https://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2050 (Article)

# Endemann's *Wörterbuch der Sotho Sprache* (1911): A Worthy Candidate for Digitisation

Elsabé Taljard, *Department of African Languages, University of Pretoria, Pretoria, South Africa (elsabe.taljard@up.ac.za) (https://orcid.org/0000-0002-4507-1633)*

Gertrud Faaβ, *Institute for Information Science and Language Technology, University of Hildesheim, Hildesheim, Germany (gertrud.faass@uni-hildesheim.de) (https://orcid.org/0000-0002-8130-617X)*

Danie Prinsloo, *Department of African Languages, University of Pretoria, Pretoria, South Africa (danie.prinsloo@up.ac.za) (https://orcid.org/0000-0003-0054-4676)*
and
Sonja Bosch, *Department of African Languages, UNISA, Pretoria, South Africa (seb@hbosch.com) (https://orcid.org/0000-0002-9800-5971)*

**Abstract:** This article re-evaluates *Wörterbuch der Sotho Sprache*, a historically significant, yet neglected Sotho–German dictionary, published in 1911 by Berlin missionary Karl Endemann. Its marginalisation stems from its choice of German as target language, outdated orthography, missionary orientation, and deviation from modern lexicographic principles. Rather than a conventional comparison with modern Sepedi dictionaries, this study positions Endemann's work within its historical and cultural context. Key lexicographic elements such as grammatical formatives, alphabetical categories, high-frequency lemmas, semantically related paradigms, and culturally significant entries are analysed in detail. The findings often reveal strengths that match or even surpass those of later Sepedi dictionaries. Despite its value, user access remains limited due to linguistic complexity and unavailability. With digitisation now permitted by the publisher, this study outlines a multi-phase strategy to enhance usability, including the use of OCR4all, an open-source tool for text recognition. While the digitisation process is not without challenges, the application of OCR4all has yielded impressive accuracy. However, despite the high-quality output, this margin of error necessitates manual verification to ensure the integrity of the digitised content as a reliable and accessible resource for modern users.

**Keywords:** ENDEMANN, SOTHO–GERMAN, CULTURAL HERITAGE, GRAMMATICAL FORMATIVES, HISTORICAL CONTEXT, LEXICOGRAPHIC PRINCIPLES, DIGITISATION, OCR4ALL, MODERN DAY USERS, ACCESSIBILITY

*Lexikos* 35 (AFRILEX-reeks/series 35: 2025): 462-481

https://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2050 (Article)



**Opsomming: Endemann se *Wörterbuch der Sotho Sprache* (1911): 'n waardige kandidaat vir digitalisering.** Hierdie artikel herevalueer *Wörterbuch der Sotho Sprache*, 'n histories betekenisvolle dog verwaarloosde Sotho–Duitse woordeboek, wat in 1911 deur die Berlynse sendeling Karl Endemann gepubliseer is. Die marginalisering daarvan spruit uit sy keuse van Duits as doeltaal, verouderde ortografie, sendingoriëntasie en afwyking van moderne leksikografiese beginsels. Eerder as 'n konvensionele vergelyking met moderne Sepedi-woordeboeke, plaas hierdie studie Endemann se werk binne sy historiese en kulturele konteks. Sleutelleksikografiese elemente soos die hantering van grammatikale formatiewe, alfabetiese kategorieë, hoëfrekwensielemmas, semanties-verwante paradigmas en kultureel betekenisvolle inskrywings word in detail ontleed. Die bevindinge openbaar dikwels sterkpunte wat ooreenstem met of selfs dié van latere Sepedi-woordeboeke oortref. Ten spyte van die waarde daarvan, bly gebruikerstoegang beperk weens linguistiese kompleksiteit en onbeskikbaarheid. Met digitisering wat nou deur die uitgewer toegelaat word, sit hierdie studie 'n multifase strategie uiteen om bruikbaarheid te verbeter, insluitend die gebruik van OCR4all, oopbronsagteware vir teksherkenning. Alhoewel die digitiseringsproses nie sonder uitdagings is nie, het die toepassing van OCR4all indrukwekkende akkuraatheid opgelewer. Ten spyte van die hoë kwaliteit uitset, noodsaak 'n foutmarge egter handmatige verifikasie om die integriteit van die gedigitiseerde inhoud te verseker as 'n betroubare en toeganklike hulpbron vir moderne gebruikers.

**Sleutelwoorde:** ENDEMANN, SOTHO–DUITS, KULTURELE ERFENIS, GRAMMATIKALE FORMATIEWE, HISTORIESE KONTEKS, LEKSIKOGRAFIESE BEGINSELS, DIGITISERING, OCR4ALL, MODERNE GEBRUIKERS, TOEGANKLIKHEID

## Introduction

In 1911, Karl Heinrich Julius Endemann, a Berlin missionary, published his dictionary of the Sotho language *Wörterbuch der Sotho Sprache*, 1911. Very little information is available regarding the actual process of data collection — in the *Vorwort* to the dictionary, Endemann (1911: VII) indicates that he collected the data in the period 1861–1873 while being active as a missionary in the then Transvaal, specifically on the missionary stations Gerlachshoop, Phatametsane, Botschabelo and Malokong. It can only be assumed that he made use of oral accounts as sources of data, since no textual material was available at that stage. Shrouded in mystery, however, is the source of his data on Setswana and Sesotho, which the dictionary also caters for. There is no evidence that speakers of these languages were present in the area in which he worked. Kosch (2012: 226) mentions that the dictionary has since its publication remained at the fringes of scholarly investigation — in their discussion of Bantu language dictionaries utilising a left-expanded article structure, Gouws and Prinsloo (2005) make no mention of Endemann's dictionary, and apart from two articles focusing on the dictionary by Kosch (2011, 2012) and references to the *Wörterbuch* in Kosch (1993) and Lombard (1970), there is no serious engagement with Endemann's *Wörterbuch* as an exceptional lexicographic endeavour. One reason is probably because the target language is German, and it is therefore not readily accessible to scholars

https://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2050 (Article)



of the African languages. It was also compiled with a specific function in mind, i.e., to assist German missionaries in their missionary field work — a function that may be difficult to reconcile with the information needs of the modern-day user of a Sepedi dictionary. The dictionary furthermore places a high demand on the dictionary using skills and grammatical knowledge of the user, and it makes use of an orthography that is long since obsolete. Nevertheless, we agree with Kosch (2011) that this dictionary contains a "huge repository of information that awaits linguistic and cultural scrutiny and insights", and to which we would like to add, a lexicographic reappraisal. In our reappraisal, we take our cue from the following quote from Tarp (2013: 294): "[H]istory should not be viewed as an ever-growing progress, but as a process with its ups and downs. In this respect, some old lexicographical works … are in some aspects extremely advanced even compared with present-day dictionaries."

