




CORE VOCABULARY OF SEPEDI-SPEAKING PRESCHOOLERS

Determining the Core Vocabulary used by Sepedi-speaking Children During Regular Preschool Activities

Rahab Mothapo,  Kerstin Tönsing  and Refilwe Morwane 

Centre for Augmentative and Alternative Communication, University of Pretoria

Twitter: @CentreforAAC

Facebook: <https://www.facebook.com/centreforaac/>

Corresponding author contact details:

Kerstin Tönsing
Room 2-27
Communication Pathology Building
University of Pretoria
Hatfield
0002
Tel: +27 12 420 4729
Email: kerstin.tonsing@up.ac.za

CORE VOCABULARY OF SEPEDI-SPEAKING PRESCHOOLERS

Abstract

Purpose: In order to provide equitable communication intervention and support services to clients from diverse cultural and linguistic backgrounds, the development of language-specific resources for assessment and intervention is needed. The purpose of the study was to develop a core vocabulary list based on language samples from Sepedi-speaking children, in order to make it available as a resource to inform vocabulary selection for augmentative and alternative communication (AAC) systems for children in need of AAC from a Sepedi language background.

Method: The speech of six typically developing Sepedi-speaking children aged 5 – 6 years was recorded using small body-worn audio recording devices. Children were recorded during their regular preschool day. The recordings were transcribed, coded and analysed.

Result: The composite transcript consisted of 17 579 words, of which 1023 were different words. The core vocabulary was determined by identifying all words that were used with a minimal frequency of 0.05%, and were used by at least half of the participants. The Sepedi core vocabulary consisted of 226 words that accounted for 88.1% of the composite sample.

Conclusion: The core vocabulary determined in this study represents a small pool of reusable linguistic elements that form the grammatical framework of the Sepedi language. As such, is a valuable resource that can be used to assist with vocabulary selection for children who require AAC from a Sepedi language background.

Keywords: Augmentative and Alternative Communication, children, core vocabulary, preschool, Sepedi, vocabulary selection.

Introduction

Children who are unable to meet their daily communication needs through speech may benefit from Augmentative and Alternative Communication (AAC). AAC systems, techniques, and strategies should result in generative, functional communication in all communication contexts with a variety of communication partners (Mirenda, 2003). When children in need of AAC are not yet literate and rely on graphic symbol-based communication systems, achieving such generative functional communication can be challenging. Graphic symbol-based systems typically contain only a limited number of symbols in order not to exceed the cognitive (and possibly also physical) demands of navigating through the system to select the desired symbol (Thistle & Wilkinson, 2013). This means that the individual using the system only has access to a limited vocabulary. Special care must therefore be taken in selecting this vocabulary in a way that will maximise its relevance and usability.

Various methods of vocabulary selection have been documented in the literature. These methods include compiling environmental inventories (Beukelman & Mirenda, 2013) and obtaining vocabulary suggestions from informants who are well-acquainted with the person in need of AAC (Trembath, Balandin, & Dark, 2006). However, vocabulary selected based on environments and informants tends to be situation-specific and noun-dominated (Bean, Cargill, & Lyle, 2019). Nouns tend to be easy to think of, and specific situations (e.g. a preschool arts and craft activity) easily elicit associations with specific people, objects, and artefacts. Situation-specific, noun-dominated vocabularies have limited generalizability and also provide limited opportunity for the development of word combinations and syntax. Core vocabulary lists have been used by AAC system developers and practitioners as another source to inform vocabulary selection in order to ensure the inclusion of words that (a) are generic to many communication situations; and (b) allow for the generation of unique novel

CORE VOCABULARY OF SPEPEDI-SPEAKING PRESCHOOLERS

sentences (Bean et al., 2019; Dada, Murphy, & Tönsing, 2017; Lund, Quach, Weissling, McKelvey, & Dietz, 2016). The concept of a core vocabulary is based on the observation that the words used by speakers without disability every day are not necessarily unique but rather consist of a small pool of frequently re-used generic words that are important for building sentences. Across a number of different studies in English, for example, it has been established that approximately 200 to 400 words represent about 80% of spoken language used by individuals of various ages (van Tilborg & Deckers, 2016). Various authors have proposed that including this core vocabulary on AAC systems would provide the person using the system the opportunity to build sentences and express a variety of messages, as these words provide the grammatical framework of a language (Balandin & Iacono, 1999; Fallon, Light, & Kramer Paige, 2001; Trembath, Balandin, & Togher, 2007; Witkowski & Baker, 2012).

Many of the resources and approaches developed in the field of AAC have emanated from high-income and mainly English-speaking countries. However, global developments in the fields of rehabilitation and speech-language pathology have highlighted the urgent need to revisit models of service provision that continue to marginalise underserved contexts (Wickenden, 2013; Wylie, McAllister, Davidson, & Marshall, 2013). Attempts have been made to develop AAC systems and services that are appropriate for, responsive to, and inclusive of persons from different countries, contexts, cultures and language backgrounds. There has been a growing realisation that merely translating AAC systems and resources such as vocabulary lists into other languages falls far short from ensuring contextual, cultural and linguistic relevance (Soto & Yu, 2014). Exploratory studies have therefore been conducted with persons with communication difficulties and their families from diverse linguistic and cultural groups in order to better understand their views and priorities regarding AAC (Amery et al., 2019; Kulkarni & Parmar, 2017; Tönsing, van Niekerk, Schlünz, & Wilken,

