Studies in African Linguistics
Volume 52 Supplement 13, 2023.

Evaluating the representativeness of the Setswana corpus using behavioural data

Naledi Kgolo-Lotshwao
University of Botswana

Thapelo J. Otlogetswe
University of Botswana

This paper presents efforts to evaluate the representativeness of the Setswana corpus data with measures that are independent of corpora. Two frequency measures were used: one sourced via a subjective frequency rating survey and another from a corpus of Setswana. Strong correlations ($r = .75$; $p < .001$) between survey ratings and corpus frequencies suggest that the corpus reflects native speaker intuitions. In addition, the study tested for frequency effects using an unprimed visual lexical decision task where participants had to judge whether a letter string on a screen is an existing word or a made-up non-word. In the analysis of reaction times, survey ratings and corpus frequencies were found to have similar correlations with reaction times, although survey ratings provided a better fit. Our study therefore makes a methodological contribution as results illustrate that in the absence of established corpus databases, participant intuitions can be used in linguistic research. This observation concurs with previous research on European languages that found that native speakers can reliably estimate the frequencies of words.

Keywords: Corpus representativeness, Setswana, frequency, subjective frequency rating survey

1. Introduction

This study presents results from behavioural psycholinguistic measures (a subjective frequency rating survey and a lexical decision task) used to assess the representativeness of an existing Setswana corpus database. A corpus is a large collection of texts (Kilgarriff & Grefenstette, 2003). Corpora may be composed of written language, spoken language, or both. Written language corpora typically consist of extracts from newspapers, academic books, popular fiction, and many other kinds of text (Burnard, 1995). Spoken language corpora usually contain spoken language collected in different contexts such as radio call-in programmes, business meetings, and informal conversations.

Many corpora that are supposed to be representative of a particular language, rely heavily on written text. For instance, 90% of British National Corpora data stems from written texts. It is therefore possible that the sampling in corpora is quite different from those of estimates of spoken, heard, or written frequency. Thus, a heavy reliance on written texts might lead to problems for psycholinguistic studies and other types of studies that require frequency information that is representative of language users' entire linguistic experience. We argue that for such studies, corpus-based measures need to be compared with results of behavioural experiments, especially for languages whose corpora are entirely based on written texts, newly established, or very small. Arppe et al. (2010) and Gilquin & Gries (2009) argue for the need to use convergent methods, combining corpus analysis with elicited data. They maintain that there is a need for converging data, including relating different frequency measures to each other. Arppe et al. (2010, p. 7) note that there is "little or no understanding of how results from these different types of data inform one another".

Moreover, the availability of a particular piece of information determines whether this factor can be studied or not (New et al., 2004). For example, psycholinguistic studies on morphological

processing require information about the respective frequency of inflectional forms. Furthermore, the accuracy of the measures in the database directly influences the accuracy of the research and the statistical reliability of the experiments done. This creates problems for under-studied and under-resourced languages, such as Setswana, where corpora or frequency databases are non-existent, very small, or still rather new and unevaluated.

Francom, LaCross, & Ussishkin (2010) also point out that corpus-based evidence is inherently limited in gauging the relative representativeness of a corpus. Therefore, external information is needed to validate corpus data (Gries, 2008, 2009). Thus, it is imperative to test the representativeness and reliability of the newly established Setswana corpus (Otlogetswe, 2010), to determine whether it reflects native speaker intuitions. This evaluation of the Setswana corpus also answers the appeal by Gilquin & Gries (2009) and Arppe et al. (2010) for the use of converging data.

To this end, we suggest an effective combination of three sets of corpus-linguistic and psycholinguistic methods that can help researchers evaluate corpus frequency information and provide insights into speakers' mental lexicon at the same time:

1) corpus-linguistic methods and evaluation.
2) subjective frequency judgment questionnaires administered to groups of native speakers; and
3) reaction-time-based measures from lexical decision experiments.

Subjective frequency ratings are tasks in which participants estimate how often they have encountered a given word (Mayberry, Hall & Zvaigzne, 2013). These rating tasks are untimed and based on native speaker intuitions. Balota, Pilotti & Cortese (2001) state that native speakers can reliably estimate words' relative frequencies. As discussed above, frequency is one of the most important and robust factors manipulated in psycholinguistics experiments. It is therefore important to have frequency lists that are as reliable and recent as possible. Balota et al. (2006) advise that researchers investigating or controlling for frequency should use more recent measures. Gaining participant intuitions then provides a current means of testing the reliability of corpora, or sourcing frequency ratings in the absence of corpora. Balota et al. (2001) concur that native speakers can reliably estimate words' relative frequencies. Participant intuitions therefore provide the most recent norms, which have been reported to make better predictors of lexical decision performance (see Balota et al., 2004; Zevin & Seidenberg, 2004).

An unprimed visual lexical decision experiment is used to test for frequency effects. A frequency effect is the observation that more frequent word forms are acquired earlier, recognised faster, and read aloud faster than forms with a lower frequency (see e.g., Baayen, Dijkstra & Schreuder, 1997). This observation suggests that high-frequency forms have strong memory traces that can facilitate their retrieval when they are encountered again, and hence speed up processing (Alegre & Gordon, 1997).

Much of the existing literature on frequency effects in experimental psycholinguistics uses the lexical decision paradigm. In this task, participants are presented with words and made-up non-words (often referred to as "nonce words" or "novel words"), either visually or auditory, and are asked to make a speeded decision as to whether what they are presented with is a valid word or not. The measures of interest are the speed (reaction/response time – RT) and accuracy of the response. Since RTs to the letter string stimuli in visual lexical decision experiments are speeded, participants rely on automatic processing in identifying the letter string. As a result, the recognition process is

automatic and not laboured and relies on lexical memory. Results of lexical decision tasks are important as they can be used to evaluate corpora and psycholinguistic models simultaneously.

