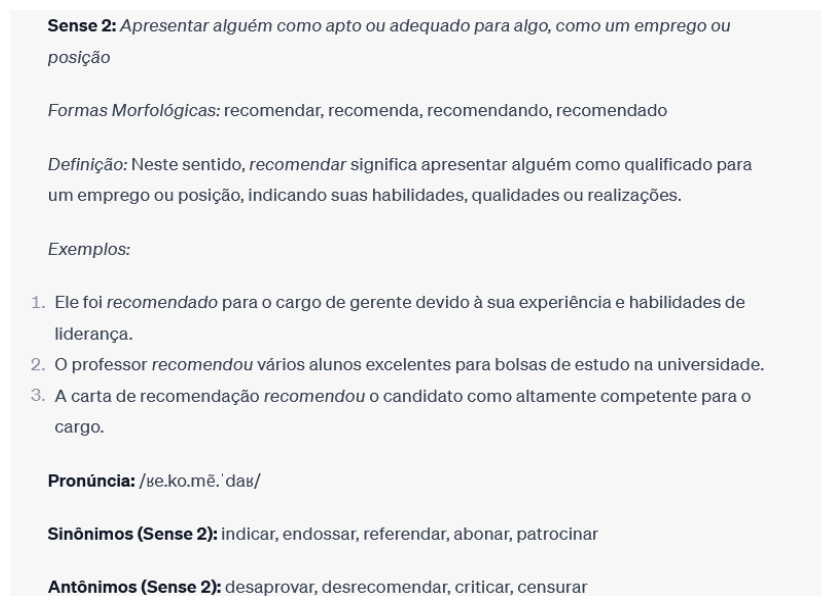


Figure 5b: ChatGPT-3.5 output for the Port. verb *recomendar* 'recommend', # 2, for the prompt seen in Figure 4

Sense 3: *Prescrever ou sugerir algo como tratamento ou solução*

Formas Morfológicas: recomendar, recomenda, recomendando, recomendado

Definição: Em termos médicos ou profissionais, *recomendar* envolve prescrever ou sugerir um tratamento, curso de ação ou solução para uma condição ou problema específico.

Exemplos:

1. O médico *recomendou* repouso e medicação para tratar a gripe.
2. O especialista *recomendou* uma dieta equilibrada e exercícios para melhorar a saúde do paciente.
3. Ela *recomendou* que ele consultasse um advogado para resolver a questão legal.

Pronúncia: /ʁe.ko.mẽ.'daʁ/

Sinónimos (Sense 3): prescrever, aconselhar, orientar, sugerir, indicar

Antónimos (Sense 3): desaconselhar, desautorizar, desaprovar, desestimular

Figure 5c: ChatGPT-3.5 output for the Port. verb *recomendar* 'recommend', # 3, for the prompt seen in Figure 4

However, and again — so clearly a serious issue — two of the three definitions are not following the COBUILD-style, despite the explicit request to do so in the prompt. An attempt was then made to even add as second sentence to the prompt: 'The most important aspect of the COBUILD-style is that for the definition, you need to use the full-sentence defining style, which consists of two clauses: the first clause illustrates the usage pattern, the second clause is a paraphrase of meaning'. But that had no effect whatsoever, despite the fact that merely mentioning COBUILD-style the very first time, resulted in a definition very close to it, see Figure 2 for the verb *strike*: 'If you strike something, you hit it forcefully or with violence'. Not as perfect as what is seen in an actual COBUILD, but close: 'If you strike someone or something, you deliberately hit them'.¹ Conclusion: ChatGPT-3.5 is unable to compile proper COBUILD-style entries for a monolingual Portuguese learners' dictionary.

What about a bilingual dictionary? The first focus is on an English to Portuguese entry (in that direction), for the lemma *bank*, with the prompt in English as seen in Figure 6.



Give me a bilingual English-Portuguese dictionary entry for the word 'bank'. Start with the frequency information at the top. Each sense should be in a numbered block. Each block starts with the part of speech and the morphological forms of the respective sense. This is followed by a sense definition in English, and a translation of that definition into Portuguese. Note that for the definition, you have to use full-sentences, not codes, and these sentences consist of two clauses: the first clause illustrates the usage pattern, the second clause is a paraphrase of meaning. Also list one or more translation equivalents in Portuguese at each sense. Add English examples with their Portuguese translations that illustrate both the use and the meaning of each particular English sense. For the example sentences, make sure to use different sentence structures, referring to different people; refer to past, present, and future situations; vary long and short example sentences; and include other elaborations so that you, dear ChatGPT, produce high-quality dictionary material for me. Also include a pronunciation of the entry using IPA, and include a list of synonyms and antonyms at the senses where such lists are relevant.

Figure 6: Prompt to obtain a COBUILD-style entry for a bilingual English-Portuguese learners' dictionary, here for the lemma *bank*. With also more information on what is meant with a 'COBUILD definition'

The lemma *bank*, a classic since first discussed seven decades ago by Noam Chomsky (1957: 95), and often used by Patrick Hanks ever after (Hanks 1987, 2000, 2013), has multiple nominal and verbal uses. ChatGPT offers three senses, two nominal ones, shown in Figures 7a and b, and one verbal one, shown in Figure 7c.

Clearly, ChatGPT did not understand what is meant with the COBUILD-style full-sentence definition (FSD).² The two clauses became two blocks, one for the usage pattern, the other for the paraphrase of meaning. Despite the fact that three attempts at defining are now made, under the heading 'Sense', the heading 'Usage Pattern' and the heading 'Definition', it is only at the third sense that a vague attempt is made at 'Definition': '*When someone banks something, they are either depositing it in a financial institution or relying on it for future use*'. Much more convoluted than what COBUILD does: '*If you bank money, you pay it into a bank*'.³

Plus, the first example for this verbal use of *bank* follows the pattern [VERB + in] and the second [VERB + on], so they don't really belong there, while COBUILD itself actually lists [VERB + with] as a verbal sense in itself: '*If you bank with a particular bank, you have an account with that bank*'.⁴

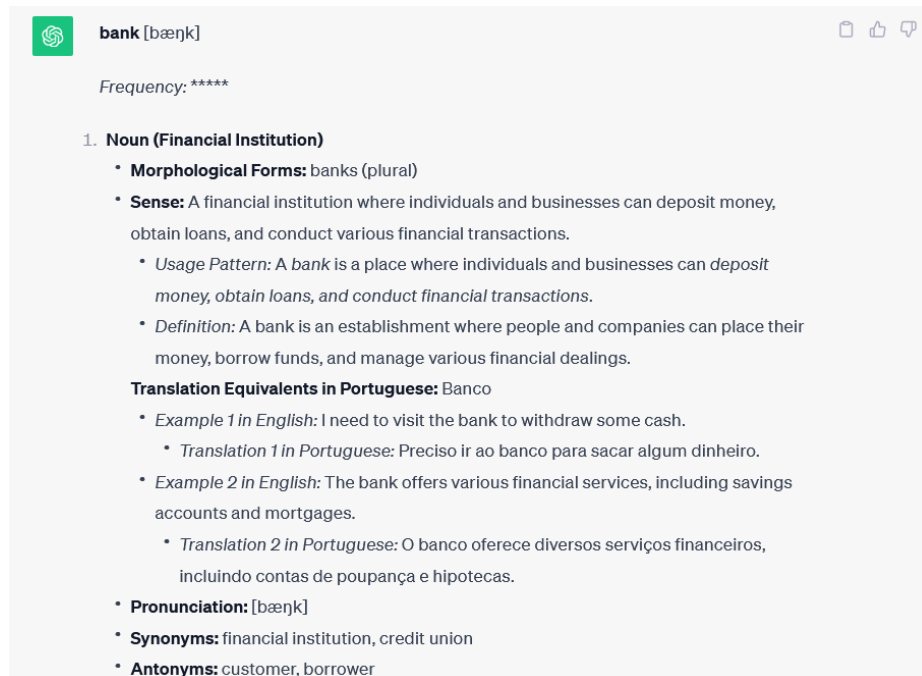


Figure 7a: ChatGPT-3.5 output for the English lemma *bank*, # 1, for the prompt seen in Figure 6

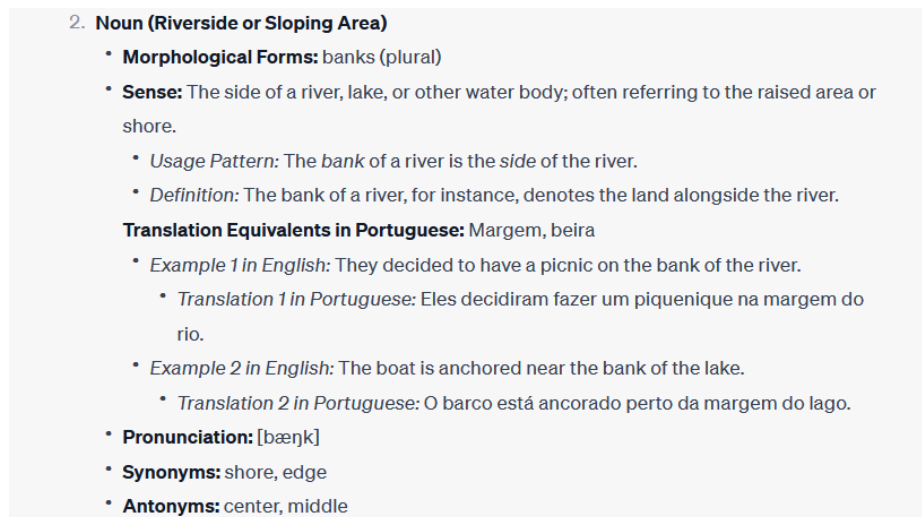


Figure 7b: ChatGPT-3.5 output for the English lemma *bank*, # 2, for the prompt seen in Figure 6

3. Verb

- **Morphological Forms:** banks (3rd person singular present), banked (past tense), banking (present participle), banks (3rd person singular present)
- **Sense:** To deposit or store something in a bank, or to rely on or store something for future use.
 - *Usage Pattern:* Someone banks something in a bank, or they bank on something for a specific purpose.
 - *Definition:* When someone banks something, they are either depositing it in a financial institution or relying on it for future use.
- Translation Equivalents in Portuguese:** Depositar, guardar, contar com
 - *Example 1 in English:* She banks her savings in a local bank.
 - *Translation 1 in Portuguese:* Ela deposita suas economias em um banco local.
 - *Example 2 in English:* He's banking on winning the competition with his skills.
 - *Translation 2 in Portuguese:* Ele está contando com sua vitória na competição com suas habilidades.
- **Pronunciation:** [bæŋk]
- **Synonyms:** deposit, rely on
- **Antonyms:** withdraw, disbelieve

Figure 7c: ChatGPT-3.5 output for the English lemma *bank*, # 3, for the prompt seen in Figure 6

Furthermore, given that a request was made for an English to Portuguese dictionary, there is also still too much in English. The next prompt became: '*Right, but what happened to the Portuguese? It is supposed to be a bilingual dictionary!*'. After that ChatGPT continued with presenting two versions, asking '*Which response do you prefer?*'. Part of the two versions is shown in Figure 8.

In the version on the left of Figure 8, the 'Definition' is now in Portuguese only, while in the version on the right of Figure 8, the synonyms and antonyms are in Portuguese only. Both additions are fine, but it had been better if both options were merged. Conclusion: As was the case for the monolingual test, it is clear that also for this direction of a bilingual dictionary, ChatGPT-3.5 is not up to the task, as it is unable to compile proper COBUILD-style entries for an English–Portuguese learners' dictionary.

1. Noun (Institution) • Morphological Forms: banks (plural) • Sense: A financial institution where individuals and businesses can deposit money, obtain loans, and conduct various financial transactions. • Usage Pattern: A bank is a place where individuals and businesses can deposit money, obtain loans, and conduct financial transactions. • Definition: Um banco é uma instituição financeira onde pessoas e empresas podem depositar dinheiro, obter empréstimos e realizar várias transações financeiras. Translation Equivalents in Portuguese: Banco • Example 1 in English: I need to visit the bank to withdraw some cash. • Translation 1 in Portuguese: Preciso ir ao banco para sacar algum dinheiro. • Example 2 in English: The bank offers various financial services, including savings accounts and mortgages. • Translation 2 in Portuguese: O banco oferece diversos serviços financeiros, incluindo contas de poupança e hipotecas. • Pronunciation: [bæŋk] • Synonyms: financial institution, credit union • Antonyms: customer, borrower	1. Noun (Financial Institution) • Morphological Forms: banks (plural) • Sense: A financial institution where individuals and businesses can deposit money, obtain loans, and conduct various financial transactions. • Usage Pattern: A bank is a place where individuals and businesses can deposit money, obtain loans, and conduct financial transactions. • Definition: A bank is an establishment where people and companies can place their money, borrow funds, and manage various financial dealings. Translation Equivalents in Portuguese: Banco • Example 1 in English: I need to visit the bank to withdraw some cash. • Translation 1 in Portuguese: Preciso ir ao banco para sacar algum dinheiro. • Example 2 in English: The bank offers various financial services, including savings accounts and mortgages. • Translation 2 in Portuguese: O banco oferece diversos serviços financeiros, incluindo contas de poupança e hipotecas. • Pronunciation: [bæŋk] • Synonyms: instituição financeira, caixa econômica • Antonyms: cliente, mutuário 2. Noun (Riverside or Sloping Area)
--	---

Figure 8: Part of the ChatGPT-3.5 output for the prompt seen in Figure 6, plus a request for more Portuguese, in two versions (left vs. right)

As a final test, the reverse was tried, producing a COBUILD-style Portuguese–English learners' dictionary, with the prompt shown in Figure 9, and the reply as seen in Figures 10a, b and c.

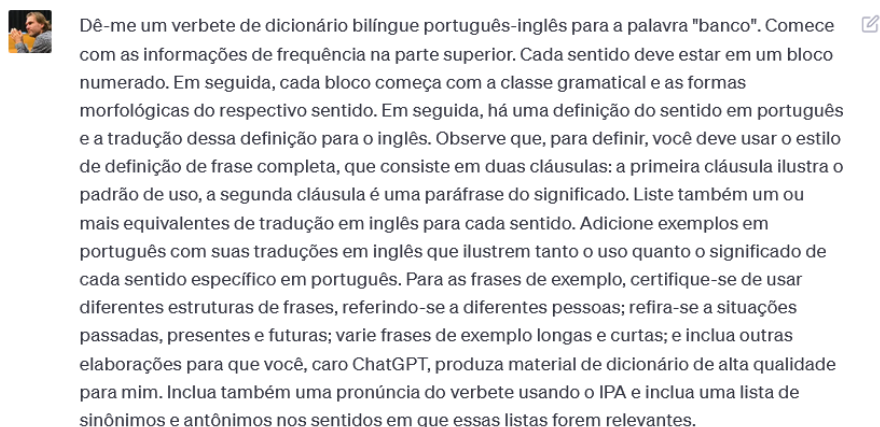
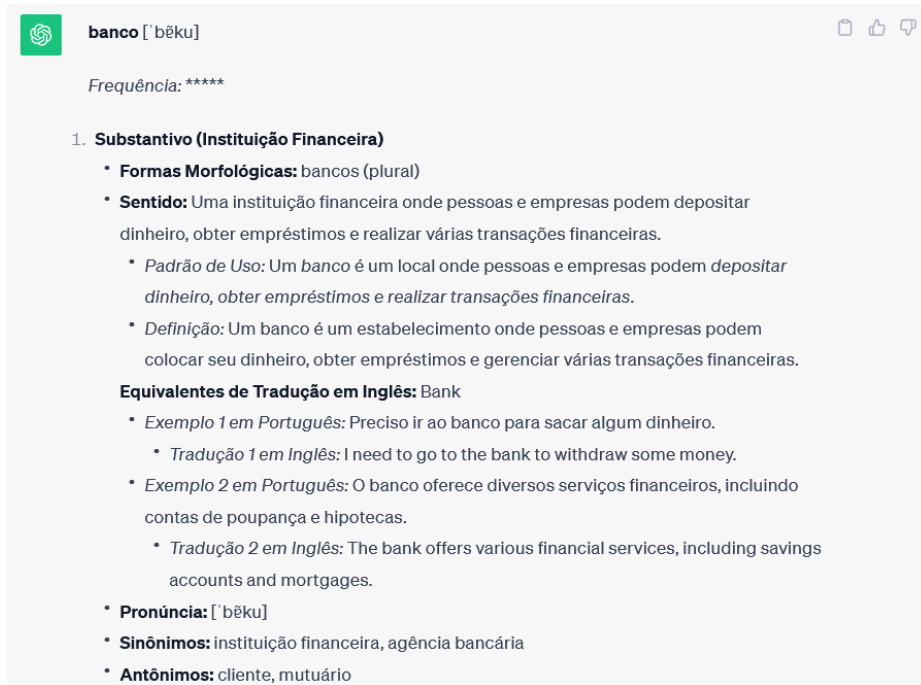


Figure 9: Prompt to obtain a COBUILD-style entry for a bilingual Portuguese–English learners' dictionary, here for the lemma *banco* 'bank'



The screenshot shows the ChatGPT-3.5 interface. At the top, the word 'banco' is followed by its phonetic transcription ['bẽku]. To the right are icons for a document, a thumbs up, and a thumbs down. Below this, the frequency is indicated as 'Frequência: *****'. The main content is organized into sections. The first section is '1. Substantivo (Instituição Financeira)'. It contains several bullet points: 'Formas Morfológicas: bancos (plural)', 'Sentido: Uma instituição financeira onde pessoas e empresas podem depositar dinheiro, obter empréstimos e realizar várias transações financeiras.', 'Padrão de Uso: Um banco é um local onde pessoas e empresas podem depositar dinheiro, obter empréstimos e realizar transações financeiras.', and 'Definição: Um banco é um estabelecimento onde pessoas e empresas podem colocar seu dinheiro, obter empréstimos e gerenciar várias transações financeiras.' Below this is 'Equivalentes de Tradução em Inglês: Bank', followed by two examples in Portuguese with their English translations. The section ends with 'Pronúncia: ['bẽku]', 'Sinônimos: instituição financeira, agência bancária', and 'Antônimos: cliente, mutuário'.

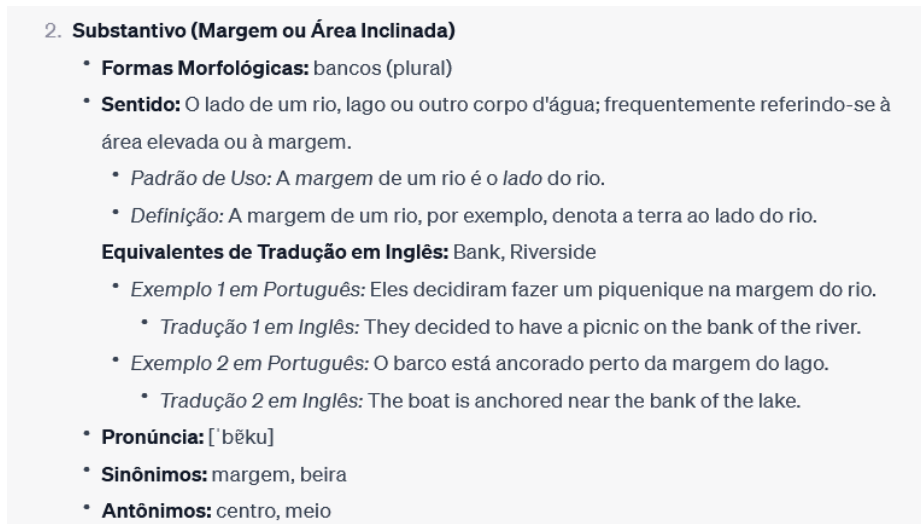
banco ['bẽku]

Frequência: *****

1. Substantivo (Instituição Financeira)

- **Formas Morfológicas:** bancos (plural)
- **Sentido:** Uma instituição financeira onde pessoas e empresas podem depositar dinheiro, obter empréstimos e realizar várias transações financeiras.
 - **Padrão de Uso:** Um banco é um local onde pessoas e empresas podem depositar dinheiro, obter empréstimos e realizar transações financeiras.
 - **Definição:** Um banco é um estabelecimento onde pessoas e empresas podem colocar seu dinheiro, obter empréstimos e gerenciar várias transações financeiras.
- **Equivalentes de Tradução em Inglês:** Bank
 - **Exemplo 1 em Português:** Preciso ir ao banco para sacar algum dinheiro.
 - **Tradução 1 em Inglês:** I need to go to the bank to withdraw some money.
 - **Exemplo 2 em Português:** O banco oferece diversos serviços financeiros, incluindo contas de poupança e hipotecas.
 - **Tradução 2 em Inglês:** The bank offers various financial services, including savings accounts and mortgages.
- **Pronúncia:** ['bẽku]
- **Sinônimos:** instituição financeira, agência bancária
- **Antônimos:** cliente, mutuário

Figure 10a: ChatGPT-3.5 output for the Portuguese lemma *banco*, # 1, for the prompt seen in Figure 9



This block continues the output from Figure 10a. It starts with '2. Substantivo (Margem ou Área Inclínada)'. It contains several bullet points: 'Formas Morfológicas: bancos (plural)', 'Sentido: O lado de um rio, lago ou outro corpo d'água; frequentemente referindo-se à área elevada ou à margem.', 'Padrão de Uso: A margem de um rio é o lado do rio.', and 'Definição: A margem de um rio, por exemplo, denota a terra ao lado do rio.' Below this is 'Equivalentes de Tradução em Inglês: Bank, Riverside', followed by two examples in Portuguese with their English translations. The section ends with 'Pronúncia: ['bẽku]', 'Sinônimos: margem, beira', and 'Antônimos: centro, meio'.

2. Substantivo (Margem ou Área Inclínada)

- **Formas Morfológicas:** bancos (plural)
- **Sentido:** O lado de um rio, lago ou outro corpo d'água; frequentemente referindo-se à área elevada ou à margem.
 - **Padrão de Uso:** A margem de um rio é o lado do rio.
 - **Definição:** A margem de um rio, por exemplo, denota a terra ao lado do rio.
- **Equivalentes de Tradução em Inglês:** Bank, Riverside
 - **Exemplo 1 em Português:** Eles decidiram fazer um piquenique na margem do rio.
 - **Tradução 1 em Inglês:** They decided to have a picnic on the bank of the river.
 - **Exemplo 2 em Português:** O barco está ancorado perto da margem do lago.
 - **Tradução 2 em Inglês:** The boat is anchored near the bank of the lake.
- **Pronúncia:** ['bẽku]
- **Sinônimos:** margem, beira
- **Antônimos:** centro, meio

Figure 10b: ChatGPT-3.5 output for the Portuguese lemma *banco*, # 2, for the prompt seen in Figure 9

3. Verbo

- **Formas Morfológicas:** banco (3ª pessoa do singular do presente), bancou (pretérito), bancando (particípio presente), banco (3ª pessoa do singular do presente)
- **Sentido:** Depositar ou armazenar algo em um banco, ou confiar ou armazenar algo para uso futuro.
 - *Padrão de Uso:* Alguém *bancou* algo em um *banco*, ou eles *bancam* algo para um propósito específico.
 - *Definição:* Quando alguém deposita algo em um banco, está armazenando em uma instituição financeira ou confiando em algo para uso futuro.
- **Equivalentes de Tradução em Inglês:** Deposit, Rely on
 - *Exemplo 1 em Português:* Ela deposita suas economias em um banco local.
 - *Tradução 1 em Inglês:* She banks her savings in a local bank.
 - *Exemplo 2 em Português:* Ele está contando com sua vitória na competição com suas habilidades.
 - *Tradução 2 em Inglês:* He's banking on winning the competition with his skills.
- **Pronúncia:** ['bẽku]
- **Sinônimos:** depositar, confiar
- **Antônimos:** sacar, desacreditar

Figure 10c: ChatGPT-3.5 output for the Portuguese lemma *banco*, # 3, for the prompt seen in Figure 9

When the output seen in Figures 10a, b and c was presented to two professors in Brazil, one a lexicographer for Portuguese (Regiani Aparecida Santos Zacarias), the other a linguist for Portuguese (who prefers to remain anonymous), they thought it all looked rather good at first. When they were asked (too) many questions, their opinion changed. While the first sense (Figure 10a), for the nominal use as a financial institution, is acceptable, the second sense (Figure 10b), for the nominal use as a sloping area, and third sense (Figure 10c), as a verb, are riddled with errors. The second block deals with the meaning 'bank of a river'. Even though this is supposed to be a sense for the lemma *banco*, the word used under 'Sense', 'Usage Pattern' and 'Definition' as well as in the two examples is a different word entirely: *margem*. At the same time, under the synonym list for *banco* one finds *margem* 'margin; riverbank; shore' and *beira* 'edge; riverbank; precipice verge'. Ironically, and as can already be deduced, only the synonyms presented refer to the bank of a river, using *banco* for that is simply not possible.

The third sense of '*banco*' supposedly presents the verbal use of this lemma. However, the sense 'to bank money' simply does not exist with the lemma *banco* in Portuguese. Now, ChatGPT not only 'created' this sense, it went as far as

inventing conjugations for *banco* to then use these conjugations as well (**bancou*, **bancam*), then continues with a convincing COBUILD-style full-sentence definition but forgets to use the lemma and instead uses the (correct) verbs *depositar* 'deposit' and *confiar* 'rely on', and finally does the same when providing the first example. Likewise for the phrasal verb 'count on', so [VERB + on], which also does not use *banco* in Portuguese, but rather *contar com* as seen in the second example of Figure 10c. When ChatGPT offers its list of synonyms, it merely repeats what it had used all along in the microstructure already. Note that there is a verb *bancar* 'fund, finance' in Portuguese, for which the morphological forms are *banca*, *bancou*, *bancando*, *banca*.⁵ Conclusion: As was the case for the other dictionary direction, ChatGPT-3.5 is once more not up to the task, as it is fully unable to compile proper COBUILD-style entries for a Portuguese–English learners' dictionary.

So, what is really going on, and what is ChatGPT-3.5 doing for bilingual lexicography? Well, the assumption is that one is actually looking at English lexical analysis through the eyes of a non-English language, here Portuguese. Reformulated, the hypothesis is that everything one sees there was created internally based on English (given the strong bias in the training data) and was merely mapped onto Portuguese since the prompts asked to go bilingual. If a further analogy is allowed: Given the overwhelming amount of English in an LLM, a model like ChatGPT-3.5 is simply incapable to 'think' in any other language, let alone analyse it linguistically in any proper way.

2.3 Low-visible Bantu

Back in 2023–2024, with the certainty that an LLM like ChatGPT is 'fantastic' for English, and the knowledge that it is basically 'fake' for bilingual lexicography, does that mean that it simply 'fails' for exotic languages? During a talk at the University of Missouri (De Schryver 2024b) various attempts were made at having ChatGPT compile proper dictionary articles for a variety of Bantu languages, amongst others Swahili (100 million speakers; spoken in more than 14 countries), Xhosa (19 million speakers; spoken in 2 countries) and Lusoga (4.5 million speakers; spoken in part of one country).