CORE VOCABULARY OF SPEPEDI-SPEAKING PRESCHOOLERS

2019). Studies have been done on appropriate symbol and picture representations, as well as on text-to-speech and other human language technology that can be incorporated into speech generating devices, giving access to various spoken languages (Babic, Slivar, Car, & Podobnik, 2015; Baker & Chang, 2006; Bhattacharya & Basu, 2009; Schlünz et al., 2017). Similarly, core vocabulary lists have been established in languages other than English, such as French (Robillard, Mayer-Crittenden, Minor-Corriveau, & Bélanger, 2014), Korean (Shin & Hill, 2016), and Mandarin (Liu & Sloane, 2006), and recently also in historically under-resourced languages such as isiZulu (Mngomezulu, Tönsing, Dada, & Bokaba, 2019). These lists have been based on primary data from conversational samples. Core vocabulary includes many structure or function words that have little meaning in themselves, but that establish grammatical relationships between the other words in a sentence (Fries, 1952). These structure words are typically not directly translatable between languages that are linguistically very different. Translation of a core vocabulary list established in one language into another (target) language would therefore not result in a list that provides the grammar framework for the target language (Trembath et al., 2007, Mngomezulu et al., 2019). The current study aimed to add another resource to the expanding repertoire of linguistically and culturally diverse AAC tools by establishing a core vocabulary list in Sepedi, one of the historically under-resourced languages in South Africa.

Sepedi is spoken as a first language by about 9.1% of the South African population, constituting an estimated 5.9 million citizens (Worldometer, 2020). It is the fifth most frequently spoken home language in South Africa (Statistics South Africa, 2012). More than half of first language Sepedi speakers reside in the Limpopo province, the most northern province of the country. According to statistics from the South African Department of Basic Education (2011), 10.7% of learners in the basic education system (Grades 1–12) spoke Sepedi as their home language in 2007 (amounting to about 1 250 400 learners). With an

CORE VOCABULARY OF SPEPEDI-SPEAKING PRESCHOOLERS

estimated population growth rate of about 20% since 2007, the current number of school-going Sepedi-speaking children is estimated at over 1.5 million. With a global incidence of severe communication disabilities of about 1.3% (Beukelman and Mirenda, 2013), it can be estimated that about 20 000 children would benefit from a graphic symbol-based AAC system that gives access to expression in Sepedi. However, it has to be noted that, in 2015, it was estimated that over 500 000 South African children with disabilities were not in school (Department of Education, 2015). Estimates therefore have to be regarded as tentative.

Education and rehabilitation services for persons with communication disabilities in South Africa have been provided predominantly in English (Dada et al., 2017; Kathard & Pillay, 2013; Tönsing, van Niekerk, Schlünz, & Wilken, 2018), even though this is the home language of only 9.6% of the population (Statistics South Africa, 2012). The prioritization of certain languages such as English and Afrikaans under colonial and Apartheid regimes has contributed to the minimal development of culturally and linguistically appropriate resources for African languages (Pascoe & Norman, 2011). In order to provide more equitable services, the development of resources in languages such as Sepedi is an urgent necessity. These efforts are undergirded both by the South African Constitution (Republic of South Africa, 1996) highlighting equitable education and healthcare access and the equal status of all 11 official languages, as well as by global trends in the speech-language profession to become more inclusive of under-served populations (Kathard & Pillay, 2013).

The morphological and orthographic structure of a language is important to consider when determining the unit of analysis suitable for a core vocabulary study (Mngomezulu et al., 2019, Shin & Hill, 2016), since a core vocabulary should consist of lexical units that can be creatively recombined to generate novel meanings. Sepedi is a synthetic, agglutinating language, meaning that it is rich in morphemes that mostly do not change structure when added to words (i.e. the morphemes are merely ‘glued’ together) (Taljard & Bosch, 2006). In

CORE VOCABULARY OF SEPEDI-SPEAKING PRESCHOOLERS

addition, Sepedi has a predominantly disjunctive orthography, meaning that single linguistic words may be represented by a number of orthographically separated units (Kosch, 2006). Most orthographic words therefore consist of only one morpheme. However, nouns, verbs, and adjectives typically consist of more than one bound morpheme (i.e. conjunctively written morphemes), specifically roots and affixes. For example, *ke a ba rata* ('I like them') consists of four orthographic words, corresponding to five morphemes namely *ke* (the first person singular concord), *a* (a present tense morpheme), *ba* (the object concord), and *rat-* (the verb stem) and *-a* (the present tense verb suffix). The linguistic structure of Sepedi seems to suggest that orthographic words would be a useful unit of analysis for a core vocabulary study – that is, the orthographic space could be used as a boundary for defining the units to be counted. However, because Sepedi contains a high number of polysemous (multi-meaning) function words (Faaß, Heid, Taljard, & Prinsloo, 2009), including heteronyms (words that are spelled the same, but have different meanings and pronunciations), additional coding would be needed to distinguish these words, as they would have different graphic representations. Also, coding could be applied to trace inflected forms of nouns, verbs, and adjectives back to the root/lemma (the latter referring to the dictionary form of the word).

The aim of this study was to identify and describe a Sepedi core vocabulary that can be used as a resource to guide vocabulary selection for a Sepedi AAC system. Specifically, the authors aimed to (1) identify the words that Sepedi-speaking children without disabilities use most frequently and commonly during regular preschool activities; and (2) to describe this core vocabulary by parts of speech as well as by differentiating structure (grammatical) and content (lexical) vocabulary.