1.1 **The Setswana corpus**. Word stimuli for the subjective frequency rating survey and single visual lexical decision task were sourced from a corpus of 7 million Setswana words/tokens (Otlogetswe, 2010), which was accessed via the corpus query system Sketch Engine (www.sketchengine.co.uk). The intended purpose of the corpus is to aid Setswana dictionary compilation and linguistic research (Otlogetswe, 2011, p. 5). This corpus is general as it is not restricted to any subject field, register, or genre. It comprises a variety of text types from both spoken and written language. 90% of the corpus comprises written texts while 10% of the corpus comprises transcribed speech. Some of the written data was collected from published educational materials and newspapers, while the spoken data mostly came from radio call-in programmes, which were later transcribed.

The data found in the corpus is mostly formal. This is a result of where it was sourced from. Even though there is spoken language in the corpora, it was collected from radio call-in programmes which are to some extent formal as the radio programmes are regulated. This proved a challenge as one cannot ascertain to what extent the corpus reflects speaker intuitions. In addition, upon searching for some stimulus items, we found some words missing from the corpus. This might be due to the size of the corpus or its design.

Another challenge that the corpus posed was that it is not morphologically annotated, and morpheme boundaries are not indicated, despite the prevalence of morphological complexity in the language. Furthermore, the Setswana corpus is not lemmatised and therefore produces frequencies for just the word form and not lemmas. As a result, each item had to be searched for individually.

2. The Present Study

2.1 Corpus frequency analysis. Experimental items consisted of 60 Setswana base verbs and their corresponding class 1 and class 9 nominalisations. The items were controlled for syllable length and number of letters. For example, the noun *tshele* 'petty gossip' has 2 syllables and 6 letters. Class 9 nouns are generally short and are usually shorter than class 1 nouns (see Table 1 below). The length of the base verb also determines the length of the derived noun, especially so for class 1 nouns as they add a prefix and suffix to the verb root.

The words were also controlled for tonal variations. Tonal distinctions carry lexical and/or grammatical meaning in Setswana. Care was taken to only include words that do not have multiple tonal contrasts (e.g., *lela* /lìlá/ "to cry", and *lela* /lìlà/ "an intestine") to avoid participants reading the word as one that was not intended for the experiment.

Each word's word-form frequency within each condition was considered in the corpus search and the design of the lexical decision experiment. A broad range of frequencies was needed for all stimulus items to establish whether there were any frequency effects for these items (e.g., word-form/surface frequency, family size, and cumulative frequency). Cumulative morpheme frequency is the sum of the frequencies of all the complex words in which a root morpheme appears (Ford et al., 2003). The morphological family size of a word is the number of other poly-morphemic words such as derivations and compounds in which it appears as a constituent. It was ensured that the presentation lists had an equal distribution of word-form and cumulative frequencies. Table 1 below summarises the frequency properties of the selected stimulus items.

Table 1: Mean frequencies and word length of stimulus items by item form (with standard deviations)

	Base Verbs	Class 1 Derivations	Class 9 Derivations
Corpus Word-Form Frequency	6.3 (1.54)	3.28 (2.19)	5.05 (2)
Corpus Cumulative Frequency	5.34 (1.33)	5.34 (1.33)	5.34 (1.33)
Morphological Family Size	2.15 (0.12)	2.15 (0.12)	2.15 (0.12)
Number of Letters	4.4 (0.95)	6.15 (1.05)	5.02 (1.11)
Number of Syllables	2.15 (0.36)	3.13 (0.34)	2.13 (0.34)

Table 1 above demonstrates that class 9 nouns are more frequent than class 1 nouns in the corpus. This observation is corroborated by the Setswana monolingual dictionary (Otlogetswe, 2012) which contains 7, 967 noun headwords, and class 9 has the highest representation, with 2,715 entries, while Class 1 has only 468 entries.

2.2 The subjective frequency rating survey. To test the representativeness of the frequency information sourced from the Setswana corpus, participant intuitions were sought. A subjective frequency rating survey was designed to determine word form frequencies for the experimental items. For this task, we adopted the procedures employed by Schreuder & Baayen's (1997) rating survey. Ethical approval was sought from and provided by the University of Botswana.

2.2.1 *Participants*. Twenty-five participants took part in the survey. The survey was completed online at www.surveygizmo.com, and a secure link of the survey was circulated via email to participants. Informed consent to participate in the survey was sought before forwarding the survey link to participants. All participants were native speakers of Setswana living in Gaborone, Botswana.

2.2.2 *Materials* The experimental items described in 2.1 above were used in the rating survey.