The aim of this article is therefore to reappraise the *Wörterbuch* within the context of existing bilingual Sepedi dictionaries, rather than to evaluate it against modern day lexicographic theory and practice. Based on our findings, we conclude that the dictionary warrants wider exposure, something that can only be attained by making it — or at least parts of it — available in digital format. Digitisation is also an important step in the preservation of this canonical work.

This article consists of three parts: an investigation of selected aspects of the macro- and microstructures of Endemann (1911), followed by a discussion of the (in)accessibility of the dictionary to modern day target users. In the third section of the article, the digitisation strategy and the necessary steps for making the dictionary available and useful for modern day users are described. We envisage a multiphase digitisation approach with different levels of accessibility to the dictionary, and different levels of true electronic features.

## Contextualisation

In her article *Innovation and compromise in K. Endemann's dictionary of the Sotho language* Kosch (2011) concentrated on shortcomings and weaknesses of the dictionary, indicating that Endemann violated some good lexicographic principles in the compilation of the dictionary. In our reappraisal of the dictionary within the context of other bilingual Sepedi dictionaries, we address the following issues: treatment of grammatical formatives, over and/or under treatment of alphabetical categories, treatment of high frequency lemmas, completing semantically related paradigms and treatment of lemmas with cultural significance. The dictionaries that have been selected to provide a contextualised assessment of Endemann's dictionary cover a timespan of almost five decades, with publication dates ranging from 1967 to 2015. These dictionaries have been, and still are, regarded as standard Sepedi reference works, and we believe that comparing Endemann's dictionary with these dictionaries will give a more balanced view of the dictionary as a lexicographic artefact and its value as a utility instrument. The dictionaries selected for this investigation are the following:

https://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2050 (Article)



De Schryver, G.-M. 2007. *Oxford Bilingual School Dictionary: Northern Sotho and English*, henceforth OBSD;

Kriel, T.J. 1967. *The New English–Northern Sotho Dictionary*, henceforth NENSD;

Kriel, T.J. 1983. *Pukuntšu woordeboek, Noord-Sotho–Afrikaans, Afrikaans–Noord-Sotho*, henceforth PWNSA83;

Kriel, T.J., E.B. van Wyk and S.A. Makopo. 1989. *Pukuntšu woordeboek, Noord-Sotho–Afrikaans, Afrikaans–Noord-Sotho*, henceforth PWNSA89;

Mojela, V.M., M.C. Mphahlele, M.R. Selokela and W.M. Mojapelo. 2015. *Sesotho sa Leboa–English Bilingual Dictionary*, henceforth SsLEBD, and

Ziervogel, D. and P.C. Mokgokong. 1975. *Groot Noord-Sotho Woordeboek*, henceforth GNSW.

**Treatment of grammatical formatives**

Grammatical formatives are usually notoriously undertreated in bilingual dictionaries in which the source language is an African language, since these formatives are typically not carriers of lexical meaning. It is therefore difficult to provide a translation equivalent in a target language which does not distinguish the same rich morphological structure.

In Tables 1 and 2, a summary of the treatment of two grammatical formatives, i.e., *a* and *sa* in six Sepedi paper dictionaries is provided. These two formatives have a high degree of homography and therefore represent a variety of different grammatical functions. Being a grammatical formative, the equivalent of which does not exist in either Afrikaans or English, it follows that instead of a translation equivalent, a paraphrase or a full description, akin to a definition, is the only option available to the lexicographer to provide the user with the kind of information that will ensure a successful lookup. Rather than refer in this regard to sense distinctions (since grammatical formatives generally do not express a 'sense' in the usual sense of the term), the focus in this discussion is on the identification of the various grammatical functions carried by *a* and *sa* in the various dictionaries.

In Table 1, the different functions of *a* as distinguished by various Sepedi paper dictionaries are listed. Where the tick mark is in brackets, it indicates that the specific function is not identified as such, but treated by means of an example.

As is evident from Table 1, Endemann succeeds in identifying most of the grammatical functions of the formative *a*, as distinguished in modern grammars and collectively treated in the other dictionaries. When compared to the other dictionaries, the only obvious omissions are the past tense function and possibly the function as an ideophone. The function of *a* as formative of the consecutive is not mentioned, but neither does any of the other dictionaries refer to this function, except Ziervogel and Mokgokong (1975: 1). The same pertains to *a* as relative pronoun of class 1 and a phonetic description of *a*. Two

https://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2050 (Article)



functions that are only treated by Endemann are those of *a* as verbal ending and as copula. The reference to *a* as negative particle is however rather dubious: it is described as a negative particle in the negation of the expression *a-e!* 'No!' (pronounced in a staccato manner) (Endemann 1911: 39).

| Treatment of the formative *-a* | | | | | | |
|---|---|---|---|---|---|---|
| | **Endemann 1911** | **De Schryver 2007** | **Kriel 1967** | **Kriel 1983** | **Kriel et al. 1989** | **Ziervogel & Mokgokong 1975** |
| s.c. class 1 | √ | √ | x | √ | √ | √ |
| s.c. class 6 | √ | √ | √ | x | √ | √ |
| o.c. class 6 | √ | √ | x | x | √ | x |
| p.c. class 6 | (√) | √ | x | √ | √ | √ |
| demonstrative class 6 | √ | √ | √ | √ | √ | √ |
| present tense morpheme | √ | √ | x | √ | √ | √ |
| question particle | √ | √ | x | x | √ | √ |
| hortative particle | √ | √ | x | √ | √ | √ |
| past tense morpheme | x | √ | x | x | √ | x |
| exclamation / interjection | √ | x | √ | √ | √ | √ |
| ideophone | x | x | √ | √ | x | (√) |
| relative pronoun class 1 | x | x | x | x | x | √ |
| relative pronoun class 6 | √ | x | x | √ | x | √ |
| possessive formative | √ | x | x | x | x | √ |
| phonetic description | x | x | x | x | x | √ |
| formative of consecutive | x | x | x | x | x | √ |
| -a as verbal ending | √ | x | x | x | x | x |
| copula | √ | x | x | x | x | x |
| negative morpheme | √ | x | x | x | x | x |

**Table 1:** Treatment of the grammatical formative a across six paper dictionaries

The treatment of the formative *a* far exceeds its treatment in the other dictionaries, not only in terms of the distinction between the different functions of this formative, but also in terms of dictionary space allocated to the treatment of *a* as lemma. Each page in this dictionary contains two columns, and three and a quarter column, spread over two pages, are dedicated to the treatment of *-a*. The only comparable treatment is that of Ziervogel and Mokgokong (1975), who use roughly one and a half column for their treatment of *a*. The encyclopaedic information that is provided by Endemann is extremely detailed, with copious examples being provided, cf. in this regard the description of *-a* as possessive formative (translated from the original German; the examples have been transcribed to follow the standard orthography):