2.3.1 Region-visible Swahili

First, and again using ChatGPT-3.5, the chatbot was questioned about some of the basics of Swahili, to see whether it gets at least the essentials right for the language. The opening prompt and answer are displayed in Figure 11.



You

Hello, do you know any Swahili?



ChatGPT

Yes, I'm familiar with Swahili. Swahili is a Bantu language spoken in various countries in East Africa, including Kenya, Tanzania, Uganda, Rwanda, Burundi, and the Democratic Republic of Congo. It's one of the official languages of Kenya, Tanzania, and Uganda, and it's also used as a lingua franca in the region. How can I assist you with Swahili?

Figure 11: ChatGPT-3.5 opening prompt and answer on the Swahili language

While the answer is only approximately correct (apart from the army, no one uses Swahili in Uganda), the next question and answer immediately delve into lexicography for Swahili, as seen in Figure 12.



You

User

How would a dictionary entry for the English word 'cat' look like in a Swahili to English translation dictionary?



ChatGPT

Title: Swahili to English Translation Dictionary

Entry: Cat

Part of Speech: Noun

Swahili Equivalent: Paka

Definition:

A small domesticated carnivorous mammal with soft fur, a short snout, and retractile claws, known for its independent and playful nature.

Example Sentence:

The cat chased after the mouse in the kitchen.

Synonyms:

- Mshikaji wa Panya
- Kifalme



Figure 12: ChatGPT-3.5 prompt and output for an entry in a Swahili–English translation dictionary, here for *cat*

As may be seen from Figure 12, right from the moment one tries out lexicography in Swahili, the chatbot already gets the dictionary direction wrong. The

question was to go from Swahili to English, but the reply came in the reverse dictionary direction, with, strangely, only synonyms (of the English word *cat*) in Swahili. Now, in passing, it would be welcome if a chatbot could start taking over from an online Swahili dictionary, such as the unidirectional Swahili to English dictionary at africanlanguages.com⁶, having been ranked by Google⁷ as the top Swahili dictionary for two decades now, even though the year before its launch, there were already 20 online dictionaries for Swahili (De Schryver 2003: 12). See Addendum B.

In any case, the direction error was pointed out, and ChatGPT then offered the revised material displayed in Figure 13.

As may be seen from Figure 13, ChatGPT basically swapped the lemma and its translation equivalent around, and (correctly) translated the English to Swahili for the definition and example sentence. All of this, together with the POS, is correct, while the first synonym is original: *mshikaji wa panya* 'mouse catcher', while the second is not clear: *kifalme* 'princess'.

Given that this reply (mostly) made sense, the next prompt asked to see more *uses* and thus more *meanings*, see Figure 14. For more on the core concept of 'mapping *meaning* onto *use*' see Hanks (2002).



You

This is not right. I wanted the direction Swahili to English.



ChatGPT

Apologies for the misunderstanding. Here's the corrected entry for 'cat' in a Swahili to English translation dictionary:

Title: Swahili to English Translation Dictionary

Entry: Paka

Part of Speech: Noun

English Equivalent: Cat

Definition:

Kiumbe mdogo wa nyama ambaye ameifanyiwa ukulima, ana manyoya laini, pua fupi, na kucha zinazoweza kuvutwa ndani, anayejulikana kwa tabia yake ya kuwa huru na ya kucheza.

Example Sentence:

Paka alimfuata panya jikoni.

Synonyms:

- Mshikaji wa Panya
- Kifalme



Figure 13: Additional ChatGPT-3.5 prompt and output for an entry in a Swahili–English translation dictionary, here for *paka* 'cat'



You

OK, makes sense. Does it have other uses and thus also other meanings?



ChatGPT

Yes, the word "paka" in Swahili primarily refers to a domestic cat, but it can also be used in other contexts or idiomatic expressions, giving rise to additional meanings or connotations. Here are a few examples:

1. **Verb Conjugation:** In Swahili, "paka" can also be used as a verb, meaning "to carry" or "to wear" when describing actions related to carrying something or wearing clothes. For example:
 - Nina paka begi langu. (I am carrying my bag.)
 - Anapaka kofia nyekundu. (He/she is wearing a red hat.)
2. **Idiomatic Expressions:** The word "paka" might be used in idiomatic expressions with meanings that extend beyond the literal translation. For instance:
 - Kuwa na paka mbili ni shida. (Literal translation: Having two cats is a problem.)
 - This expression might mean that having too many of something can lead to complications or difficulties.
3. **Metaphorical Use:** Sometimes "paka" can be used metaphorically to describe someone's behavior or characteristics, often implying slyness or cunning similar to the characteristics attributed to cats in various cultures.

Overall, while "paka" primarily means "cat" in Swahili, its usage can extend to various contexts and idiomatic expressions, allowing for different interpretations and meanings.

Figure 14: Second additional ChatGPT-3.5 prompt and output for an entry in a Swahili–English translation dictionary, here for *paka* 'cat'

As may be seen from Figure 14, this reply went horribly wrong. Instead of seeing an entry with multiple (numbered) senses, all the earlier information has been dispensed with. As for the three new numbered paragraphs, the first makes no sense at all, as it mixes up the noun *paka* 'cat' with the verb *-paka* 'smear, spread, apply' — which is anything but another sense of the noun *cat*; the second paragraph is an idiomatic expression with the noun *cat* — which is good; while the third paragraph talks about metaphorical uses in English, but there is not a single usage example in either language — so not useful. The wrap-up at the end is correct, but does not add anything new.

Courtesy obliges, one needs to end with the verb of communication *-pendekeza* 'recommend'. Figure 15 shows the usual prompt (which was also used for Portuguese in Figures 4 and 9), but now translated into Swahili, to obtain a COBUILD-style entry for a monolingual Swahili learners' dictionary. For the translation, here and further on, see Addendum C.



You

Nipe ingizo la kamusi ya Kiswahili-Kiingereza kwa lugha mbili kwa neno '-pendekeza'. Anza na maelezo ya marudio hapo juu. Kila maana inapaswa kuwa katika block iliyohesabiwa. Kisha kila kizuizi huanza na sehemu ya hotuba na maumbo ya kimofolojia ya maana husika. Hii inafuatwa na ufafanuzi wa maana katika Kiswahili, na tafsiri ya ufafanuzi huo kwa Kiingereza. Kumbuka kwamba ili kufafanua, inabidi utumie mtindo wa kufafanua sentensi kamili, ambao una vishazi viwili: kifungu cha kwanza kinaonyesha muundo wa matumizi, kifungu cha pili ni kifungu cha maana. Pia orodhesha tafsiri moja au zaidi zinazolingana katika Kiingereza katika kila maana. Ongeza mifano ya Kiswahili pamoja na tafsiri zao za Kiingereza zinazoonyesha matumizi na maana ya kila maana ya Kiswahili. Kwa sentensi za mfano, hakikisha unatumia miundo tofauti ya sentensi, ukirejelea watu tofauti; rejea hali zilizopita, za sasa, na zijazo; tofautisha sentensi ndefu na fupi za mifano; na ujumuishe maelezo mengine ili wewe, mpendwa ChatGPT, unitayarishie nyenzo za ubora wa juu za kamusi. Pia jumuishe matamshi ya ingizo kwa kutumia IPA, na ujumuishe orodha ya visawe na vinyume kwenye hisi ambapo orodha kama hizo zinafaa.

Figure 15: ChatGPT-3.5 prompt to obtain a COBUILD-style entry for a monolingual Swahili learners' dictionary, here for the verb *-pendekeza* 'recommend'

The reply by ChatGPT looked decidedly weird, as may be seen from Figure 16 and the explanation that follows.



ChatGPT

-pendekeza

Marudio: -pendekeza (Kitenzi)

1. Kitenzi - Tendeka

Maana: Kutoa maoni au mapendekezo ya kufanya kitu fulani kwa sababu unaamini kuwa ni bora au inafaa kufanywa.

Tafsiri: recommend

Visawe: kupendekeza, kushauri

Vinyume: kukataa, kupinga

Mifano:

- *Daktari alimpendekeza mgonjwa apumzike zaidi.*
(The doctor recommended that the patient rest more.)
- *Mwalimu alitupendekeza tufanye utafiti wa ziada kwa ajili ya mtihani ujao.*
(The teacher recommended that we do extra research for the upcoming exam.)

Matamshi: /pende'keza/

Visawe na Vinyume:

- Visawe: kupendekeza, kushauri
- Vinyume: kukataa, kupinga



Figure 16: ChatGPT-3.5 output for the Swahili lemma *-pendekeza* 'recommend', for the prompt seen in Figure 15

From Figure 16, it is quickly clear that what ChatGPT-3.5 produces is of little value: The lemma is repeated twice, the part of speech (*kitenzi* 'verb') is repeated twice, the synonym and antonym slots are repeated twice, and the first so-called synonym simply takes the lemma and puts it into class 15, which is nothing special, as all verbs in any Bantu language belong to class 15, and thus start with *ku-*. The definition does not follow the COBUILD-style, and even so, turns the verb into the noun *mapendekezo* 'recommendation' to explain the verb *-pendekeza* 'recommend', which is bad lexicography. And in terms of layout, there is also no reason for the number 1, at the first and only *kitenzi* 'verb' block. Conclusion: ChatGPT-3.5 can simply not handle Swahili lexicography.

2.3.2 Nation-visible Xhosa

Given that Xhosa is a much smaller Bantu language compared to Swahili, basically only spoken in South Africa and recognized as a minority language in Zimbabwe, the approach this time was to try to make it super-simple, by merely asking for a small number of random short dictionary entries for Xhosa, as seen in Figure 17.



You

Please give me a short list of Xhosa dictionary entries.



ChatGPT

Certainly! Here are some common Xhosa dictionary entries:

1. "Molo" - Hello
2. "Enkosi" - Thank you
3. "Mhlekezzi" - Sir / Mr.
4. "Mhlekezzi" - Madam / Mrs.
5. "Unjani?" - How are you?
6. "Ndiphilile" - I am fine
7. "Ndiyavuya ukukwazi" - Nice to meet you

8. "Hamba kahle" - Goodbye
9. "Indlu" - House
10. "Inkosana" - Child
11. "Ibhokisi" - Box
12. "Iziko" - Museum
13. "Indawo" - Place
14. "Umntu" - Person
15. "Umntwana" - Child
16. "Umlilo" - Fire
17. "Inja" - Dog
18. "Inyoni" - Bird
19. "Igqirha" - Doctor
20. "Isango" - Door



Figure 17: ChatGPT-3.5 prompt and reply to obtain a small number of random short dictionary entries for Xhosa

Figure 17 hardly displays a list of dictionary entries, only just a random vocabulary list. Unfortunately, it is abundantly clear that ChatGPT cannot even pick 20 frequent words from Xhosa, as item number 18, *inyoni* 'bird', is from the neighbouring and bigger language Zulu (27 million speakers; spoken in at least 6 countries), see Figure 18.

inyoni ★★ *noun* 9/10 (*pl. izinyoni*) ■ bird
♦ Silalel' izinyoni zitshiloza. *We listen to the birds chirping.*
► **kwelenyoni** ■ topmost part of a tree
♦ Wathi lapho esephezulu kwelenyoni, wase elula isandla ukuze abambe igatsha elithele kakhulu izithelo. *When he was at the top of the tree, he stretched out his hand to grab a branch that was full of fruit.*

Figure 18: Dictionary entry for *inyoni* 'bird' in the *Oxford Zulu School Dictionary* (De Schryver 2010: 96)

Rather, the word for bird in Xhosa is *intaka*, see Figure 19.

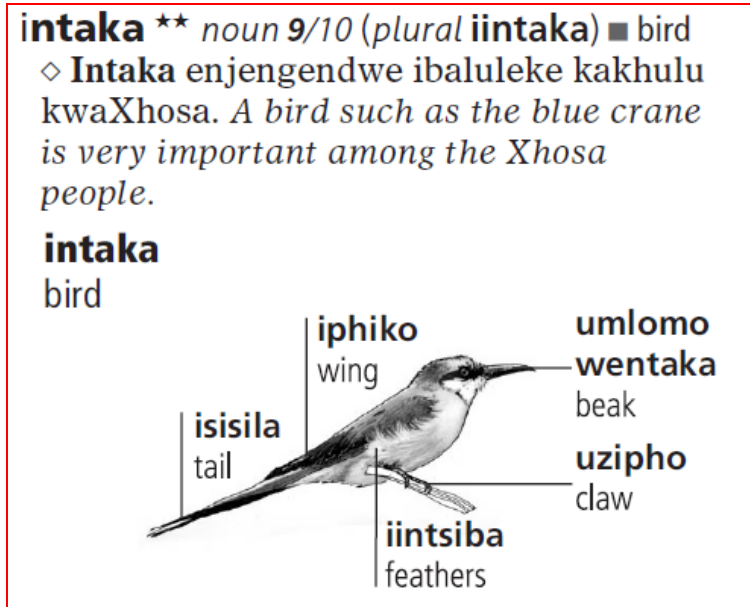


Figure 19: Dictionary entry for *intaka* 'bird' in the *Oxford Xhosa School Dictionary* (De Schryver and Reynolds 2014: 89)

Even several translations are debatable: item 10, *inkosana*, is translated as 'child', but see Figure 20, where it is prince.

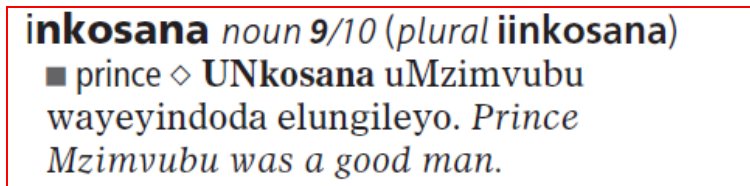


Figure 20: Dictionary entry for *inkosana* 'prince' in the *Oxford Xhosa School Dictionary* (De Schryver and Reynolds 2014: 87)

Item 12, *iziko*, is translated as 'museum', but see Figure 21, where it is '1. fireplace; 2. centre; institution'.

iziko ★★ *noun* 5/6 (plural **amaziko**)
1 ■ fireplace ◇ Lumka, **iziko** lisheshushu. *Be careful, the fireplace is still hot.* **2** ■ centre; institution ◇ **IZiko** leeLwimi likhuthaza zonke iilwimi. *The Centre for Languages encourages the use of all languages.*

Figure 21: Dictionary entry for *iziko* '1. fireplace; 2. centre; institution' in the *Oxford Xhosa School Dictionary* (De Schryver and Reynolds 2014: 119)

While the translation for museum is actually a loanword, namely *imyuziyam*, see Figure 22.

museum ★ (say myoo-zee-uhm) *noun* (plural **museums**) ■ imyuziyam ◇ In this **museum** you can see many interesting things from Xhosa history. *Kule myuziyam uya kubona izinto ezininzi ezinika umdla ezivela kwimbali yamaXhosa.*

Figure 22: Dictionary entry for *museum* in the *Oxford Xhosa School Dictionary* (De Schryver and Reynolds 2014: 418)

Therefore, and as an overall conclusion so far, with ChatGPT-3.5 English not only does much better than other big languages (like Portuguese), but small languages (like Zulu) take over even smaller ones (like Xhosa).

2.3.3 Local-visible Lusoga

The next level, and still within Bantu, concerns what happens to even smaller languages of the primarily oral type, i.e. Bantu languages for which there is hardly any written documentation, no matter the actual number of sometimes impressively high number of speakers. Lusoga is such a case, spoken by up to 4.5 million speakers in the southeast of Uganda. The prompt was kept even more basic than any of the previous ones, just asking for the top frequent words in the language, as displayed in Figure 23.



You

User

Can you give me a list of top frequent words used in Lusoga?



ChatGPT

Lusoga is a Bantu language spoken in Uganda, primarily by the Basoga people. Here are some common Lusoga words:

1. "Wasuze otya" - How are you? (informal)
2. "Wasibye otya" - How are you? (formal)
3. "Wee" - Yes
4. "Nedda" - No
5. "Webale" - Thank you
6. "Nkikola" - I am doing
7. "Ndiga" - I am eating
8. "Lala bulungi" - Good night
9. "Oli wa?" - Who are you?
10. "Gyebale ko" - I am sorry
11. "Nsonyiwa" - Welcome
12. "Mugende gwe?" - Where are you going?
13. "Gwe nna?" - Who, me?
14. "Omukyala" - Woman
15. "Omuntu" - Person
16. "Obwavu" - Sickness
17. "Embeera" - Situation
18. "Obubonero" - Example
19. "Amagezi" - Knowledge
20. "Obusungu" - Wisdom

These are some common words in Lusoga that you may encounter in everyday conversations.

Figure 23: ChatGPT-3.5 prompt and reply to obtain a list of the frequent words in Lusoga

Firstly, this list has moved away from a list of words, to mostly a list of phrases. Secondly, an analysis of the ChatGPT-3.5 output by Minah Nabirye, a native speaker of Lusoga, reveals that only 7 out of the 20 are correct, with two of those correct ones actually only in the neighbouring and much bigger language Luganda (20 million speakers; spoken in Uganda immediately to the west of Lusoga, as well as in Rwanda). See Table 1 for the details.

Table 1: Analysis of the ChatGPT-3.5 output of so-called 'frequent words in Lusoga'

ChatGPT-3.5	ChatGPT-3.5 translation	Language	Status	Lug(anda)	Lus(oga)	English
1 <i>Wasuze otya</i>	How are you? (informal)	Lug/Lus	Wrong translation	-	-	Good morning
2 <i>Wasibye otya</i>	How are you? (formal)	Lug/Lus	Wrong form	<i>Osiibye otya?</i>	<i>Osiibye otya?</i>	-
3 <i>Wee</i>	Yes	??	Unknown form	-	-	-
4 <i>Nedda</i>	No	Lug	Correct	-	-	-
5 <i>Webale</i>	Thank you	Lug/Lus	Wrong form	<i>Weebale</i>	<i>Weebale</i>	-
6 <i>Nkikola</i>	I am doing	Lug	Correct	-	-	-
7 <i>Ndiga</i>	I am eating	Lug	Wrong translation	-	-	Sheep
8 <i>Lala bulungi</i>	Good night	Lug/Lus	Wrong form	<i>Sula bulungi</i>	<i>Sula bulungi</i>	-
9 <i>Oli wa?</i>	Who are you?	Lug/Lus	Wrong translation	-	-	Where are you?
10 <i>Gyebale ko</i>	I am sorry	Lug	Wrong translation	<i>Gyebaleko</i>	<i>Gyebaleku</i>	Thank you for your work
11 <i>Nsonyiwa</i>	Welcome	Lug	Wrong translation	-	<i>Nsonhiwa</i>	Forgive me
12 <i>Mugende gwe?</i>	Where are you going?	??	Unknown form	<i>Mugende (*gwe)</i>	<i>Mugye (*iwe)</i>	You go(pl) (gwe/iwe = you)
13 <i>Gwe nna?</i>	Who, me?	??	Unknown form	-	-	-
14 <i>Omukyala</i>	Woman	Lug/Lus	Correct	-	-	-
15 <i>Omuntu</i>	Person	Lug/Lus	Correct	-	-	-
16 <i>Obuwavu</i>	Sickness	Lug/Lus	Wrong translation	-	-	Poverty
17 <i>Embeera</i>	Situation	Lug/Lus	Correct	-	-	-
18 <i>Obubonero</i>	Example	Lug/Lus	Correct	-	-	-
19 <i>Amagezi</i>	Knowledge	Lug/Lus	Correct	-	-	-
20 <i>Obusungu</i>	Wisdom	Lug/Lus	Wrong translation	-	-	Anger

That means that all in all, there are merely 5 correct Lusoga forms out of 20, or thus one quarter (25%). Conclusion: If three quarters (75%) of the returned material is erroneous on the most basic of the basic lexicographic tasks, there is simply no point in trying anything else with ChatGPT-3.5 for Lusoga, even though it is spoken by a respective number of as many as 4.5 million speakers.

3. What to do next?

The main take-away from this article is that the lexicographic prompt introduced by De Schryver (2023b), being a single prompt with which to compile an entire English-style COBUILD learners' dictionary, may indeed work well for English, but is not good enough for other languages. At least, not using ChatGPT-3.5. The field of GenAI chatbots, however, has moved on since then, so a serious attempt must first be made to see how later out-of-the-box models like GPT-4o or Gemini-2.5 perform the job on other languages than English. Second, an attempt should be made to move from an out-of-the-box model, to one where one starts customising the model. These two aspects are precisely the topics of a follow-up article (see De Schryver 2025).

PS.

In this day and age, and as the secretary of the exam committee for African Studies at UGent where the author of this publication is tasked with checking

all BAs, MAs and PhDs on their correct use of GenAI, the reader needs to be informed that GenAI was *not* used for any of the writings in this article, except for some of the sub-headings. Addendum D, available online (click [here](#)), displays the full chat between the author and OpenAI's GPT-4o in this respect. In itself, it provides an interesting perspective on using GenAI as a writing assistant. Research on using AI as a writing assistant in our field started with the work of Henrik Køhler Simonsen, first published as Leroyer and Simonsen (2019), and from then on, year after year, as Simonsen (2020, 2021, 2022, 2023). Subsequent to the release of ChatGPT, other colleagues followed (Huete-García and Tarp 2024, Tarp and Nomdedeu-Rull 2024).

Reading through the entire chat of Addendum D, available online (click [here](#)), reveals how good GPT-4o has become as a lexicographer; it is like consulting a wide variety of six different dictionaries, going in all sort of directions: (1) magnitudes, scales, sizes, (2) demographic speaker populations, (3) socio-technological reach, visibility, language power, (4) digital inclusion, corpus resources, language equity, (5) localness without grandeur, geographic nuance, judgmental rings, and (6) progression from hyper to local, capturing visibility and rank, balancing clarity, tone and structure. Simply stunning, and it is unlikely that any human lexicographer can do this in the fractions of seconds that GPT-4o spends on each of those levels. For the interested reader, even some screen recordings exist, from which the varied speed of the GenAI replies can be monitored. That said, one of the adjudicators observed all of the following, with which I fully agree:

That's not a fair point. It is, in my opinion overly simplistic and reductive to use generation speed as the sole criterion. What about training time or about the time it takes to critically check the responses? If one argues at this level, one could also point to the questionable use of resources in training and running current LLMs: not only does this speed come at the price of enormous resource consumption (electricity and water), but the language training material was also most likely acquired in violation of copyright laws. When praising the speed of LLM's responses, I believe it is important not to remain silent about these issues. (Anonymous adjudicator for this article, to whom sincere thanks are due.)

Endnotes

1. See <https://www.collinsdictionary.com/dictionary/english/strike>, sense 3.
2. For more on the FSD, see Hanks (1987), Barnbrook (2002), Lew (2023), and Rundell (2024).
3. See <https://www.collinsdictionary.com/dictionary/english/bank>, finance and storage, sense 3.
4. See <https://www.collinsdictionary.com/dictionary/english/bank>, finance and storage, sense 4.
5. See <https://www.definify.com/word/bancar>.
6. See <https://africanlanguages.com/swahili/>.
7. See <https://www.google.com/>.

References

- Barnbrook, G. 2002. *Defining Language: A Local Grammar of Definition Sentences* (Studies in Corpus Linguistics 11). Amsterdam: John Benjamins.
- Chomsky, N. 1957. *Syntactic Structures* (Janua Linguarum 4). The Hague: Mouton.
- De Schryver, G.-M. 2003. Online Dictionaries on the Internet: An Overview for the African Languages. *Lexikos* 13: 1-20.
- De Schryver, G.-M. 2010. *Oxford Bilingual School Dictionary: Zulu and English / Isichazamazwi Sesikole Esinezilimi Ezimbili: IsiZulu NesiNgesi, Esishicilelwe abakwa-Oxford*. Cape Town: Oxford University Press Southern Africa.
- De Schryver, G.-M. 2023a. *Contemporary Lexicography: A Case Study in AI [Keynote Lecture]*. Paper presented at the Inaugural Conference of the Association of Lexicography for the Americas: South, Central, Caribbean, and Mexico (Americalex-S), São Paulo, Brazil, 20–25 October 2023. <https://youtu.be/watch?v=yAxzTx7A2LU>
- De Schryver, G.-M. 2023b. Generative AI and Lexicography: The Current state of the Art Using ChatGPT. *International Journal of Lexicography* 36(4): 355-387 + ten pages of supplementary material online.
- De Schryver, G.-M. 2024a. *Customising LLMs for Lexicography [Keynote Lecture]*. Paper presented at the 8th International Workshop on Cognitive Aspects of the Lexicon (CogALex-VIII), co-located with the Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), Turin, Italy, 20 May 2024.
- De Schryver, G.-M. 2024b. *Generative Artificial Intelligence & Three Lexicographic Fs: FANTASTIC for Monolingual English Dictionaries, FAKE for Translation Dictionaries, FAIL for Exotic Dictionaries [Invited Talk]*. Paper presented at the Department of English, University of Missouri, Columbia, MO, USA, 5 April 2024. https://youtu.be/zUI0GrPrH_0
- De Schryver, G.-M. 2024c. The Road towards Fine-tuned LLMs for Lexicography. Krek, S. (Ed.). 2024. *Book of Abstracts of the Workshop 'Large Language Models and Lexicography' @ EURALEX 2024, 8 October 2024, Cavtat, Croatia: 6-11*. Ljubljana: ELEXIS Association.
- De Schryver, G.-M. 2025. Customised GPTs for Lexicography. *Lexikos* 35(2): 397-422.
- De Schryver, G.-M. forthcoming-a. 100 Scholarly Echoes of Generative AI in Lexicography.
- De Schryver, G.-M. forthcoming-b. Citing the Future: The Legacy of Patrick Hanks.
- De Schryver, G.-M. and M. Reynolds. 2014. *Oxford Bilingual School Dictionary: IsiXhosa and English / Oxford isiXhosa-isiNgesi English-isiXhosa Isichazi-magama Sesikolo*. Cape Town: Oxford University Press Southern Africa.
- Fuertes-Olivera, P.A. 2024. Making Lexicography Sustainable: Using ChatGPT and Reusing Data for Lexicographic Purposes. *Lexikos* 34(1): 123-140.
- Hanks, P. 1987. Definitions and Explanations. Sinclair, J.M. (Ed.). 1987. *Looking Up: An Account of the COBUILD Project in Lexical Computing and the Development of the Collins COBUILD English Language Dictionary: 116-136*. London: Collins ELT.
- Hanks, P. 2000. Do Word Meanings Exist? *Computers and the Humanities* 34(1–2): 205-215.
- Hanks, P. 2002. Mapping Meaning onto Use. Corréard, M.-H. (Ed.). 2002. *Lexicography and Natural Language Processing. A Festschrift in Honour of B.T.S. Atkins: 156-198*. s.l.: Euralex.
- Hanks, P. 2013. *Lexical Analysis: Norms and Exploitations*. Cambridge, MA: The MIT Press.