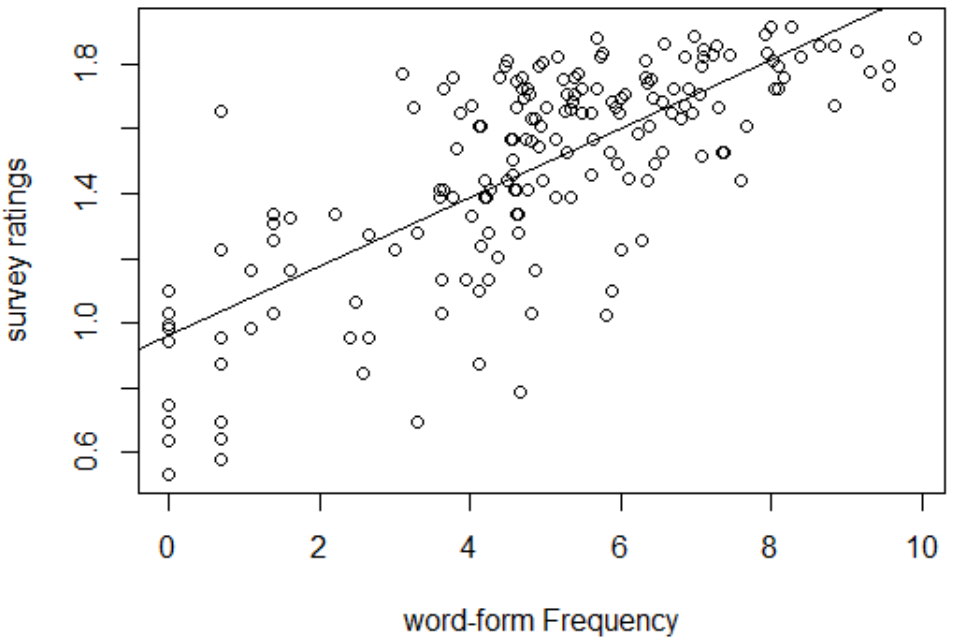
2.2.3 *Experimental procedure* Participants were asked to indicate on a seven-point scale how frequent they thought a word was used in Setswana. The scale ranged from 1 'extremely low' to 7 'extremely high'. Some studies require participants to rate frequency or familiarity using a slider, whose left side corresponds to less familiar and right corresponds to very familiar – with the measures ranging from 0-100 (e.g., Francom et al., 2010). This slider method proves too broad for many participants to use for rankings. Francom et al. in the review of their results note that the use of the 7-point scale, as in the present study, is much more user-friendly.

Participants were informed prior to the survey that they would know the words to be presented but that the words would differ in terms of frequency of usage. They were informed that their task was to rank how frequent in their opinion a particular word occurred in Setswana. The survey was accessed over the internet and participants completed it at their convenience. The frequency rating task lasted approximately 5–10 minutes.

2.2.4 *Results***.** A correlational approach was adopted for the analysis of results (see e.g., Francom et al., 2010 for similar analyses). Ten of the low-ranking word forms from the survey did not appear in the corpus at all. In order to include such items with a zero corpus word-form frequency, we added 1 to each of these item frequencies. Previous studies have utilised this method (e.g., Schreuder and Baayen, 1997; Alegre & Gornon, 1999). Further, we carried out analyses on log frequency (natural log) rather than absolute frequency. It has been found that log frequency is more linearly related to recognition times while the relationship with absolute frequency is nonlinear (Schreuder and Baayen, 1997, p. 122).

Figure 1 below shows that word-form frequency measures from the Setswana corpus align well with results from the subjective frequency rating survey ($r = .75$; $p < .001$).

Figure 1: Scatterplot with Regression Line for Logged Ratings ~ Logged Word-Form Frequency



Overall, words that have a high corpus frequency were highly rated by participants, while low corpus frequency words were given low ratings by survey participants. This suggests that the corpus frequencies match native speaker intuitions.

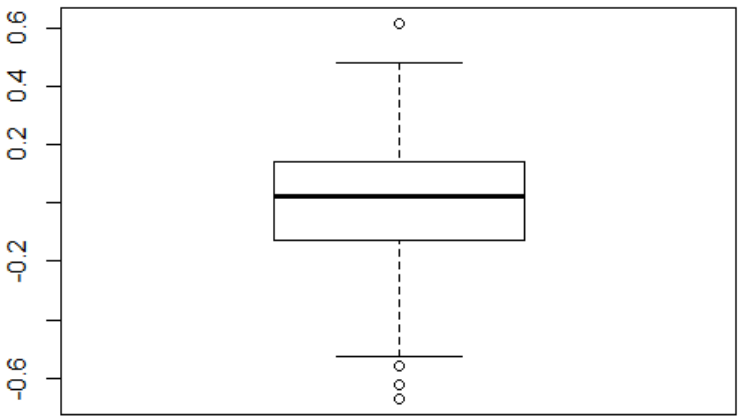
However, a visual inspection of Figure 1 above shows that there were discrepancies between corpus word-form frequencies and survey ratings for some items, where some items were rated highly in the survey even though they have a low corpus frequency – and vice versa. Therefore, it was worth checking if there were outlier items. One way of measuring deviations is via residualisation. Residuals are the differences between the observed and expected values (Baayen, 2008; Field, Miles & Field, 2012). In other words, residuals indicate how incorrect study predictions are when researchers make predictions regarding the relationship between variables. They show how far each observation on a scatterplot with a regression line is from the fitted regression line vertically (Y-axis). Deviations can be in two directions: negative or positive. For instance, residuals

may be higher or lower than expected given the frequency of the corpus. Observations with a larger residual value are further from the fitted regression line than those with a smaller residual value.

To test which items show a divergence between corpus frequency and psycholinguistic measures, and are hence not captured well by the regression model, a linear model with corpus-based word-form frequencies as a predictor for subjective frequency ratings was fitted. A linear model specifies a linear relationship between a dependent variable (e.g., reaction time) and independent variables (e.g., word-form frequency or word length). In an illustration of effects in regression analyses, the intercept is where the regression line crosses the y-axis. The fixed effects of the independent variables specify the change to the main intercept for the given factor. Also, of relevance in regression modelling are 'slopes': i.e., the distance that the regression line covers on the x-axis for a given change on the y-axis (see Baayen, 2008; Balling, 2008a, 2008b).

For the current analysis, a regression model was fitted, and residuals were calculated to determine deviations from the regression line. The residuals were plotted as in Figure 2 below:

Figure 2: Boxplot showing discrepancies between Survey Ratings and Corpus Word-Form Frequencies



A positive residual value (at the top of the boxplot) indicates that the survey frequency rating was higher than expected given a lower corpus frequency (see Table 2 below). In contrast, negative residuals (seen at the bottom of the boxplot) show that the survey rating was lower than expected given a high corpus frequency. Therefore, there were more items (3) with a low survey rating even though they have a high corpus word-form frequency. From the boxplot above, items with the shown residual values were identified. These items are shown in Table 2 below.