> Genitive particle, more correctly possessive particle, since in Sotho, and in the Bantu languages in general, the particular relationship is always a possessive relationship and indicates belonging, for which reason Sotho and Bantu generally do not form a genitive object (Genetivus objectivus); where it (the term) is currently used, it is a language corruption of which Europeans are guilty which must be rooted out — The subject pronoun of the logical governing noun ("regens") is prefixed to the possessive particle. This situation leads to the possessive particle being identical to the verbal root *a* and indeed in it lies the idea of direction towards

https://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2050 (Article)



the following governed noun ("rectum"). Its high tone correlates with the emphasis of the connection with the governed noun. Here, the prefixed pronoun acquires a relative character — The following may serve as examples of the possessive construction according to the different noun classes:

*mohlanka wa morena* "servant of the Lord"
*batho ba (= ba-a) kgoši* "people of the chief"
*modumo wa lewatle* "roaring of the sea"
*melao ya mmuši* "commandments of the ruler"
*botho bja (= bo-a) tate* "humanity of (i.e., my) father"
*leina la (= le-a) ngwana* "name of the child"
*mala a kgomo (= a-a)* "intestines of the cow"
*seedi sa (= se-a) letšatši* "light of the sun"
*nku ya (= e-a) mohumi* "sheep of the rich person"
*diphoofolo tša (= di-a) naga* "wild animals of the veld"

(Endemann 1911: 39)

The second formative *sa* was selected because it appears towards the end of the alphabet. Endemann's dictionary was compiled according to the so-called traditional method, which means that compilation was based on the lexicographer's intuition without proper attention being paid to lexicographic planning and the formulation of a strategy for inclusion and/or exclusions of lemmas, cf. Kosch (2011: 112-113) and Prinsloo and De Schryver (2007: 179). Entering lemmas as the lexicographer comes across them leads to a number of macrostructural inconsistencies, such as over-treating the initial sections or alphabetical categories and under-treating categories dealing with the letters appearing towards the end of the alphabet (Prinsloo and De Schryver 2007: 186). Apart from checking the extent of the treatment of *sa* as grammatical formative, focusing on *sa* can therefore also provide initial insight to which extent Endemann succumbed to lexicographer's fatigue when compared to other dictionaries compiled using the same compilation strategy. Compare the table below, and once again, tick marks appearing in brackets indicate that the specific function is not identified, but treated by means of an example:

| Treatment of the formative -*sa* | | | | | | |
|---|---|---|---|---|---|---|
| | **Endemann 1911** | **De Schryver 2007** | **Kriel 1967** | **Kriel 1983** | **Kriel et al. 1989** | **Ziervogel & Mokgokong 1975** |
| s.c. class 7 | x | √ | x | √ | x | √ |
| p.c. class 7 | x | √ | √ | √ | √ | √ |
| aspectual prefix | √ | √ | √ | √ | √ | √ |
| negative morpheme | √ | √ | x | (√) | √ | √ |
| verb stem | √ | √ | √ | √ | √ | √ |
| (adverb) | x | x | √ | x | x | x |
| (formative of past tense) | x | x | x | √ | x | x |
| ideophone / interjection | x | x | x | √ | √ | √ |

**Table 2:** Treatment of the grammatical formative sa across six paper dictionaries

https://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2050 (Article)



The first obvious seeming omission is the non-treatment of *sa* as (consecutive) subject concord of class 7. This is an unfortunate omission, since this function is probably the most prominent one of this formative. Although Endemann (1911: 456) does not distinguish *sa* as possessive particle, it does figure in his treatment of *-a* as base for possessive concords, where an illustrative example is also provided — see the excerpt above. It is however highly unlikely that a user will look up the lemma *-a* when looking for the meaning of *sa*. He does not treat *sa* as ideophone, but neither do De Schryver (2007) or Kriel (1967). Kriel's (1967: 355) distinction of *sa* as adverb is debatable, as is Kriel's (1983: 267) distinction of *sa* as a formative of the past tense. As a whole, Endemann's treatment of the formative *sa* compares well with that of the other dictionaries. There does not seem to be any evidence of lexicographer's fatigue in his treatment of this particular formative — this preliminary conclusion is further explored in the next paragraph.

**Over and/or under treatment of alphabetical categories**

In order to test the preliminary conclusion regarding over and/or under treatment of alphabetical categories in Endemann's dictionary further, a random selection of pages from both Endemann (1911) and Ziervogel and Mokgokong (1975) was made, and the number of lemmas treated on each page was recorded. The results are reflected in Figure 1:

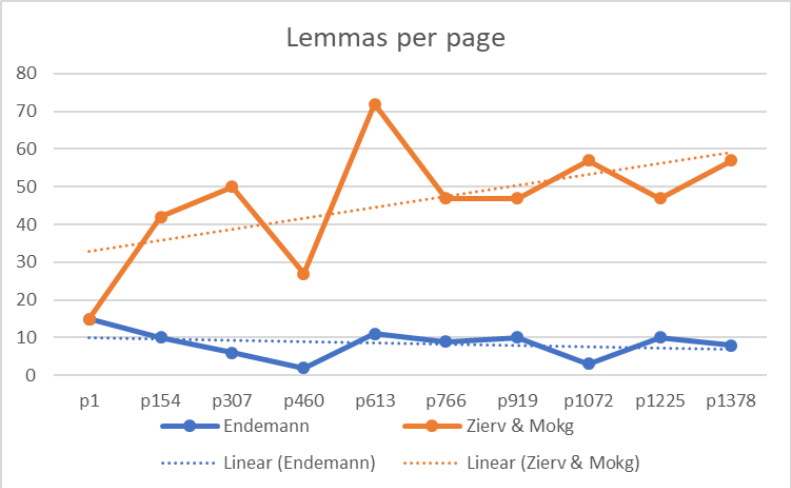


**Figure 1:** Lemmas per page in Endemann (1911) and Ziervogel and Mokgokong (1975)

As can be seen from Figure 1, the number of lemmas per page in Endemann's dictionary stays relatively stable as one moves through the alphabetical catego-

https://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2050 (Article)



ries. The number of lemmas treated in Ziervogel and Mokgokong however, follows a steadily increasing trajectory. The stability in the number of lemmas per page in Endemann (1911) is in sharp contrast with Prinsloo and De Schryver's (2007: 186) findings with regard to, for example, Kriel's (1983) dictionary. They indicate that the number of articles on page 2 of the dictionary is 22, whereas this number is 75 on page 281, thus indicating an extremely reduced treatment of lemmas as the lexicographer moved towards the end of the alphabet.