- Huete-García, Á. and S. Tarp. 2024. Training an AI-based Writing Assistant for Spanish Learners: The Usefulness of Chatbots and the Indispensability of Human-assisted Intelligence. *Lexikos* 34: 21-40.
- Jakubíček, M. and M. Rundell. 2023. The End of Lexicography? Can ChatGPT Outperform Current Tools for Post-editing Lexicography? Medved', M., M. Měchura, C. Tiberius, I. Kosem, J. Kallas, M. Jakubíček and S. Krek (Eds.). 2023. *Proceedings of the eLex 2023 Conference: Electronic Lexicography in the 21st Century*: 518-533. Brno: Lexical Computing.
- Leroyer, P. and H.K. Simonsen. 2019. Google Translate: trussel eller redning for oversættelsesordbøger? [Google Translate: Threat or Salvation for Translation Dictionaries?]. *LexicoNordica* 26: 95-115.
- Lew, R. 2023. ChatGPT as a COBUILD Lexicographer. *Humanities and Social Sciences Communications* 10(704): 10 pp.
- Lew, R. 2024. Dictionaries and Lexicography in the AI Era. *Humanities and Social Sciences Communications* 11(426): 8 pp.
- Lew, R., B. Ptasznik and S. Wolfer. 2024. The Effectiveness of ChatGPT as a Lexical Tool for English, Compared with a Bilingual Dictionary and a Monolingual Learner's Dictionary. *Humanities and Social Sciences Communications* 11(1324): 10 pp.
- McKean, E. and W. Fitzgerald. 2023. The ROI of AI in Lexicography [Keynote Lecture]. 2023. *Proceedings of the 16th International Conference of the Asian Association for Lexicography: "Lexicography, Artificial Intelligence, and Dictionary Users"*: 18-27. Seoul: Yonsei University.
- Nichols, W. 2023. Invisible Lexicographers, AI, and the Future of the Dictionary [Keynote Lecture]. Medved', M., M. Měchura, C. Tiberius, I. Kosem, J. Kallas, M. Jakubíček and S. Krek (Eds.). 2023. *Book of Abstracts of the eLex 2023 Conference: Electronic Lexicography in the 21st Century*: 3-3. Brno: Lexical Computing.
- Pavvydytė, K. 2024. *ChatGPT generuojami iliustraciniai laukai: turinio ir kokybės analizė bei palyginimas su BLKŽ [Illustrative Sentences Generated by ChatGPT: The Analysis of Its Content, Quality, and Comparison with the Dictionary of the Standard Lithuanian Language]*. MA Dissertation. Vilnius: Vilnius University.
- Ptasznik, B., S. Wolfer and R. Lew. 2024. A Learners' Dictionary Versus ChatGPT in Receptive and Productive Lexical Tasks. *International Journal of Lexicography* 37(3): 322-336.
- Rees, G.P. and R. Lew. 2024. The Effectiveness of OpenAI GPT-generated Definitions Versus Definitions from an English Learners' Dictionary in a Lexically orientated Reading Task. *International Journal of Lexicography* 37(1): 50-74.
- Rundell, M. 2024. Obituary: Patrick Wyndham Hanks. *International Journal of Lexicography* 37(2): 135-142.
- Simonsen, H.K. 2020. Augmented Writing: Nye muligheder og nye teorier [Augmented Writing: New Possibilities and New Theories]. Sandström, C., U.-M. Forsberg, C. af Hållström-Reijonen, M. Lehtonen and K. Ruppel (Eds.). 2020. *Nordiska studier i lexikografi 15, Rapport från 15 konferensen om lexikografi i Norden, Helsingfors, 4-7 juni 2019 [Nordic Studies in Lexicography 15, Report from the 15th Nordic Lexicography Conference, Helsinki, 4-7 June 2019]*: 307-315. Helsinki: Nordic Association for Lexicography.
- Simonsen, H.K. 2021. AI Writers in Language Learning. Chang, M., N.-S. Chen, D.G. Sampson and A. Tlili (Eds.). 2021. *Proceedings of the 21st International Conference on Advanced Learning Technologies (ICALT 2021), Online, 12-15 July 2021*: 238-240. Los Alamitos, CA: IEEE.

- Simonsen, H.K.** 2022. AI Text Generators and Text Producers. Chang, M., N.-S. Chen, M. Dascalu, D.G. Sampson, A. Tlili and S. Trausan-Matu (Eds.). 2022. *Proceedings of the 22nd International Conference on Advanced Learning Technologies (ICALT 2022), Bucharest, Romania, 1–4 July 2022*: 218-220. Los Alamitos, CA: IEEE.
- Simonsen, H.K.** 2023. Bilinguale ordbøger vs. AI-skriveassistenter? [Bilingual Dictionaries vs. AI Writing Assistants?]. *LexicoNordica* 30: 117-136.
- Tarp, S. and A. Nomdedeu-Rull.** 2024. Who Has the Last Word? Lessons from Using ChatGPT to Develop an AI-based Spanish Writing Assistant. *Círculo de Lingüística Aplicada a la Comunicación [Circle of Linguistics Applied to Communication]* 97: 309-321.
- Zhong, A.** 2025. Prompts for Language Learners: A Practical Guide to Using DeepSeek as a Dictionary. *Lexikos* 35(1): 274-285.

Addendum A: Using ChatGPT-3.5 — For Portuguese Lexicography [translations into English]

- Translation of Figure 1:

Please give me a COBUILD-style monolingual dictionary entry for 'strike'. Start with the frequency information at the top. Each sense should be in a numbered block. Each block then starts with the part of speech and the morphological forms of the respective sense. This is followed by a sense definition and sense examples that illustrate both the use and the meaning of each particular sense. For the example sentences, make sure to use different sentence structures, referring to different people; refer to past, present, and future situations; vary long and short example sentences; and include other elaborations so that you, dear ChatGPT, produce high-quality dictionary material for me.

- Translation of Figure 3:

Thanks, but actually I wanted to see a monolingual dictionary entry for the Portuguese entry '*greve*' (which is 'strike' in English).

I apologize for the confusion earlier. Here is a monolingual dictionary entry for the word '*greve*' in Portuguese:

greve (feminine noun)

Frequency: The word *strike* is among the 10,000 most frequent words in Portuguese.

Definition: A *strike* is a stoppage organized by workers in which they stop working as a form of protest, usually to demand better working conditions, higher wages, or other rights from employers.

Examples:

1. Factory workers went on *strike* to demand a pay rise.
2. The teachers decided to go on *strike* because of the lack of investment in education.
3. The truckers' *strike* caused supply shortages across the country.

Note: The word 'strike' is often used in the context of labour protests and is a right guaranteed by labour laws in many countries. It can be carried out by various professional sectors, including teachers, doctors, drivers, and others, as a means of putting pressure on employers or governments to meet workers' demands.

- Translation of Figure 4:

Please give me a COBUILD-style monolingual dictionary entry for '*recomendar*'. Start with the frequency information at the top. Each sense should be in a numbered block. Each block then starts with the part of speech

and the morphological forms of the respective sense. This is followed by a sense definition and sense examples that illustrate both the use and the meaning of each particular sense. For the example sentences, make sure to use different sentence structures, referring to different people; refer to past, present, and future situations; vary long and short example sentences; and include other elaborations so that you, dear ChatGPT, produce high-quality dictionary material for me. Also include a pronunciation of the entry using IPA, and include a list of synonyms and antonyms at the senses where such lists are relevant.

Note that a final line was added to the prompt, as was done at the end of the initial study as well (De Schryver 2023b); highlighted in grey above.

- Translation of Figure 5a:

recomendar (verb)

Frequency: The word *recomendar* is among the 2,000 most frequent words in Portuguese.

Sense 1: *To express a favourable opinion about something or someone; to advise or suggest something to someone*

Morphological forms: recommend, recommends, recommending, recommended

Definition: When you recommend something, you express a positive opinion about it, often by advising or suggesting that someone else do it.

Examples:

1. I *recommend* this restaurant; the food is delicious.
2. She *recommended* a good book for me to read on vacation.
3. They are *recommending* that everyone wears masks to protect against the virus.

Pronunciation: /ʁe.ko.mẽ.'daʁ/

Synonyms (Sense 1): advise, suggest, indicate, endorse, approve

Antonyms (Sense 1): advise against, disapprove of, discourage

- Translation of Figure 5b:

Sense 2: *Presenting someone as fit or suitable for something, such as a job or position*

Morphological forms: recommend, recommends, recommending, recommended

Definition: In this sense, to recommend means to present someone as qualified for a job or position, indicating their skills, qualities or achievements.

Examples:

1. He was *recommended* for the position of manager because of his experience and leadership skills.
2. The professor *recommended* several excellent students for scholarships at the university.
3. The letter of recommendation *recommended* the candidate as highly competent for the job.

Pronunciation: /ʁe.ko.mẽ.'daʁ/

Synonyms (Sense 2): indicate, endorse, refer, endorse, sponsor

Antonyms (Sense 2): disapprove, disrecommend, criticize, censure

- Translation of Figure 5c:

Sense 3: *To prescribe or suggest something as a treatment or solution*

Morphological forms: recommend, recommends, recommending, recommended

Definition: In medical or professional terms, recommending involves prescribing or suggesting a treatment, course of action or solution for a specific condition or problem.

Examples:

1. The doctor *recommended* rest and medication to treat the flu.
2. The specialist *recommended* a balanced diet and exercise to improve the patient's health.
3. She *recommended* that he consult a lawyer to resolve the legal issue.

Pronunciation: /ʁe.ko.mẽ.'daʁ/

Synonyms (Sense 3): prescribe, advise, guide, suggest, indicate

Antonyms (Sense 3): advise against, disallow, disapprove of, discourage

- Translation of Figure 9:

Give me a bilingual Portuguese–English dictionary entry for the word 'banco'. Start with the frequency information at the top. Each sense should be in a numbered block. Each block then starts with the part of speech and the morphological forms of the respective sense. This is followed by a sense definition in Portuguese, and translation of that definition into English. Note that to define, you have to use the full-sentence defining style, which consists of two clauses: the first clause illustrates the usage pattern, the second clause is a paraphrase of meaning. Also list one or more translation equivalents in English at each sense. Add Portuguese examples with their English translations that illustrate both the use and the meaning of each particular Portuguese sense. For the example sentences, make sure to use different sentence structures, referring to different people; refer to past, present, and future situations; vary long and short example sentences; and include other elaborations so that you, dear ChatGPT, produce high-quality dictionary material for me. Also include a pronunciation of the entry using IPA, and include a list of synonyms and antonyms at the senses where such lists are relevant.

The grey highlight added on top of an extra sentence in the prompt here, marks the attempt to make sure a COBUILD-style definition, so a full-sentence definition (FSD), is obtained by ChatGPT-3.5.

- Translation of Figure 10a:

bank ['bẽku]

Frequency: *****

1. Noun (Financial Institution)

- **Morphological forms:** banks (plural)
- **Sense:** A financial institution where people and companies can deposit money, obtain loans and carry out various financial transactions.
 - *Usage Pattern:* A bank is a place where people and companies can deposit money, obtain loans and carry out financial transactions.
 - *Definition:* A bank is an establishment where people and businesses can put their money, get loans and manage various financial transactions.
- **English Translation Equivalents:** Bank
 - *Example 1 in Portuguese:* I need to go to the bank to withdraw some money.
 - *Translation 1 in English:* I need to go to the bank to withdraw some money.
 - *Example 2 in Portuguese:* The bank offers various financial services, including savings accounts and mortgages.
 - *Translation 2 in English:* The bank offers various financial services, including savings accounts and mortgages.
- **Pronunciation:** ['bẽku]
- **Synonyms:** financial institution, bank branch
- **Antonyms:** customer, borrower

- Translation of Figure 10b:

2. Noun (Bank or Sloping Area)

- **Morphological forms:** banks (plural)
- **Sense:** The side of a river, lake or other body of water; often referring to the raised area or the margin.
 - *Usage Pattern:* The margin of a river is the side of the river.
 - *Definition:* The margin of a river, for example, denotes the land next to the river.
- **Translation Equivalents in English:** Bank, Riverside
 - *Example 1 in Portuguese:* They decided to have a picnic on the margin of the river.
 - *Translation 1 in English:* They decided to have a picnic on the bank of the river.
 - *Example 2 in Portuguese:* The boat is anchored near the margin of the lake.
 - *Translation 2 in English:* The boat is anchored near the bank of the lake.
- **Pronunciation:** ['bẽku]
- **Synonyms:** margin, edge
- **Antonyms:** centre, middle

- Translation of Figure 10c:

3. Verb

- **Morphological forms:** *banco (3rd person singular present tense), *bancou (past tense), *bancando (present participle), *banco (3rd person singular present tense)
- **Sense:** To deposit or store something in a bank, or to entrust or store something for future use.
 - *Usage Pattern:* Someone *banks something in a bank, or they *bank something for a specific purpose.
 - *Definition:* When someone deposits something in a bank, they are storing it in a financial institution or entrusting it for future use.
- **English Translation Equivalents:** Deposit, Rely on
 - *Example 1 in Portuguese:* She deposits her savings in a local bank.
 - *Translation 1 in English:* She banks her savings in a local bank.
 - *Example 2 in Portuguese:* He's counting on you winning the competition with your skills.
 - *Translation 2 in English:* He's banking on winning the competition with his skills.
- **Pronunciation:** ['bẽku]
- **Synonyms:** deposit, entrust
- **Antonyms:** withdraw, discredit

Note that the stars (*) were added to this translation, to mark (some of) the various errors in Portuguese. See the main text for the explanation.

<http://lexikos.journals.ac.za>; <https://doi.org/10.5788/35-2-2108> (Volle uitgawe / Full volume)
<https://lexikos.journals.ac.za>; <https://doi.org/10.5788/35-2-2103> (Article)

Addendum B: Online Swahili–English Dictionary / Kamusi ya Kiswahili–Kiingereza

- Without support, the following online Swahili–English dictionary — having been free and top-ranked by Google⁷ for over two decades — will have to be taken down (<https://africanlanguages.com/swahili/>).

The screenshot shows a web browser window with the URL <https://africanlanguages.com/swahili/>. A pink banner at the top reads: "Dear Swahili dictionary user, Unfortunately, we are out of funds for hosting costs for this site and urgently need donations to keep it running. Please consider donating at PayPal.me/davidjoffe. Thank you! If you would like to sponsor this site (for example, with a retail advertising banner here), [Contact Us](#)". Below this, a grey box contains links: "New 3 Oct 2023: New article on [Generative AI and Lexicography](#)", "[Song Trein na Pretoria](#)", and "[Creative Gardening with Indigenous Plants](#)". A white box with a blue border contains a PayPal donation form with the text: "Please donate now with Payfast to help keep this Swahili dictionary online and going - thank you. - David Joffe". The form has fields for "Amount (Rands)" with a value of 80, "(R80 = ~\$4.30)", and "Quantity" with a value of 1, and a green "Pay Now" button. Below the form, a grey box says "This dictionary needs support to help pay our costs, please help support this site:" with buttons for "PayPal donations" and "Bitcoin donations". A blue header bar reads "Online Swahili - English Dictionary" with a link "Change page language: [Kiswahili](#) | [English](#)". Below this is a search bar with the text "search / tafuta" and a "Search" button. A grey box contains the text: "Welcome to africanlanguages.com/swahili/ (now more mobile-friendly! please give the mobile version a try)", "If you would like to advertise on this page, please [contact us](#).", and a link "Click here to try or buy a **downloadable Windows version** of this dictionary » (Sales help support this dictionary)". At the bottom, a grey box reads: "News 13 Oct 2014: This dictionary now also returns results from the Microsoft Swahili terminology list, below the regular results!"

← → ↻ 🔒 <https://africanlanguages.com/swahili/> 90% ☆ 🔍 Search

Dear Swahili dictionary user,
Unfortunately, we are out of funds for hosting costs for this site and urgently need donations to keep it running.
Please consider donating at PayPal.me/davidjoffe. Thank you!
If you would like to sponsor this site (for example, with a retail advertising banner here), [Contact Us](#)

New 3 Oct 2023: New article on [Generative AI and Lexicography](#)
[Song Trein na Pretoria](#)
[Creative Gardening with Indigenous Plants](#)

Please donate now with Payfast to help keep this Swahili dictionary online and going - thank you. - David Joffe
Amount (Rands)
(R80 = ~\$4.30)
Quantity:
[Pay Now](#)

This dictionary needs support to help pay our costs, please help support this site: [PayPal donations](#) [Bitcoin donations](#)

Online Swahili - English Dictionary
Change page language: [Kiswahili](#) | [English](#)

[Search](#)

Welcome to africanlanguages.com/swahili/ (now more mobile-friendly! please give the mobile version a try)
If you would like to advertise on this page, please [contact us](#).
[Click here to try or buy a downloadable Windows version of this dictionary »](#)
(Sales help support this dictionary)

News 13 Oct 2014: This dictionary now also returns results from the Microsoft Swahili terminology list, below the regular results!

Addendum C: Using ChatGPT-3.5 — For Swahili Lexicography [here the translations into English]

- Translation of Figure 15:

Give me a bilingual Swahili–English dictionary entry for the word '-pendekeza'. Start with the frequency information at the top. Each sense should be in a numbered block. Each block then starts with the part of speech and the morphological forms of the respective sense. This is followed by a sense definition in Swahili, and translation of that definition into English. Note that to define, you have to use the full-sentence defining style, which consists of two clauses: the first clause illustrates the usage pattern, the second clause is a paraphrase of meaning. Also list one or more translation equivalents in English at each sense. Add Swahili examples with their English translations that illustrate both the use and the meaning of each particular Swahili sense. For the example sentences, make sure to use different sentence structures, referring to different people; refer to past, present, and future situations; vary long and short example sentences; and include other elaborations so that you, dear ChatGPT, produce high-quality dictionary material for me. Also include a pronunciation of the entry using IPA, and include a list of synonyms and antonyms at the senses where such lists are relevant.

- Translation of Figure 16:

-pendekeza

Repeat: -pendekeza (Verb)

1. Verb - To do

Meaning: To give an opinion or recommendation to do something because you believe it is good or appropriate to do it.

Translation: recommend

Synonyms: to recommend, advise

Antonyms: to reject, to oppose

Examples:

- *The doctor recommended that the patient rest more.*
(The doctor recommended that the patient rest more.)
- *The teacher recommended that we do extra research for the upcoming exam.*
(The teacher recommended that we do extra research for the upcoming exam.)

Pronunciation: /pende'keza/

Synonyms and Antonyms:

- Synonyms: to recommend, advise
- Antonyms: to reject, oppose

Customised GPTs for Lexicography

Gilles-Maurice de Schryver, *BantUGent & LT³,
Ghent University, Ghent, Belgium;
and Department of African Languages,
University of Pretoria, Pretoria, South Africa
(gillesmaurice.deschryver@UGent.be)
(<https://orcid.org/0000-0001-7272-9878>)*

Abstract: About a year and a half ago, lexicographers realised that the enthusiasm they had when using ChatGPT-3.5 for English lexicography, could not be carried over to other languages, not even big ones like Portuguese. A year ago, the lexicographic community was made aware that ChatGPT-3.5 could not be used at all for small languages, and especially not for oral ones like Lusoga. In mid-2025, we notice that the models have improved drastically. Both Gemini-2.5 and especially GPT-4o perform miracles on languages like Portuguese, and for undocumented languages like Lusoga, customising GPT-4o by feeding it corpora as well as entire dictionaries is a major step forward, down to tackling neologisms. We are not yet there in full but moving forward.

Keywords: GENERATIVE ARTIFICIAL INTELLIGENCE (GENAI), LARGE LANGUAGE MODELS (LLMs), GENERATIVE PRE-TRAINED TRANSFORMER (GPT), CUSTOMISED GPTs, LEXICOGRAPHY, DICTIONARIES, LEXICOGRAPHERS, PORTUGUESE, LUSOGA

Obufunze: GPT emmumbe eya GenAI mw'isomo ly'okuwandiika amawanika.

Mu ibbanga nga ely'omwaka mulala n'ekitundu eliweile, abawandiisi b'amawanika baatandiika okuteebeleza nti eitutumu lye baatuukaku nga bakozeza ChatGPT-3.5 ku lulimi Olungeleza, tilisobola kwongezebwayo kukola ku nnimi dhindi ng'Olupootigo. Omwaka oguweile, abawandiisi b'amawanika bano beene b'amala b'ayolekebwa nti ChatGPT-3.5 tesobola kukozesebwa ku nnimi dhitamanhiibwa inho, nailala ng'edho edhikaasinga okuba obw'omu ndhogela ng'Olusoga. Ebyafaayo by'ensonga dhino dhombi waigulu byandhulwa mu kitundu ky'olupapula luno ekisooka. Kino n'ekitundu eky'okubili ekilaga ebibailewo okutuuka mu magatigati g'omwaka 2025, ekilaga nti enteeko ya tekilogiya akozesebwa mu ChatGPT-3.5, ayongeile okwiluulibwamu. Kino kiidhie oluvainhuma lw'okugezesebwa okulaze nti, olwaleelo luno, enkola dha Gemini-2.5 ni GPT-4o dhiyumbwilwa inho mu ngeli ye dhikola ku lulimi ng'Olupootigo okusingila ilala. Bwe kiila ku nnimi edhitali mu buwandiike ng'Olusoga, okuliikiliza GPT-4o ebiwandiiko n'ebigambo ebiva mu itu ly'Olusoga, awalala n'ebyo ebiva mu nkenga dh'Eiwanika ly'Olusoga, kusenvwilwa kinene inho kw'aba nti kusobozesa n'ennondoola y'ebigambo ebiyaaka ebyakaingila mu lulimi luno. Waile ng'enkola eyeefaanaanhiliza kw'eyo eliwo buti ku Lungeleza n'Olupootigo ekaali kutuukibwaku mu Lusoga, aye, olugendo olukiluubilila lutandiikiibwaku.

Ebigambo ebikulu: KALIMAGEZI-KISIBUKILA, OKUTETENKELEZA KW'ENNIMI ENNHINGI, KISIBUKILA AKAALI KUKENGULWA, GPT EMMUMBE, EISOMO LY'OKUWANDIIKA AMA-WANIKA, AMAWANIKA, ABAWANDIISI B'AMAWANIKA, OLUPOOTIGO, OLUSOGA

1. Out-of-the-Box GPTs

1.1 Chatbot Arena

Even though Chatbot Arena¹ now exists where one can currently pair any two of about two-hundred LLMs — and thus choose from models like **Gemini** (Google), **GPT** (OpenAI), **Claude** (Anthropic), **DeepSeek** (High-Flyer), **Grok** (xAI), **Qwen** (Alibaba), **LeChat** (Mistral), **Hunyuan** (Tencent), **Command R+** (Cohere), **Gemma** (Google), **GLM** (Zhipu), **Step** (StepFun), **LLaMA** (Meta), **Yi** (01.AI), **Amazon Nova** (Amazon), etc. — OpenAI's GPT models still have the strongest conversational abilities, content creation, and summarisation, while Google's Gemini models are top for real-time web searches, complex reasoning, and research tasks. OpenAI's GPT and Google's Gemini models will therefore be used in this article.

1.2 Semi-visible Portuguese

The availability of the Chatbot Arena thus means that any of the prompts designed in earlier lexicographic research may now be re-used to test the quality of the output of over 200 models, to see which one works best for which type of dictionary, which language, etc. While this clearly provides for near-endless fun, all of these, as offered, are still out-of-the-box (see De Schryver 2025). Apart from the obvious improvement to have easy access to so much variety, one should check whether some of the recent LLMs are still as bad for non-English as those from a year and a half ago, as was the case for Portuguese when the then ChatGPT-3.5 was used at the end of October 2023 (De Schryver 2023b).

As examples, Figures 1 to 4 show a comparison between the Gemini-2.5 model and the GPT-4o model, for the prompts shown at the top-right in each case, to get to proper dictionary articles for: (1) the verb of communication *recommend* in an English–Portuguese learners' dictionary, (2) the lemma *table* in an English–Portuguese learners' dictionary, (3) the lemma *table* in a Portuguese–English learners' dictionary, with the prompt in English, and (4) the lemma *table* in a Portuguese–English learners' dictionary, with the prompt in Portuguese. The full texts offered by both the Gemini and GPT models, received as-is and thus left untouched, are presented in Addendum A, available online, as Figures 1 to 4 only show the opening sections. Click [here](#) for Addendum A.