Table 2: Outlier items for Survey Ratings & Word-Form Frequency shown with their residual value

Item	Gloss	Item Class	Residual value	Explanation
tshebo	'gossiping'	Class 9	6.16e-01	Mostly colloquial
morwesi	'one who puts on something on the head or foot (old ritual for infants)'	Class 1	-6.71e-01	The frequency is of personal names
mogorogi	'one who arrives'	Class 1	-6.21e-01	The frequency is of personal names
thago	'manner of kicking, skill in kicking'	Class 9	-5.59e-01	Restricted to sports reporting

None of the deviant items is a base verb. Only class 1 and class 9 items attracted outlier residual values. Correlations between survey ratings and corpus word-form frequency were re-run to see how filtering out the deviant items above affects the relationship between the two variables. With the four outlier items removed, the overall correlation coefficient between survey ratings and word-form frequency improved: $r = .78$, $p < .001$ (compared to the $r = .75$, $p = .000$ before residualising and removing deviant items). Further, correlations were split by item type to see where improvements were by morphological type (Table 3).

Table 3: Correlations of logged Corpus Word-Form Frequency and Logged Survey Ratings split by morphological type without deviant items

Item Class	r-value and p-value
Base verb	.506, $p < .001$
Class 1 nouns	.716, $p < .001$
Class 9 nouns	.787, $p < .001$

As noted in the table above, correlations for the base verbs remain the same as before residualisation as no items were removed from this item type. The correlations for the noun derivations in contrast have improved as the deviant items have been removed.

It is interesting to note that noun derivations have a better correlation than base verbs. To explain this pattern, a new hypothesis was formulated to explain how participants may rate a base form compared to a complex form. For complex forms, participants focus on the stimulus form whereas in rating the base form, more forms become activated, which may inflate the ratings for these base forms. Previous studies on nouns have reported a similar observation. Schreuder & Baayen (1997) found that the frequency of the unseen plural word-forms influences the recognition and ratings of simple nouns: Nouns that have high-frequency plurals were responded to faster in a lexical decision task and rated higher in a subjective frequency rating survey than nouns with a lower plural frequency. Taft (1979) reported similar results for English as did Baayen, Dijkstra & Schreuder (1997) for Dutch. It appears that for simple forms, reaction times and ratings activate more than the presented stimulus form while the opposite is not true for complex forms. This would explain why class 1 and class 9 derivations have better correlations than base verbs.

Means of each morphological type's residuals were calculated to see whether participants consistently rated base verbs higher than expected (Table 4).

Table 4: Means of residuals split by morphological type

Item Class	Mean
Base verb	0.05
Class 1 nouns	-0.02
Class 9 nouns	-0.01

The results in Table 4 confirm the prediction discussed above as the mean of the base verb residual is positive. This indicates that base verbs were in general rated higher than one would expect, unlike the noun derivations which have a negative mean residual.

2.2.5 *Discussion* Correlations from the subjective frequency rating survey show that frequency ratings are well-aligned with corpus frequency. That is, words that have a high frequency in the Setswana corpus were also ranked highly by participants in the subjective frequency rating survey. This indicates that the corpus frequency information is reliable as it tallies with the intuitions of participants. This therefore establishes the Setswana corpus as a useful tool for further psycholinguistic Setswana studies.

However, some words such as *tshebo* 'gossiping' show a higher frequency rating by participants while showing low word-form frequency in the corpus data. This item is mostly used colloquially and was not frequent in the corpus, which has data that is rather formal owing to its data sources. In day-to-day usage, the phrasal verb *go seba* 'to gossip' is more common. Similarly, some words were given low ratings by participants even though the corpus showed that they are highly frequent. *thago* 'manner of kicking, skill in kicking' is an example of such words. This word is often used in sports reporting in newspapers. Since the corpus has a substantial amount of newspaper text, this may explain the high frequency of this word. Participants may themselves not use this word in their day-to-day language despite it being highly frequent in the corpus, and therefore rated it lowly. Although *morwesi* and *mogorogi* are morphologically viable and exist in the corpus probably as old personal names, these are not words that native speakers derive for everyday use. This explains the disparity between the ratings and corpus frequencies. These deviations show interesting usage patterns and the diachronic state of the language.

Further, strong correlations between corpus word-form frequencies and survey ratings indicate that the scale used in the survey (1 'extremely low' to 7 'extremely high') and the way the question was posed were effective. Participants understood that what was required of them was to make judgements based on the word-form frequency and not the whole spectrum of the lexeme.

The frequency measures from the corpus analysis and the subjective frequency rating survey were taken as design variables for the next analysis in the lexical decision task discussed below.

2.3 Unprimed visual lexical decision task This experiment presents an unprimed lexical decision task measuring RTs and accuracy rates, analysed with linear mixed effects regression modelling using the statistical software package R (R Development Core Team, 2011).

2.3.1 *Participants* Eighty-three native Setswana speakers (27 males, 56 females; mean age: 25; age range: 20–41) took part in this task. All participants were both physically and mentally unimpaired and had normal or corrected-to-normal vision. They were all naive to the real aim of the experiment. Written consent was sought from all participants for taking part in the study.

Participants were recruited from staff and students at the University of Botswana in Gaborone, Botswana. They were paid an equivalence of £1 for taking part in the experiment. All participants

were Setswana native speakers, with 91.5% of the participants having studied Setswana as a subject in school. In addition, 46.6% of the participants indicated that they use Setswana every day in their day-to-day lives; 47.1% indicated sometimes and 6.3% said they rarely used Setswana in their day-to-day interactions. The study population was 66.6% female.

2.3.2 *Materials*

Experimental items

The same experimental items described in 2.1 above were used in the lexical decision experiment.