The results obtained from this experiment are quite remarkable, considering the resources that were at the lexicographer's disposal during compilation of the dictionary, and also, the fact that he was simultaneously learning the language as he was collecting the data. Endemann collected the data used in his dictionary during the 12 years that he spent in South Africa — from 1861 to 1873 — doing missionary work in Sekhukhuniland and Botshabelo near Middelburg (Kosch 2011: 112). To say that he had no access to the electronic and technological resources available to modern-day lexicographers would be stating the obvious. Furthermore, Kosch (2011: 113) points out that the collection of the data was not carried out with the express purpose of compiling a dictionary, that Endemann had no formal lexicographic training, and that his data was probably mostly based on orally collected data, since very few written texts were available at that stage. Even so — as is clear from Figure 1 above — he managed to avoid one of the pitfalls to which even modern-day lexicographers are no strangers, cf. Prinsloo and De Schryver (2002) for the over treatment of the alphabetical category **K** in the WAT.

**Treatment of high frequency lemmas**

Inclusion of lemmas based on their frequency of use is a feature of modern corpus-based lexicography. When the lemma lists of corpus-based dictionaries are compared to frequency lists drawn from a corpus, these dictionaries usually fall seriously short, in that high frequency lemmas are not included. Compare De Schryver and Prinsloo (2000: 294-297) for a detailed discussion of macrostructural deficiencies regarding the omission of high frequency lemmas in selected Sepedi dictionaries. Including lemmas based on frequency considerations rests on the assumption that users will look up high frequency words, an assumption which has been substantiated by De Schryver and Joffe (2004: 190). Including high frequency items therefore increases the probability that dictionary users' look-up needs are met. For the purpose of this discussion, a frequency list was extracted from a 7.5 million-word Sepedi corpus. The assumption was that a comparison of the topmost frequent items on the frequency list with the lemma list of the dictionary would not deliver significant results, since (a) many of the top frequency items are homographic grammatical formatives, and (b) being so frequent, it is highly unlikely that these items would not cross the lexicographer's path. The probability of entering these lemmas into the dictionary is therefore high, even when the lemma list is compiled based on the intuition of the lexicographer. This assumption is borne out by the results of a small experiment in

https://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2050 (Article)



which the top 100 nouns and verbs were extracted from the frequency list. On this list 67 items are nouns, 33 are verbs. Since the frequency list is a raw one, i.e., an unlemmatised list, it also contains plural nouns. For the purpose of the experiment, plural nouns were counted as being treated if they appear in the article of the corresponding singular noun. Of the 67 nouns, only four found on the frequency list are not treated in Endemann's dictionary, i.e., *baithuti* 'learners' (frequency rank 150), *Afrika* 'Africa' (frequency rank 188), *sekolo* 'school' (frequency rank 199) and *tšhelete* 'money' (frequency rank 202). Of the 33 verbs, only *-rile* 'said' is not treated, although it is referred to in the encyclopaedic information provided for the lemma *-re* 'say'.

Consequently, the top 100 derived verb stems were manually isolated from the frequency list. The lowest frequency rank included in the list is 1034, thus drilling down much deeper than isolating the first 100 most frequent nouns and verbs. Since the corpus is a raw corpus, i.e., not annotated with part-of-speech, only items that are non-ambiguous were included in the list. Three categories of treatment are distinguished: items that are treated, items that are not treated, but are referred to within the article by means of an example or as part of the grammatical information provided, and items that are not treated in the dictionary at all. Compare Figure 2 in this regard:

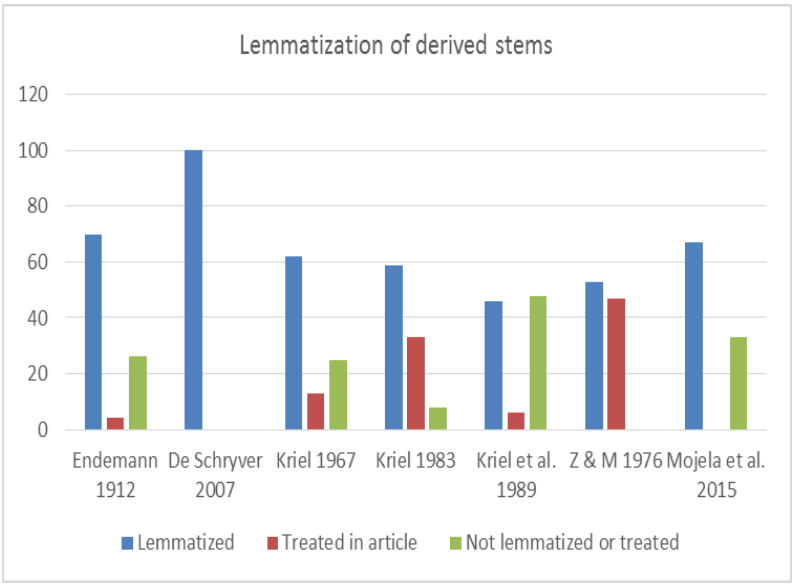


**Figure 2:** Treatment of derived verbs stems across seven dictionaries

From Figure 2 above, it is clear that Endemann's dictionary outperforms five of the other dictionaries with regard to the treatment of high frequency items. The full score obtained by De Schryver's dictionary is due to the fact that this is a

https://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2050 (Article)



corpus-based dictionary, and the selected derived verb stems all made the cut-off point, frequency wise, for inclusion in the dictionary. In terms of lemmatisation of high frequency items, the dictionary of Mojela et al. (2015) seems to compare well with De Schryver (2007) and Endemann (1911). Endemann (1911) treats 70 of the most frequent items; Mojela et al. (2015) treat 67; however, the actual treatment of the lemmas in these two dictionaries is vastly different, and this is mostly due to the over-simplified microstructure of the dictionary of Mojela et al. (2015). For the lemma *-makatša* 'surprise' for example, the latter provides the following information: part of speech, which is incorrectly indicated as a noun, and a translation equivalent 'surprise', incorrectly spelled as 'suprise'. Endemann provides grammatical information by indicating that it is the causative form of the base form *-makala*, with 'amaze' (*in Erstaunen setzen*) as translation equivalent. With regard to the lemma *-amogela*, Mojela et al. once again provide the part of speech, and also three translation equivalents, i.e., 'receive' (incorrectly spelled as 'receice'), 'welcome' and 'accept'. Endemann again provides grammatical information by indicating that it is an applicative form of the base form *amoga*, followed by four translation equivalents. i.e., 'relieve someone of a load', 'receive hospitably', 'catch (something which is thrown)' and 'echo', followed by a usage example and its idiomatic translation, followed by a literal translation of the example. Generally speaking, Endemann's dictionary therefore compares very well with existing Sepedi dictionaries with regard to treatment of high frequency items.

