Katsuo Tamaoka*, Shaoyun Yu, Jingyi Zhang, Koji Miwa and Masatoshi Koizumi

Constructing a web-accessible lexical database for core Tongan vocabulary

https://doi.org/10.1515/flin-2024-2053
Received March 4, 2024; accepted October 17, 2024; published online December 11, 2024

Abstract: According to UNESCO. 2020. *World atlas of languages*, 3rd edn. United Nations Educational, Scientific and Cultural Organization. Available at: https://en.wal.unesco.org, the Tongan language is classified as 'potentially vulnerable.' This study aimed to identify a core Tongan vocabulary and develop a web-searchable database to support the preservation and accessibility of the language. Due to the lack of a large-scale corpus for calculating lexical frequencies, we utilized vocabulary from the *Students' English-Tongan and Tongan-English Dictionary* (Thompson, Richard H. & 'Ofa Thompson. 1992. *The students' English-Tongan and Tongan-English dictionary*. Tonga: Faletohi 'Otumotu Angl'ofa'), a resource commonly used in Tongan schools. This dictionary compiles words used in both Tongan and English in everyday life. By examining processing speed and accuracy rates in lexical decision tasks, we identified a core Tongan vocabulary. Lexical decision tasks conducted on 4,013 words from the dictionary demonstrated a mean accuracy of 95.40 % and a rapid mean reaction time of 575 ms. A comparative analysis with an equal number of high-frequency words in English (Balota, David A., Melvin J. Yap, Michael J. Cortese, Keith A. Hutchison, Brett Kessler, Bjorn Loftis, James H. Neely, Douglas L. Nelson, Greg B. Simpson & Rebecca Treiman. 2007. The English lexicon project. *Behavior Research Methods* 39. 445–459) and French (Ferrand, Ludovic, Boris New, Marc Brysbaert, Emmanuel Keuleers, Patrick Bonin, Alain Méot, Maria Augustinova & Christophe

***Corresponding author: Katsuo Tamaoka**, Ph.D. & Prof, School of Foreign Languages, Shanghai University, 99 Shangda Road, Baoshan, Shanghai, 200444, China; and Graduate School of Humanities, Nagoya University, Furo-cho, Chikusa-ku, Nagoya-shi, Aichi-ken, 464-8601, Japan, E-mail: ktamaoka@gc4.so-net.ne.jp. https://orcid.org/0000-0003-1192-0358
Shaoyun Yu, Hong Kong Polytechnic University, Hong Kong, China, E-mail: shaoyun.yu@polyu.edu.hk. https://orcid.org/0000-0003-1903-0640
Jingyi Zhang, University of Miyazaki, Miyazaki, Japan, E-mail: jingyizhang@cc.miyazaki-u.ac.jp. https://orcid.org/0009-0006-7322-5632
Koji Miwa, Nagoya University, Nagoya, Japan, E-mail: kojimiwa@nagoya-u.jp. https://orcid.org/0000-0002-0890-9265
Masatoshi Koizumi, Tohoku University, Sendai, Japan, E-mail: koizumi@tohoku.ac.jp. https://orcid.org/0000-0003-0719-7128

Pallier. 2010. The French Lexicon Project: Lexical decision data for 38,840 French words and 38,840 pseudowords. *Behavior Research Methods* 42. 488–496) revealed that Tongan words were processed more quickly, though with an average accuracy approximately 2 % lower than the English and French words. Despite this slight difference in accuracy, the overall comprehension level for Tongan words remained high, with accuracy exceeding 95 %. Given the faster processing speeds observed, we propose that these words represent a core vocabulary in Tongan. To facilitate access to this vocabulary, we developed a web-based search engine (https://tonganlex.vercel.app/), which provides detailed information on letter (grapheme) count, phonemes, moras, parts of speech, meanings, as well as lexical decision times and accuracies. This study contributes to the accessibility of the Tongan language by making essential lexical information readily available online.