Methods

Participants

Three boys and three girls (six children in total), ranging in age from 5;3 (years; months) to 6;8 ($M = 6;3$ and $SD = 7$ months) who spoke Sepedi as a first language were recruited from the reception grades (Grade R) of three preschools where Sepedi was the language of instruction in a semi-rural area in the Limpopo province which is the northern part of South Africa. The participant selection involved that participants' had to (1) be between the ages of 5;0 and 6;11; (2) have no speech and language impairments or any other developmental impairment or delay; (3) have attended the preschool for at least three months prior to the study and attend school at least three days per week; and (4) speak Sepedi as home language. After receiving approval from the respective schools to carry out the study, the school teachers of the three Grade R classes were requested to identify a boy and a girl each who, in their opinion, met the selection criteria. Teachers supplied the parents of these children with detailed information letters and consent forms. Parents of all six children gave permission for their child to participate in the study, and also completed a questionnaire to provide background information. The study was then explained to each child individually using child-friendly language and pictures to support comprehension. Each child was also asked a series of questions to ensure that he/she understood all procedures and their right to withdraw at any time. The child was then given an opportunity to give or decline assent to participate. All six children assented to participate. Verbal assent was also obtained from each child every morning of data collection, before the child was fitted with the recording equipment.

Materials

Small digital voice recorders with lapel microphones were used to collect speech samples. The voice recorders were inserted into custom-made body-worn pouches that were able to fit around the participants' waists and the microphones were attached to the top part of the

CORE VOCABULARY OF SPEPEDI-SPEAKING PRESCHOOLERS

participants' jerseys/shirts using the microphone clip. The recorded audio files were loaded from the recorders onto a laptop computer. Transcriptions were conducted using the System for Analysing Language Samples (SALT) software (Miller & Iglesias 2012). Headphones (Sony Wireless NFC Headphones with Noise Cancelling) were used to listen to the playback of the audio files during transcription.

Data collection procedures

The researcher met the participants at their respective preschools and, after obtaining assent, fitted each participant with the voice recorder in the body-worn pouch and the lapel microphone. The participants then returned to their classrooms. Teachers agreed to monitor the participants' comfort with the recording equipment and to remove it whenever they felt it was unsafe or inappropriate to wear it, or when participants requested it to be removed. Teachers were also asked to behave as they would normally in the classroom and not to alter their behaviour towards the children in an attempt to entice the child to become more talkative. At the end of the preschool day, the researcher again came to the preschool to remove the recording equipment. Recording continued on consecutive days until 3000 orthographic words (including unintelligible words, phrases, and sentences) were recorded per participant. The total time taken for all participants ranged from 07 hours 47 minutes to 21 hours 26 minutes, and the number of days on which recordings were done ranged from 2 to 4. The recordings were taken during regular preschool activities. In all three preschools, these included activity time, meals, morning rings and reading time.

Transcription and analysis

The first author and two trained research assistants transcribed the recordings using standard Sepedi orthography. The first 20 minutes of recording were not analysed, to counter any novelty effects. Any references the children made to the recording equipment or process were also not transcribed. Transcriptions were done by following SALT conventions as well as a

CORE VOCABULARY OF SPEPEDI-SPEAKING PRESCHOOLERS

set of predetermined transcription rules developed for the study, based on Trembath et al. (2007). Language samples were transcribed into the SALT program (Miller & Iglesias, 2012). Individual files were created for each participant.

Transcription reliability was enhanced by cross-checking the transcript with the voice recording and correcting it for each participant. The transcripts were checked by a different person (first author or research assistant) from the one who had transcribed the data before. Procedures followed were similar to those implemented by Ronski et al. (2010) in their study.

The individual transcripts were then combined into one composite file. The first author coded the composite transcript according to the pre-developed coding rules in order to identify code switching; trace inflected forms of nouns, verbs and adjectives back to the root/lemma; and to distinguish between heteronyms (words that are spelled the same, but have different meanings and pronunciations). In order to determine coding reliability, 20% of each participant's transcribed language sample was randomly selected and was coded a second time by a research assistant who was provided with the coding rules. Inter-coder agreement was calculated by dividing all agreements by the sum of agreements and disagreements (with disagreements including omitted codes, added codes, and codes that differed between the two coders) and multiplying the result by 100. The percentage of agreement between coders ranged from 92.2% to 95.5% per participant, with an average of 94.1 (SD = 0.89).

The SALT programme was used to determine tokens (total number of words) after removal of unintelligible words, phrases, and sentences. The programme was also used to determine types (total number of different words) occurring in the composite sample, the type-token-ratio (TTR) and also the number of occurrences of each word, with inflected nouns, verbs, and adjectives being counted under their root/lemma. A score of 6 was

CORE VOCABULARY OF SPEPEDI-SPEAKING PRESCHOOLERS

allocated if all six participants used the word, whereas a score of 1 meant that only one participant used the word. All words that occurred with a frequency of 0.05% or more and were used by at least three participants (commonality score of 3 or above) were classified as core vocabulary. The frequency and commonality score criteria for words to be regarded as core are somewhat arbitrary. There is no scientific justification for using a commonality score of ≥ 3 (50%) and a frequency count of $\geq 0.05\%$ as criteria for the inclusion of words in core vocabulary (Shin & Hill, 2016). However these criteria for determining a core vocabulary have been used in previous studies (Trembath et al., 2007; Boenisch & Soto, 2015; Mngomezulu et al., 2019).