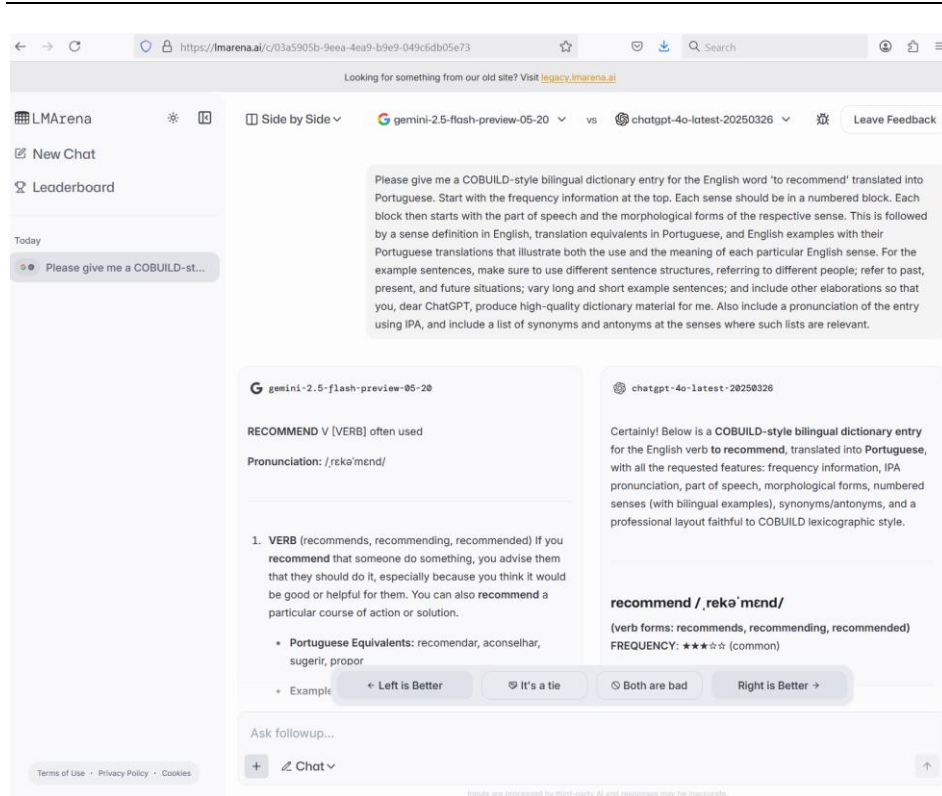


Figure 1: Using Chatbot Arena to compare any two models for lexicographic purposes, here Google's *gemini-2.5-flash-preview-05-20* vs. OpenAI's *chatgpt-4o-latest-20250326*, for the verb of communication *recommend* in an English–Portuguese learners' dictionary

<http://lexikos.journals.ac.za>; <https://doi.org/10.5788/35-2-2108> (Volle uitgawe / Full volume)
<https://lexikos.journals.ac.za>; <https://doi.org/10.5788/35-2-2104> (Article)

400 Gilles-Maurice de Schryver

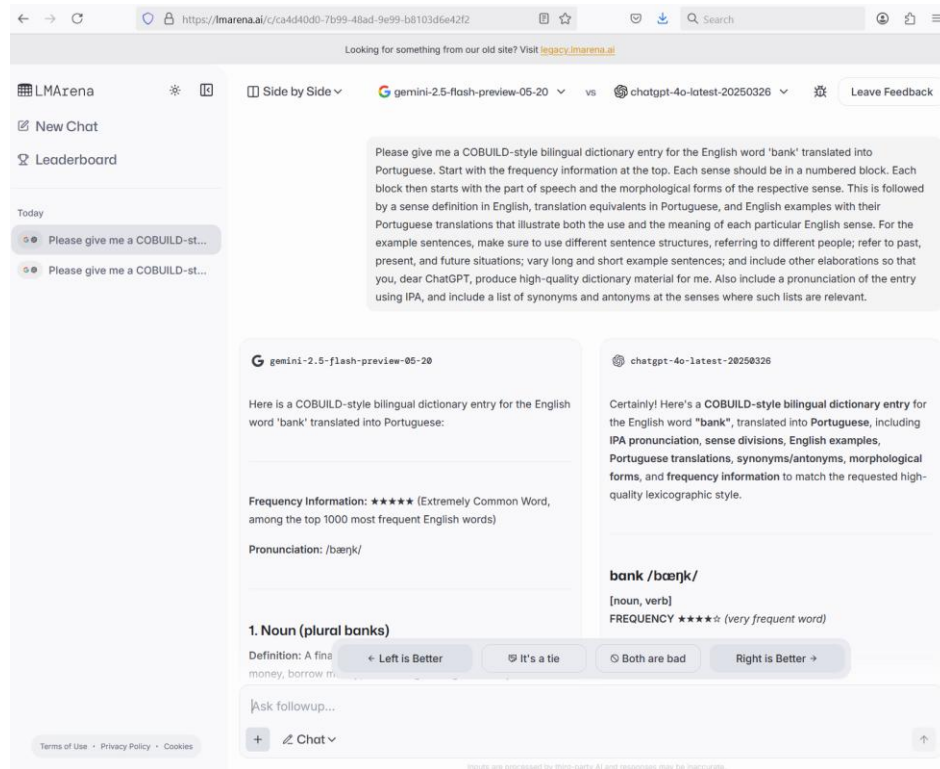


Figure 2: Using Chatbot Arena to compare any two models for lexicographic purposes, here Google's *gemini-2.5-flash-preview-05-20* vs. OpenAI's *chatgpt-4o-latest-20250326*, for the lemma *table* in an English–Portuguese learners' dictionary

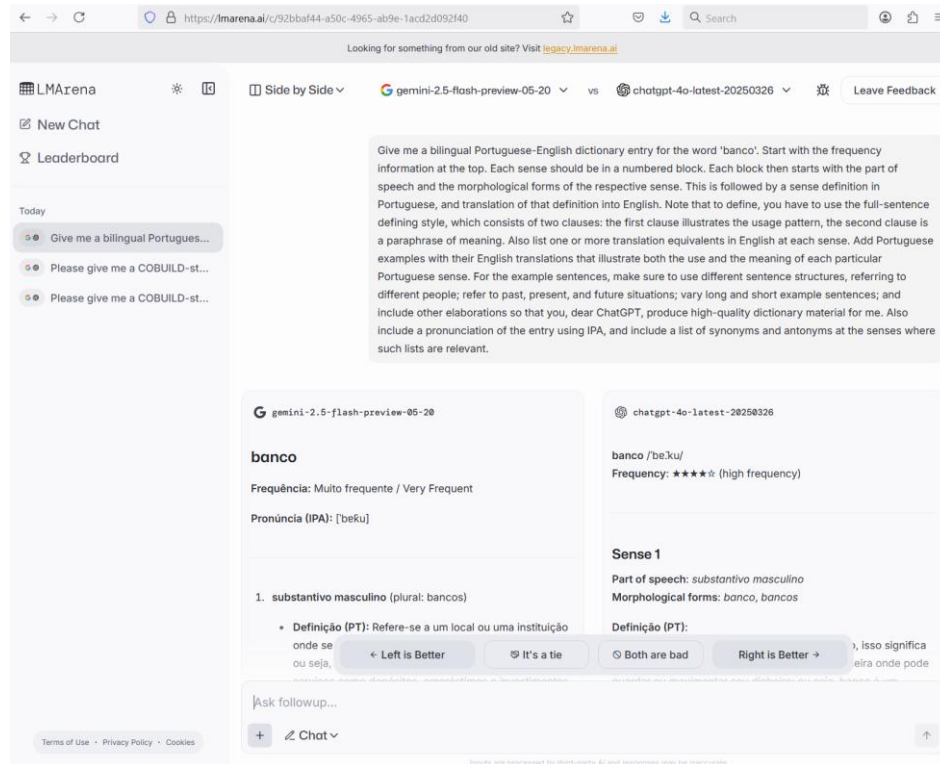


Figure 3: Using Chatbot Arena to compare any two models for lexicographic purposes, here Google's *gemini-2.5-flash-preview-05-20* vs. OpenAI's *chatgpt-4o-latest-20250326*, for the lemma *table* in a Portuguese–English learners' dictionary, with the prompt in English

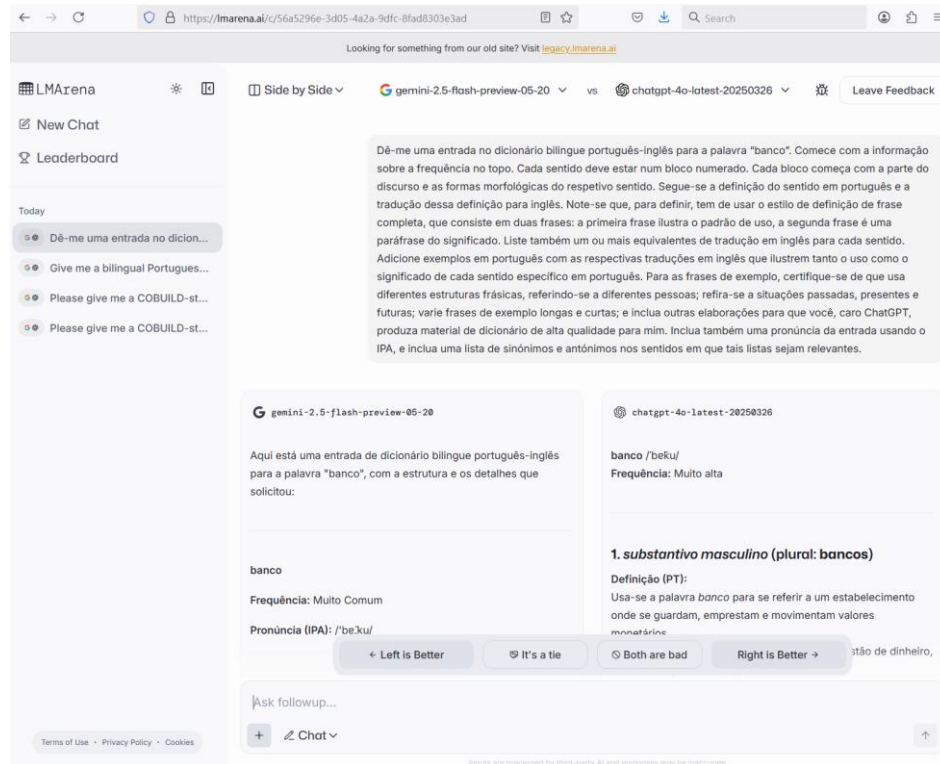


Figure 4: Using Chatbot Arena to compare any two models for lexicographic purposes, here Google's *gemini-2.5-flash-preview-05-20* vs. OpenAI's *chatgpt-4o-latest-20250326*, for the lemma *table* in a Portuguese–English learners' dictionary, with the prompt in Portuguese

Astonishingly, all the problems ChatGPT-3.5 had when trying to treat non-English like Portuguese for lexicographic purposes (see De Schryver 2025) are simply gone. This is most evident from the treatment of the lemma *table* when prompting GPT-4o, no matter the direction of the dictionary, no matter whether the prompt is in English or in Portuguese. As of June 2025 no more hallucinations are evident, and only noun senses are given, and thus no more inventions of non-existent verbal uses of Portuguese *banco*. Given that there are four prompts, submitted to two LLMs, there are a total of 8 answers in Addendum A, available online, numbered A1 to A8. Click [here](#) for Addendum A. More interesting summaries now follow for each of the eight dictionary articles produced via GenAI.

In A1, Gemini-2.5 for *recommend*, we immediately note excellent COBUILD-style full-sentence definitions (FSDs). On top, at sense 3, following the FSD, a second sentence starts with '*This usage often implies that ...*' — brilliant, as indeed, meaning is the result of usage, with meaning potentials only triggered by the

context, as Patrick Hanks reminded us throughout his life (Hanks 1979, 1988, 2000, 2002, 2012, Hanks and Franklin 2019, Hanks and Može 2019, Grefenstette and Hanks 2023). The Portuguese translations of the examples at sense 3 are also brilliant, as the various translation equivalents offered are indeed used (the underlinings, both full and broken up, were added by the present author). Lastly, the separation into three uses, and thus three senses, makes perfect sense.

In A2, GPT-4o for *recommend*, and thus even though the same verb is treated, the three uses/senses are not the same as in A1. In a way, the uses here are even further apart, as is also obvious from the additions of [VERB + that] at sense 2, and [VERB + itself, be + VERB] at sense 3. All of this implicitly refers to Patrick Hanks's Corpus Pattern Analysis (Hanks 2004, Hanks and Pustejovsky 2005, Hanks 2008, Hanks et al. 2018, Hanks and Ma 2021), and more specially his (unfinished) *Pattern Dictionary of English Verbs*². As impressive are the Usage Notes, which fully delve into linguistics, and the Related entries linking the verb to the noun and adjective.

In A3, Gemini-2.5 for *bank*, the first sense (as a financial institution) and second sense (as the sloping land alongside a river) mention that there are 'no direct antonyms' with a reason why — smart. At the second sense the mistake is not made to include **banco* as one of the translation equivalents, and as to the synonym 'levee' the context is added between brackets: '*for a raised bank to prevent flooding*' — very good. At the third sense (as a mass or pile of something) the translation equivalents are provided with context — excellent. At the fourth sense (as the verb *to bank*) the definition is sadly not COBUILD-style, but the mistake is not made to include **banco* as a translation equivalent. The fifth sense (of an aircraft or vehicle) is even far-more detailed than as seen in the actual COBUILD.³ The sixth sense (a row or series of similar items) correctly illustrates two of the translation equivalents. At the seventh sense, we see all sorts of correct compounds. At the eighth sense the phrasal verb *bank on* is treated, but alas no FSD. The examples here use two different translation equivalents.

In A4, GPT-4o for *bank*, the first sense (as a financial institution) *does* include antonyms — even smarter than Gemini-2.5. At sense two (as the side or edge of a river) only *margem* is (correctly) offered as a translation equivalent. Sadly, at the third sense (as the verb *to bank*), no proper FSD, and all examples use only one of two possible translation equivalents. At the fourth sense (a row or series of similar things) the third example uses a translation equivalent not listed; the other two (correctly) use two different ones. All is well with the fifth sense (the noun in aviation or driving). Idem for the sixth sense (the verb in aviation or driving), and seventh sense (as a mass or layer of something). The 'Summary Usage Note' is most-excellent and reminds the user of the highly polysemous nature of this lemma, as it is used both literally and metaphorically, across various fields, pointing out that the resulting translation equivalents in Portuguese truly depend on context. Brilliant lexicography.

In A5, Gemini-2.5 for Portuguese *banco* (prompt in English), at the first sense (as a financial institution) the translation equivalent and use in example sen-

tences are all correct. No antonym is provided, although we know, thanks to GPT-4o under A4 sense 1 that there are antonyms (namely *borrower*, *debtor*). At the second sense (as furniture for sitting), three translation equivalents are provided and two are used in the examples (*bench*, *seat*). At the third sense (as a sandbank, or school of fish), four translation equivalents are provided and all four are used in the examples (*sandbank*, *school*, *bank*, *shoal*) — good lexicography in a learners' dictionary. At the fourth sense (as a blood or organ bank), only bank is offered as a translation equivalent, but all examples combine it (once each with *blood*, *organ*, *gene*, *tissue*). None of senses 2 to 4 include antonyms; but GPT-4o does for all, see the next example.

In A6, GPT-4o for Portuguese *banco* (prompt in English), the definition at the first sense (as a financial institution) is a good attempt at an FSD, which is always tougher for a noun than for a verb. Two translation equivalents are offered, but only one is used in the three examples (*bank*). As antonyms of this use of *banco*, Portuguese *cofre* 'safe' and *colchão (em uso figurado para guardar dinheiro em casa)* 'mattress (figuratively for storing money at home)' are offered — very good. At the second sense (as a bench in a park), the FSD is near-excellent. Two translation equivalents are offered, of which only one is used in all three examples (*bench*). Here too, the antonyms are head-on, Portuguese *chão* 'floor' and *cadeira individual* 'single chair'. At the third sense (as a substitute bench), the FSD is again near-excellent. While two translation equivalents are offered, again only one is used in all three examples (*bench*), though here it makes more sense, as the expanded *substitute bench* is of course rarer. The antonyms are again very good, Portuguese *titular* 'starter' and *escalado* 'line-up'. At the fourth sense (as a database), we see the usual near-FSD, two translation equivalents of which one used in the three examples (*database*), while the antonyms are original, Portuguese *ausência de dados* 'no data' and *sistema manual* 'manual system'. At the fifth sense (as a sandbank), again a near-FSD, three translation equivalents, but in the three examples only one of them (twice *sandbank*, plus even *bank*, which is not among the translation equivalents), and as antonyms the debatable Portuguese *canal* 'canal' and *profundidade* 'depth'. The concluding summary at the end, focusing on the polysemy of Portuguese *banco*, found with financial, physical, sportive, digital, and geological uses — listed in the order of the five senses — is most precise and helpful. That note continues with the claim that this GenAI entry is 'offering a comprehensive bilingual reference for language learners and lexicographers alike'. Compared with the ridiculously short entry for *banco* as offered by Collins,⁴ this is undoubtedly better, near-perfect even. Near-perfect, as the FSD does indeed contain the two parts, but is followed by a semi-colon and then a more traditional definition on top, making the FSD a bit too long. Other than that, let GPT-4o compile an entire dictionary in this format overnight, for the full lexicon, and we might have (the draft of) a good bilingual Portuguese–English learners' dictionary right away.

In A7, Gemini-2.5 for Portuguese *banco* (prompt in Portuguese), even more is presented in Portuguese as a result of prompting in Portuguese. At the first

sense (as a financial institution), the definition consists of two sentences, so not an FSD, but if the two had been merged and shortened, it comes close. Of the two translation equivalents, all four examples use one (*bank*). Two synonyms are offered, no antonyms. At the second sense (as a bench in a park), we see the same type of non-FSD. Of the three translation equivalents that are offered, two are used in the four examples (*bench*, *pew*). Three synonyms are offered, no antonyms. At the third sense (as a sandbank), again a non-FSD. Of the four translation equivalents that are offered, three are used in the three examples (*sandbank*, *mudflat*, *shoal*). Three synonyms are offered, no antonyms. At the fourth sense (as a blood or data bank), another non-FSD. Of the four translation equivalents that are offered, only one is used in the four examples (*bank*). Four synonyms are offered, no antonyms. At the fifth sense (as a school or shoal of fish), the same issue with the non-FSD is evident. Of the two translation equivalents that are offered, only two are used in the three examples (*school*, *shoal*). Only one synonym is offered, no antonyms.

In A8, GPT-4o for Portuguese *banco* (prompt in Portuguese), there is again more in Portuguese, so here too the result of the Portuguese prompt. The first sense (as a financial institution) presents a definition in the FSD-style, use vs. meaning, followed by a second sentence that is a variation on the meaning only. Surely, the two sentences could have been merged, to achieve the FSD-goal of a learners' dictionary. One translation equivalent is offered and used in all three examples (*bank*). Two synonyms follow, with this time an indication that there are no antonyms in this sense (even though GPT-4o offered some for this sense in A4 and A6). The second sense (as a long seat) presents a similar FSD cum extra sentence. Two translation equivalents are offered, but only one is used in all three examples (*bench*). Three synonyms follow, and two antonyms (*cadeira* 'chair' and *poltrona* 'armchair'). The third sense (as a subaquatic formation) again has the pseudo-FSD and then a second sentence on a reformulation of the meaning. Four translation equivalents are offered, of which two are used in two examples (*sandbank*, *bank*) and then one not listed under the translation equivalents (*school*). Three synonyms follow, and two antonyms (*mar profundo* 'deep sea' and *leito rochoso* 'bedrock'). The fourth sense (as a vehicle seat) once again has the pseudo-FSD and second sentence. Two translation equivalents are offered, and both are used in all three examples (*car seat*, *seat*). Two synonyms follow, and two antonyms (*espaço livre* 'free space' and *chão do veículo* 'vehicle floor'). The fifth sense (as a collection of information) repeats the pseudo-FSD plus extra sentence on meaning. Two translation equivalents are offered, and both are used in all three examples (*database*, *data bank*). Two synonyms follow, and two antonyms (*informação dispersa* 'scattered information' and *fonte não estruturada* 'unstructured source'). At the very end, a pronunciation is offered for the lemma (both in European and Brazilian Portuguese, although the same here), followed by an invitation to treat GPT-4o as a professional lexicographer, as it invites the user with: 'If you'd like examples of compositions with '*banco*' (such as *data bank*, *square bank*, *image bank*), I can draw up more specific blocks. Would you like to continue?'.

Even a cursory look — but especially the more detailed study just presented — of the lexicographic output of Gemini-2.5 vs. GPT-4o, makes clear that for lexicography GPT-4o is the better tool. Comparing the output under A6 (with the prompt in English) vs. the output under A8 (with the prompt in Portuguese) further indicates that prompting in English remains better than prompting in another language. It is therefore not surprising that over the past two years and a half, lexicographers not only started with but also mainly stuck by ChatGPT, as seen from the detailed study of a sample of 100 contributions on 'GenAI in lexicography' analysed in De Schryver (forthcoming-a, forthcoming-b). In any case, for large languages like Portuguese there is thus no need anymore to customise an LLM, at least not for general lexicographic purposes.

2. Customised GPTs

The necessary move for large non-English languages — so from a 'hyper-visible' to a 'semi-visible' uptake in technology like GenAI — has been made: Recent LLM models are now also achieving top lexicographic performances for languages such as Portuguese with out-of-the-box tools. Of course these same tools won't have improved much for low-visible languages such as the Bantu languages, and very likely not for a local-only visible language such as Lusoga: As far as we know, no extra data was added anywhere online in recent years for that language, so for exotic oral languages like Lusoga it becomes important to now know whether a tool like GPT-4o can be used to take the next step for them, namely *customisation*.

2.1 Local-visible Lusoga

Since the release of GPT-4o on 13 May 2024, customisation has been put into the hands of users. Within days of the release of GPT-4o a serious attempt was made to customise this model by trying the impossible: Could it compile dictionary articles for *neologisms* in the oral — so extremely under-documented — Bantu language Lusoga, by feeding GPT-4o everything that exists on Lusoga, viz. a corpus of 3.7 million tokens (De Schryver and Nabirye 2022), and a full comprehensive monolingual dictionary in XML, the *e-Eiwanika ly'Olusoga* (Nabirye et al. 2012)⁵? The outcome, presented on 20 May 2024 as a keynote at the CogALex workshop of LREC-COLING 2024 in Turin (De Schryver 2024a), was tantalising: whereas GPT-4o generated only nonsense out-of-the-box, it started to make sense with this customisation.

Indeed, following an opening prompt in which GPT-4o was merely informed that the word *ente* is 'cow' in Lusoga; an out-of-the-box generation of a definition for it produced utter gibberish: the reply seen in Figure 5 is meaningless.



Figure 5: Trying the out-of-the-box GPT-4o, asking for a Lusoga definition of *ente* 'cow'

A team of three therefore joined forces — programmer David Joffe from South Africa, native Lusoga speaker Minah Nabirye from Uganda, and lexicographer Gilles-Maurice de Schryver from Belgium — and messaged the built-in GPT builder to create a custom GPT, which they called 'Lusoga Linguist'. After agreeing on a profile picture, the refinement of the context could start, with Figure 6 showing one of the goals.

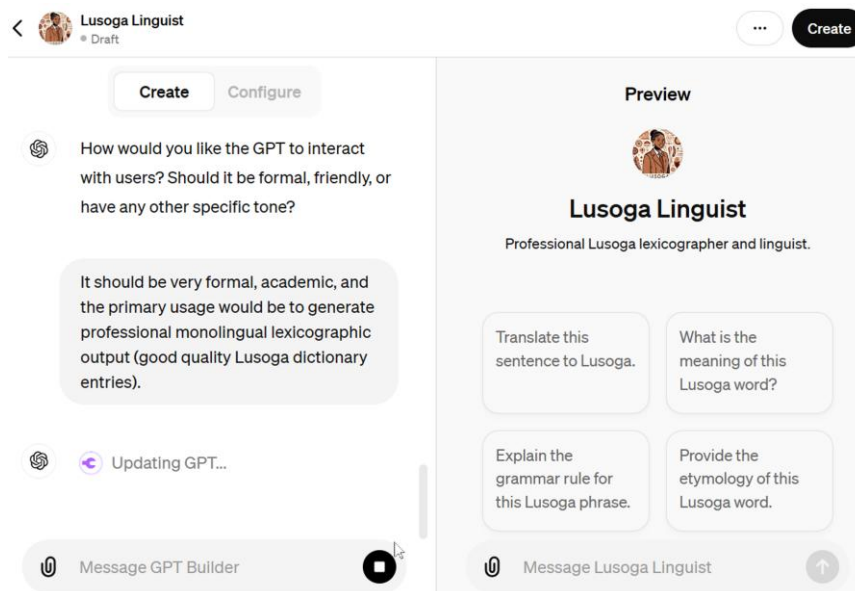


Figure 6: Work on refining the goal of the custom GPT for Lusoga lexicography has begun

After that, a first attempt was made at adding Lusoga corpus files to the 'Knowledge base', as seen in Figure 7.

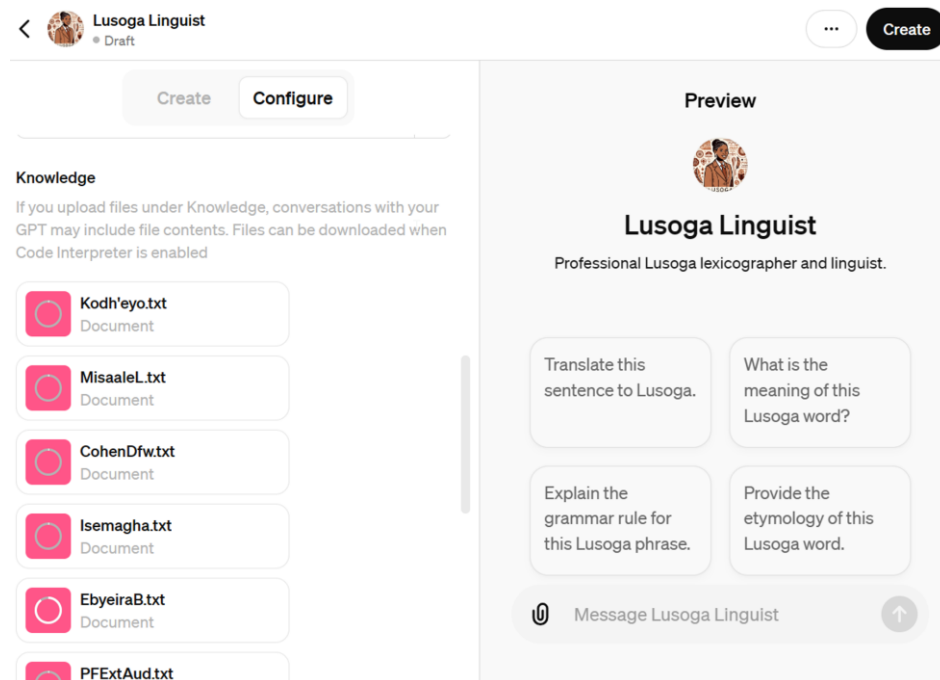


Figure 7: Adding a Lusoga sample corpus to the 'Knowledge base' of GPT-4o

Adding very many files did not work, so a single word file (of about 500KB) with 10,000 lines of corpus texts was added. In there the word *ente* 'cow' occurred 36 times. Once done, GPT-4o was instructed to '*study the uploaded corpus of Lusoga text*' to then use this knowledge before moving to the '*task of generating professional monolingual Lusoga dictionary entries*', as seen in Figure 8.

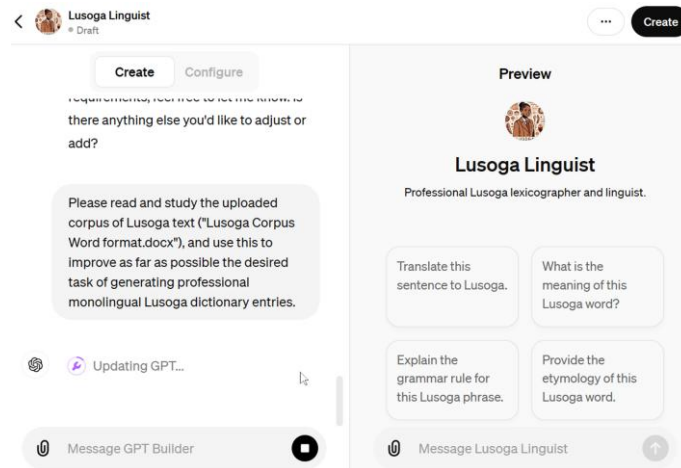


Figure 8: Instructing GPT-4o to study the submitted sample Lusoga corpus

The initial prompt was then repeated, as seen in Figure 9, together with the output.

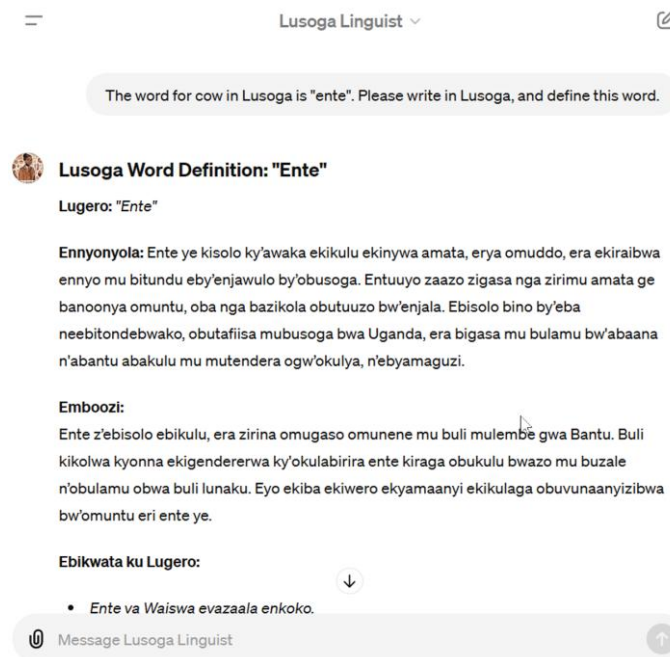


Figure 9: A customised GPT-4o tries to define Lusoga *ente* 'cow', using a sample corpus

The output seen in Figure 9 reads more like a story than a dictionary. Plus, most is in the neighbouring and much larger language Luganda, with little in Lusoga. The opening of the definition translates as '*A cow is an animal that drinks milk and eats grass ...*' and the last line of the screenshot has an example that translates as '*The cow of Waiswa gives birth to a chicken*'. Further prompting in which GPT-4o was asked to only use Lusoga (and no Luganda) and to compile a proper dictionary (rather than a story) failed to improve the result. The decision was therefore taken to load samples of the monolingual Lusoga dictionary into the 'Knowledge base' of GPT-4o. That work was published as a paper dictionary (Nabirye 2009) as well as an e-dictionary (Nabirye et al. 2012), and for the latter both as a downloadable dictionary app as well as a free online dictionary. See Addendum B, available online (click [here](#)).