Measures were taken to randomise the items: Prior to designing the DMDX (Forster and Forster, 2003) file for each set, each set of items was randomised via www.random.org. The items were subjected to a Latin Square design with three presentation lists, with each variant of the lexical item (base verb, class 1, and class 9 nouns) appearing in one of the lists and each list containing the same number of base verbs, class 1 deverbative nouns, and class 9 deverbative nouns. This ensured that each participant saw only one form of an experimental item to avoid semantic associations between words. To avoid order effects and fatigue effects, the items were presented in pseudo-randomised order so that the order of items did not provide semantic priming by not having semantically related words and words from the same condition follow one another. Moreover, the list of items was presented in two different orders to groups of participants. This was done to avoid a bias to words appearing last in the task.

Each participant saw 60 experimental items in the lexical decision task as with the survey. The 60 items were made up of twenty words from each of the three conditions. Each group therefore saw all types of word forms, each word appeared in all three forms, distributed over the three presentation lists.

Fillers

The lexical decision task also included filler or distractor words, i.e., experimental items were interspersed with a list of items that were unrelated to the features under investigation. This increased the number of unrelated items in the study. Adding fillers prevents participants from becoming aware of the phenomenon being investigated (as this might bias their reaction).

The current experiment used two types of filler word-forms: word fillers and nonce word fillers. Fillers were identical for all groups of participants. Sixty nonce word fillers corresponding to the experimental items were used. The list of fillers consisted of twenty items with class 1 *mo-* prefix and the *-i* suffix. Another twenty items mirrored the base verbs and had the terminative verb final vowel *-a*. The last twenty mimicked class 9 nouns and ended with the suffix *-o*.

100 word-fillers were further used in the study to distract participants from the real items of the study. These word fillers comprised adjectives, base verbs, derived verbs, and simple nouns. In addition, a hundred non-word fillers corresponding to these fillers were created and used in the experiment. The created nonce words followed the phonotactics of Setswana, thus co-occurrence restrictions in the language were maintained. All the nonce words were controlled for number of letters and syllable length, varying between two to four syllables.

The same fillers were seen by all participants in the same order in the experiment. Stimuli were presented in a pseudo randomised order such that no more than three of a given type of item appeared after each other and also no more than three items that were either real words or nonce words followed one another in the presentation of the experiment. Six pseudo-randomised lists were made and presented to participants, such that participants saw words in different orders.

Each participant therefore saw 60 stimulus items made up of twenty items from the three conditions, together with 260 fillers.

2.3.3 Procedure

Experimental Procedure

For the visual lexical decision task, participants were visually presented with a mixture of real words and nonce words on a computer screen, one stimulus item at a time. Of interest to the study was the response times of the participants in deciding about the stimuli, together with error rates in correctly identifying stimuli as a word or not.

The experimental task was administered using the DMDX software (Forster and Forster, 2003). Participants were required to indicate whether the presented stimulus was a word or not, as quickly and as accurately as they could by pressing a button for either 'YES' or 'NO'. Participants were tested individually in a dedicated, quiet room.

The basic procedure of the experiment was explained to participants before they agreed to take part. To ensure that participants understood the task and applied instructions as intended, the experiment was preceded by a series of 10 practice trials which consisted of real and nonce words equally. The stimulus was presented in Calibri font 14. All target words were written in lower case and centred on the computer screen. A trial began with a visual ready signal ("+") presented on the screen for 500 milliseconds (ms) and then followed by the stimulus item which remained on the screen until the participant's response. Participants had 2000ms in which to respond before the programme timed out and moved onto the next item. The next trial was initiated 2000ms later. The entire experimental session lasted about 20 - 30 minutes.

Analysis Procedure

In the analysis of results, certain predictors of reaction time which are known to influence recognition times and affect processing in the mental lexicon such as frequency, word length, and cumulative frequency were noted.

The study items had two frequency measurements, one corpus-based measurement and one survey-based measurement, which were found to be highly correlated with each other. Both measurements were never included in the same model, but only the one that resulted in a converging model with a better fit than the other one. Analysis of Variance (ANOVA) could not be used to compare two models that contained the two different frequency measures as these models were not nested. Hence, Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) were used for comparisons when both models converged. The AIC and BIC are information-theoretic criteria of comparing models. They are used to evaluate the fit of a given model. AIC measures show how much variance is left unexplained by the model so that a lower AIC means a model is considered to be closer to the truth and that more variance is being accounted for. BIC assesses the overall fit of a model and allows the comparison of both nested and non-nested models (see e.g., Hastie et al., 2009; Cunnings, 2012).

Both word-form frequency and survey ratings predictor variables involved a scale and as such were centred. In other words, the predictor variables were rescaled by subtracting from each individual value of a predictor the predictor's overall mean (see Jaeger, 2011, p. 20). Centring thereby reduces collinearity (when predictors have strong correlations, see Belsley et al., 1980). As already noted, survey ratings and corpus frequencies are highly correlated. In an ideal instance, as Baayen (2008, p. 181) states, predictors should be orthogonal, that is, uncorrelated. If they are highly correlated, it becomes difficult to tease apart the explanatory values of such predictors.

RTs and percentage of correct responses by participants were taken as dependent variables in a mixed effects linear regression analysis. A distinction was made in the analysis between two types of predictors: design factors and additional predictors (covariates) listed in (b) above. The design factor was ITEM_FORM (base verbs, class 1, and class 9), SURFACE FREQUENCY. The maximal mixed linear models in the sense of Barr et al. (2013) were adopted. For the logged RTs, mixed linear regressions model were fitted with the *lmer* program of the *lme4* package (Bates, Maechler & Bolker, 2013) available in the R programming environment for statistical computing (R Development Core Team, 2013). For the accuracy data, mixed logit models which involved a binary correct-vs.-incorrect/timed-out distinction were fitted to the data (Jaeger, 2008; Barr et al., 2013). For both RT and accuracy models, the maximal converging random effect structure was used, with by-subject and by-item random intercepts and slopes for the design factor ITEM_FORM (Jaeger, 2008 and Barr et al., 2013).