Each word in the core vocabulary was then classified as a content or structure word. Content words are those that carry lexical meaning (e.g. nouns, verbs and adjectives). Structure words (also termed function words) fulfil a grammatical function, as they create the grammatical structure that conveys how the lexical words relate to each other (Shi et al., 2006). In Sepedi these include concords, conjunctions, and prepositions.

Each word in the core vocabulary was also classified by parts of speech according to the classification provided in the Oxford Pukuntšu ya Sekolo dictionary (de Schryver et al., 2007). Where necessary, the grammar books by Poulos and Louwrens (1994) and van Wyk, Groenewald, Prinsloo, Kock and Taljard (1992) were also consulted. The Sepedi 'part-of-speech tagger' demonstration (de Pauw & de Schryver, 2007) available online at <https://www.aflat.org/sothotag> was also consulted at times.

Results

The number of words collected per participant varied from 2719 to 2978. This amounted to 17 569 tokens. A total of 1023 types were found and the TTR was 0.06. When the frequency and commonality criteria were applied, 226 words were designated as core vocabulary. The frequency counts of the 226 words were summed, and amounted to 88.1%. This constitutes

CORE VOCABULARY OF SPEPEDI-SPEAKING PRESCHOOLERS

the coverage of the core vocabulary – meaning that 88.1% of the words used during conversations were core words. The remaining 797 words were designated as fringe words. Although these words were considerably higher in number when considering number of different words, their coverage only amounted to 11.9%. A list of the 100 most frequently occurring core words with English translations is provided in the appendix.

The classification of the core vocabulary into content and structure words resulted in the identification of 144 content and 82 structure words. Although the core vocabulary therefore contained many more content words, these content words were only used with a frequency of 32.7%, whereas structure core words covered 55% of the speech sample.

The number of different core words falling into the different parts of speech, the proportion of each part of speech category within the total core vocabulary and frequency with which each part of speech category in the core vocabulary appeared in the sample were calculated. Results are displayed in Table 1 (insert Table 1 about here). From this table, it is apparent that the 18 different concord words occurred with a high frequency, accounting for nearly 25% of the words used in the sample. A total of 83 different verbs, 49 different nouns, and 24 different pronouns also occurred with a high frequency. A total of 15 different verbal affixes (prefixes and suffixes) occurred with a frequency of about 12% in the sample. These affixes are written disjunctively from the rest of the verb and were therefore counted separately in this study. They included negative morphemes (e.g. *ga*), aspectual prefixes (e.g. *sa*), future and present tense morphemes (e.g. *tla* and *a*) amongst others. Each of these affixes modifies the meaning of the verb. Interjections, conjunctions, prepositions, adverbs, adjectives, and locative particles all occurred less frequently, and accounted for about 11% of the total sample.

Regarding content vocabulary, most content words seemed relatively generic, and not context-specific (e.g., *selo* – “thing”, *bona* – “see”, *nyaka* – “want”). However, a few words

CORE VOCABULARY OF SPEPEDI-SPEAKING PRESCHOOLERS

did seem to reflect the preschool context (e.g., *raloka* – “play”, *sekolo* – “school”, and *ngwala* – “write”). The words *ja*-“eat” and *toilet* (code switch from English) may have reflected specific regular preschool activities.

Discussion

The parameters of the speech sample identified (tokens, types, and TTR) are similar to those found in other studies with similar participant numbers and speech sample lengths. Boenisch and Soto’s (2015) study, for example, yielded a TTR of ~0.06 on a composite sample of 19 885 words from eight English second language speakers, while Trembath et al.’s (2007) composite sample of 18 000 words (collected from six English-speaking Australian children) contained 1,411 unique words (TTR = ~0.08). The study by Mngomezulu et al. (2019) on speech samples from six isiZulu-speaking children identified the most frequently used morphemes or formatives rather than the most frequently used orthographic words. Still, the TTR of ~0.06 on a composite sample comprising 20,137 formatives is similar to that found in the current study. These TTR values suggest that, across languages, it is possible to identify words that are re-used often in conversations in order to consider including them on AAC systems.

The size and coverage of the core vocabulary identified are also similar to the parameters of core vocabularies identified in some other studies that used similar criteria to identify the core vocabulary and also relied on analyses of spoken corpora collected during a range of naturally-occurring activities. Trembath et al. (2007), for example, determined a core of 263 words, which accounted for 79.8% of the total sample in Australian preschool children speaking English. Boenisch and Soto’s (2015) monolingual English participants used 200 words for 78.7% of their recorded communication, while Robillard et al.’s (2014) monolingual French participants made use of 216 words for 80.15% of their communication. The isiZulu-speaking participants in the study by Mngomezulu et al. (2019) used 213

CORE VOCABULARY OF SEPEDI-SPEAKING PRESCHOOLERS

formatives (or morphemes) covering 88% of their communication. It is interesting to note that a slightly higher coverage was found in Sepedi and isiZulu (nearly 90%). These two languages belong to the same linguistic family of African languages, which may explain the similarities in the core vocabulary parameters. However, Crestani, Clendon, and Hemsley (2010) for example, found that a mere 173 words covered 80% of the words that Australian English-speaking children used during narrative tasks. The more structured elicitation context (story retelling, personal narratives elicited from standard pictures and a scripted narrative task) may have contributed to a smaller core vocabulary.