**Completion of semantically related paradigms / lexical sets**

Rundell (2015: 302) points out that it is common practice for dictionaries to cover all members of any clearly defined set of lexical items, citing days of the week and signs of the Zodiac as examples of such lexical sets. We are aware of the questions raised by Swanepoel (2010: 429 et seq.) with regard to the identification of such sets, the first being how they are to be identified. Swanepoel's discussion is aimed at the identification and utilisation of lexical sets in the crafting of better definitions, but since this is not the focus of our investigation, we follow Atkins and Rundell (2008: 123) who define a lexical set as "groups of words that share a common element of meaning such as days of the week, or months of the year, or birds, trees, flowers, and metals." By implication, lexical sets are based on numerous sense relations such as synonymy and hyponymy, or the fact that they belong to the same semantic domain or field, see Swanepoel (2010: 429). The days of the week is a prototypical example of a lexical set, therefore we used that as a first topic to ascertain to what extent Endemann was aware of the standard lexicographic practice of completing semantically related paradigms. Once again, we used the selected Sepedi dictionaries listed above as a benchmark. The outcome of this first small investigation was rather surprising. All dictionaries lemmatise the names of all seven weekdays, except Endemann (1911). None of the weekdays appear in his dictionary. However, the time frame dur-

https://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2050 (Article)



ing which his data collection took place needs to be considered — it could be that the Western concept of a week as a time unit consisting of seven days had at that stage not yet been conceptually established in the Sepedi-speaking community. It is also significant that the Sepedi word for 'week' does not appear in the dictionary. As Kosch (2011: 114) remarks: "the dictionary was compiled … to capture the uncorrupted forms of the language from a time before contact with Europeans". Furthermore, it is unlikely that the names of the weekdays had at the time of the compilation of the dictionary been standardised, considering that the first official attempt at standardising the orthography, terminology and spelling rules of Sepedi was only undertaken in 1930 (Kosch 1993: 23).

A second lexical set, i.e., the four seasons was consequently investigated, also with interesting results. The four official terms as provided in the Terminology and Orthography no 4 of 1988 are *marega* 'winter', *seruthwana* 'spring', *selemo* 'summer' and *lehlabula* 'autumn'. Apart from Endemann (1911), all six the other dictionaries treat these four terms; Endemann only treats three of them and provides the following translation equivalents: *marega* 'winter', *selemo, lehlofo* 'spring, early summer' and *lehlabula, hwetla* 'late summer, autumn'.[1] There are however, considerable differences with regard to the translation equivalents provided by the other dictionaries. The only term that all dictionaries agree on is the one for 'winter' Kriel (1967), Kriel (1983) and Kriel et al. (1989) provide 'autumn' as a translation equivalent of *seruthwana*, whereas the other dictionaries give 'spring' as an equivalent. There seems to be considerable overlap between *selemo* and *seruthwana*, with both having 'spring', 'early summer' and 'ploughing time' (the latter in the case of Kriel et al. (1989) and Ziervogel and Mokgokong (1975)) as translation equivalents. It may very well be that the distinction of four seasons is a case of cultural projection, where the Western conceptualisation of four seasons is projected onto the Sepedi linguistic system, leading to a mismatch between what could be described as an indigenous knowledge system and the treatment of its members in modern Sepedi dictionaries. In indigenous knowledge systems, seasons are based on weather patterns, and in the case of agriculture-based cultures on agricultural activities in a given area, and therefore do not necessarily correspond to the Western, meteorologically based notion of four seasons. In Nyungar, an aboriginal language spoken in Southwest Australia, 6 seasons are distinguished: season of the young *birak* (dry and hot, burning time), of adolescence *bunuru* (hottest part of the year), of adulthood *djeran* (cooler weather begins), of fertility *makuru* (coldest and wettest season of the year), of conception *djilba* (mixture of wet days with increasing number of clear, cold nights and pleasant warm days) and of birth *kambarang* (longer dry periods) (http://www.bom.gov.au/iwk/calendars/nyoongar.shtml); the ancient Hindu calendar also distinguishes six seasons: spring, summer, monsoon, autumn, pre-winter and winter (https://www.javatpoint.com/seasons-in-india), and in countries close to the equator, a distinction is made between the wet season and the dry season. The possibility that Endemann's treatment of terms referring to only three seasons (winter, spring and early summer, and late summer and autumn) could in actual fact be a better reflection of the African conceptual

https://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2050 (Article)



system regarding seasonal distinctions should therefore not be summarily dismissed. As a matter of interest, Colenso's *Zulu–English Dictionary* (1905), compiled around the same time as Endemann's, only includes entries for summer (*ihlobo*) and winter (*ubusika*). However, under the entry for *ihlobo* (summer), the season is further subdivided into five distinct parts. Like contemporary Sepedi dictionaries, modern Zulu dictionaries also adopt the Western concept of four distinct seasons.

A third lexical set that was perused was that of body parts, specifically facial features. There is no fixed definition as to what exactly constitutes facial features, so for the purpose of this study we selected the following: *sefahlogo/sefahlego* 'face', *phatla* 'forehead', *leihlo* 'eye', *nko* 'nose', *molomo* 'mouth', *thama/lerama* 'cheek', *ntšhi* 'eyelashes, eyebrows' and *seledu* 'chin'. All dictionaries, except De Schryver (2007), treat all these items; De Schryver misses out on the last three.[2] This is again due to the fact that his dictionary is frequency-based; however, including these lemmas even though they did not make the cut-off point frequency-wise, should have been considered. Once again, Endemann holds his own when compared to other Sepedi dictionaries with regard to the completion of lexical sets, despite the lacunae discussed above.

**Treatment of culturally significant lemmas**

Endemann's (1911) treatment of lemmas is characterised by an extremely rich sense distinction and detailed definitions of lemmas, the latter especially with regard to culturally-bound lemmas. Apart from providing a translation equivalent, an extended paraphrase of meaning is provided in some cases of lemmas where a cultural element is present. Since Ziervogel and Mokgokong (1975) is the only other dictionary that is more or less comparable to Endemann (1911) with regard to comprehensiveness, the treatment of the lemmas as provided in the former is provided by way of comparison. A few illustrative examples will suffice. (The examples from Endemann (1911) have been translated from the original German and transcribed to follow the standard orthography; only translation equivalents and paraphrases of meaning are provided):

1. *lesiba*

   name of a musical instrument; it is a flat bow strung with a string, at one end a long quill is attached, this end of the bow is put in the mouth and the sound is produced by means of the quill, by inhaling and exhaling through the mouth; one hears two tones (Endemann 1911: 473).

   kind of musical instrument on which one blows (Ziervogel and Mokgokong 1975: 1177).

2. *mongala*

   a protester, someone who resists laws and customs, an uncircumcised person (who resists circumcision), a deserter (from circumcision), who is forever

https://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2050 (Article)



banned (as a result of his resistance against customs), an instigator (Endemann 1911: 339).