Keywords: Tongan language; core Tongan vocabulary; lexical decision; web-accessible search engine; lexical processing

1 Introduction

The Tongan language, an Austronesian language, is spoken in the Kingdom of Tonga, a Polynesian nation located in the South Pacific. Beyond its borders, Tongan is also spoken by migrant Tongan communities dispersed across Australia, New Zealand, the United States, and various other countries. The Kingdom of Tonga encompasses 171 islands with a population of 104,108 as of August 2024, with approximately 79 % of the population residing on the principal island of Tongatapu. Despite its rich cultural significance and widespread usage, Tongan faces the classification of 'potentially vulnerable' according to the World Atlas of Languages (UNESCO 2020), which encompasses 8,324 languages. Tongan serves as a vital component of Tongan culture and identity, playing an integral role in various cultural practices. The designation of Tongan as 'potentially vulnerable' underscores the necessity of implementing measures to protection and promote the language's vitality and continued usage.

Core vocabulary lists are essential tools in identifying commonly used words that frequently appear across various texts (Carter 2012; Hunt and Beglar 2005; Nation 2001; Nation and Waring 1997). These lists can be directly utilized by language learners or assist teachers and textbook authors in selecting materials that are suitable for specific groups of students. Typically, large-scale corpora are used to identify core vocabulary based on word frequency occurrences. For example, Brezina and Gablasova (2015) identified a core English vocabulary consisting of 2,494 words, known as the General Service List, which covers between 80.1 % and 81.7 % of the text in the source corpora. However, Tongan lacks such a large corpus.

Therefore, in the present study, we used lexical decision speeds and accuracies as a substitute method to identify core vocabulary in Tongan.

The first study aimed to ascertain whether the Tongan language serves as the primary language for Tongan individuals. In many cases, speakers of minority languages integrate major languages such as English, Spanish, or French into their daily lives, with these dominant languages often serving as the primary medium of instruction in educational settings. In Tonga specifically, English serves as the foundational language of instruction from elementary through high school, extending to university education as well. Consequently, the Tongan language may transition into a secondary role, with various levels of language processing – such as phonological, lexical, phrasal, sentential, and pragmatic – becoming increasingly influenced by English. Thus, the first study employed a questionnaire survey (see Appendix A) targeting native Tongan speakers to explore their self-perceived frequency of use in daily life and perceived proficiency levels in both Tongan and English languages.

The second study aimed to determine whether the selected words for lexical decisions constitute fundamental "core vocabulary" in the Tongan language. The core vocabulary refers to the set of words that occur most frequently across a wide range of texts and contexts, making them essential for basic communication within a language. The concept of core vocabulary is rooted in Zipf's law (Zipf 1932, 1935), which suggests that a small number of highly frequent words account for a large portion of any written or spoken text. Specifically, the top 2,000 most frequently used words in a language typically make up approximately 80 % of most texts (Ferrer-i-Cancho 2005; Linders and Louwerse 2023; Piantadosi 2014). In many of the world's major languages, such as Chinese, English, French, Japanese, and Spanish, large corpora have been developed, allowing for the identification of the frequently-used core vocabulary using the large corpora. However, for minority languages like Tongan, the lack of a large corpus makes it challenging to establish a comprehensive list of word frequencies. In this study, we addressed this limitation by drawing on nearly all entries in the basic Tongan dictionary, *the Students' English-Tongan and Tongan-English Dictionary* (Thompson and Thompson 1992), which is widely used by students up to the high school level. Although Tongan has a smaller vocabulary size compared to major world languages (Kongaika and Fisher 2019; Thompson and Thompson 1992; Tu'Inukuafe and Holding 1992), and this dictionary includes a limited number of lexical items, it remains a primary educational resource in Tonga.

The significance of frequently used vocabulary lies in its impact on processing efficiency. Various studies have shown that words with higher frequency are processed more quickly – a phenomenon known as the "word frequency effect" (Barry et al. 2001; Brysbaert et al. 2011, 2016, 2017; 2018; Coltheart et al. 2001; Cop et al. 2015; Hino and Lupker 1998; Kuperman and van Dyke 2013; Monaghan et al. 2017;

Monsell et al. 1989; Taft 1979; Yonelinas 2002). We adopted this psychological phenomenon as an indicator of core fundamental words. To validate the Tongan words as core vocabulary, we employed a lexical decision task to measure the accuracy and speed of processing these words. Furthermore, an equivalent number of words were selected from the English Lexicon Project (Balota et al. 2007) and the French Lexicon Project (Ferrand et al. 2010) based on their highest frequencies. We hypothesized that if no significant differences in lexical processing efficiency (accuracy and speed) were observed between Tongan and English/French, these words could be identified as part of the fundamental Tongan core vocabulary. The identification of these Tongan core words holds promise for investigating phonetic, lexical, phrasal, or sentential processing in both monolingual and bilingual contexts.