Like Boenisch and Soto (2015), the current study relied primarily on the orthographic space to identify the units of the core vocabulary, but also counted inflected forms of verbs, nouns and adjectives under the root or lemma. Although Sepedi is a synthetic language (high morpheme-to-linguistic word ratio), its disjunctive orthography results in a low ratio of morphemes to orthographic words. The orthographic space therefore provided a useful method of separating units. In contrast, isiZulu with its conjunctive orthography, required orthographic words to be separated into individual morphemes in order to arrive at useful reusable core vocabulary units (Mngomezulu et al., 2019). Taken together, these studies illustrate that linguistic and orthographic structure of a language need to be considered in deciding on the unit of analysis to be used in a core vocabulary study.

The frequent use of a relatively small number of structure core words emphasizes the importance of these words in spoken Sepedi. As has been noted by other authors of core vocabulary studies (Boenich & Soto, 2015, Robillard et al., 2014), these words are necessary for the production of sentences, but tend to be omitted from AAC systems when their vocabulary is selected by informants (Bean et al., 2019). While the resulting noun-dominated AAC systems may be useful for expressing single-word messages that can be interpreted by knowledgeable partners within context, the expression of more complex and decontextualized

CORE VOCABULARY OF SEPEDI-SPEAKING PRESCHOOLERS

messages is dependent on access to a variety of word classes, including structure vocabulary to produce grammatically correct sentences that express the relationships between content words.

The presence of various parts of speech in the Sepedi core vocabulary further highlights that children aged 5-6 years use different parts of speech frequently. The core vocabulary contained parts of speech identified in previous English core vocabulary studies, such as verbs, nouns, conjunctions, interjections, adverbs, adjectives, pronouns, and prepositions (e.g. Boenisch & Soto, 2015; Trembath et al., 2007). However, it also contained parts of speech that do not have equivalents in English, such as concords and verbal affixes. These frequently used parts of speech are specific to various African languages belonging to the same language family as Sepedi (e.g. Sesotho and Setswana) but are not directly translatable to English. Their inclusion in the core vocabulary highlights that language-specific studies are needed to identify a core vocabulary that provides the grammatical framework allowing for the generation of novel utterances.

The identification of the Sepedi core vocabulary is intended as a resource for the selection of vocabulary for graphic symbol-based AAC systems for children in need of AAC who come from a Sepedi language background. To date, no other published AAC vocabulary resources or AAC vocabulary sets exist in this language. One aim of including core words in an AAC system is to foster the acquisition of expressive grammar and syntax. However, there are clearly more questions to be answered to guide the construction of such a system. For one, graphic representations would need to be developed for structure vocabulary that does not have translation equivalents in languages for which graphic symbol representations exist. Furthermore, the core vocabulary should not be seen as the only or ultimate resource to guide vocabulary selection. Fringe vocabulary (typically selected via informants or environmental inventories) allows for the expression of specific and personalised information that is

CORE VOCABULARY OF SPEPEDI-SPEAKING PRESCHOOLERS

reflective of the child's context, personality, interests and preferences. Also, the current core vocabulary list is a snapshot of the most commonly and frequently used vocabulary of Sepedi-speaking children aged 5-6. While it provides a robust grammatical framework (because children have a relatively mature grammar at this age), it does not provide individualised guidance on how the appropriate proportions of core and fringe vocabulary could be represented, organised and expanded over time. These questions would need to be addressed in order for the system to minimise learning demands, maximise ease of production, and maximise appropriate language coverage for children at different stages of language development. The needs and skills of partners and the communication demands of different environments would also need to be taken into account. Appropriate methods of expanding the system to allow for communication and language development would also need investigation, as would appropriate scaffolding or teaching methods. Although it is unlikely that these questions have a one-size-fits-all answer, further studies could help to better understand aspects of this complex process.

Limitations

Although three different sites were used for the study, the sample comprised of only six participants. Also, the sites were relatively homogenous (preschools from the same area), the time span of collecting data was relatively short (two to four days per child) and children were similar in age (five to six years). This introduces a limitation concerning the extent to which the core vocabulary can be regarded as completely representative and to what extent it can be generalised to the larger population. Participant reactivity remained an unavoidable factor, as with all observational designs, and this may have affected the internal validity. The children appeared to have conversed freely about various topics but one should note that they may still have changed their behaviour in response to the presence of the recorders. The noise in the classroom may have affected the accuracy of the transcriptions. One solution would

CORE VOCABULARY OF SEPEDI-SPEAKING PRESCHOOLERS

have been to collect supplementary visual data, for example, by means of video recordings. However, practicalities and privacy concerns may be more difficult to navigate with these methods. Transcription reliability could have been more rigorously determined by letting an independent transcriber transcribe the audio recordings, and by calculating the percentage of agreement with the first transcription. The frequency and commonality score criteria for words to be regarded as core are somewhat arbitrary. There is no scientific justification for using a commonality score of ≥ 3 (50%) and a frequency count of $\geq 0.05\%$ as criteria for the inclusion of words in core vocabulary (Shin & Hill, 2016). There may be other methods of analysis, such as grouped frequency counts (Shin & Hill, 2016), which represent a more objective way of defining core versus fringe vocabulary.

Conclusions

Vocabulary selection remains an important but challenging task for AAC team members who support young children using picture-based AAC systems (Bean et al., 2019). The Sepedi core vocabulary list of 226 words determined in this study can be used as one source that speech language therapists and others can draw on to select functional and developmentally appropriate vocabulary that will support communication interactions across contexts and also ensure that expressive language skills can develop. Sepedi home language speakers represent the fifth largest group in South Africa, and the list is therefore expected to have clinical application to a sizeable population of children who require AAC. The study furthermore illustrates how the orthographic and linguistic structure of a language needs to be taken into account when analysing speech samples with the purpose of establishing a core vocabulary.