Not treated in Ziervogel and Mokgokong (1975).

3. *sedimo*

oracle; ghostly being that can be heard, but not seen. An animal that is slaughtered so that its entrails can be used for divination, is also called thus (Endemann 1911: 273).

offering, sacrifice (to spirits) (Ziervogel and Mokgokong 1975: 156).

4. *sekgapa*

name of a musical instrument, consisting of and bow and string, the resonance is created by means of a calabash that is fastened in the middle of the bow, on which the string rests, and it gives two tones (Endemann 1911: 209).

calabash, stringed musical instrument, † bass. (Ziervogel and Mokgokong 1975: 604).

5. *tšilo*

millstone; (i.e., a fist stone, a hewn, fist-sized, rounded stone, with which grain is ground on a grinding stone) (Endemann 1911: 595).

grinding stone, mill (Ziervogel and Mokgokong 1975: 1218).

Even though not extensive, these examples do give a sense of the cultural richness to be found in Endemann (1911), making the dictionary a valuable cultural artefact, and as such, worth not only of preserving, but also of making it accessible for modern day users.

**Inaccessibility of Endemann (1911)**

Despite comparing favourably with existing Sepedi dictionaries, Endemann (1911) is not accessible to modern day users. There are several reasons for this, the least not being that it has been out of hard-copy print for more than a century. The publication rights currently reside with De Gruyter Mouton (Verlag) and an e-version of the dictionary was published in 2012 and is available at a cost of €229[3], which makes it quite prohibitively expensive for South African users. The format of the dictionary is an unsearchable pdf-version, which is not particularly useful. A second reason for the inaccessibility of the dictionary is the language pair (Northern) Sotho–German, which is a rather unusual one, and it narrows the pool of potential target users substantially. The biggest stumbling block for the modern-day dictionary user is however the orthography used by Endemann and, as a result, the unusual ordering of alphabetical categories. In the introduction of the dictionary (Endemann 1911: 1-8), a detailed description of the orthography is provided in which the use of specific orthographic sym-

https://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2050 (Article)



bols to represent certain sounds is explained, but this information is provided in German only and requires an in-depth knowledge of phonetics. Without having read the introduction, the hapless user would seek in vain for the word *-bala* in an alphabetical stretch B following A; the bilabial fricative [β] is represented by the symbol v, and is therefore to be found between the alphabetical stretches U and X (the semi-vowel [w] is represented as ŏa and is to be found under O), where X is used to represent the velar fricative [ɣ], in the modern orthography represented by g. These challenges are exacerbated by the fact that no index appears in the front matter. Lastly, the left-expanded article structure is problematic, especially from a user-perspective, and in the case of disjunctively written languages, more so in the case of nouns than of verbs. In a left-expanded lemmatisation procedure, the stem is the alphabetical point of reference, similar to a purely stem-based lemmatisation procedure. The main difference between left-expanded and stem-based lemmatisation strategies is that in the case of the former, the prefixal element is included as part of the lemma sign, albeit in a slot preceding the stem. Compare the excerpt from the alphabetical stretch E in Figure 3 by way of illustration:

*le-ėma*, Pl. *ma-*, „Horde, Stamm, Parteilichkeit (S.-S.); Zaun vom *kχoro* (N.)“ Der Sinn des W. ist der eines Dinges, das individuell (für sich) Stellung einnimmt.

*se-ėma*, Pl *li-*, „Einfriedigung, Dornzaun vom *kχoro*, Pfahlhütte mit offenen Wänden (Laube) im *kχoro;* Spruch, Sprichwort (also „stehende Redensart“), Wort (einzelnes), Meinung; Versammlung von stehenden Menschen (S.-S.)“. *χo sava ligma* „witzeln“, *χo sava ligma kà* „bewitzeln“.

*se-ėma* (S.-S.) „Dickleibigkeit, Wassersucht“. Vergl. *ima.*

*ėmàėma*, Iterat. v. *ema*, „immer stehen bleiben, zögern, säumen“.

*se-emahale* (S.-S.) „Ding, welches unbeweglich feststeht“ (z. B. Grenzstein, Denkmal) Von *ema* + *hale* (= *χale* „lange“)

**Figure 3:** Excerpt from the alphabetical stretch E (Endemann 1911: 60).

Note that there is vertical alignment on *e*, which is a feature of a true left-expanded article structure (Gouws and Prinsloo 2005: 43). In order to look up

https://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2050 (Article)



any noun, the user first needs to identify the stem of the noun, and as Gouws and Prinsloo (2005) and Van Wyk (1995: 89) point out, stem identification is problematic in itself, and requires sound morphophonological knowledge of the user, since the prefix morphology of nouns is not only irregular, but is also subject to fairly complex morphological rules. As a result, "it is simply not possible for either the user or the lexicographer to determine unambiguously what the form of the isolated stem is" (Gouws and Prinsloo 2005: 33). This can lead to inconsistency and deviation from the stated editorial policy with regard to lemmatisation. In Endemann (1911) for example, verb stems to which the object concord of the first-person singular *N-* has been prefixed, are not treated at all, whereas reflexive verb stems with the prefix *i-* are treated under the letter I, which does not represent a left-expanded article structure, where lemmas are treated under the first letter of the stem. Nouns in class 9 are particularly problematic: nouns which display a nasal prefix *N-* are treated under the relevant nasal, e.g. *nku* 'sheep' under N, *mpho* 'gift' under M and *ngaka* 'doctor' under NG, whereas nouns that do not display a visible class prefix are lemmatised under the first letter of the noun, e.g. *tau* 'lion' under T and *kgomo* 'cow' under KG, despite the fact that the stem of the latter is actually *-gomo*. These inconsistencies impede the successful retrieval of information from the dictionary.

## Digitisation of the dictionary

In the last section of our article, we map out the steps we follow for the digitisation of the dictionary. The ideal option would be to transform the complete paper dictionary to a completely digital dictionary, displaying the features of a true electronic dictionary. However, the researchers have been granted permission by the publisher, De Gruyter Brill, to publish only parts of the dictionary electronically, without any cost. Most of the challenges outlined above can potentially be solved by transforming the original dictionary to a fully-fledged electronic dictionary. Digitisation is the first step in this transformation process.

There is currently no digital or searchable version of the dictionary, which requires the use of OCR (Optical Character Recognition) to extract all the dictionary entries. Furthermore, the dictionary features an obsolete, rather idiosyncratic, phonetically-based orthography, which needs to be converted to the modern orthography. We decided to use OCR4all (https://www.ocr4all.org/, Reul et al. 2019), an open source tool that is free to use. OCR4all offers the following functions:

— Segmentation: this process involves dividing a scanned document into its constituent parts, such as separating text from images or distinguishing individual characters, words, and lines. Segmentation ensures that the OCR system accurately identifies the structure of the document.