After uploading the sample with monolingual Lusoga dictionary examples, GPT-4o was instructed to study those sample entries, in the hope that it would 'understand' the lexicographic composition. Figure 10 displays the start of the conversation on this level, while Figure 11 displays the end of the analysis by GPT-4o. Recall that all the data submitted for training of the model is fully in Lusoga, not in any other language, so the task is clearly hard. On the other hand, there is a good lexical overlap between Lusoga and Luganda, a language for which there are considerable amounts of data online.

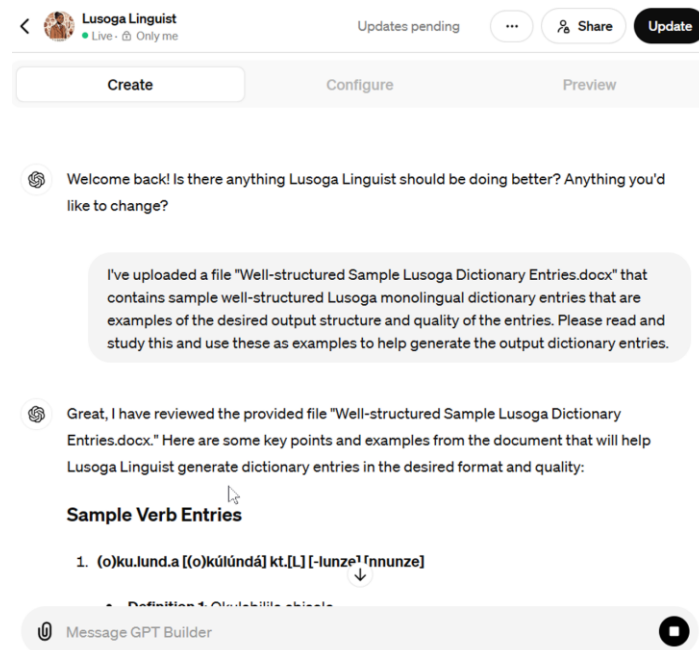


Figure 10: After adding a sample of monolingual Lusoga dictionary entries to the 'Knowledge base', asking GPT-4o to study the structure of that

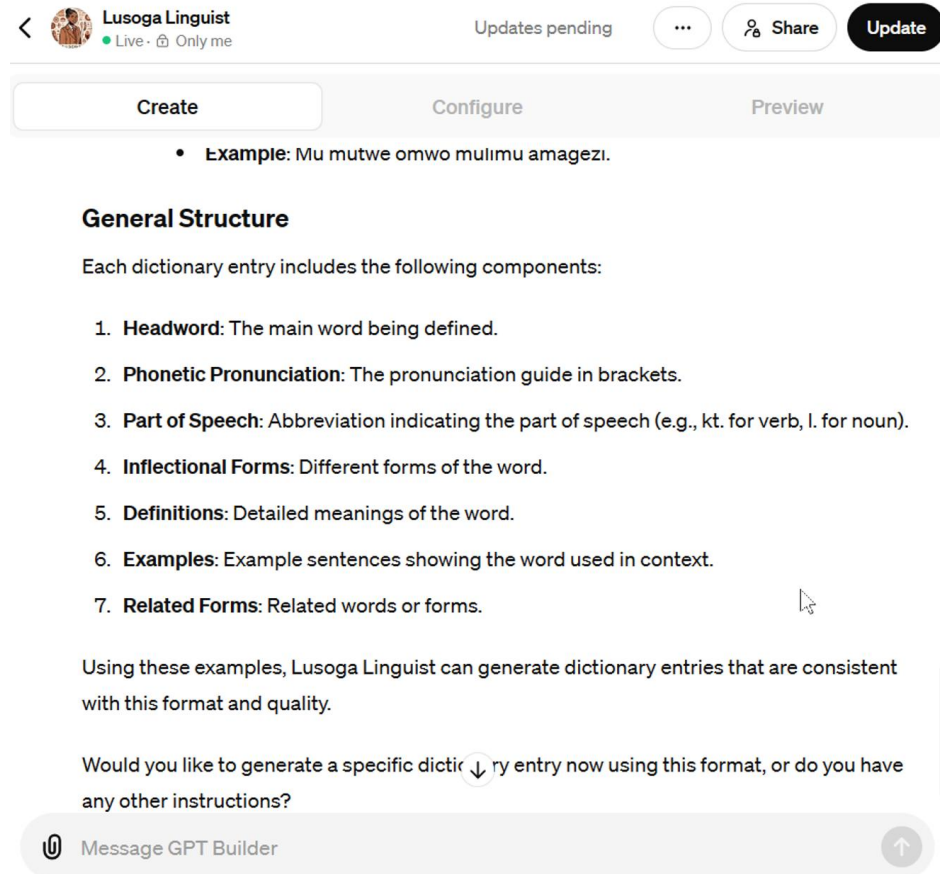


Figure 11: GPT-4o has 'learned' the lexicographic composition of entries in the monolingual Lusoga dictionary

Equipped with a sample Lusoga corpus, and a sample of Lusoga dictionary entries, GPT-4o was asked for a definition of *omukono* 'hand/arm' (not being among the sample dictionary entries), as seen in Figure 12. The overall appearance of Figure 12 is only slightly more dictionary-like. Plus, one continues to notice both Luganda and Lusoga, but whenever it is Lusoga it is still not coherent. Actually, but still surprisingly, only sense 2 is (1) written in Lusoga, and (2) makes somewhat sense, as the translation is '*One of the *drafts of the body ...*'. Is the custom GPT-4o starting to learn Lusoga, or simply lucky in making things up? Two efforts were undertaken to make the custom GPT-4o much better.

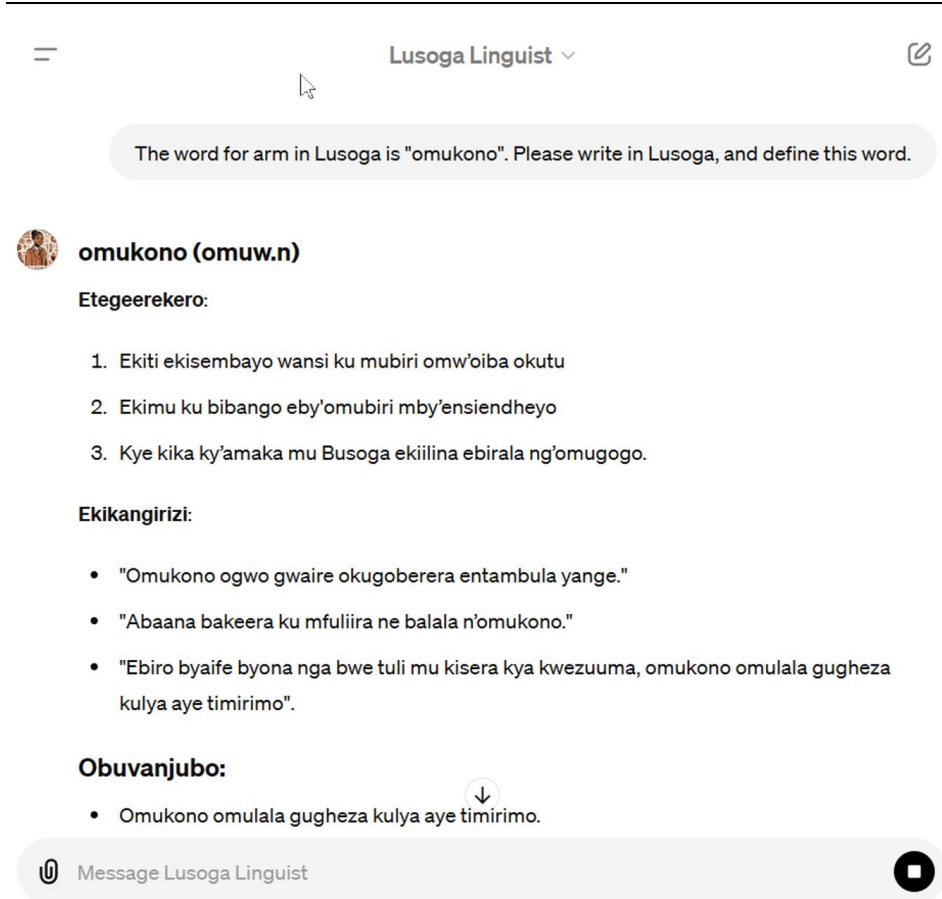


Figure 12: Asking the first version of a customised GPT-4o for the Lusoga entry *omukono* 'hand/arm'

First, given that up to 20 documents may be added to the 'Knowledge base', with each not containing more than 2 million LLM tokens and not larger than 512MB, all the files of the sample corpus were taken away and were now replaced with the full Lusoga corpus of 3.7 million word tokens, cut up in ten parts.⁶ That corpus was the result of about two decades of corpus building by Minah Nabirye (De Schryver and Nabirye 2018, 2022), and enlarging it even further is a truly heroic effort.⁷ Figure 13 shows how the 3.7 million Lusoga word tokens were added to the 'Knowledge base', and Figure 14 prompts with the same question as in Figure 12, but now assuming that the full corpus 'taught' the customised GPT-4o model more about the Lusoga language.

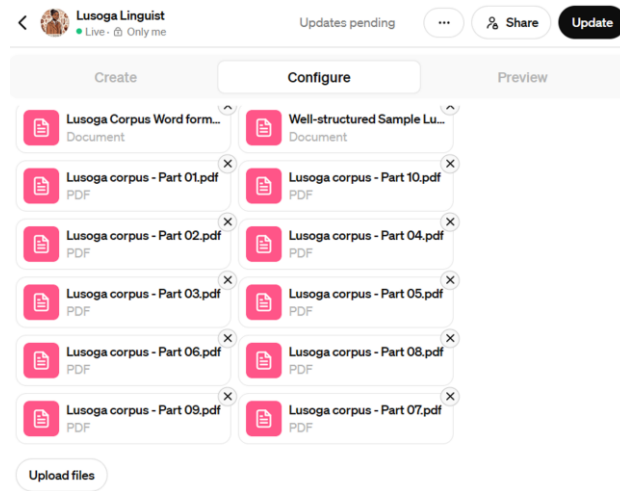


Figure 13: Adding all 3.7m word tokens of Lusoga to the 'Knowledge base' of the customised GPT-4o

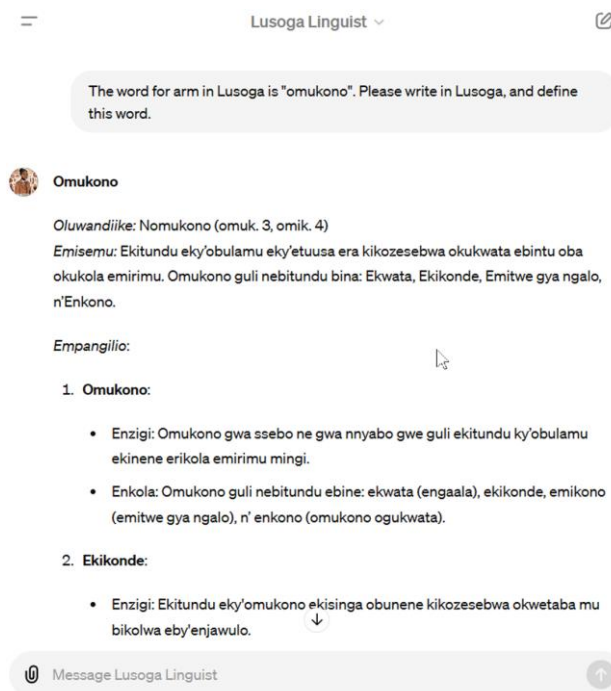


Figure 14: Asking the re-customised GPT-4o for the Lusoga entry *omukono* 'hand/arm'

While still far from correct, Figure 14 suddenly comes close to resembling real Lusoga and the analysis of Lusoga. On the first line, for instance, while the singular prefix is (obviously) correct (*omu-* in class 3), for the plural *omi-* in class 4 is suggested, where it should be *emi-* in class 4. Then, the definition of *omukono* 'hand/arm' contains not only too much Luganda, but also Swahili, and translates as '*The part of the body which is close and is used to hold things and to do work*'. A rather 'funny' definition, but it is valid: it is a close body part, and hands are used to hold things and also perform work. Compare this to the actual entry in the *e-Eiwanika ly'Olusoga*, as seen in the first screenshot of Addendum B, available online (click [here](#)).

This result logically led to the second effort, namely the addition to the 'Knowledge base' of the XML-file of the entire 600-page *e-Eiwanika ly'Olusoga*, in other words the full monolingual Lusoga dictionary, as seen in Figure 15. (The 'Well-structured Sample' was taken out.)

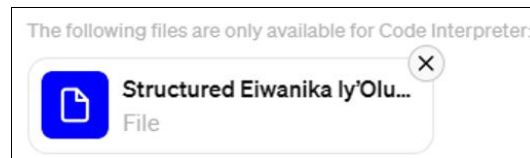


Figure 15: Adding the XML-file of the full monolingual *e-Eiwanika ly'Olusoga* to the 'Knowledge base' of the customised GPT-4o

This was immediately followed by an instruction to the custom GPT-4o to study this full dictionary, as seen in Figure 16.

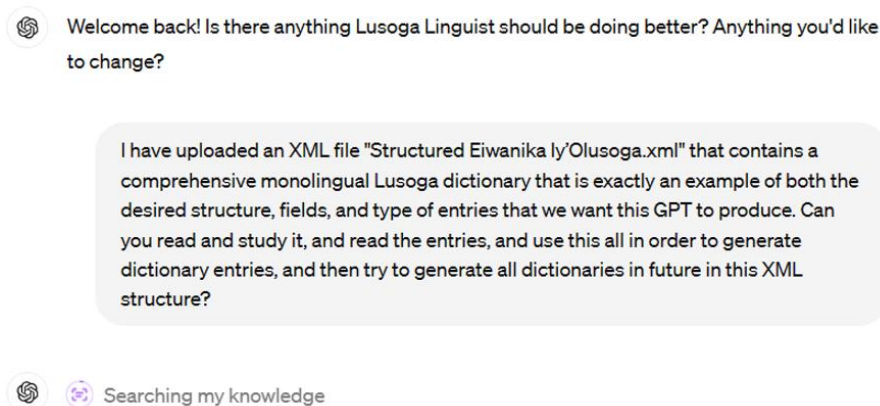


Figure 16: Adding the XML-file of the entire monolingual *e-Eiwanika ly'Olusoga* to the 'Knowledge base' of the customised GPT-4o

Working with ChatGPT, Barrett (2023: 10) had noticed that even their GPT-4 model performed some stunts at times, namely to blatantly copy over existing dictionary contents wholesale (in his case for definitions of *virus*). In order to be sure that the customised GPT-4o model does not simply go on a copy-and-paste venture now that the full *e-Eiwanika ly'Olusoga* is in the 'Knowledge base', an additional difficulty was added, namely to focus on neologisms. The lemma *omuyumo* 'entertainer' was chosen, as it is a known neologism, first described in the academic literature in De Schryver and Nabirye (2022: 843), and *not* in the monolingual Lusoga dictionary, but *in* the Lusoga corpus. In Figure 17 one sees the reply to the question to compile a monolingual dictionary article for *omuyumo* 'entertainer'.

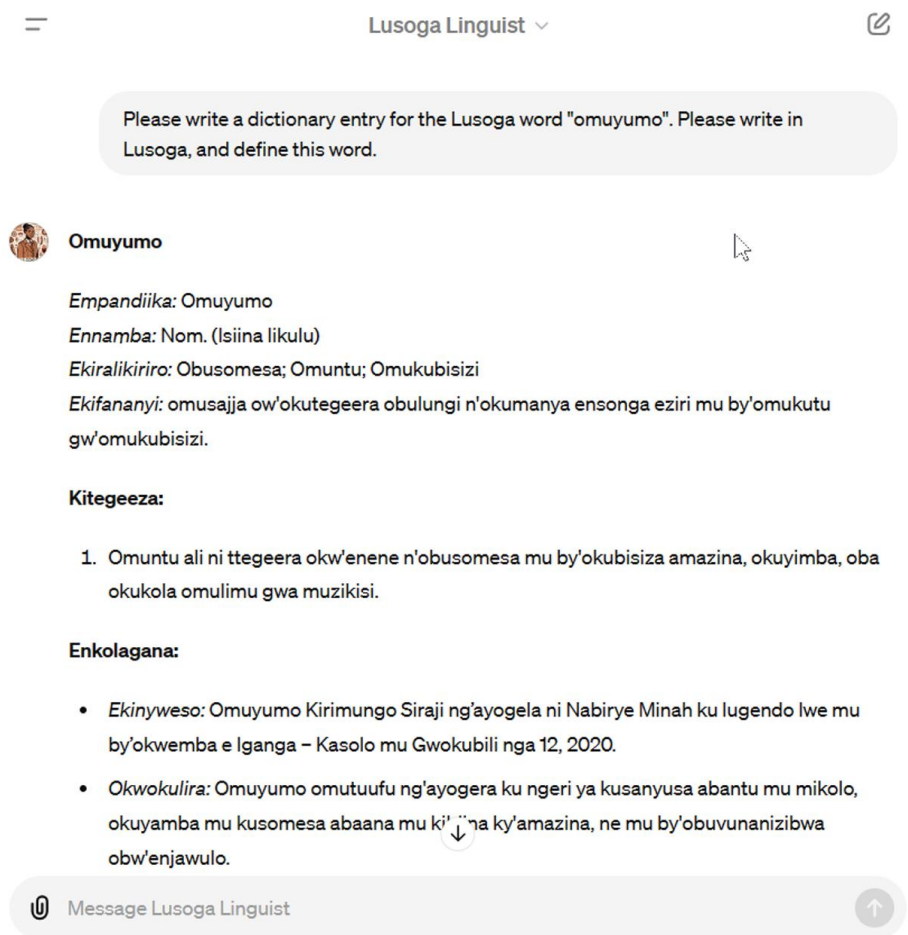


Figure 17: Asking the re-re-customised GPT-4o to compile a dictionary entry for Lusoga *omuyumo* 'entertainer'

While the opening section of Figure 17 still makes no lexicographic sense, the first bullet at the bottom provides an example that is fully grammatical, coherent and on-topic Lusoga. The translation of this example sentence is: '*The entertainer known as Kirimungo Siraj is in the process of speaking with Nabirye Minah about his Singing Journey at Iganga, Kasolo on 12-02-2020*'. This sentence does not appear in the corpus, and is clearly an excellent summary of one of the transcribed recordings made in 2020 for a book that came out four years later, ironically called *Obuyumu mu Busoga 'Entertainment in Busoga'* (Nabirye 2024). Unfortunately, while the example at the next bullet is again grammatical, coherent, and on-topic, it is couched in Luganda: '*A true entertainer who is speaking about ways of entertaining people at a function, who helps to teach the truth to children in class and for different types of responsibilities*'.

While there is clearly still a long way to go to get to proper lexicography, and while the neighbouring language Luganda still creeps in, seeing that GPT-4o can actually start learning a brand-new language and generate some dictionary material in it is promising, and, as far as we know, has not yet been tried for any other language within the field of lexicography.⁸

Moreover, when asked to explain where the data for the generated examples come from, GPT-4o is able to point to the various corpus sections from where it sourced and summarised information. For instance, for the example in Luganda at the last bullet in Figure 17, Figure 18 adds the Lusoga sources — in other words, here the example is based on material taken from the Lusoga corpus, but strangely 'translated' from Lusoga into Luganda.

- Ekinyweso: Omuyumo Kirimungo Siraji ng'ayogela ni Nabirye Minah ku lugendo lwe mu by'okwemba e Iganga – Kasolo mu Gwokubili nga 12, 2020.
- Okwokulira: Omuyumo omutuufu ng'ayogera ku ngeri ya kusanyusa abantu mu mikolo, okuyamba mu kusomesa abaana mu kibiina ky'amazina, ne mu by'obuvunanizibwa obw'enjawulo 【21:0†source】
 【21:1†source】 【21:2†source】 【21:3†source】 【21:5†source】
 【21:6†source】 【21:7†source】 【21:10†source】 .

Figure 18: The custom GPT-4o reveals the Lusoga sources used for the last example in Figure 17

Experimenting more with trying to discover what was used to come up with certain definitions and/or examples, at some point later on, while working on the neologism *obukurisitu* 'Christianity', the custom GPT-4o claims all of what is seen in Figure 19, where the most important claim is: '*Each reference helps trace*

back the information used in the dictionary entry to its original location in the uploaded documents. This ensures that the definition provided is well-supported by the source material'. And a bit later yet again, when working on the neologism *omukwampa* 'the direction of', the custom GPT-4o even goes as deep as to claim all that is seen in Figure 20, where the summary is: 'The references are not from a single contiguous line or section but rather from different parts of the text that discuss related topics, such as boundary demarcations, geographical descriptions, and historical or cultural contexts'. All of this has indeed a lot to do with the meaning of the neologism *omukwampa* 'the direction of', so if all true, this means that a customised LLM stops being the full black box feared by lexicographers, as it becomes somewhat less opaque. At the same time, customisation for an unknown exotic language such as Lusoga is not good enough, so we need more.

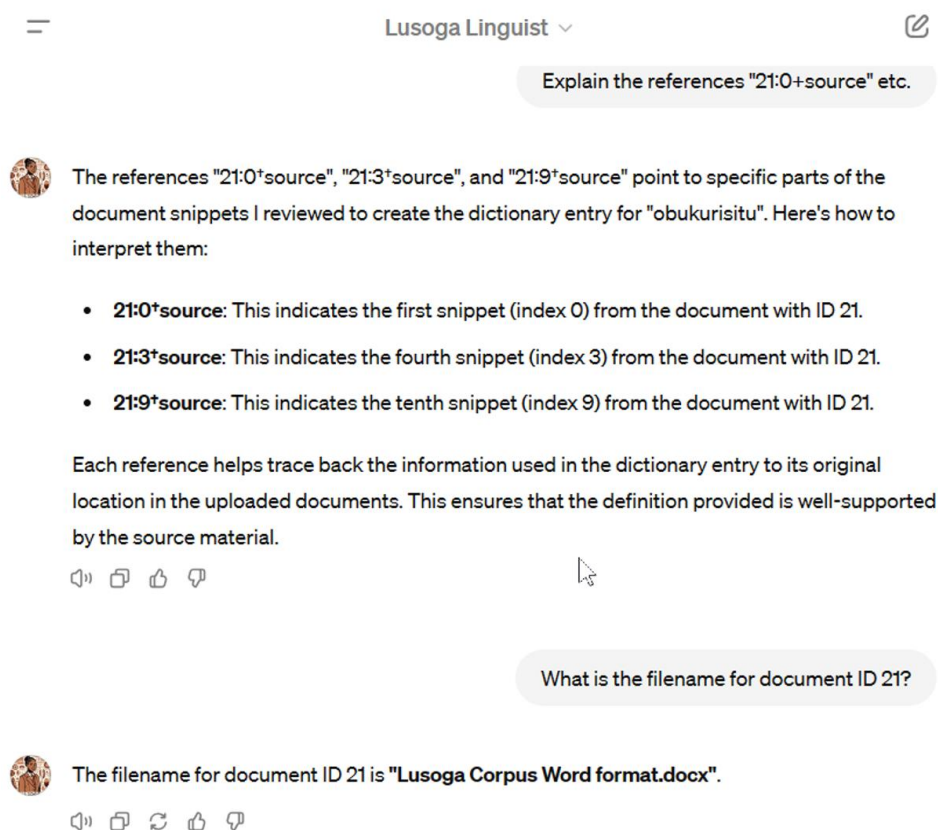


Figure 19: The custom GPT-4o reveals the Lusoga sources used for the Lusoga lemma *obukurisu* 'Christianity'

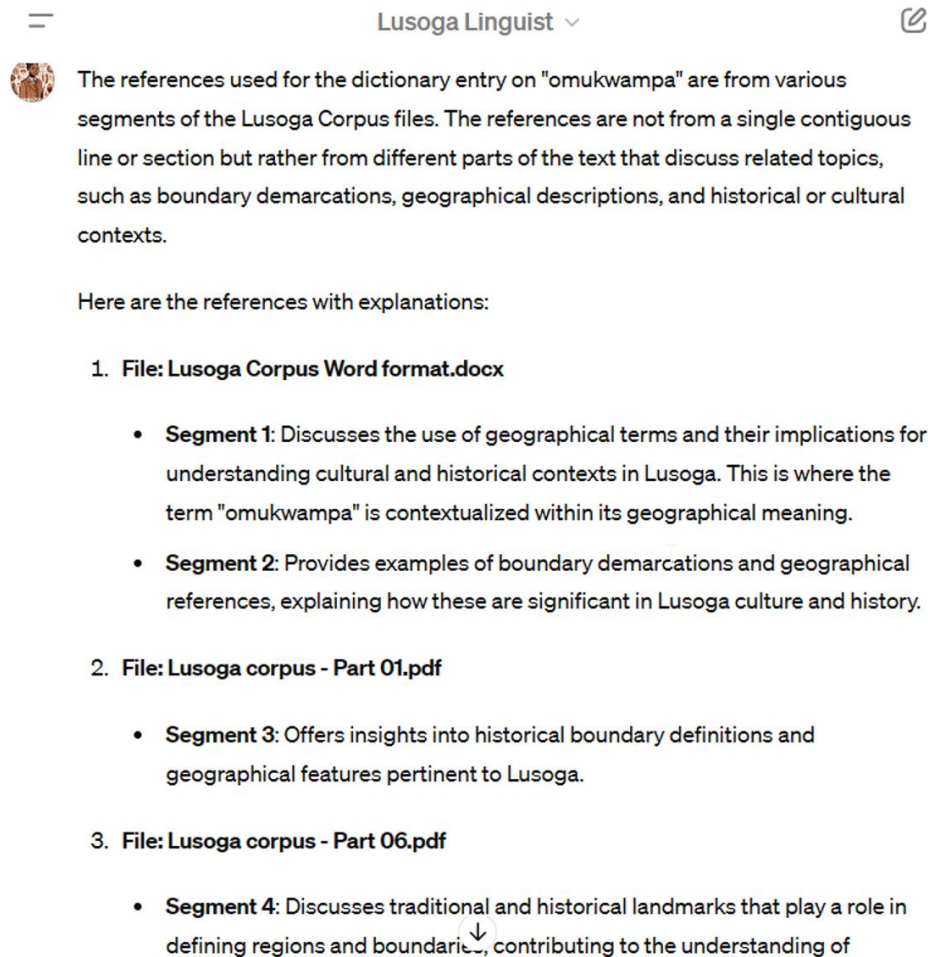


Figure 20: The custom GPT-4o reveals the Lusoga sources used for the Lusoga lemma *omukwampa* 'the direction of'

3. Epilogue

Armed with all the facts presented so far, and having proceeded from an out-of-the-box GPT (see De Schryver 2025), to a customised GPT for lexicographic purposes, it is now necessary to take the next step for under-resourced languages, namely the fine-tuning of an LLM to (1) compile a dictionary, (2) act as a dictionary in its own right, or even (3) chat with the LLM as if it were a lexicographer. The differences between the three types are summarised in Table 1, adapted from De Schryver (2024b: 9).