Acknowledgements

The financial assistance of the National Research Foundation (NRF) of South Africa (grant no. TTK150617119597) towards this research is herewith acknowledged. Opinions expressed and conclusions arrived at are those of the authors and are not necessarily to be attributed to the NRF. The authors would like to thank the children who participated in the study and their parents, the principals and other school staff who provided access to the premises and assistance with logistical arrangements, as well as the research assistants.

Declaration of interest

The authors declare no conflict of interest. The authors further report that they are responsible for the content and writing of the paper.

References

- Amery, R., Wunungmurra, J. G., Gondarra, J., Gumbula, F., Raghavendra, P., Baker, R., Theodoros, D., Amery, H., Massey, L., & Lowell, A. (2019). Yolŋu with Machado-Joseph disease: Exploring communication strengths and needs. *International Journal of Speech-Language Pathology, Early Online*, 1–12.
- Babic, J., Slivar, I., Car, Z., & Podobnik, V. (2015). Prototype-driven software development proceeb for augmentative and alternative communication applications. *Proceedings of the 13th International Conference on Telecommunications, ConTEL 2015*, 1–8.
- Baker, B. R., & Chang, S. K. (2006). A Mandarin language system in augmentative and alternative communication (AAC). *International Journal of Computer Processing of Languages, 19*(04), 225–237.
- Balandin, S., & Iacono, T. (1999). Crews , Wusses , and Whoppas : Core and fringe vocabularies of Australian meal-break conversations in the workplace. *Augmentative and Alternative Communication, 15*, 95-109.
- Bean, A., Cargill, L. P., & Lyle, S. (2019). Framework for selecting vocabulary for preliterate children who use augmentative and alternative communication. *American Journal of Speech-Language Pathology, 28*(3), 1000–1009.
- Beukelman, D. R., & Mirenda, P. (2013). *Augmentative and Alternative Communication: Supporting children and adults with complex communication needs* (4th ed.). Baltimore, MD: Paul H. Brookes.
- Bhattacharya, S., & Basu, A. (2009). Design of an iconic communication aid for individuals in india with speech and motion impairments. *Assistive Technology, 21*(4), 173–187.
- Boenisch, J., & Soto, G. (2015). The oral core vocabulary of typically developing English-speaking school-aged children: Implications for AAC practice. *Augmentative and*

CORE VOCABULARY OF SPEPEDI-SPEAKING PRESCHOOLERS

Alternative Communication, 31(1), 77–84.

Crestani, C.A. M., Clendon, S. A., & Hemsley, B. (2010). Words needed for sharing a story: implications for vocabulary selection in augmentative and alternative communication.

Journal of Intellectual & Developmental Disability, 35(4), 268–278.

Dada, S., Murphy, Y., & Tönsing, K. (2017). Augmentative and alternative communication practices: a descriptive study of the perceptions of South African speech-language

therapists. *Augmentative and Alternative Communication*, 33(4), 189–200.

Department of Basic Education. (2011). *The status of the language of learning and teaching (LOLT) in South African public schools*. Pretoria, South Africa.

Department of Education. (2015). *Report on the implementation of Education White Paper 6 on inclusive education*. Pretoria, South Africa: Author.

Faaß, G., Heid, U., Taljard, E., & Prinsloo, D. (2009). Part-of-Speech tagging of Northern Sotho: Disambiguating polysemous function words. *Proceedings of the EACL 2009 Workshop on Language Technologies for African Languages-AfrLaT 2009*, Greece, 38–45.

Fallon, K. A., Light, J. C., & Kramer Paige, T. (2001). Enhancing vocabulary selection for preschoolers who require augmentative and alternative communication (AAC).

American Journal of Speech-Language Pathology, 10(1), 81–94.

Fries, C. C. (1952). *The structure of English*. New York: Harcourt Brace.

Kathard, H., & Pillay, M. (2013). Promoting change through political consciousness: A South African speech-language pathology response to the World Report on Disability.

International Journal of Speech-Language Pathology, 15(1), 84–89.

Kosch, I. M. (2006). *Topics in morphology in the African language context* (1st ed.). Pretoria, South Africa: Unisa Press.

Kulkarni, S. S., & Parmar, J. (2017). Culturally and linguistically diverse student and family

CORE VOCABULARY OF SPEPEDI-SPEAKING PRESCHOOLERS

- perspectives of AAC. *Augmentative and Alternative Communication*, 33(3), 170–180.
- Liu, C., & Sloane, Z. (2006). Developing a core vocabulary for a Mandarin Chinese AAC system using word frequency data. *International Journal of Computer Processing of Oriental Languages*, 19(4), 285–300.
- Lund, S., Quach, W., Weissling, K., McKelvey, M., & Dietz, A. (2016). Assessment with children who need augmentative and alternative communication (AAC): Clinical decisions of AAC specialists. *Language, Speech, and Hearing Services in Schools*, 48, 56–68.
- Miller, J., & Iglesias, A. (2012). Systematic Analysis of Language Transcripts (SALT), Student version 2012 [Computer software]. Middleton, WI: SALT Software. LLC.
- Mirenda, P. (2003). Toward functional augmentative and alternative communication for students with autism: manual signs, graphic symbols, and voice output communication aids. *Language, Speech, and Hearing Services in Schools*, 34(3), 203–216.
- Mngomezulu, J. (2017). *Determining an AAC core vocabulary for Zulu-speaking preschool children* (Master's thesis). University of Petoria. Pretoria, South Africa.
- Mngomezulu, J., Tönsing, K. M., Dada, S., & Bokaba, B. (2019). Determining a Zulu core vocabulary for children who use augmentative and alternative communication. *Augmentative and Alternative Communication*, 35(4), 274–284.
- Pascoe, M., & Norman, V. (2011). Contextually relevant resources in speech-language therapy and audiology in South Africa - are there any? *The South African Journal of Communication Disorders*, 58, 2–5.
- Republic of South Africa. (1996). Constitution for the Republic of South Africa (Act No 108 of 1996).
- Robillard, M., Mayer-Crittenden, M., Minor-Corriveau, C., & Bélanger, R. (2014). Monolingual and bilingual children with and without primary language impairment:

CORE VOCABULARY OF SPEPEDI-SPEAKING PRESCHOOLERS

Core vocabulary comparison. *Augmentative and Alternative Communication*, 30(3), 267–278.

- Romski, M., Sevcik, R. A., Adamson, L. B., Cheslock, M., Smith, A., & Barker, R. M. (2010). Randomized comparison of augmented and nonaugmented language intervention for toddlers with developmental delays and their parents. *Journal of Speech Language and Hearing Research*, 53(2), 350–364.
- Schlünz, G. I., Gumede, T., Wilken, I., Van Der Walt, W., Moors, C., Calteaux, K., Tönsing, K., Van Niekerk, K. (2017). Applications in accessibility of text-to-speech synthesis for South African languages: Initial system integration and user engagement. In *ACM International Conference Proceeding Series* (Vol. Part F1308).
- Shi, R., Werker, J. F., & Cutler, A. (2006). Recognition and representation of function words in English- learning infants. *Infancy*, 10(2), 187–198.
- Shin, S., & Hill, K. (2016). Korean word frequency and commonality study for augmentative and alternative communication. *International Journal of Language and Communication Disorders*, 51(4), 415–429.
- Soto, G., & Yu, B. (2014). Considerations for the provision of services to bilingual children who use augmentative and alternative communication. *Augmentative and Alternative Communication*, 30(1), 83–92.
- Statistics South Africa. (2012). *Census 2011: Provinces at a glance*. Pretoria, South Africa: Author.
- Taljar, E., & Bosch, S. E. (2006). A comparison of approaches to word class tagging: Disjunctively vs. conjunctively written Bantu languages. *Nordic Journal of African Studies*, 15(4), 428–442.
- Thistle, J. J., & Wilkinson, K. M. (2013). Working memory demands of aided augmentative and alternative communication for individuals with developmental disabilities.

CORE VOCABULARY OF SPEPEDI-SPEAKING PRESCHOOLERS

Augmentative and Alternative Communication, 29(3), 235–245.

Tönsing, K. M., Van Niekerk, K., Schlünz, G. I., & Wilken, I. (2018). AAC services for multilingual populations: South African service provider perspectives. *Journal of Communication Disorders*, 73(March), 62–76. doi:10.1016/j.jcomdis.2018.04.002

Tönsing, K. M., Van Niekerk, K., Schlünz, G.I., & Wilken, I. (2019). Multilingualism and augmentative and alternative communication in South Africa – Exploring the views of persons with complex communication needs. *African Journal of Disability*, 8, a507.

Trembath, D., Balandin, S., & Dark, L. (2006). Why any old words won't do: The importance of vocabulary selection. *Acquiring Knowledge in Speech, Language and Hearing*, 8(3), 117–119.

Trembath, D., Balandin, S., & Togher, L. (2007a). Vocabulary selection for Australian children who use augmentative and alternative communication. *Journal of Intellectual & Developmental Disability*, 32(4), 291–301.

Van Tilborg, A., & Deckers, S. R. J. (2016). Vocabulary selection in AAC: Application of core vocabulary in atypical populations. *Perspectives of the ASHA Special Interest Groups*, 1(4), 125–138.

Wickenden, M. (2013). Widening the SLP lens: How can we improve the wellbeing of people with communication disabilities globally. *International Journal of Speech-Language Pathology*, 15(1), 14–20.

Witkowski, D., & Baker, B. (2012). Addressing the content vocabulary with core: Theory and practice for nonliterate or emerging literate students. *Perspectives on Augmentative and Alternative Communication*, 21, 74–81.

Worldometer (2020). South African population (live). Retrieved from <https://www.worldometers.info/world-population/south-africa-population/>

Wylie, K., McAllister, L., Davidson, B., & Marshall, J. (2013). Changing practice:

CORE VOCABULARY OF SPEPEDI-SPEAKING PRESCHOOLERS

Implications of the World Report on Disability for responding to communication disability in under-served populations. *International Journal of Speech-Language Pathology*, 15(1), 1–13.