— Recognition: after segmentation, recognition is the core OCR function where the tool analyses the segmented parts and converts the text from image format into machine-readable text. This is where characters and words are identified based on patterns.

https://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2050 (Article)



— Ground Truth Production (GTP): this refers to creating a manually corrected version of the recognised text to serve as a reference or "ground truth." This step ensures high accuracy by comparing the OCR output with the verified, corrected version, enabling further improvements in OCR performance.

For the digitisation of the Endemann dictionary, the following steps were followed:

Step 1: Segmentation for OCR processing. The scanned pages of the dictionary are divided into individual sections (characters, words, lines) in preparation of the OCR process. During segmentation each page is divided into two columns, and the texts to be recognised are marked in advance for several pages. Information such as page numbers and headings which form part of the access structure, are ignored. Segmentation is initially performed on 10 to 20 pages, and after several rounds of training, the system suggests the parts of the text to be recognised with high accuracy.

Step 2: Training of a Custom Recognition Model. In this phase, a specialised OCR model is created, adapted to the unique font used in the dictionary. We start with a pre-trained OCR model from OCR4all, i.e., the "default/deep3_fraktur19/4," which is suitable for recognising Fraktur fonts commonly used in historical texts. This is followed by incremental improvement by means of Ground Truth Production (GTP). First, the pre-trained model is used to produce the initial OCR output. This is followed by manual correction of the initial 10 to 20 pages to correct recognition errors, specifically those related to the specialised phonetic symbols and unique fonts used in the dictionary. The OCR model is then retrained, using the corrected output, refining its accuracy incrementally with each cycle of corrections. This iterative process, also called 'bootstrapping' assists the model to adapt to the specific characteristics of the dictionary, leading to improved accuracy over time.

Step 3: Generation of a full-text output. A .txt file is generated, containing the complete digital format of the dictionary. Segmentation information is also exported — segmentation details such as the structure of pages, words and characters are exported in .xml format, thus preserving the layout and organization of the dictionary for future use.

Step 4: Post-processing. This entails the application of shell scripting to clean and refine the OCR4all-produced output text, ensuring accuracy and formatting consistency. Shell scripting refers to writing a series of commands for the shell to execute. A shell script is essentially a file that contains multiple instructions, which can automate tasks that would otherwise be performed manually by typing commands one by one. In the context of post-processing for dictionary digitisation, shell scripting can be used to:

— Automate corrections of formatting issues.
— Perform batch replacements of specific characters or patterns (e.g., correct-

https://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2050 (Article)



ing OCR errors).
— Organise and structure the output text.
— Extract specific data or manipulate files (like splitting or merging).
— Add page numbers that had to be ignored in the OCR-process.
— Separate the lemmas from the rest of the article (see 1 in Figure 5)
— Re-join words that are separated by a hyphen (see 2 in Figure 5)

Shell scripting is particularly useful for processing large amounts of text data efficiently and repetitively, making it ideal for refining the dictionary's digital text after completion of the OCR process. Figure 4 presents an illustration of the way in which entries are organised while Figure 5 displays the output resulting from the OCR process.

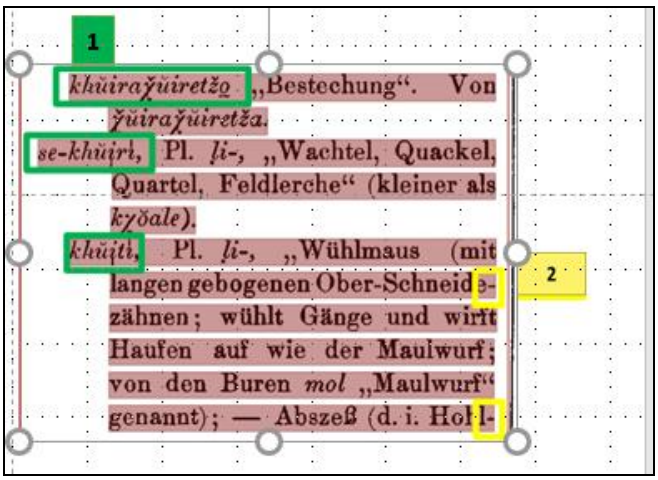


**Figure 4:** Organizing the entries

khiirahuiretšo "Bestechung". Von huirahuiretša.

se-khiiri , Pl. di-, "Wachtel, Ouackel, Ouartel, Feldlerche" (kleiner als kgoale).

khiiti , Pl. di-, "Wühlmaus (mit langen gebogenen Ober-Schneidezähnen; wühlt Gänge und wirft Haufen auf wie der Maulwurf; von den Buren mol "Maulwurf" genannt); - Abszeß (d. i. Hohlgang); - Viehweide (Feld, wo die Wühlmaus baut?)".

**Figure 5:** Output resulting from OCR

https://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2050 (Article)



The digitisation process is not flawless, but the overall results are surprisingly good. There is for example an overall error rate of 0.2% in the OCR results, therefore each article should be checked before copying. This error rate is calculated by OCR4all and would probably need to be further evaluated. Also, the model does not recognise curved brackets and articles grouped together by means of these brackets are erroneously split into multiple articles. Compare the following example shown in Figure 6, and the OCR-results in Figure 7:

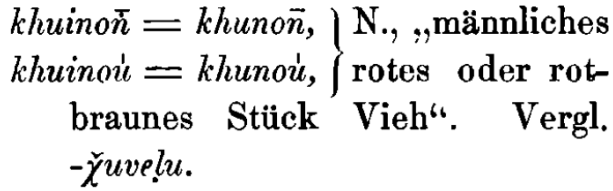
*khuinoñ* = *khunoñ*, } N., „männliches
*khuinoù* = *khunoù*, } rotes oder rot-
braunes Stück Vieh". Vergl.
*-χuveḷu*.

**Figure 6:** Articles grouped together by means of curved brackets by Endemann

khuinon = khunon, N., "männlicheskhuinou =
khunou, rotes oder rotbraunes Stück Vieh". Vergl. -
hubedu.

**Figure 7:** Resulting OCR output

After completion of the OCR process a number of working groups were formed. The German–English working group consists of German speakers who collect data from the OCR files and who also add English translations. The English–Sotho working group is composed of speakers of the Sotho languages and dialects and they are responsible for quality assurance. This working group can comment on Endemann's interpretation of the Sotho data. The third working group, the Sotho–English–Afrikaans group is tasked with translation of the English text into Afrikaans. All data is stored in pre-database tables to ensure an automated transfer into the database which will be utilised by the planned web-interface.