Table 1: Comparing models for lexicography: Out-of-the-box GPT vs. customised GPT vs. fine-tuned GPT

Feature	Out-of-the-box GPT	Customised GPT	Fine-tuned GPT
Modifies core model	No	No	Yes
Training required	No	No, leverages pre-trained model	Yes, on a specific dataset
User input	Text prompt	Instructions and/or relevant documents	New data for the model to learn from
Complexity	Least complex, easy to use	Rather complex, potentially no-code/low-code	More complex, requires expertise
Cost	Least expensive	More expensive	A lot more expensive
Output control	Limited control over the generated text	More control over the direction of the output	High degree of control over the generated text
Suitability for 'exotic lexicography'	Limited, may not be familiar with 'exotic language'	Can be adapted to the 'exotic language' through prompts and instructions	More directly addresses the 'exotic language' through fine-tuning

The most important difference is listed in the first line of Table 1: one will need to modify the core model. In keeping with seeking a true challenge, as hinted at in the last line of Table 1: one could proceed with an exotic language about which the model has initially no clue. This is definitely a task for a future endeavour.⁹ But whatever the outcomes, they should be reported on, there and then.

PS.

Addendum C, available online (click [here](#)), gives an example of how the addition of the LLM Gemini to the search engine Google¹⁰ has turned that search engine into an intriguing GenAI chatbot. Search for 'broken uganda' today, and see how all data available on the Internet on this topic is first analysed and summarised, in real time, before listing the usual hyperlinks.

Endnotes

1. See <https://lmarena.ai/>.
2. See <https://pdev.org.uk/>.
3. See <https://www.collinsdictionary.com/dictionary/english/bank>, other verb uses.
4. See <https://www.collinsdictionary.com/dictionary/portuguese-english/banco>.
5. See <https://menhapublishers.com/dictionary/>.
6. Note that each token in an LLM like GPT-4o is either a whole word, part of a word, or punctuation, so a Lusoga corpus of 3.7 million word tokens has many more millions of LLM tokens.
7. Ironically, a massive effort to enlarge the Lusoga corpus was indeed undertaken starting at the end of June 2024, but after nearly three months of non-stop data recording and transcrip-

- tion it was all in vain, as the material was stolen and the fieldworker himself suffered quite an impact in Uganda (De Schryver forthcoming-b).
8. More generally, and outside the field of lexicography, LLMs have (1) been trained to translate low-resource languages, such as for Inuktitut (40 thousand speakers; spoken in northern Canada) — Elsner and Needle (2023), or (2) been asked to learn new languages, such as for Zhuang (16 million speakers; spoken in southern China) — Zhang et al. (2024), or even (3) been taught languages with rare scripts, such as for Dzonghka (640,000 speakers; spoken in western Bhutan; Tibetan script), Santali (7.6 million speakers; spoken in India, Bangladesh and Nepal; Ol Chiki script), Nko (millions of speakers; spoken in West Africa; N'Ko script), Tamasheq (900,000 speakers; spoken in Mali and Burkina Faso; Tifinagh script), and Tigrinya (9.9 million speakers; spoken in Eritrea and Ethiopia; Ge'ez script) — Li et al. (2025).
 9. That endeavour had actually been kickstarted, see Endnote 7. But the fieldworker was broken into, both literally (his hut was emptied) and sadly also literally (his skull and brain were broken up with machetes). After the fieldworker miraculously came fully back to life, he decided it was time to release his talk on 'Broken in Uganda', dating from August 2023.
 10. See <https://www.google.com/>.
 11. See <https://www.youtube.com/>.

References

- Barrett, G.** 2023. *Defin-o-Bots: Challenging A.I. to Create Usable Dictionary Content*. Paper presented at the 24th Biennial Conference of the Dictionary Society of North America, Boulder, CO, USA, 31 May–3 June 2023.
- De Schryver, G.-M.** 2023a. *Broken in Uganda: Uganda's National Curriculum Development Centre (NCDC)*. Paper presented at the Third Biannual Conference of the Language Association of Eastern Africa (LAEA), Makerere University, Kampala, Uganda, 15–16 August 2023, Kampala.
<https://youtu.be/wYL6ZNlNy5o>
- De Schryver, G.-M.** 2023b. *Contemporary Lexicography: A Case Study in AI [Keynote Lecture]*. Paper presented at the Inaugural Conference of the Association of Lexicography for the Americas: South, Central, Caribbean, and Mexico (Americalex-S), São Paulo, Brazil, 20–25 October 2023.
<https://youtu.be/watch?v=yAxzTx7A2LU>
- De Schryver, G.-M.** 2024a. *Customising LLMs for Lexicography [Keynote Lecture]*. Paper presented at the 8th International Workshop on Cognitive Aspects of the Lexicon (CogALex-VIII), co-located with the Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), Turin, Italy, 20 May 2024.
- De Schryver, G.-M.** 2024b. The Road towards Fine-tuned LLMs for Lexicography. Krek, S. (Ed.). 2024. *Book of Abstracts of the Workshop 'Large Language Models and Lexicography' @ EURALEX 2024, 8 October 2024, Cavtat, Croatia*: 6–11. Ljubljana: ELEXIS Association.
- De Schryver, G.-M.** 2025. Out-of-the-Box GPTs for Lexicography. *Lexikos* 35(2): 362–396.
- De Schryver, G.-M.** forthcoming-a. 100 Scholarly Echoes of Generative AI in Lexicography.
- De Schryver, G.-M.** forthcoming-b. Citing the Future: The Legacy of Patrick Hanks.
- De Schryver, G.-M. and M. Nabirye.** 2018. Corpus-driven Bantu Lexicography, Part 1: Organic Corpus Building for Lusoga. *Lexikos* 28: 32–78.

- De Schryver, G.-M. and M. Nabirye.** 2022. Towards a Monitor Corpus for a Bantu Language. A Case Study of Neology Detection in Lusoga. Klosa-Kückelhaus, A., S. Engelberg, C. Möhrs and P. Storjohann (Eds.). 2022. *Dictionaries and Society. Proceedings of the XX EURALEX International Congress, 12–16 July 2022, Mannheim, Germany*: 833-849. Mannheim: IDS-Verlag.
- Elsner, M. and J. Needle.** 2023. Translating a Low-resource Language Using GPT-3 and a Human-readable Dictionary. Nicolai, G., E. Chodroff, F. Mailhot and Ç. Çöltekin (Eds.). 2023. *Proceedings of the 20th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*: 1-13. Toronto: Association for Computational Linguistics.
- Grefenstette, G. and P. Hanks.** 2023. Competing Views of Word Meaning: Word Embeddings and Word Senses. *International Journal of Lexicography* 36(2): 211-219.
- Hanks, P.** 1979. Meaning and Grammar. Hanks, P. (Ed.), T.H. Long (Managing ed.), L. Urdang (Ed. director) and others (Eds.). 1979. *Collins Dictionary of the English Language*: xxxi-xxxv. London: William Collins Sons & Co.
- Hanks, P.** 1988. Typicality and Meaning Potentials. Snell-Hornby, M. (Ed.). 1988. *ZüriLEX '86 Proceedings: Papers read at the EURALEX International Conference, University of Zürich, 9–14 September 1986*: 37-47. Tübingen: A. Francke.
- Hanks, P.** 2000. Do Word Meanings Exist? *Computers and the Humanities* 34(1–2): 205-215.
- Hanks, P.** 2002. Mapping Meaning onto Use. Corréard, M.-H. (Ed.). 2002. *Lexicography and Natural Language Processing. A Festschrift in Honour of B.T.S. Atkins*: 156-198. s.l.: Euralex.
- Hanks, P.** 2004. Corpus Pattern Analysis. Williams, G. and S. Vessier (Eds.). 2004. *Proceedings of the Eleventh EURALEX International Congress, EURALEX 2004, Lorient, France, July 6–10, 2004*: 87-97. Lorient: Faculté des Lettres et des Sciences Humaines, Université de Bretagne Sud.
- Hanks, P.** 2008. Lexical Patterns: From Hornby to Hunston and Beyond. Bernal, E. and J. Decesaris (Eds.). 2008. *Proceedings of the XIII EURALEX International Congress (Barcelona, 15–19 July 2008)* (Sèrie Activitats 20): 89-129. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra.
- Hanks, P.** 2012. How People Use Words to Make Meanings: Semantic Types Meet Valencies. Thomas, J. and A. Boulton (Eds.). 2012. *Input, Process and Product: Developments in Teaching and Language Corpora*: 54-69. Brno: Masaryk University Press.
- Hanks, P., I. El Maarouf and M. Oakes.** 2018. Flexibility of Multiword Expressions and Corpus Pattern Analysis. Sailer, M. and S. Markantonatou (Eds.). 2018. *Multiword Expressions: Insights from a Multilingual Perspective* (Phraseology and Multiword Expressions 1): 93-119. Berlin: Language Science Press.
- Hanks, P. and E. Franklin.** 2019. Do Online Resources Give Satisfactory Answers to Questions about Meaning and Phraseology? Corpas Pastor, G. and R. Mitkov (Eds.). 2019. *Computational and Corpus-based Phraseology. Third International Conference, Europhras 2019, Malaga, Spain, September 25–27, 2019, Proceedings* (Lecture Notes in Artificial Intelligence (LNAI), subseries of Lecture Notes in Computer Science (LNCS) 11755): 159-172. Berlin: Springer.
- Hanks, P. and W. Ma.** 2021. Meaning and Grammar in the Light of Corpus Pattern Analysis. *International Journal of Lexicography* 34(1): 135-149.
- Hanks, P. and S. Može.** 2019. The Way to Analyse 'way': A Case Study in Word-specific Local Grammar. *International Journal of Lexicography* 32(3): 247-269.
- Hanks, P. and J. Pustejovsky.** 2005. A Pattern Dictionary for Natural Language Processing. *Revue Française de linguistique appliquée* 10(2): 63-82.

- Li, Y., Z. Zhao and C. Scarton.** 2025. It's All about In-context Learning! Teaching Extremely Low-resource Languages to LLMs. Christodoulopoulos, C., T. Chakraborty, C. Rose and V. Peng (Eds.). 2025. *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*: 29532-29547. Suzhou: Association for Computational Linguistics.
- Nabirye, M.** 2009. *Eiwanika ly'Olusoga. Eiwanika ly'aboogezi b'Olusoga n'abo abenda okwega Olusoga* [A Dictionary of Lusoga. For Speakers of Lusoga, and for Those Who Would Like to Learn Lusoga] (Linguistics Series 1). Kampala: Menha Publishers.
- Nabirye, M.** 2024. *Obuyumu mu Busoga [Entertainment in Busoga]* (History Series 3). Kampala: Menha Publishers.
- Nabirye, M. and G.-M. de Schryver.** 2013. Digitizing the Monolingual Lusoga Dictionary: Challenges and Prospects. *Lexikos* 23: 297-322.
- Nabirye, M., G.-M. de Schryver, D. Joffe and M.J. MacLeod.** 2012. *e-Eiwanika ly'Olusoga [Digital Lusoga Dictionary]* (Linguistics Series 3). Kampala / Cape Town: Menha Publishers / TshwaneDJe HLT.
- Zhang, C., X. Liu, J. Lin and Y. Feng.** 2024. Teaching Large Language Models an Unseen Language on the Fly. Ku, L.-W., A. Martins and V. Srikumar (Eds.). 2024. *Findings of the Association for Computational Linguistics: ACL 2024*: 8783-8800. Bangkok: Association for Computational Linguistics.

A Systematic Literature Review of Over Three Decades of Dictionary Use Research in *Lexikos*

Balázs Fajt, *Centre for Foreign Languages and Language Exams,
Budapest University of Economics and Business,
Budapest, Hungary*
(fajt.balazs@uni-bge.hu)
(<https://orcid.org/0000-0003-4983-3962>)

Mátyás Bánhegyi, *Centre for Foreign Languages and Language Exams,
Budapest University of Economics and Business, Budapest, Hungary*
(banhegyi.matyas@uni-bge.hu)
(<https://orcid.org/0000-0003-1011-0345>)
and

Katalin P. Márkus, *Department of English Linguistics, Károli Gáspár
University of the Reformed Church in Hungary, Budapest, Hungary*
(p.markus.kata@kre.hu)
(<https://orcid.org/0000-0001-6826-7631>)

Abstract: This study presents a systematic literature review of empirical research on dictionary use published in the journal *Lexikos* over its 35-volume history (1991–2025). The paper employs the PRISMA methodology. Based on 44 relevant studies identified, the review analyses trends in research focus, methodology, educational context, geographic distribution, and investigated languages. Findings reveal a growing interest in dictionary use, particularly since 2012, with a strong emphasis on English-language contexts and higher education. Methodologically, mixed-methods and cross-sectional designs dominate, while longitudinal research remains notably absent. Asia and Europe are the most represented regions, although studies span diverse global contexts. The review highlights underexplored areas such as primary education, non-English dictionary use, and longitudinal perspectives. The review also calls for future research that addresses these gaps, explores dictionary use at different language proficiency levels, and examines the pedagogical integration of digital dictionary tools. The findings provide valuable insights for researchers aiming to investigate dictionary use and dictionary-based learning.

Keywords: DICTIONARY USE, SYSTEMATIC REVIEW, USER BEHAVIOUR, PRISMA, METHODOLOGY, *LEXIKOS*, LANGUAGE EDUCATION, EMPIRICAL STUDIES, QUANTITATIVE RESEARCH, QUALITATIVE RESEARCH, MIXED METHODS RESEARCH

Zusammenfassung: Eine systematische Literaturübersicht über mehr als drei Jahrzehnte Wörterbuchgebrauchsforschung in *Lexikos*. In dieser Studie wird eine systematische Literaturübersicht über die empirische Forschung zum Wörterbuchgebrauch vorgestellt,

die in der Zeitschrift *Lexikos* im Laufe ihrer 35 Bände umfassenden Geschichte (1991–2025) veröffentlicht wurde. Die Studie basiert auf der PRISMA-Methode. Auf der Grundlage von 44 identifizierten relevanten Studien werden Trends in Bezug auf Forschungsschwerpunkte, Methodik, Bildungskontext, geografische Verteilung und untersuchte Sprachen analysiert. Die Ergebnisse zeigen ein wachsendes Interesse an der Wörterbuchnutzung, insbesondere seit 2012, mit einem starken Schwerpunkt auf englischsprachigen Kontexten und Hochschulbildung. Methodisch gesehen dominieren gemischte Methoden und Querschnittsstudien, während Längsschnittstudien nach wie vor nicht vorhanden sind. Asien und Europa sind die am stärksten vertretenen Regionen, obwohl die Studien verschiedene globale Kontexte umfassen. Der Bericht hebt bisher wenig erforschte Bereiche wie die Grundschulbildung, die Verwendung nicht-englischer Wörterbücher und die Längsschnittperspektive hervor. Die Übersicht fordert außerdem zukünftige Forschungsarbeiten, die diese Lücken schließen, den Wörterbuchgebrauch auf verschiedenen Sprachniveaus untersuchen und die pädagogische Integration von digitalen Wörterbuchwerkzeugen untersuchen. Die Ergebnisse liefern wertvolle Erkenntnisse für Forscher, die die Wörterbuchnutzung und das wörterbuchgestützte Lernen untersuchen wollen.

Stichwörter: WÖRTERBUCHBENUTZUNG, SYSTEMATISCHE ÜBERPRÜFUNG, BENUTZERVERHALTEN, PRISMA, METHODIK, LEXIKOS, SPRACHUNTERRICHT, EMPIRISCHE STUDIEN, QUANTITATIVE FORSCHUNG, QUALITATIVE FORSCHUNG, MIXED METHODS RESEARCH

1. Introduction

Dictionaries have long been regarded as indispensable tools in language learning and teaching that provide learners with access to a wealth of lexical, grammatical and cultural information (Karras 2019). Their importance spans various educational contexts from traditional classroom environments through digital learning to even self-directed learning scenarios (Verhun et al. 2024). Over the years, the study of dictionary use has evolved significantly, drawing the interest of researchers in fields such as second language acquisition (SLA), lexicography, and educational technology. Early studies focused primarily on the utilisation of print dictionaries, and investigated how users interacted with monolingual (Barnhart 1962, Quirk 1973), bilingual (Tomaszczyk 1979), and specialized dictionaries (Bergenholtz and Tarp 1995) to enhance their vocabulary (Bogaards 2010, Laufer and Hill 2000, Laufer 1993, Laufer and Hadar 1997), grammar (Bogaards and Van der Kloot 2001, Lew and Dziemianko 2006), reading comprehension (Ard 1982, Bensoussan et al. 1984, Knight 1994, Nesi and Meara 1991, Tono 2001), and writing skills (Bogaards 1996, Dziemianko 2010, Hartmann 2008, Laufer and Hadar 1997). These studies laid the groundwork for understanding the behaviours, strategies, and challenges learners face when using dictionaries as learning aids. With technological advancements, the focus of research has broadened to include electronic, digital and online dictionaries (Dziemianko 2010, Rundell 2015, Taylor and Chan 1994), which offer more dynamic and user-friendly interfaces compared to their print counterparts. These digital tools have introduced features such as audio pronunciations, visual aids, and context-sensitive

or fuzzy search options, thereby fundamentally changing the way users interact with lexical information. Consequently, research began to investigate the effectiveness of digital dictionaries, exploring their impact on language learning outcomes (Dziemianko 2012, Lew and Galas 2008), user preferences (Müller-Spitzer and Koplenig 2014), and the cognitive processes involved (Miller 1986, Okuyama and Igarashi 2007) in dictionary consultation. The shift toward digital dictionaries has raised new questions about how learners access, interpret, and utilize the information provided (Bogaards 1998, Carstens 1995, Scholfield 1999), and this prompted ongoing investigations into the pedagogical implications of integrating dictionary use into language instruction (Chi 2003, Hartmann 1999, Lew and Galas 2008).

In this diverse research landscape, the study of dictionary use remains a dynamic field, reflecting the continuous evolution of both language learning practices and lexicographical innovations. As such, understanding how dictionaries are utilized in various learning contexts is relevant not only for language educators but also for lexicographers aiming to develop more effective reference tools (Atkins and Varantola 1997, Lew and De Schryver 2014, Miller 1986, Müller-Spitzer and Koplenig 2014). The growing body of such research underscores the importance of conducting systematic literature reviews (e.g., Ge et al. 2024) and meta-analyses (e.g., Zhang et al. 2021) to map the current state of knowledge, identify trends, and highlight areas that require further investigation.

Within the field of linguistics, *Lexikos* has established itself over the past three decades as a leading forum for lexicographical scholarship. With 35 volumes published to date, the journal is set to commemorate its 35th anniversary in 2025 as a specialized and influential journal dedicated to the advancement of lexicographic research. It has consistently featured a diverse collection of studies focusing on dictionary use, including user preferences, lookup behaviours, and digital dictionary developments. Despite the existence of numerous reviews and meta-analyses on dictionary use, there is a noticeable gap in research that specifically examines the contributions of *Lexikos* to this body of knowledge. Therefore, the present study seeks to fill this gap by systematically reviewing the literature on dictionary use (and users) published in *Lexikos*. By focusing exclusively on this journal, this study aims to uncover trends, patterns, and areas that may warrant further exploration. Through utilizing the PRISMA methodology, this review synthesizes the findings from *Lexikos* in order to provide a detailed overview of lexicographical research conducted over the past three decades.

Methodology

The present study employs the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) methodology (Page et al. 2021) to systematically identify, select, and synthesize the literature on dictionary use published in *Lexikos*. The methodology is comprised of several key steps, namely

defining inclusion and exclusion criteria, identifying sources, screening for eligibility, and analysing the data.

A comprehensive search was conducted in *Lexikos* on 15 April 2025: 35 volumes published from the journal's inception in 1991 to 2025 were scrutinised. Altogether there were 1028 electronic items (including articles, forewords, editorials, full volumes in a separate publication announcement, front matters and tables of contents), out of which 915 were articles. However, the search targeted (research) articles that investigated any aspect of dictionary use only, including but not limited to dictionary consultation behaviours, digital vs. print dictionary usage, the impact of dictionaries on language learning outcomes, user preferences, and the pedagogical implications of dictionary use. When searching for articles, we decided to manually examine each volume of *Lexikos* to ensure that no relevant studies were overlooked. This involved systematically going through the titles and abstracts of every article in each volume. By manually reviewing the content, we aimed to capture studies that could not be identified through keyword searches alone, thereby enhancing the thoroughness and accuracy of the study selection process.

Inclusion criteria were established to identify articles that specifically focused on dictionary use. The criteria were as follows:

1. published in *Lexikos*: only articles published in the journal *Lexikos* were included to ensure a focused analysis of research trends within the journal;
2. focus on dictionary use: studies must involve theoretical or empirical research related to any form of dictionary use, user behaviour, or dictionary-based language learning, including both qualitative and quantitative as well as mixed methods studies;
3. language: articles published in Afrikaans, Dutch, English, French or German (the language of contributions in *Lexikos*) were all included in the analysis.

Exclusion criteria were designed to eliminate studies that were not relevant to the review's objectives. Articles were excluded if their primary focus was not on dictionary use.

As explained earlier, the initial search identified all articles published in *Lexikos*. After that, relevant titles, abstracts and keywords of all the identified articles were screened for relevance based on the above inclusion and exclusion criteria. In cases where the title, abstract and keywords were insufficient to determine eligibility, the full text of the article was examined. The screening was conducted by all three authors to ensure consistency. Any disagreements regarding article eligibility were resolved through discussion until a consensus was reached. The results of the selection process were recorded in a PRISMA flow diagram (see Figure 1), detailing the number of articles identified, screened, assessed for eligibility, and included in the final review. This flow diagram provides transparency and replicability, as recommended by the PRISMA guidelines (Page et al. 2021).

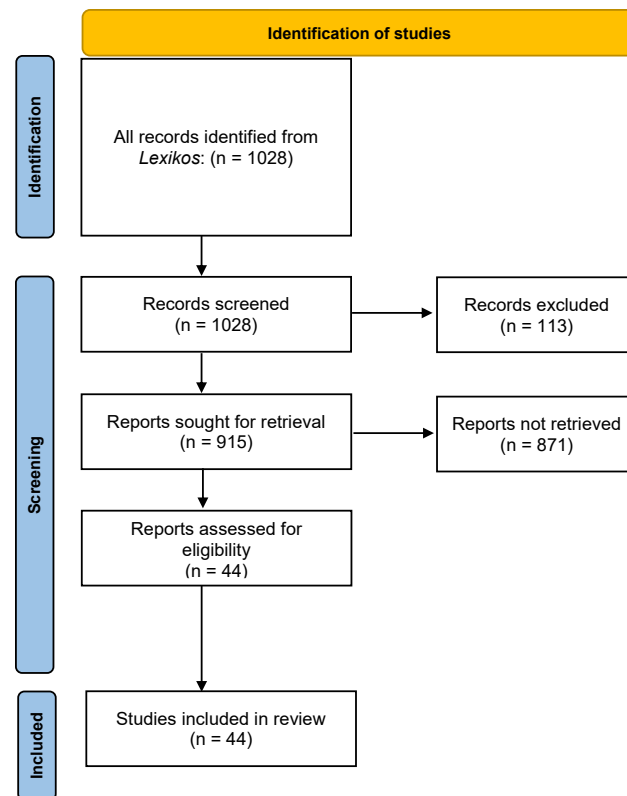


Figure 1

Data from the included studies were extracted systematically using the following coding sheet. Key information extracted from each study included:

- bibliographic details: author(s), year of publication, and volume number of *Lexikos*;
- research focus: the specific aspect(s) of dictionary use addressed in the study (e.g., user behaviour, etc.);
- study design and methodology: type of study (e.g., qualitative, quantitative, or mixed methods), number of participants, country from which the participants/respondents were drawn, languages investigated, sample size, and data collection methods.

The coding process was carried out by all authors to ensure accuracy and consistency. Any discrepancies in data extraction were discussed and resolved.

As a next step, a narrative synthesis approach was employed to analyse the findings of the studies included. Accordingly, the extracted data were organized thematically based on the research focus and methodology, which allowed for the identification of common trends, and evolving themes in dictionary use

over time. Descriptive statistics (e.g., n = number of items) were also used to summarize the frequency of different research topics, methodologies, and findings across the studies.

Finally, to ensure the robustness of the review, the quality of the included studies was assessed using a standardized checklist adapted to the context of dictionary use research. This checklist evaluated factors such as the clarity of research objectives, appropriateness of study design and methodology, transparency of data collection and analysis, and the validity of conclusions drawn. Studies were not excluded based on quality assessment.

Results

Regarding the number of published papers on dictionary use (n = 44, see Appendix), the analysis of publication trends — based on the corpus — clearly demonstrates evolving and growing scholarly interest in dictionary use studies over the examined time period (see Figure 2).