Adding all or most of the additional predictors in (1) to the model with the design factors and the maximal converging random effect structure would not have been ideal as this would have led to non-convergence. Adding predictors incrementally was not a viable alternative as this has been shown to be anti-conservative (Harell, 2001). Hence, another method was used to determine which control predictors to include in the final model: a series of minimally extended models was first created by adding each of the additional control predictors on its own to the model with the design factors ITEM_FORM and SURFACE FREQUENCY and the maximal random effect structure. In the creation of the minimally extended models, by-item and by-subject random slopes for the design factors were included, but not for the additional predictors as including them led to non-convergence, and it is not essential to specify random effects for control predictors – at least when there are no systematic interactions (Barr et al., 2013, p. 33). The model that only involved the design factors and the maximal effect structure were then compared to each of the minimally extended models with one additional predictor, using pairwise ANOVAs.

As the models included random slopes, the pvals-function of R (that is, Markov Chain Monte Carlo (MCMC) sampling) could not be used to determine whether a given predictor had a significant effect in the final mixed models for RTs. Instead, the chi-squared statistic (χ2-value) are provided and the p–value from the ANOVA that compared (i) the full model with the design factor(s) and any additional predictors to (ii) a model in which the respective unresidualised predictor had been removed. This allowed the evaluation of the contribution of the respective predictor to the final model.

Due to the Latin Square design of the lexical decision experiment, each item was presented three ways with each participant seeing three forms of items, but each item only in one form (base verbs, class 1, and class 9 derivations). Therefore, ITEM_FORM is a within subject variable for both items and participants. Thus, ITEM_FORM was included in the slopes of both ITEM and SUBJECT in all analyses.

2.3.4 *Results* Prior to the analysis of lexical decision times, the data of two participants was removed. One of these two participants had high error rates and time-outs whereas the second participant had many RTs below 200ms. RTs below 200ms are taken as erroneous button presses in lexical decision tasks: these experiments require at least 200ms or more for visual uptake and response (Baayen, 2008). Consequently, the analyses include RTs ranging between 200ms – 2000ms. For the RT-analysis, the data of the remaining 81 participants was cleaned-up to remove incorrect (i.e., erroneous word/non-word responses) or RTs not within the time-out limit of 2000ms. These incorrect responses and timeouts accounted for 11% of the experimental items. All RTs were logarithmically transformed. Mean reaction times are shown in Table 5 below.

Table 5: Overall means and error rates for the visual lexical decision task

Item Form	Mean RT	SD	Error rates
Base Verbs	839.79	235.18	4.7%
Class 1 derivation	969.10	291.09	10.1%
Class 9 derivation	896.10	266.51	8.5%

As Table 5 shows, higher frequency items elicited shorter reaction times. Base verbs had the highest corpus frequencies and highest survey ratings. The lexical decision task reaction times reflect this as Base verbs have the shortest response times of all stimulus items.

As pointed out before, this study had two word-form frequency measures: one corpus-based and the other rating-based. We ran correlations between these measures and RT. The correlation between RT and survey ratings was $r = -0.33$, $p<.001$, while between RT and corpus frequency it was $r = -0.34$, $p<.001$. These negative correlations observed indicate that highly rated/frequent stimulus items elicited shorter reaction times.

Results given above indicate that there are two effects:

1) Word-form frequency effect: Items with a high word-form frequency (both corpus-based and rating-based) attract shorter RTs as shown by the figures above and the negative correlations.
2) Item form/type effect: Base verbs are reacted to faster than the two classes of deverbative nouns. Class 9 nouns have a shorter mean RT than class 1 deverbative nouns. Error rates as shown in Table 5 also show a similar pattern: Base verbs have the highest accuracy scores, followed by class 9 and class 1 nominalisations gaining the highest error rates. These classes differ and this difference might be due to frequency as their RTs and error rates correspond to their frequencies.

2.3.5 *Discussion* Mean RTs showed a frequency effect for all stimulus forms. Further, the modelling results show that the processing of both noun derivations is influenced by survey frequency ratings. Taken together, these results indicate that the frequency of the full form, that is, surface frequency, influences the processing of stimulus forms.

The results for verbs can be understood in two ways: In one way, a frequency effect for these items is rather unexpected. These items are regular. They are formed by a root and a final vowel (e.g., *dumel-a* ‘agree’; *bolay-a* ‘kill’). This is a regular pattern in the verbal morphology of Setswana. The final vowel *-a* is the default terminative vowel of verbs in Setswana. Previous studies such as Clahsen et al. (1997) and Clahsen et al. (2004) found no frequency effects for regular items.

However, some researchers argue that highly regular forms can be stored. Such words leave a strong memory trace due to their high frequency of occurrence. Alegre and Gordon (1999) refer to a 'threshold' in explaining lexical processing. They argue that words that have a frequency count of more than six in a million occurrences should exhibit a frequency effect due to forming a strong memory trace. Other researchers such as Wurm and Baayen (2002) have argued that this threshold is even lower than 6 occurrences in a million. Carteret (2003) argues for 8-9 per million. Base Verbs used in this study had a higher threshold, the minimum being 16 appearances in a million.

For both noun derivations, a surface frequency effect was also observed. This suggests that the combined occurrence of a stem and derivational affixes is stored and accessed as a whole unit. In other words, a surface frequency indicates that derived forms are stems with stored full word-form representations. This finding concurs with results observed in previous studies (e.g., Clahsen et al., 1997; Meunier & Segui, 1999). Clahsen et al. (2003) found a surface frequency in an unprimed lexical decision task with German diminutives and nominalisations.