## Conclusion and future work

The wealth of lexical, cultural and linguistic information contained in Endemann's *Wörterbuch der Sotho Sprache* (1911) makes it an invaluable resource for resource scarce languages, and it is therefore imperative that this resource should not only be preserved, but that it should also be made accessible to a wider tar-

https://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2050 (Article)



get audience, especially to the speakers of these languages. There are still some unresolved questions with regard to the dictionary itself, due to the lack of historical information regarding its actual compilation. The use of the term 'Sotho' is one such issue. It is not clear what the exact reference of this term is, whether it refers to what is traditionally known as Northern Sotho (Sesotho sa Leboa), encompassing a number of dialects spoken in the southern and central parts of the Limpopo province, or rather to Sepedi, a dialect spoken in Endemann's area of missionary work.

Digitisation is but the first step towards producing a fully-fledged, electronic dictionary containing selected articles from the dictionary. By utilising full electronic features of such dictionaries, many of the challenges outline above, e.g. restricted access and difficult access structure can be addressed. The next step is the creation of a database, containing the proposed microstructure of the electronic version. The digitised data in Endemann's dictionary will be used as a basis with the addition of further data categories, e.g. English and Afrikaans translation equivalents, and selected grammatical information.

## Endnotes

1. Examples from Endemann are transcribed into the standard orthography and are also provided with translation equivalents in English.
2. However, even though these three items are not treated in the central list, they are treated in the back matter of the dictionary by way of illustrations and bilingual captions.
3. See https://www.lehmanns.de/shop/geisteswissenschaften/32384232-9783111418513-woerterbuch-der-sotho-sprache-sued-afrika

## References

### Dictionaries

**Colenso, J.W.** 1905. *Zulu–English Dictionary.* Second Edition. Pietermaritzburg: Shuter & Shooter.

**De Schryver, G.-M.** 2007. *Oxford Bilingual School Dictionary: Northern Sotho and English*. Cape Town: Oxford University Press Southern Africa.

**Endemann, K.** 1911. *Wörterbuch der Sothosprache*. Abhandlungen des Hamburgischen Kolonialinstituts, Band VII. (Reihe B. Völkerkunde, Kulturgeschichte und Sprachen Band 4). Hamburg: L. Friedrichsen & Co.

**Kriel, T.J.** 1967. *The New English–Northern Sotho Dictionary*. King William's Town: Educum.

**Kriel, T.J.** 1983. *Pukuntšu woordeboek, Noord-Sotho–Afrikaans, Afrikaans–Noord-Sotho.* Third edition. Pretoria: J.L. van Schaik.

**Kriel, T.J., E.B. van Wyk and S.A. Makopo.** 1989. *Pukuntšu woordeboek, Noord-Sotho–Afrikaans, Afrikaans–Noord-Sotho*. Fourth, revised and expanded edition. Pretoria: J.L. van Schaik.

**Mojela, V.M., M.C. Mphahlele, M.R. Selokela and W.M. Mojapelo.** 2015. *Sesotho sa Leboa–English Bilingual Dictionary*. Cape Town: Phumelela Books.

**Ziervogel, D. and P.C. Mokgokong.** 1975. *Groot Noord-Sotho Woordeboek*. Pretoria: J.L. van Schaik.

https://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2050 (Article)



## Other sources

**Atkins, B.T.S. and M. Rundell.** 2008. *The Oxford Guide to Practical Lexicography*. Oxford/New York: Oxford University Press.

**De Schryver, G.-M. and D. Joffe.** 2004. On How Electronic Dictionaries are Really Used. Williams, G. and S. Vessier (Eds.). 2004. *Proceedings of the Eleventh EURALEX International Congress, EURALEX 2004, Lorient, France, July 6–10, 2004*: 187-196. Lorient: Faculté des Lettres et des Sciences Humaines, Université de Bretagne Sud.

**De Schryver, G.-M. and D.J. Prinsloo.** 2000. Electronic Corpora as a Basis for the Compilation of African-language Dictionaries, Part 1: The Macrostructure. *South African Journal of African Languages* 20(4): 291-309.

**Gouws, R.H and D.J. Prinsloo.** 2005. Left-expanded Article Structures in Bantu with Special Reference to isiZulu and Sepedi. *International Journal of Lexicography* 18(1): 25-46.

**Kosch, I.** 1993. German-speaking Pioneers in African Linguistics and Literature with Special Reference to Northern Sotho. *South African Journal of African Languages* 13 (Supplement 2): 2-5.

**Kosch, I.** 2011. Innovation and Compromise in K. Endemann's Dictionary of the Sotho Language (1911). *South African Journal of African Languages* 31(1): 110-120.

**Kosch, I.** 2012. Challenges of Predictability and Consistency in the First Comprehensive Sotho Dictionary. *Lexikos* 22: 226-242.

**Lombard, D.P.** 1970. Bantoetaalstudie: Dictionaries in Northern Sotho. *Bantu Education Journal* 16(3): 12-13.

**Prinsloo, D.J. and G.-M. de Schryver.** 2002. Designing a Measurement Instrument for the Relative Length of Alphabetical Stretches in Dictionaries, with Special Reference to Afrikaans and English. Braasch, A. and C. Povlsen (Eds.). 2002. *Proceedings of the Tenth EURALEX International Congress, EURALEX 2002, Copenhagen, Denmark, August 13–17, 2002*: 483-494. Copenhagen: Center for Sprogteknologi, Københavns Universitet.

**Prinsloo, D.J. and G.-M. de Schryver.** 2007. Crafting a Multidimensional Ruler for the Compilation of Sesotho sa Leboa Dictionaries. Mojalefa, M.J. (Ed.). 2007. *Rabadia Ratšhatšha: Studies in African Language Literature, Linguistics, Translation and Lexicography:* 177-201. Stellenbosch: SUN PReSS.

**Reul, C., D. Christ, A. Hartelt, N. Balbach, M. Wehner, U. Springmann, C. Wick, C. Grundig, A. Büttner and F. Puppe.** 2019. OCR4all — An Open-source Tool Providing a (Semi-)Automatic OCR Workflow for Historical Printings. *Applied Sciences* 9(22): 4853.

**Rundell, M.** 2015. From Print to Digital: Implications for Dictionary Policy and Lexicographic Conventions. *Lexikos* 25: 301-322.

**Swanepoel, P.** 2010. Improving the Functionality of Dictionary Definitions for Lexical Sets: The Role of Definitional Templates, Definitional Consistency, Definitional Coherence and the Incorporation of Lexical Conceptual Models. *Lexikos* 20: 425-449.

**Tarp, S.** 2013. Old Wisdom: The Highly Relevant Lexicographical Knowledge Obtainable from a Specialized Dictionary from 1774. *Lexikos* 23: 394-413.

**Van Wyk, E.B.** 1995. Linguistic Assumptions and Lexicographical Traditions in the African Languages. *Lexikos* 5: 82-96.