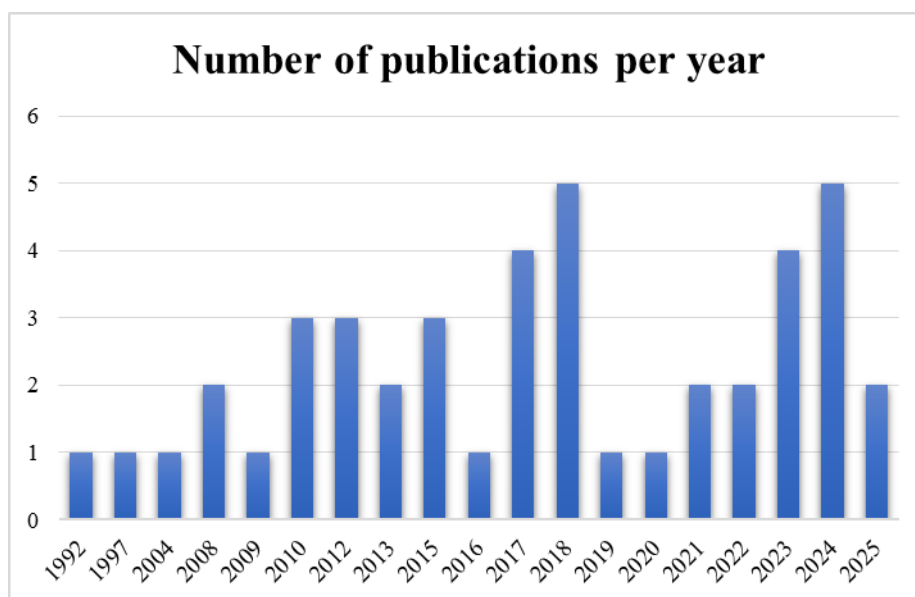


Figure 2

Initially, from 1992 to 2009, research papers in this field appeared sporadically and remained at a relatively low number typically ranging between one and two studies per year. A noticeable shift occurred in 2012, when the frequency of publications on dictionary use began to increase. This development coincided with the rapid spread of online dictionary use, supported by the growing accessi-

bility, speed, and reliability of internet connections. A particularly significant event was Macmillan's announcement in 2012 that, from 2013 onwards, it would cease publishing dictionaries in book form and instead would focus exclusively on digital resources (EURALEX 2012). This shift symbolized a broader trend: web-based dictionaries were becoming increasingly popular and dominant, opening up new research avenues, as the use of digital and online dictionaries involves fundamentally different practices compared to paper dictionaries.

The upward trajectory peaked in the years 2017 and 2018, which represented the highest number of scholarly contributions during this growth period. Following this peak, there was a temporary decrease in the rate of published works, probably due to the COVID-19 situation, which prevented the conducting of many studies building on formerly used research methodologies. However, the data reveal a recent surge in research activities, particularly in 2023 and 2024: this suggests renewed scholarly interest and potentially new directions or methodological advancements in the field.

The language distribution of the analysed studies further highlights a clear dominance of English: nearly all publications (43 out of 44) were written in English, and only one paper appeared in Afrikaans. This dominance underscores the status of English as the primary language of academic communication in the field of lexicographical and dictionary use research. Given that English is widely regarded as the global lingua franca as well as the language of science (O'Neil 2018), this overwhelming preference is hardly surprising. It suggests not only the international scope of the field but also the global scholarly community's sustained interest in issues related to dictionary usage. Moreover, the use of a global lingua franca, such as the English language, indubitably facilitates the wider dissemination and influence of these studies across diverse academic contexts.

From a research methodological point of view, the reviewed studies point to a distinct inclination towards certain research paradigms. However, before examining these in detail, it is important to acknowledge several previous works closely related to our topic. One notable systematic literature review is *Lexikos at Eighteen* by De Schryver (2009), which focused exclusively on articles published in *Lexikos*. More recently, Ge et al. (2024) conducted a scoping review of studies on dictionary use and language learning, offering a broader, systematic overview across various journals. In addition to these major reviews, two other studies have provided valuable insights into dictionary use in specific national contexts: Margalitadze and Meladze (2023) explored dictionary use in Georgia, while P. Márkus and Dringó-Horváth (2023) investigated the Hungarian context. Although these works are closely related to the topic of dictionary use, we did not include or categorize them within our dataset, as they are primarily descriptive in nature and/or synthesize findings from earlier studies, rather than present new empirical research.

The rest of papers may be classified into different categories. Mixed-methods designs are particularly prevalent, featuring in 12 studies, which signals a marked tendency among researchers to involve both qualitative and quantitative approaches for gaining a more holistic understanding of dictionary use (habits).

This preference is consistent with findings from broader review studies in the field. In a recent meta-review of mixed-methods research throughout the 2020s, Riazi and Farsani (2024) cite the work of Amini Farsani et al. (2021), who — drawing their conclusion on the basis of an analysis of 3,814 empirical articles published between 2009 and 2018 in 18 leading applied linguistics journals — found that, after quantitative studies (42.6%), mixed-methods research constituted the second most frequent methodological approach (25.9%). This growing reliance on mixed methods likely reflects the complex, context-dependent nature of dictionary use, which often requires both statistical generalisation and in-depth contextual insights.

Regarding the quantitative research paradigm, a total of 8 studies in our corpus were quantitative in nature. Notably, the dataset also includes four replication studies (Chen and Liu 2023, Ekwa Ebanéga and Moussavou 2008, Jewalani et al. 2019, Li et al. 2024). Although replication studies remain relatively rare in applied linguistics, their presence in our corpus reflects a growing emphasis on methodological rigour and the validation of previous findings. This trend aligns with the increasing recognition of the value of replication studies in the field. In fact, a recent special issue of *Studies in Second Language Acquisition* was dedicated entirely to the topic, highlighting the field's urgent need for more replication studies. As noted in the introduction to that issue (see McManus 2024), replication not only strengthens the empirical foundations of linguistic research but also contributes significantly to a deeper and more reliable understanding of key issues. On the other hand, 10 studies (Al-Ajmi 2008, Benzehra and McCreary 2010, Boonmoh 2012, Campoy-Cubillo 2021, Chaves-Yuste and Márquez-Cruz 2025, Gavriilidou et al. 2024, Kim 2017, Kim 2018, Milić et al. 2018, Pujol and Corrius 2013) explicitly utilized either a pre-test and post-test design or an experimental-group and control-group design. These designs allowed researchers to assess the immediate effects of dictionary use interventions on specific language learning outcomes, such as vocabulary acquisition, comprehension, or test performance.

Regarding purely qualitative research methods in the sample, 6 studies were identified. These studies employ various research methodologies with distinct features and purposes. Law (2024) uses qualitative content analysis to examine the effectiveness of secondary school EFL textbooks for dictionary use training. Furthermore, using the think-aloud protocol method, Vrbinc, Farina and Vrbinc (2022) compare dictionary look-up behaviours. In a descriptive study, Xue (2017) investigates how learner's dictionaries represent the cultural dimension of meaning for EFL learners. Through trends analysis De Schryver (2009) provides a comprehensive overview of *Lexikos'* contributions to key themes and developments in lexicography. Relying on log file analysis, De Schryver, Joffe, Joffe and Hillewaert (2006) study dictionary users' look-up behaviour. In addition, with the help of self-reflections and offering insights into dictionary use experience, Mdee (1997) compares university students' dictionary use and needs. The above studies present different approaches to lexicography research: qualitative content analysis and descriptive research are more systematic in nature, whereas think-

aloud protocols and self-reflections are more introspective. On the other hand, trends analysis and log file analysis are more focused on temporal changes and usage patterns, respectively.

An analysis of the applied research designs reveals clear methodological preferences and highlights notable research patterns. Cross-sectional studies seem dominant, appearing in 34 studies: these clearly indicate researchers' preference for gathering data at a single, specific point in time to gain immediate insights and to conduct comparative analyses of dictionary use. Interestingly, among the papers examined, no explicitly identified longitudinal studies were found. This gap supports Nesi's (2013) observation, who claims that most studies merely take a "snapshot" of dictionary user skills. This suggests that research which tracks participants and observes dictionary usage patterns over extended periods is notably absent from this scholarly domain. This enables the identification of future research directions.

Overall, the examined studies display considerable variability in participant numbers. The sample sizes range widely from small-scale studies involving as few as 10 or 13 participants (generally qualitative studies) to large-scale research with over 300 participants. More specifically, small sample sizes (fewer than 50 participants) are relatively common, with several studies ($n=11$) reporting groups of 10–45 participants, and a moderate number of studies ($n=15$) involved between 50 and 150 participants. Finally, large-scale studies involving more than 150 participants are less frequent but notable ($n=8$), with reported sample sizes of between 168 and 509 participants. This distribution highlights two trends: while small to moderately sized context-specific studies dominate, there is a noteworthy presence of large-sample research, which may allow for more robust statistical analyses and greater generalizability. Nevertheless, the overall variability in sample sizes suggests a lack of standardization in research designs within the field.

The analysis of educational levels targeted by the reviewed studies reveals distinct trends and research preferences. Studies clearly prioritise higher education contexts, with 31 studies explicitly focusing on university-level learners. This suggests that dictionary research predominantly explores issues pertinent to adult learners. This might be explained by easier access to participants or the importance of dictionary use at advanced proficiency levels (cf. Béjoint 1981, Battenburg 1989, Kosem 2008, Nesi 2000). On the other hand, research explicitly focusing on younger learners is noticeably scarce. Only two studies examined dictionary use at the primary education level: one involved mainstream and intercultural primary school pupils aged 10–12 (grades 5–6) (Gavriilidou et al. 2024), while the other study (Van der Merwe 2012) concentrated on primary learners in Grade 4. Additionally, one study extended the scope of research across educational levels: a study by Margalitadze and Meladze (2023) examined dictionary use across primary, secondary and higher education levels, and another focused concurrently on secondary and higher education (Tono 1992).

An analysis of the countries from which the participants/respondents of the analysed studies were drawn highlights distinct geographic areas of research.

China stands out as the most frequently examined context, with seven studies. Following China, Poland appears prominent with four studies, which indicates focused and consistent attention on dictionary use in Polish contexts. Thailand and South Africa each feature in three studies, and a moderate level of interest is apparent in countries like Spain ($n = 3$), Indonesia ($n = 2$), and Korea ($n = 2$), each represented by two studies. At a continental comparison, Asia emerges as the most represented region, with countries such as China, Thailand, Indonesia, Korea, Japan, Taiwan, Jordan, and Kuwait (with the latter two countries classified under 'Asia' in the current study) contributing to the body of research. **Europe** also shows substantial representation, led by Poland and Spain, and supplemented by individual studies from Germany, Greece, Hungary, Serbia, and Slovenia. This points to a well-established tradition of lexicographic inquiry within the European academic landscape. Furthermore, Africa, while less represented overall, still shows a notable presence with three studies from Republic of South Africa and one each from Zimbabwe and Namibia. This suggests growing scholarly engagement with dictionary use on the African continent, particularly in multilingual or second-language learning contexts. It is also important to note that *Lexikos* — the journal analysed in this study — is the only journal in Africa exclusively devoted to lexicography, which likely contributes to the relatively higher visibility of African research in this field. North America and Oceania are less prominently featured, with the United States being the only representative from North America, and no studies originating from Oceania in the reviewed sample. South America is absent: this points to potential gaps in regional representation and opportunities for future research.

The analysis of languages investigated in the reviewed studies underscores the clear dominance of English, reflecting its global status and importance in educational and lexicographical research. English was the primary focus in 26 studies, which indicates widespread scholarly attention towards understanding dictionary use in English-language contexts (cf. Nied Curcio 2022). This focus on English is likely to be due to the international relevance of English and its role as a lingua franca and a language and medium of instruction. Beyond the English language, several studies explored bilingual or multilingual dictionary use. These studies investigated dictionary interactions between English and other languages including e.g. Serbian, Chinese, Arabic, Japanese, German, French, and Swahili. These studies reflect an interest in examining dictionary use in linguistically diverse settings and highlight cross-linguistic issues, challenges and opportunities. In addition, studies examined dictionary use related solely to specific languages like Greek, Georgian, Bahasa Indonesian, Afrikaans, Ndebele, and Gabonese. Though individually less represented, these studies illustrate curiosity about dictionary use concerning languages beyond the dominant global language of English.

Conclusion

This systematic literature review set out to examine dictionary use research published in *Lexikos*, a specialized journal of lexicography that has played a

pivotal role in shaping the discourse around dictionary-related studies over the past three decades. By analysing 44 empirical papers from 1991 to 2025 using the PRISMA framework, the review has revealed clear trends, methodological patterns, and emerging gaps in the field. The findings demonstrate that scholarly interest in dictionary use has grown significantly over time, with a notable increase in publications from 2012 onwards, coinciding with the rise of digital and online dictionaries. This shift has brought about methodological diversification, particularly in the increased use of mixed-methods designs. Despite this methodological expansion, cross-sectional research seems to dominate the field and there is scarcity of longitudinal research on dictionary use. In terms of research contexts, the field shows a focus on higher education, with minimal attention given to primary and secondary school learners. Geographically, the reviewed studies predominantly originate from Asia and Europe, with a strong emphasis on English-language contexts, both in terms of the language of publication and the language of investigation.

Importantly, the review identifies a noticeable lack of replication studies and large-scale investigations, as well as insufficient reporting of participant numbers in some cases, which are factors that can limit the generalizability and transparency of findings. Additionally, while the reviewed studies cover a wide range of topics such as user behaviour, digital innovations, and pedagogical applications, there remains untapped potential for deeper exploration into cognitive processes, long-term learning effects, and classroom-based interventions. As a final note, we encourage future authors in this field to place more emphasis on the consistent use of research methodology and terminology, as clearer conceptual frameworks and methodological transparency will strengthen the validity and impact of dictionary use research.

References

- Al-Ajmi, H.** 2008. The Effectiveness of Dictionary Examples in Decoding: The Case of Kuwaiti Learners of English. *Lexikos* 18: 15-26.
<https://doi.org/10.5788/18-0-474>
- Alzi'abi, S.E.** 2017. Guessing Verb-Adverb Collocations: Arab EFL Learners' Use of Electronic Dictionaries. *Lexikos* 27: 50-77.
<https://doi.org/10.5788/27-1-1394>
- Amini Farsani, M., H.R. Jamali, M. Beikmohammadi, B. Daneshvar Ghorbani and L. Soleimani.** 2021. Methodological Orientations, Academic Citations, and Scientific Collaboration in Applied Linguistics: What Do Research Synthesis and Bibliometrics Indicate? *System* 100(3): 102547.
<https://doi.org/10.1016/j.system.2021.102547>
- Ard, J.** 1982. The Use of Bilingual Dictionaries by ESL Students While Writing. *ITL Review of Applied Linguistics* 58: 1-27.
- Atkins, B.T.S. and K. Varantola.** 1997. Monitoring Dictionary Use. *International Journal of Lexicography* 10(1): 1-45.
- Barnhart, C.** 1962. Problems in Editing Commercial Monolingual Dictionaries. Householder, F.W. and S. Saporta (Eds.). 1962. *Problems in Lexicography*: 161-181. Bloomington: Indiana University Press.

- Battenburg, J.D.** 1989. *A Study of English Monolingual Learners' Dictionaries and their Users*. West Lafayette, Indiana, USA: Purdue University.
- Béjoint, H.** 1981. The Foreign Student's Use of Monolingual English Dictionaries: A Study of Language Needs and Reference Skills. *Applied Linguistics* II(3): 207-222.
- Bensoussan, M., D. Sim and R. Weiss.** 1984. The Effect of Dictionary Usage on EFL Test Performance Compared with Student and Teacher Attitudes and Expectations. *Reading in a Foreign Language* 2(2): 262-276.
- Benzehra, R. and D.R. McCreary.** 2010. The OEAD: New Perspectives on English–Arabic Dictionaries for English-Speaking Users and Users' Surveys. *Lexikos* 20: 1-25.
<https://doi.org/10.5788/20-0-130>
- Bergenholtz, H. and S. Tarp (Eds.).** 1995. *Manual of Specialised Lexicography. The Preparation of Specialised Dictionaries*. Amsterdam/Philadelphia: John Benjamins.
- Beyer, H.L. and A.K. Faul.** 2010. Namibian University Entrants' Concepts of 'a Dictionary'. *Lexikos* 20: 644-679.
<https://doi.org/10.5788/20-0-160>
- Bogaards, P.** 1996. Dictionaries for Learners of English. *International Journal of Lexicography* 9(4): 277-320.
- Bogaards, P.** 1998. Using Dictionaries: Which Words Are Looked Up by Foreign Language Learners? Atkins, B.T.S. and K. Varantola (Eds.). 1998. *Studies of Dictionary Use by Language Learners and Translators*: 151-157. Tübingen: Niemeyer.
- Bogaards, P.** 2010. Dictionaries and Second Language Acquisition. Dykstra, A. and T. Schoonheim (Eds.). 2010. *Proceedings of the XIV EURALEX International Congress, 6–10 July 2010, Leeuwarden*: 99-123. Ljouwert: Fryske Akademy.
- Bogaards, P. and W.A. van der Kloot.** 2001. The Use of Grammatical Information in Learners' Dictionaries. *International Journal of Lexicography* 14(2): 97-121.
<https://doi.org/10.1093/ijl/14.2.97>
- Boonmoh, A.** 2010. The Use of Pocket Electronic Dictionaries by Thai University Students. *Lexikos* 20: 680-690.
<https://doi.org/10.5788/20-0-161>
- Boonmoh, A.** 2012. E-dictionary Use under the Spotlight: Students' Use of Pocket Electronic Dictionaries for Writing. *Lexikos* 22: 43-68.
<https://doi.org/10.5788/22-1-997>
- Boonmoh, A.** 2021. Use of Dictionaries and Online Tools for Reading by Thai EFL Learners in a Naturalistic Setting. *Lexikos* 31(1): 239-258.
<https://doi.org/10.5788/31-1-1645>
- Çağlayan Özkan, A.** 2025. The Effectiveness of the "Real-Life-Based School Dictionary" as an Electronic Dictionary. *Lexikos* 35(1): 143-156.
<https://doi.org/10.5788/35-1-2006>
- Campoy-Cubillo, M.C.** 2021. Fostering Learners' Online Dictionary Skills through Active Dictionary Rubrics. *Lexikos* 31(1): 487-510.
<https://doi.org/10.5788/31-1-1663>
- Carstens, A.** 1995. Language Teaching and Dictionary Use: An Overview. *Lexikos* 5: 105-116.
- Chan, A.Y.W.** 2012. Advanced Cantonese ESL Learners' Use of a Monolingual Dictionary for Language Production. *Lexikos* 22: 109-138.
<https://doi.org/10.5788/22-1-1000>

- Chan, A.Y.W.** 2017. The Effectiveness of Using a Bilingualized Dictionary for Determining Noun Countability and Article Selection. *Lexikos* 27: 183-213.
<https://doi.org/10.5788/27-1-1399>
- Chaves-Yuste, B. and M. Márquez-Cruz.** 2025. Digital Lexicography in ESP: Enhancing Vocabulary Acquisition and Reading Comprehension in Students of Commerce Degree. *Lexikos* 35(1): 109-128.
<https://doi.org/10.5788/35-1-2004>
- Chen, Y. and S. Liu.** 2023. A Further Look into the Use of a Dictionary APP in EFL Writing: A Replication Study. *Lexikos* 33(1): 324-349.
<https://doi.org/10.5788/33-1-1820>
- Chi, M.L.A.** 2003. *An Empirical Study of the Efficacy of Integrating the Teaching of Dictionary Use into a Tertiary English Curriculum in Hong Kong. Vol. IV. Research Reports* (Ed. G. James). Hong Kong: Language Centre, Hong Kong University of Science and Technology, Language Center.
- Claassen, D.** 2015. 'n Evaluasie van *Pharos Aanleerderwoordeboek vir skole/Learner's Dictionary for Schools* en *Longman-HAT Afrikaans Dictionary and Grammar for English Speakers* in 'n Afrikaanse taalverwerwingskursus op universiteitsvlak. *Lexikos* 25: 31-55.
<https://doi.org/10.5788/25-1-1290>
- De Schryver, G.-M.** 2009. *Lexikos* at Eighteen: An Analysis. *Lexikos* 19: 372-403.
<https://doi.org/10.5788/19-0-445>
- De Schryver, G.-M., D. Joffe, P. Joffe and S. Hillewaert.** 2006. Do Dictionary Users Really Look Up Frequent Words? — On the Overestimation of the Value of Corpus-based Lexicography. *Lexikos* 16: 67-83.
<https://doi.org/10.5788/16-0-648>
- Dziemianko, A.** 2010. Paper or Electronic? The Role of Dictionary Form in Language Reception, Production and the Retention of Meaning and Collocations. *International Journal of Lexicography* 23(3): 257-273.
- Dziemianko, A.** 2012. On the Use(fulness) of Paper and Electronic Dictionaries. Granger, S. and M. Paquot (Eds.). 2012. *Electronic Lexicography*: 319-341. Oxford: Oxford University Press.
- Ekwa Ebanéga, G.-M. and F.T. Moussavou.** 2008. A Survey of the Dictionary Use of Gabonese Students at Two South African Universities. *Lexikos* 18: 349-365.
<https://doi.org/10.5788/18-0-491>
- EURALEX.** 2012. *End of Print Dictionaries at Macmillan*.
<https://euralex.org/2012/11/05/end-of-print-dictionaries-at-macmillan/> [10 April 2025].
- Fajt, B.** 2024. Exploring Dictionary Preferences: A Comparative Study of EFL and GFL Learners in Hungarian Higher Education. *Lexikos* 34(1): 419-436.
<https://doi.org/10.5788/34-1-1945>
- Fallianda, F.** 2020. A Survey of Indonesian Students' Use of Dictionaries. *Lexikos* 30(1): 609-628.
<https://doi.org/10.5788/30-1-1607>
- Gavrilidou, Z., A. Markos and E. Konstantinidou.** 2024. How Can We Raise Strategic Dictionary Use in the Classroom: The Effect of a Dictionary Awareness Program on Dictionary Use Strategies. *Lexikos* 34(1): 99-122.
<https://doi.org/10.5788/34-1-1882>
- Ge, X., S. Zhang, H. Xu and X. Zhang.** 2024. A Scoping Review of Studies into Dictionary Use and Language Learning. *Lexikos* 34(1): 331-354.
<https://doi.org/10.5788/34-1-1940>

- Hadebe, S.** 2004. Improving Dictionary Skills in Ndebele. *Lexikos* 14: 89-104.
<https://doi.org/10.5788/14-0-684>
- Hartmann, R.R.K. (Ed.).** 1999. *Dictionaries in Language Learning. Recommendations, National Reports and Thematic Reports from the TNP Sub-project 9: Dictionaries*. Thematic Network Project in the Area of Languages, Berlin Free University.
- Hartmann, R.R.K.** 2008. Twenty-five Years of Dictionary Research: Taking Stock of Conferences and Other Lexicographic Events Since LEXeter '83. Bernal, E. and J. DeCesaris (Eds.). 2008. *Proceedings of the XIII EURALEX International Congress, 15-19 July 2008, Barcelona, Spain*: 131-148. Barcelona: Universitat Pompeu Fabra.
- Jewalani, A.P., M. Laksman-Huntley and H.Y. Anjarningsih.** 2019. Lexical Understanding of Native Bahasa Indonesia Speakers through Word Association to Improve Dictionary Definitions. *Lexikos* 29: 36-54.
<https://doi.org/10.5788/29-1-1511>
- Karras, I.** 2019. The Use of Dictionaries in English as Foreign Language Teaching and Learning. Osuchowska, D. and L. Harmon (Eds.). 2019. *Contents, Use, Usability: Dictionaries from the Perspective of a Translator and a Language Teacher*: 49-62. Bern: Peter Lang International Academic Publishing.
- Kim, S.** 2017. "Minimum Input, Maximum Output, Indeed!" Teaching Collocations Through Collocation Dictionary Skills Development. *Lexikos* 27: 265-286.
<https://doi.org/10.5788/27-1-1403>
- Kim, S.** 2018. A Lexicographic Approach to Teaching the English Article System: Help or Hindrance? *Lexikos* 28: 196-220.
<https://doi.org/10.5788/28-1-1462>
- Knight, S.** 1994. Dictionary Use While Reading: The Effects on Comprehension and Vocabulary Acquisition for Students of Different Verbal Abilities. *The Modern Language Journal* 78(3): 285-299.
<https://doi.org/10.2307/330108>
- Kosem, I.** 2008. Dictionaries for University Students: A Real Deal or Merely a Marketing Ploy? Bernal, E. and J. DeCesaris (Eds.). 2008. *Proceedings of the XIII EURALEX International Congress, 15-19 July 2008, Barcelona, Spain*: 1575-1584. Barcelona: Universitat Pompeu Fabra.
- Laufer, B.** 1993. The Effect of Dictionary Definitions and Examples on the Use and Comprehension of New L2 Words. *Cahiers de Lexicologie* 63(2): 131-142.
- Laufer, B. and L. Hadar.** 1997. Assessing the Effectiveness of Monolingual, Bilingual, and "Bilingualized" Dictionaries in the Comprehension and Production of New Words. *The Modern Language Journal* 81(2): 189-196.
- Laufer, B. and M. Hill.** 2000. What Lexical Information Do L2 Learners Select in a CALL Dictionary and How Does It Affect Word Retention? *Language Learning & Technology* 3(2): 58-76.
- Law, Wai-on.** 2024. Dictionary Use Training in Secondary School EFL Textbooks in Taiwan. *Lexikos* 34(1): 77-98.
<https://doi.org/10.5788/34-1-1880>
- Lew, R. and G.-M. de Schryver.** 2014. Dictionary Users in the Digital Revolution. *International Journal of Lexicography* 27(4): 341-359.
- Lew, R. and A. Dziemianko.** 2006. A New Type of Folk-inspired Definition in English Monolingual Learners' Dictionaries and its Usefulness for Conveying Syntactic Information. *International Journal of Lexicography* 19(3): 225-242.

- Lew, R. and K. Galas.** 2008. Can Dictionary Skills Be Taught? The Effectiveness of Lexicographic Training for Primary School Level Polish Learners of English. Bernal, E. and J. DeCesaris (Eds.). 2008. *Proceedings of the XIII EURALEX International Congress, 15-19 July 2008, Barcelona, Spain*: 1273-1285. Barcelona: Universitat Pompeu Fabra.
- Lew, R., M. Grzelak and M. Leszkowicz.** 2013. How Dictionary Users Choose Senses in Bilingual Dictionary Entries: An Eye-Tracking Study. *Lexikos* 23: 228-254.
<https://doi.org/10.5788/23-1-1213>
- Li, L. and H. Xu.** 2015. Using an Online Dictionary for Identifying the Meanings of Verb Phrases by Chinese EFL Learners. *Lexikos* 25: 191-209.
<https://doi.org/10.5788/25-1-1295>
- Li, L., H. Wang and H. Xu.** 2024. Adaptation and Validation of the Strategy Inventory for Electronic Dictionary Use (S.I.E.D.U.) for Chinese Learners. *Lexikos* 34(1): 245-268.
<https://doi.org/10.5788/34-1-1914>
- Liang, P. and D. Xu.** 2018. An Empirical Study of EFL Learners' Dictionary Use in Chinese-English Translation. *Lexikos* 28: 221-244.
<https://doi.org/10.5788/28-1-1463>
- Margalitadze, T. and G. Meladze.** 2023. How to Solve Problems in Dictionary Use: The Georgian Case. *Lexikos* 33(2): 63-78.
<https://doi.org/10.5788/33-2-1840>
- McManus, K.** 2024. Replication Studies in Second Language Acquisition Research: *Definitions, Issues, Resources, and Future Directions: Introduction to the Special Issue. Studies in Second Language Acquisition* 46(5): 1299-1319.
<https://doi.org/10.1017/S0272263124000652>
- Mdee, J.S.** 1997. Language Learner's Use of a Bilingual Dictionary: A Comparative Study of Dictionary Use and Needs. *Lexikos* 7: 94-106.
<https://doi.org/10.5788/7-1-974>
- Milić, M., T. Glušac and A. Kardoš.** 2018. The Effectiveness of Using Dictionaries as an Aid for Teaching Standardization of English-based Sports Terms in Serbian. *Lexikos* 28: 262-286.
<https://doi.org/10.5788/28-1-1465>
- Miller, G.A.** 1986. Dictionaries in the Mind. *Language and Cognitive Processes* 1(3): 171-185.
<https://doi.org/10.1080/01690968608407059>
- Müller-Spitzer, C. and A. Koplenig.** 2014. Online Dictionaries: Expectations and Demands. Müller-Spitzer, C. (Ed.). 2014. *Using Online Dictionaries*. Lexicographica Series Maior 145: 143-188. Berlin: Walter de Gruyter.
- Müller-Spitzer, C., M.J. Domínguez Vázquez, M. Nied Curcio, I.M. Silva Dias and S. Wolfer.** 2018. Correct Hypotheses and Careful Reading Are Essential: Results of an Observational Study on Learners Using Online Language Resources. *Lexikos* 28: 287-315.
<https://doi.org/10.5788/28-1-1466>
- Nesi, H.** 2000. *The Use and Abuse of EFL Dictionaries: How Learners of English as a Foreign Language Read and Interpret Dictionary Entries*. Tübingen: Max Niemeyer.
- Nesi, H.** 2013. Researching Users and Uses of Dictionaries. Jackson, H. (Ed.). 2013. *The Bloomsbury Companion to Lexicography*: 62-74. London: Bloomsbury Academic.
- Nesi, H. and P. Meara.** 1991. How Using Dictionaries Affects Performance in Multiple-choice EFL Tests. *Reading in a Foreign Language* 8(1): 631-643.