The third study culminated in the development of a web-accessible search engine dedicated to the core 4,013 Tongan words (https://tonganlex.vercel.app/). This innovative platform integrates crucial linguistic features such as processing times and accuracies for the lexical decision task, word meanings, parts of speech, and counts of graphemes, phonemes, and moras for each word. Furthermore, it provides a detailed analysis of reaction times for lexical decisions based on the aforementioned linguistic parameters. One of the distinctive linguistic traits of the Tongan language is its verb-initial order, wherein the typical sentence structure follows a pattern of verb, subject, and object (VSO) (Churchward 1953; Custis 2004; Dixon 1979, 1994; Otsuka 2000, 2005a, 2005b). This characteristic sets Tongan apart linguistically, as noted in the World Atlas of Language Structures Online by Dryer and Haspelmath (2013), which highlights that only about 9 % of the 1,377 languages surveyed exhibit verb-initial order, including VSO (95 languages) and VOS (25 languages) structures. Traditional syntactic studies have predominantly focused on subject-first languages, particularly those with SVO or SOV orders. Consequently, the verb-initial structure of Tongan offers a unique opportunity to explore syntactic theory and syntactic processing in less commonly studied VSO languages. Using the Tongan lexicon database, researchers in psycholinguistic studies (e.g., Kiyama et al. 2013, 2016; Koizumi et al. 2014; Koizumi et al. 2020; Tamaoka et al. 2024; Yano et al. 2017) can control Tongan lexical stimuli for constructing sentences to run processing experiments. Additionally, Tongan language features significant phonetic characteristics, such as vowel sequences (Anderson and Otsuka 2006; Taumoefolau 1998, 2002), primary and secondary stress on vowels (Garellek and White 2015), intonation (Kuo and Vicenik 2012), phonological encoding (Tamaoka et al. 2022), and the use of the definitive accent (Garellek and Tabain 2020).

Given the relatively limited research conducted on verb-initial languages and the definitive accent, the Tongan word search engine represents a significant contribution towards understanding and analyzing Tongan lexical characteristics. Although the search engine itself does not directly address the definitive accent

and verb-initial order, the comprehensive database it provides is essential for constructing various psycholinguistic and bilingual studies. This database includes critical information on Tongan words, such as letter (grapheme) counts, phonemes, moras, parts of speech, meanings, and lexical decision times. By offering these resources, the search engine facilitates a deeper exploration of Tongan language structure and processing mechanisms, including the phonological aspects and the effects of its verb-initial order and the phonetic features, both of which are unique features of the language.

2 Study 1 – Tongan and English proficiency among native Tongan speakers

Tonga was under British protection from 1900 until 1970, after which it gained independence while remaining a member of the British Commonwealth. Despite growing up with Tongan as their mother tongue, many Tongan children receive a significant portion of their education in English, leading to a relatively high level of proficiency in English among Tongan speakers. The pervasive influence of globalization has further elevated the importance of English in Tonga, both socially and economically (Otsuka 2007). Thus, the first study employed a questionnaire survey targeting native Tongan speakers to assess their usage patterns and subjective proficiencies in both Tongan and English languages. The aim was to determine the percentage of language use and perceived proficiency levels among participants in these two languages.

2.1 Participants

All 48 native Tongan speakers were recruited from students at a government teacher training institute, the Tonga Institute of Higher Education, located in the capital city of Nuku‘alofa on the island of Tongatapu. These participants had completed 12 years of school education (graduated from senior high school) and were thus proficient in both Tongan and English. Therefore, while these participants are not fully representative of the general Tongan population, they reflect a more educated segment of Tongan speakers. It is important to note that at the time of the survey, these students were enrolled at this institute, which, in 2023, merged into the newly established Tonga National University in Nuku‘alofa. All participants received monetary compensation for their involvement and provided written informed consent. Informed consent was obtained from all participants. The questionnaire

survey was conducted in accordance with the Declaration of Helsinki. The actual survey questions are attached to the end of this paper (see Appendix A).