3. General Discussion

The present study set out to evaluate the representativeness of the Setswana corpus. This was done by ascertaining whether subjective frequency ratings are well aligned with objective corpus word-form frequencies, and testing which of the two frequency measures is a better predictor of recognition time and accuracy: objective corpus frequencies or subjective survey frequency ratings (indicating the intuition of participants). Furthermore, this study answered the call made by Arppe et al. (2010) and Gilquin & Gries (2009) to use converging frequency measures. Frequencies sourced from a fairly new corpus database and those sourced through participant intuitions via a subjective frequency rating survey were analysed. Therefore, this study has demonstrated that it is possible to employ psycholinguistic methods to evaluate the representativeness of corpora.

A subjective frequency rating survey was conducted to ascertain the representativeness of the Setswana corpus data. Results from this experiment indicate that subjective frequency ratings are well aligned with the corpus word-form frequencies. This is in line with previous studies which have reported a high correlation between subjective ratings and corpus frequencies (e.g., Schreuder & Baayen, 1997). This strong correlation shows that participant intuitions align well with corpus frequencies. This result attests to the representativeness and reliability of the existing Setswana corpus. Therefore, this corpus may be used in further (psycho)linguistic studies (e.g., Kgolo & Eisenbeiss, 2015; Ciaccio, Kgolo, & Clahsen, 2020).

There were points of divergence between the subjective ratings and the corpus frequencies, where participants rated an item low while it was of high frequency in the corpus, or vice versa. These points of divergence can be informative for corpus creators and researchers interested in language change or sociolinguistic variation.

In addition, the present study used another behavioural measure, in addition to the participant survey, the unprimed visual lexical decision task. In the analysis of reaction times, survey ratings were found to be a better fit to RT in mixed models. This makes a methodological contribution as results illustrate that in the absence of established corpus databases, participant intuitions can be used. This observation concurs with Balota, Pilotti & Cortese (2001) who also found that native speakers can reliably estimate frequencies of words. Several previous studies have also reported that subjective frequency or familiarity outperform corpus frequency in predicting lexical processing (Connine et al., 1990; Williams & Morris, 2004). Such results are encouraging for linguistic work

on languages that are under-resourced. Running rating or familiarity tasks is low cost as no special software is required.

Furthermore, a frequency effect was observed in the lexical decision task. This means that words that participants encounter and use more frequently were responded to and recognised faster than less frequent words. The word frequency effect is taken to indicate that language processing and representation is susceptible to statistical properties of linguistic experience (Davis, van Casteren, and Marslen-Wilson, 2003).

4. Conclusion

This paper has demonstrated, with converging evidence, that the existing Setswana corpus is representative of the Setswana language. We have thus provided corpus-external evidence that validates the corpus data, as it tallies with native speaker intuitions. Another way of looking at these results is to say that they demonstrate that speaker intuitions in general are accurate in predicting corpus frequencies. Given the importance of word frequency to many linguistic fields (e.g., language processing), having measures for ascertaining frequency is quite pertinent.

References

Alegre, Maria. & Gordon, Peter. 1997. Rule-based versus associative processes in derivational morphology. Brain and Language, 68, 347-354. https://doi.org/10.1006/brln.1999.2066

Arppe, Antti., Gilquin, Gaëtanelle, Hilpet, Martin, & Zeschel, Arne 2010. Cognitive Corpus Linguistics: Five Points of Debate on Current Theory and Methodology, Corpora, 5: 1-2. https://doi.org/10.3366/cor.2010.0001

Baayen, R. Harald., Dijkstra, Ton., and Schreuder, Robert. 1997. Singulars and plurals in Dutch: Evidence for a parallel dual route model. Journal of Memory and Language, 37, 94–117. https://doi.org/10.1006/jmla.1997.2509

Balling, Laura Winther. 2008a. Morphological Effects in Danish Auditory Word Recognition. Ph.d.-afhandling, Aarhus Universitet.

Balling, Laura Winther. 2008b. A brief introduction to regression designs and mixed-effects modelling by a recent convert. Copenhagen Studies in Language 36, pp. 175-192.

Balota, David A., Pilotti, Maura, & Cortese, Michael J. 2001. Subjective frequency estimates for 2,938 monosyllabic words. Memory & Cognition, 29, 639-647. https://doi.org/10.3758/BF03200465

Balota, David A., Cortese, Michael J., Sergent-Marshall, Susan, Spieler, Daniel H., & Yap, Melvin J. 2004. Visual Word Recognition of Single-Syllable Words, Journal of Experimental Psychology: General, 133(2), 283-316. https://doi.org/10.1037/0096-3445.133.2.283

Balota, Daniel A., Yap, Melvin J., & Cortese, Michael J. 2006. Visual word recognition: The journey from features to meaning (A travel update). In M. Traxler & M. A. Gernsbacher (Eds). Handbook of psycholinguistics (2nd edition).Bane, M. 2008. Quantifying and Measuring Morphological Complexity. Proceedings of the 26th West Coast Conference on Formal Linguistics, 69-76.

Barr, Dale J., Levy, Rodger, Scheepers, Christoph, Tily, Harry J. 2013. Random effects structure in mixed-effects models: Keep it maximal. Journal of Memory and Language. 59, 457–474. doi: 10.1016/j.jml.2012.11.001

Bates, Douglas, Maechler, Martin, & Bolker, Benjamin. 2013. lme4: Linear Mixed-effects Models using S4 Classes. R Package Version 0.999999-2, URL http://lme4.r-forge.r-project.org/

Belsley, David A., Kuh, Edwin and Welsch, Roy E. 1980: Regression Diagnostics. Identifying Influential Data and Sources of Collinearity. New York: John Wiley & Sons.

Burnard, Lou. 1995 (ed). Users' Reference Guide for the British National Corpus, version 1.0. Oxford: Oxford University Computing Services.