- Nied Curcio, M.** 2022. Dictionaries, Foreign Language Learners and Teachers. New Challenges in the Digital Era. Klosa-Kückelhaus, A., S. Engelberg, Ch. Möhrs and P. Storjohann (Eds.). 2022. *Dictionaries and Society: Proceedings of the XX EURALEX International Congress, 12–16 July 2022, Mannheim, Germany*: 71–84. Mannheim: IDS-Verlag.
- Okuyama, Y. and H. Igarashi.** 2007. Think-aloud Protocol on Dictionary Use by Advanced Learners of Japanese. *The JALT CALL Journal* 3(1–2): 45–58.
- O'Neil, D.** 2018. English as the Lingua Franca of International Publishing. *World Englishes* 37(2): 146–165.
- Page, M.J., J.E. McKenzie, P.M. Bossuyt, I. Boutron, T. Hoffmann, C.D. Mulrow et al.** 2021. The PRISMA 2020 Statement: An Updated Guideline for Reporting Systematic Reviews. *BMJ* 372(71): 1–9.
<https://doi.org/10.1136/bmj.n71>
- P. Márkus, K. and I. Dringó-Horváth.** 2023. Developing Dictionary Skills through Monolingual and Bilingual English Dictionaries at Tertiary-level Education in Hungary. *Lexikos* 33(1): 350–381.
<https://doi.org/10.5788/33-1-1821>
- Ptasznik, B.** 2022. After the Digital Revolution: Dictionary Preferences of English Majors at a European University. *Lexikos* 32(1): 220–249.
<https://doi.org/10.5788/32-1-1721>
- Ptasznik, B.** 2023. Evaluating the Usefulness of the Learning Tools in Monolingual Online Dictionaries for Learners of English: Gauging the Preferences of Polish Students of English. *Lexikos* 33(1): 109–124.
<https://doi.org/10.5788/33-1-1799>
- Ptasznik, B. and R. Lew.** 2018. Which Learning Tools Accompanying the Paid Online Version of LDOCE Do Advanced Learners of English Find Useful? *Lexikos* 28: 448–460.
<https://doi.org/10.5788/28-1-1474>
- Pujol, D. and M. Corrius.** 2013. The Role of the Learner's Native Culture in EFL Dictionaries: An Experimental Study. *Lexikos* 23: 565–584.
<https://doi.org/10.5788/23-1-1230>
- Quirk, R.** 1973. The Social Impact of Dictionaries in the UK. McDavid, R.I. and A.R. Duckert (Eds.). 1973. *Lexicography in English*: 76–88. New York: New York: Academy of Sciences.
- Riazi, A.M. and M.A. Farsani.** 2024. Mixed-methods Research in Applied Linguistics: Charting the Progress through the Second Decade of the Twenty-first Century. *Language Teaching* 57(2): 143–182.
<https://doi.org/10.1017/S0261444823000332>
- Rundell, M.** 2015. From Print to Digital: Implications for Dictionary Policy and Lexicographic Conventions. *Lexikos* 25: 301–322.
<https://doi.org/10.5788/25-1-1301>
- Schierholz, S.J.** 2015. Methods in Lexicography and Dictionary Research. *Lexikos* 25: 323–352.
<https://doi.org/10.5788/25-1-1302>
- Scholfield, P.** 1999. Dictionary Use in Reception. *International Journal of Lexicography* 12(1): 13–34.
- Taylor, A. and A. Chan.** 1994. Pocket Electronic Dictionaries and Their Use. Martin, W., W. Meijs, M. Moerland, E. ten Pas, P. van Sterkenburg and P. Vossen (Eds.). 1994. *EURALEX '94 Proceedings. Papers Submitted to the 6th EURALEX International Congress on Lexicography in Amsterdam, The Netherlands*: 598–605. Amsterdam: EURALEX.

- Tomaszczyk, J.** 1979. Dictionaries: Users and Uses. *Glottodidactica* 12: 103-119.
- Tono, Y.** 1992. The Effect of Menus on EFL Learners' Look-up Processes. *Lexikos* 2: 230-253.
<https://doi.org/10.5788/2-1-1136>
- Tono, Y.** 2001. *Research on Dictionary Use in the Context of Foreign Language Learning: Focus on Reading Comprehension*. Lexicographica Series Maior 106. Tübingen: Max Niemeyer.
- Van der Merwe, M.F.** 2012. A Study of the Use of the *HAT Afrikaanse Skoolwoordeboek* by Primary School Children. *Lexikos* 22: 352-366.
<https://doi.org/10.5788/22-1-1012>
- Verhun, T., M. Davydovych, K. Parnus, M. Kosheleva, I. Biliak and O. Khorosh.** 2024. Algorithms for Optimizing Vocabulary Acquisition in Language Learning. *Proceedings of the International Conference on Applied Innovation in IT* 12(1): 105-114.
- Vrbinc, M., D.M.T.Cr. Farina and A. Vrbinc.** 2022. Who is Sharpest at Looking Up *sharp*? Comparing Two Parallel Groups of Dictionary Users. *Lexikos* 32(1): 368-391.
<https://doi.org/10.5788/32-1-1750>
- Xue, M.** 2017. Representing the Cultural Dimension of Meaning in Learner's Dictionaries — From the Perspective of Chinese EFL Learners in L2 Reception. *Lexikos* 27: 578-596.
<https://doi.org/10.5788/27-1-1416>
- Zhang, S., H. Xu and X. Zhang.** 2021. The Effects of Dictionary Use on Second Language Vocabulary Acquisition: A Meta-analysis. *International Journal of Lexicography* 34(1): 1-38.
<https://doi.org/10.1093/ijl/ecaa010>
- Zou, D.** 2016. Comparing Dictionary-induced Vocabulary Learning and Inferencing in the Context of Reading. *Lexikos* 26: 372-390.
<https://doi.org/10.5788/26-1-1345>

APPENDIX

Paper	Research design and number of participants (n)	Target group	Country	Language investigated
Chaves-Yuste, B. and M. Márquez-Cruz (2025)	experiment (n=90)	first year university students	Spain	English
Çağlayan Özkan, A. (2025)	quantitative (questionnaire + achievement test) (n=83)	secondary school students (5th and 7th graders)	Turkey	Turkish
Ge, X., S. Zhang, H. Xu and X. Zhang (2024)	scoping review	—	—	—
Fajt, B. (2024)	quantitative (n=371)	English and German majors at university	Hungary	English, German
Li, L., H. Wang and H. Xu (2024)	quantitative (n=494)	English majors in their third year (probably around C1)	China	English
Gavriilidou, Z., A. Markos and E. Konstantinidou (2024)	quantitative (n=150)	primary school pupils (aged 10–12) (native)	Greece	Greek
Law, Wai-on (2024)	qualitative content analysis	—	Taiwan	English
Margalitadze, T. and G. Meladze (2023)	analyses the situation of dictionary use in Georgia	Primary, secondary and higher education	Georgia	—
P. Márkus, K. and I. Dringó-Horváth (2023)	analyses the situation of dictionary use in Hungary	—	Hungary	—
Chen, Y. and S. Liu (2023)	mixed methods (n=62)	English majors (C1)	China	English
Ptasznik, B. (2023)	mixed methods (n=318)	English majors (C1–C2)	Poland	English
Vrbinc, M., D.M.T.Cr. Farina and A. Vrbinc (2022)	qualitative (n=17) (think-aloud protocol)	English majors (C1–C2)	Slovenia	English
Ptasznik, B. (2022)	mixed methods (n=168)	English majors (B2–C1)	Poland	English
Campoy-Cubillo, M.C. (2021)	mixed methods (n=75)	English majors (B1)	Spain	English
Boonmoh, A. (2021)	experiment (?) in a case study (n=14)	university students (A2–B1)	Thailand	English
Fallianda, F. (2020)	quantitative (n=88)	university students (A2–B1)	Indonesia	English

Jewalani, A.P., M. Laksman-Huntley, and H.Y. Anjarningsih (2019)	observational study, experiment (n=45)	native speakers	Indonesia	Bahasa Indonesia
Ptasznik, B. and R. Lew (2018)	experiment (n=114)	university students (B2–C1)	Poland	English
Müller-Spitzer, C., M.J. Domínguez Vázquez, M. Nied Curcio, I.M. Silva Dias and S. Wolfer (2018)	observational study (mixed methods: different quantitative and qualitative methods via the examination) (n=42)	university students (A2–B1)	Germany, but the students were from Portugal, Italy and Spain	English
Milić, M., T. Glušac and A. Kardoš (2018)	quantitative (n=255)	university students (B1)	Serbia	Serbian, English
Liang, P. and D. Xu (2018)	mixed methods (n=50) (observational and qualitative)	university students (B1–B2)	China	Chinese, English
Kim, S. (2018)	quasi experiment (n=43)	university students (B1–B2)	Korea	English
Xue, M. (2017)	descriptive (qualitative)	—	China	—
Kim, S. (2017)	experiment (n=59)	university students (A2–C1)	Korea	English
Chan, A.Y.W. (2017)	mixed methods (n=30) (quantitative + test)	university students (B2–C1)	China	English
Alzi'abi, S.E. (2017)	mixed methods (n=82)	university students (B2–C1)	Jordan	English
Zou, D. (2016)	mixed methods (n=104) (experiment + qualitative analysis with codes)	university students (B2)	China	English
Schierholz, S.J. (2015)	descriptive	—	Germany	—
Li, L. and H. Xu (2015)	mixed methods (n=32)	university students (B1)	China	English
Claassen, D. (2015)	qualitative (n=2 dictionaries)	—	Republic of South Africa	Afrikaans
Pujol, D. and M. Corrius (2013)	experiment (n=100)	university students (B1)	Spain	English
Lew, R., M. Grzelak and M. Leszkowicz (2013)	experiment (n=10)	higher education (5 participants: B2–C1; 5 participants: A2–B1)	Poland	English
Van der Merwe, M.F. (2012)	experiment (200)	native	Republic of South Africa	Afrikaans home language and Afrikaans first additional language
Chan, A.Y. (2012)	mixed methods (n=31)	university students (C1)	China	English

Boonmoh, A. (2012)	mixed methods (n=13) (think-aloud protocol, observation, and retrospective interviews)	university students (B1–B2)	Thailand	English
Beyer, H L. and A.K. Faul (2010)	quantitative (n=509)	first year students in higher education	Namibia	English
Boonmoh, A. (2010)	mixed methods	university students (B1–B2)	Thailand	English
De Schryver, G.-M. (2009)	descriptive investigating trends in <i>Lexikos</i> (qualitative)	—	—	—
De Schryver, G.-M., D. Joffe, P. Joffe and S. Hillewaert (2006)	log life analysis	—	Republic of South Africa	Swahili/English
Ekwa Ebanéga, G.-M. and F.T. Moussavou (2008)	quantitative (n=100)	native	Republic of South Africa	Gabonese
Al-Ajmi, H. (2008)	experiment (n=54)	university students (B2–C1)	Kuwait	English
Hadebe, S. (2004)	descriptive	—	Zimbabwe	Ndebele
Mdee, J.S. (1997)	qualitative (n=56) (self-reflections of participants)	university students (B1–C1)	Tanzania/ United Kingdom	German/French
Tono, Y. (1992)	experiment (n=239) (test)	secondary school students, university students (B1–B2)	Japan	English, Japanese

Publikasieaankondigings / Publication Announcements

Edward Finegan and Michael Adams. *The Cambridge Handbook of the Dictionary*. 2024, 801 pp. ISBN 978-1-108-83624-1 (Hardback), ISBN 978-1-108-81887-2 (Paperback), ISBN 978-1-108-86443-5 (eBook). Cambridge/New York, NY: Cambridge University Press. Price: US\$ 170.00 (Hardback), US\$ 170.00 (eBook). (Review in *Lexikos* 35(1).)

Pedro A. Fuertes-Olivera. *A Guide to Practical Online Lexicography*. 2025, 334 pp. ISBN: 978-1032731544 (Hardback), 9781003426998 (eBook). London/New York: Routledge. Price: US\$ 180.00 (Hardback).

Yongwei Gao. *A Dictionary of Contemporary English Neologisms and Trendy Words*. 2024, 406 pp. ISBN: 978-7-5213-5810-0 (Paperback). Beijing: Foreign Language Teaching and Research Press. Price \$9.32. (Review in *Lexikos* 35(1).)

Laura Giacomini and Valentina Piunno (Editors). *Patterns of Meaning in Lexicography and Lexicology*. 2025, XI + 319 pp. ISBN 978-3-11-148075-6 (Hardback), ISBN 978-3-11-154594-3 (eBook), ISSN 0175-9264. Lexicographica Series Maior 167. Berlin/Boston: Walter de Gruyter. Price: €109.95 (Hardback, eBook).

Theresa Kruse (Herausgeber). *Ein fachsprachliches elektronisches Lernwörterbuch. Konzeption: Prototypenstellung und Evaluation am Beispiel der Mathematik*. 2025, XIX + 356 pp. ISBN 978-3-11-138949-3 (Hardback), ISBN 978-3-11-159135-3 (eBook), ISSN 0175-9264. Lexicographica Series Maior 168. Berlin/Boston: Walter de Gruyter. Preis: €119.95 (Hardback, eBook).

Du Ruiqing and Jiang Yajun. *New Century Chinese–English Dictionary*. Third Edition. 2025, XXXVIII + 2490 pp. ISBN 978-7-5213-5418-8 (Hardback). Beijing: Foreign Language Teaching and Research Press. Price \$54.43. (Review in *Lexikos* 35(1).)

Stephen Turton. *Before the Word was Queer: Sexuality and the English Dictionary, 1600–1930*. 2024, xv + 333 pages. ISBN 978-1-316-51873-1 (Hardback). Cambridge: Cambridge University Press. Price: US\$ 94.67 (Hardback). (Review in *Lexikos* 35(1).)

VOORSKRIFTE AAN SKRYWERS

(Tree asseblief met ons in verbinding (lexikos@sun.ac.za) vir 'n uitvoeriger weergawe van hierdie instruksies of besoek ons webblad: <http://lexikos.journals.ac.za/>)

A. REDAKSIONELE BELEID

1. Aard en inhoud van artikels

Artikels kan handel oor die suiwer leksikografie of oor implikasies wat aanverwante terreine, bv. linguistiek, algemene taalwetenskap, terminologie, rekenaarwetenskap en bestuurskunde vir die leksikografie het.

Bydraes kan onder een van die volgende rubrieke geklassifiseer word:

(1) **Artikels:** Grondige oorspronklike wetenskaplike navorsing wat gedoen en die resultate wat verkry is, of bestaande navorsingsresultate en ander feite wat op 'n oorspronklike wyse oorsigtelik, interpreterend, vergelykend of krities evaluerend aangebied word.

(2) **Resensieartikels:** Navorsingsartikels wat in die vorm van 'n kritiese resensie van een of meer gepubliseerde wetenskaplike bronne aangebied word.

Bydraes in kategorieë (1) en (2) word aan streng anonieme keuring deur onafhanklike akademiese vakgenote onderwerp ten einde die internasionale navorsingsgehalte daarvan te verseker.

(3) **Resensies:** 'n Ontleding en kritiese evaluering van gepubliseerde wetenskaplike bronne en produkte, soos boeke en rekenaarprogramme.

(4) **Projekte:** Besprekings van leksikografiese projekte.

(5) **Leksikonotas:** Enige artikel wat praktykgerigte inligting, voorstelle, probleme, vrae, kommentaar en oplossings betreffende die leksikografie bevat.

(6) **Leksikovarja:** Enigeen van 'n groot verskeidenheid artikels, aankondigings en nuusvrystellings van leksikografiese verenigings wat veral vir die praktiserende leksikograaf van waarde sal wees.

(7) **Ander:** Van tyd tot tyd kan ander rubrieke deur die redaksie ingevoeg word, soos Leksikoprogrammatuur, Leksiko-opname, Leksikobibliografie, Leksikonuus, Lexikofokus, Leksiko-eerbewys, Leksikohuldeblyk, Verslae van konferensies en werksessies.

Bydraes in kategorieë (3)–(7) moet almal aan die eise van akademiese geskrifte voldoen en word met die oog hierop deur die redaksie gekeur.

2. Wetenskaplike standaard en keuringsprosedure

Lexikos is deur die Departement van Hoër Onderwys van die Suid-Afrikaanse Regering as 'n gesubsidieerde, d.w.s. inkomstegenererende navorsingstydskrif goedgekeur. Dit verskyn ook op die *Institute of Science Index* (ISI).

Artikels sal op grond van die volgende aspekte beoordeel word: taal en styl; saaklikheid en verstaanbaarheid; probleemstelling, beredenering en gevolgtrekking; verwysing na die belangrikste en jongste literatuur; wesenlike bydrae tot die spesifieke vakgebied.

Manuskripte word vir publikasie oorweeg met dien verstande dat die redaksie die reg voorbehou om veranderinge aan te bring om die styl en aanbieding in ooreenstemming met die redaksionele beleid te bring. Outeurs moet toesien dat hulle bydraes taalkundig en stilisties geredigeer word voordat dit ingelewer word.

3. Taal van bydraes

Afrikaans, Duits, Engels, Frans of Nederlands.

4. Kopiereg

Nóg die Buro van die WAT nóg die African Association for Lexicography (AFRILEX) aanvaar enige aanspreeklikheid vir eise wat uit meewerkende skrywers se gebruik van materiaal uit ander bronne mag spruit.

Outeursreg op alle materiaal wat in *Lexikos* gepubliseer is, berus by die Direksie van die Woordeboek van die Afrikaanse Taal. Dit staan skrywers egter vry om hulle materiaal elders te gebruik mits *Lexikos* (AFRILEX-reeks) erken word as die oorspronklike publikasiebron.

5. Oorspronklikheid

Slegs oorspronklike werk sal vir opname oorweeg word. Skrywers dra die volle verantwoordelikheid vir die oorspronklikheid en feitelike inhoud van hulle publikasies. Indien van toepassing, moet besonderhede van die oorsprong van die artikel (byvoorbeeld 'n referaat by 'n kongres) verskaf word.

6. Gratis oordrukke en eksemplare

Lexikos is sedert volume 28 slegs elektronies beskikbaar op <http://lexikos.journals.ac.za>. Geen oordrukke of eksemplare is dus beskikbaar nie.

7. Uitnodiging en redaksionele adres

Alle belangstellende skrywers is welkom om bydraes vir opname in *Lexikos* te lewer en verkieslik in elektroniese formaat aan die volgende adres te stuur: lexikos@sun.ac.za, of Die Redakteur: LEXIKOS, Buro van die WAT, Postbus 245, 7599 STELLENBOSCH, Republiek van Suid-Afrika.

B. VOORBEREIDING VAN MANUSKRIP

Die manuskrip van artikels moet aan die volgende redaksionele vereistes voldoen:

1. Lengte en formaat van artikels

Manuskrip moet verkieslik in elektroniese formaat per e-pos of op rekenaarskyf voorgelê word in sagteware wat versoenbaar is met MS Word. Die lettersoort moet verkieslik 10-punt Palatino of Times Roman wees. Bydraes moet verkieslik nie 8 000 woorde oorskry nie.

Elke artikel moet voorsien wees van 'n opsomming van ongeveer 200 woorde en ongeveer 10 sleutelwoorde in die taal waarin dit geskryf is, sowel as 'n opsomming en sleutelwoorde in Engels. Engelse artikels van Suid-Afrikaanse oorsprong moet 'n opsomming en sleutelwoorde in Afrikaans hê, terwyl Engelse artikels van buitelandse oorsprong 'n tweede opsomming en sleutelwoorde in enigeen van die aangeduide tale mag gee. As die outeur dit nie doen nie, sal die redaksie 'n Afrikaanse vertaling voorsien. Maak seker dat die opsomming in die tweede taal ook 'n vertaling van die oorspronklike titel bevat.

2. Grafika

Figure, soos tabelle, grafieke, diagramme en illustrasies, moet in 'n gepaste grootte wees dat dit versoen kan word met die bladspieël van *Lexikos*, naamlik 18 cm hoog by 12 cm breed. Die plasing van grafika binne die teks moet duidelik aangedui word. Indien skryftekens of grafika probleme oplewer, mag 'n uitdruk van die manuskrip of 'n e-pos in .pdf-formaat aangevra word.

3. Bibliografiese gegewens en verwysings binne die teks

Kyk na onlangse nommers van *Lexikos* vir meer inligting. Buiten in spesiale gevalle moet verwysings na *Lexikos*-artikels tot twee of drie per artikel beperk word. Uitsonderings moet met die redakteur van *Lexikos* uitgeklaar word. Dit word gedoen om die status van *Lexikos* in verskeie internasionale indekse te behou.

4. Aantekeninge/voetnote/eindnote

Aantekeninge moet deurlopend in die vorm van boskritef genommer en aan die einde van die manuskrip onder die opskrif **Eindnote** gelys word.

INSTRUCTIONS TO AUTHORS

(For a more detailed version of these instructions, please contact us (lexikos@sun.ac.za)
or refer to our website: <http://lexikos.journals.ac.za/>)

A. EDITORIAL POLICY

1. Type and content of articles

Articles may treat pure lexicography or the implications that related fields such as linguistics, general linguistics, terminology, computer science and management have for lexicography.

Contributions may be classified in any one of the following categories:

- (1) **Articles:** Fundamentally original scientific research done and the results obtained, or existing research results and other facts reflected in an original, synoptic, interpretative, comparative or critically evaluative manner.
- (2) **Review articles:** Research articles presented in the form of a critical review of one or more published scientific sources. Contributions in categories (1) and (2) are subjected to strict anonymous evaluation by independent academic peers in order to ensure the international research quality thereof.
- (3) **Reviews:** An analysis and critical evaluation of published scientific sources and products, such as books and computer software.
- (4) **Projects:** Discussions of lexicographical projects.
- (5) **Lexiconotes:** Any article containing practice-oriented information, suggestions, problems, questions, commentary and solutions regarding lexicography.
- (6) **Lexicovaria:** Any of a large variety of articles containing announcements and press releases by lexicographic societies which are of particular value to the practising lexicographer.
- (7) **Other:** From time to time other categories may be inserted by the editors, such as Lexicosoftware, Lexicosurvey, Lexicobibliography, Lexiconews, Lexicofocus, Lexicohonour, Lexicotribute, Reports on conferences and workshops.

Contributions in categories (3)-(7) must all meet the requirements of academic writing and are evaluated by the editors with this in mind.

2. Academic standard and evaluation procedure

The Department of Higher Education of the South African Government has approved *Lexikos* as a subsidized, i.e. income-generating research journal. It is also included in the *Institute of Science Index* (ISI).

Articles will be evaluated on the following aspects: language and style; conciseness and comprehensibility; problem formulation, reasoning and conclusion; references to the most important and most recent literature; substantial contribution to the specific discipline.

Manuscripts are considered for publication on the understanding that the editors reserve the right to effect changes to the style and presentation in conformance with editorial policy. Authors are responsible for the linguistic and stylistic editing of their contributions prior their submission.

3. Language of contributions

Afrikaans, Dutch, English, French or German.

4. Copyright

Neither the Bureau of the WAT nor the African Association for Lexicography (AFRILEX) accepts any responsibility for claims which may arise from contributing authors' use of material from other sources.

Copyright of all material published in *Lexikos* will be vested in the Board of Directors of the Woordboek van die Afrikaanse Taal. Authors are free, however, to use their

material elsewhere provided that *Lexikos* (AFRILEX Series) is acknowledged as the original publication source.

5. Originality

Only original contributions will be considered for publication. Authors bear full responsibility for the originality and factual content of their contributions. If applicable, details about the origin of the article (e.g. paper read at a conference) should be supplied.

6. Free offprints and copies

Lexikos is only available electronically on <http://lexikos.journals.ac.za> from volume 28 onward. No offprints or copies are available.

7. Invitation and editorial address

All interested authors are invited to submit contributions, preferably in electronic format, for publication in *Lexikos* to: lexikos@sun.ac.za, or

The Editor: LEXIKOS
Bureau of the WAT
P.O. Box 245
7599 STELLENBOSCH, Republic of South Africa

B. PREPARATION OF MANUSCRIPTS

Manuscripts of articles must meet the following editorial requirements:

1. Format and length of articles

Manuscript should preferably be submitted in electronic format by email or on a disk, in software compatible with MS Word. The typeface used should preferably be 10-point Palatino or Times Roman. Contributions should not exceed **8 000 words**.

Each article must be accompanied by **abstracts** of approximately 200 words and approximately 10 **keywords** in the language in which it is written, as well as in **English**. English articles of South African origin should carry an abstract and keywords in Afrikaans, whilst English articles of foreign origin should carry a second abstract and keywords in any of the other languages mentioned. In cases where this is not done, the editors will provide an Afrikaans version. Ensure that the abstract in the second language also contains a **translation of the original title**.

2. Graphics

Figures such as tables, graphs, diagrams and illustrations should be in an appropriate size to be well accommodated within the page size of *Lexikos*, namely 18 cm high by 12 cm wide. The locations of figures within the text must be clearly indicated. If orthographic marks or graphics used in the text prove problematic, a printout of the manuscript or an email in .pdf format may be requested.

3. Bibliographical details and references in the text

Examine recent issues of *Lexikos* for details. Self-references to *Lexikos* should be limited to two or three per article, except in exceptional circumstances. Exceptions should be cleared with the editor of *Lexikos*. This is done to preserve the status of *Lexikos* in various international indices.

4. Notes/footnotes/endnotes

Notes must be numbered consecutively by superscript numbers and grouped together at the end of the manuscript under the heading **Endnotes**.