CORE VOCABULARY OF SEPEDI-SPEAKING PRESCHOOLERS

Appendix

The 100 most frequent Sepedi core words with English translations

Words	Frequency	Commonality	Part of speech	English translation or translation approximate
o	5.95	6	concord	you/her/him/it
cn	4.33	6	noun	child's name
ke	4.14	6	concord	I
go	3.01	6	concord	it/there/you
ke	2.53	6	copulative particle	is/are
a	2.37	6	present tense morpheme	-no translation-
nna	2.17	6	pronoun	I/myself/me
le	1.91	6	conjunction	with/and
wena	1.82	6	pronoun	you
e	1.77	6	concord	he/she/it/they
ba	1.71	6	concord	they/them/of
ga	1.66	6	negative morpheme	do(es) not
re	1.63	6	concord	us/we
ka	1.58	6	preposition	with/about/through
ya	1.57	6	verb	go
bona	1.43	6	verb	see
tlo	1.33	6	future morpheme	shall/will
ya	1.30	6	concord	he/she/it/of
a	1.20	6	concord	he/she/them/of
se	1.14	6	negative morpheme	won't/will not
wa	1.09	6	concord	of/you
re	0.98	6	verb	say
mo	0.97	5	demonstrative particle	here
tla	0.97	6	verb	come
ka	0.85	6	pronoun	mine
ye	0.79	6	pronoun	this one
dira	0.73	6	verb	do/make
ah	0.71	6	interjection	ah
nyaka	0.69	6	verb	search/look for/want
a	0.68	6	hortative particle	your(s)
tn	0.67	6	noun	teacher's name
ee	0.67	6	interjection	yes
le	0.63	6	concord	you
akere	0.59	6	interjection	isn't it
eng	0.59	6	noun	what
wo	0.59	6	pronoun	this
gago	0.59	6	pronoun	your(s)

CORE VOCABULARY OF SPEPEDI-SPEAKING PRESCHOOLERS

Words	Frequency	Commonality	Part of speech	English translation or translation approximate
ngwala	0.57	6	verb	write
mang	0.56	6	noun	who
nto	0.51	6	noun	thing/something
so	0.50	6	adverb	like this
dula	0.48	6	verb	sit down/live/stay
ngwana	0.47	6	noun	child
na	0.46	6	verb	had/have
tšea	0.45	6	verb	take
botša	0.44	6	verb	tell/inform
di	0.42	6	concord	they/them
aowa	0.42	6	interjection	no
mo	0.42	6	concord	him/her
tlaleya	0.42	6	verb	tell on
heh	0.41	6	interjection	what
la	0.40	6	concord	you (plural)/of
betha	0.40	6	verb	hit/beat
ka	0.40	6	potential morpheme	can/could
mmata	0.36	6	noun	friend
ja	0.35	6	verb	eat
motho	0.35	6	noun	person/human being
gape	0.34	6	adverb	again
kae	0.34	6	adverb	where
tseba	0.34	6	verb	know
ha-eh	0.32	6	interjection	no
gore	0.31	6	conjunction	so that
tša	0.31	6	concord	they/of
kgona	0.29	6	verb	can/be able to
se	0.28	6	pronoun	this
yena	0.28	6	pronoun	her/him/she/he
bolela	0.27	6	verb	speak/talk/tell
lena	0.27	6	pronoun	you (plural)
kua	0.27	6	locative particle	over there
ngwe	0.24	6	adjective	another
tše	0.24	6	pronoun	these ones
rena	0.24	6	pronoun	we/ours
swara	0.24	5	verb	hold
fa	0.23	5	verb	give
kgale	0.22	5	noun	long ago
ge	0.22	5	conjunction	when/while
sa	0.22	5	concord	he/she/it/of

CORE VOCABULARY OF SPEPEDI-SPEAKING PRESCHOOLERS

Words	Frequency	Commonality	Part of speech	English translation or translation approximate
tsena	0.22	5	verb	enter/go into
ngwanenyana	0.21	5	noun	little girl
leina	0.20	5	noun	name
eh	0.20	5	interjection	eh
sa	0.20	6	aspectual prefix	still
be	0.19	5	verb	must be/must become
bea	0.19	6	verb	put
dlala	0.19	4	verb	play
kwa	0.19	4	verb	hear/feel
maaka	0.19	5	noun	lies
gafa	0.19	5	verb	be mad/be crazy
gagwe	0.19	5	pronoun	hers/his
gona	0.19	5	pronoun	there
yela	0.19	5	pronoun	that one
bo	0.18	6	concord	it
ebile	0.18	5	conjunction	then
ga	0.18	5	locative particle	at
kgopela	0.18	6	verb	ask for/request
mara	0.18	6	conjunction	but
selo	0.18	6	noun	thing
tla	0.18	6	future morpheme	shall/will
thoma	0.17	5	verb	begin/start
tsamaya	0.17	6	verb	go

CORE VOCABULARY OF SEPEDI-SPEAKING PRESCHOOLERS

Table 1

Parts of Speech Occurring in the Core Vocabulary with Corresponding Number of Different Words and Frequency Counts

Parts of speech	NDW	Proportion in core (in terms of NDW)	No. of occurrences in sample	Frequency of occurrences %	Most frequently used word	Approximate English translation (where possible)
Concords	18	8%	4,324	24.61	o	you/her/him/it
Verbs	83	36.7%	3,214	18.29	ya	go
Nouns	49	21.7%	2,173	12.37	eng	what
Verbal affixes	15	6.6%	2,121	12.07	go	-
Pronouns	24	10.6%	1,717	9.77	nna	I
Interjections	14	6.2%	673	3.83	ah	ah
Conjunctions	7	3.1%	518	2.95	le	and/together with
Prepositions	1	0.4%	278	1.58	ka	with/about
Adverbs	6	2.6%	238	1.35	so	like this
Adjectives	6	2.6%	123	.7	ngwe	other

CORE VOCABULARY OF SPEPEDI-SPEAKING PRESCHOOLERS

Parts of speech	NDW	Proportion in core (in terms of NDW)	No. of occurrences in sample	Frequency of occurrences %	Most frequently used word	Approximate English translation (where possible)
Locative particles	3	1.3%	106	.6	kua	there
Total	226	100%	15,485	88.14		