Carteret, Cathie. 2003. Regular and Irregular verb inflection in the French Mental Lexicon: A Dual-Mechanism perspective. Doctoral Thesis, University of Essex, Colchester, UK.

Ciaccio, Laura A., Kgolo, Naledi. & Clahsen, Harald. 2020. Morphological decomposition in Bantu: a masked priming study on Setswana prefixation, Language, Cognition and Neuroscience, 35:10, 1257-1271. https://doi.org/10.1080/23273798.2020.1722847

Clahsen, Harald, Eisenbeiss, Sonja., and Sonnenstuhl, Ingrid. 1997. Morphological structure and the processing of inflected words. Theoretical Linguistics, 23(3), 201–249. https://doi.org/10.1515/9783110197815.355

Clahsen, Harald, Sonnenstuhl, Ingrid & Blevins, James P. 2003. Derivational Morphology in the German Mental Lexicon: A Dual-Mechanism Account. In: Baayen, H. & Schreuder, R. (Eds.). Morphological Structure in Language Processing. Mouton De Gruyter: Berlin, pp. 125-155. DOI:10.1515/9783110910186.125

Clahsen, Harald., Hadler, Meike, and Weyerts, Helga. 2004. Speeded production of inflected words in children and adults. Journal of Child Language, 31, 687–712. https://doi.org/10.1017/s0305000904006506

Connine, Cynthia M., Mullennix, John, Shernoff, Eve, & Yelen, Jennifer. 1990. Word Familiarity and Frequency in Visual and Auditory Word Recognition. Journal of Experimental Psychology: Learning Memory Cognition, 16 (6), 1084–1096.

Cunnings, Ian. 2012 An overview of mixed-effects statistical models for second language researchers. Second Language Research, 28 (3). pp. 369-382. https://doi.org/10.1177/0267658312443651

Davis, Matthew H., Van Casteren, Maarten & Marslen-Wilson, William D. 2003. Frequency Effects In Processing Inflected Dutch Nouns: A Distributed Connectionist Account. Trends In Linguistics Studies And Monographs, 151, 427-462.

Ford, Michael, Marslen-Wilson, William, & Davis, Matthew. 2003. Morphology and frequency: Contrasting methodologies. In R. Baayen & R. Schreuder (Ed.), Morphological Structure in Language Processing (pp. 89-124). Berlin, New York: De Gruyter Mouton.

Forster, Kenneth I. and Forster, Jonathan C. 2003. DMDX: A Windows Display with Millisecond Accuracy. Behaviour Research Methods, Instruments, and Computers, 35, 116-124.

Francom, Jerid, LaCross, Amy, & Ussishkin, Adam. 2010. How specialized are specialized corpora? Behavioral evaluation for corpus representativeness for Maltese. Proceedings: Language Resource Evaluation Conference, Malta.

Gilquin, Gaëtanelle & Gries, Stefan Th. 2009. Corpora and Exeprimental Methods: A State of the Art Review. Corpus Linguistics & Linguistic Theory, 5: 1-26. https://doi.org/10.1515/CLLT.2009.001

Gries, Stefan Th. 2008. Dispersions and Adjusted Frequencies in Corpora. International Journal of Corpus Linguistics 13(4), 403–437. https://doi.org/10.1075/ijcl.13.4.02gri

Gries, Stefan Th. 2009. What is Corpus Linguistics? Language and Linguistics Compass 3, 1–17. https://doi.org/10.1111/j.1749-818X.2009.00149.x

Harrell, Frank E. 2001. Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis. New York: Springer.

Kilgarriff, Adam & Grefenstette, Gregory. 2003. Introduction to the Special Issue on Web as Corpus. Computational Linguistics 29 (3):333-347. https://doi.org/10.1162/089120103322711569

Kgolo, Naledi & Eisenbeiss, Sonja. 2015 The role of morphological structure in the processing of complex forms: evidence from Setswana deverbative nouns, Language, Cognition and Neuroscience, 30:9, 1116-1133. https://doi.org/10.1080/23273798.2015.1053813

Mayberry, Rachel I., Hall, Matthew L., & Zvaigzne, Meghan T. 2013. Subjective frequency estimates for 432 signs in relation to age of ASL exposure, Behavioural Research Methods. 46, 526–539.

New, Boris, Brysbaert, Marc, Segui, Juan, Ferrand, Ludovic, & Rastle, Kathleen. 2004. The processing of singular and plural nouns in French and English. Journal of Memory & Language, 51, 568-585. https://doi.org/10.1016/j.jml.2004.06.010

Otlogetswe, Thapelo J. 2010. Setswana Sketch Engine Corpus. URL: http://www.sketchengine.co.uk/

Otlogetswe, Thapelo J. 2011. Text Variability Measures in Corpus Design for Setwana Lexicology. Cambridge: Cambridge Scholars Press.

Otlogetswe, Thapelo J. 2012. Tlhalosi ya Medi ya Setswana. Gaborone: Medi Publishing.

Meunier, Fanny and Segui, Juan. 1999 Frequency effects in auditory word recognition: The case of suffixed words. Journal of Memory and Language, 41 327-344. https://doi.org/10.1006/jmla.1999.2642

R Development Core Team. 2013. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing: http://www.R-project.org

Schreuder, Robert and Baayen, R. Harald. 1997. How complex simplex words can be. Journal of Memory and Language, 37, 118–139. https://doi.org/10.1006/jmla.1997.2510

Zevin, Jason D., & Seidenberg, Mark S. 2004. Age of acquisition effects in reading aloud: Tests of cumulative frequency and frequency trajectory. Memory & Cognition, 32, 31-38. https://doi.org/10.3758/BF03195818

Naledi Kgolo-Lotshwao
Department of English
University of Botswana
kgolon@ub.ac.bw

Thapelo J. Otlogetswe
Faculty of Humanities
University of Botswana
otlogets@ub.ac.bw