2.2 Survey questions and results

The survey queried participants about the frequency of daily use of Tongan and English, utilizing a scale ranging from 0 to 100 %. A one-way ANOVA (analysis of variance) with repeated measures was employed to analyze the daily use percentage of both languages. The results indicated that the mean percentage of Tongan use in daily life was 79.98 % (*SD* = 15.35 %), whereas the mean percentage for English use was 23.92 % (*SD* = 15.26 %). Notably, participants were asked about their language usage independently, thus the total did not sum to 100 %. This marked difference of 56.06 % between Tongan and English usage was statistically significant [$F(1, 47) = 138.91$, $p < 0.001$, $\eta_p^2 = 0.75$]. Table 1 provides a breakdown of the means and standard deviations.

Furthermore, subjective proficiency judgments for the four language skills of speaking, listening, reading, and writing in both Tongan and English were assessed using a 0-to-6 point scale (0 representing 'none' and 6 indicating 'very high'). Table 1 presents the means and standard deviations of these proficiency scores. A two-way ANOVA with repeated measures was conducted on the scores obtained from the 0-to-6 point scale, considering both the language difference (Tongan vs. English) and language skills (speaking, listening, reading, and writing). The analysis revealed a highly significant main effect of language difference between Tongan and English [$F(1, 47) = 56.211$, $p < 0.001$, $\eta_p^2 = 0.55$], indicating that overall proficiency in Tongan language skills (average of the four skills; $M = 5.52$, $SD = 0.50$) was perceived to be superior to overall proficiency in English language skills (average of the four skills; $M = 4.59$, $SD = 0.88$). Additionally, a significant main effect of language skills was observed [$F(3, 141) = 4.62$, $p < 0.01$, $\eta_p^2 = 0.09$]. Among the four skills, reading

Table 1: Means and standard deviations of used-percentages and the four skills.

Language	Used percentage		The four skills (0-to-6 point scale)							
			Speaking		Listening		Reading		Writing	
	M	SD	*M*	SD	*M*	SD	*M*	SD	*M*	SD
Tongan	79.98 %	15.35 %	5.60	0.64	5.50	0.74	5.56	0.60	5.31	0.78
English	23.92 %	15.26 %	4.38	0.91	4.58	1.09	4.81	0.98	4.60	1.11
Difference	56.06 %	***	1.22	***	0.92	***	0.75	***	0.71	***

***$p < 0.001$.

proficiency was rated the highest in both Tongan (M = 5.56, SD = 0.60) and English (M = 4.81, SD = 0.98). Conversely, writing proficiency was rated the lowest in Tongan (M = 5.31, SD = 0.78), while speaking proficiency was rated the lowest in English (M = 4.38, SD = 0.91). This pattern indicates a significant interaction between language difference and language skills [$F(3, 141) = 4.09$, $p < 0.01$, $\eta_p^2 = 0.08$], primarily driven by the larger discrepancy of 1.22 in speaking proficiency between Tongan (M = 5.60, SD = 0.64) and English (M = 4.38, SD = 0.91) compared to differences observed in other skills.

2.3 Discussion

The subjective assessments of daily language use and self-reported proficiency consistently indicated a high level of perceived proficiency in Tongan, contrasted with a lower perceived proficiency in English. While minority languages often experience a decline in daily functional use, supplanted by major languages, Tongan speakers, particularly those living in the Kingdom of Tonga, view Tongan as their primary language, as reflected in their self-perceived assessments.

3 Study 2 – Identifying the core Tongan vocabulary

The Students' English-Tongan and Tongan-English Dictionary (Thompson and Thompson 1992) serves as a primary educational resource for students in Tonga up to the high school level, albeit containing a limited number of lexical items. As part of this study, a lexical decision task was administered for the words included in this dictionary. Participants were tasked with determining whether presented sequences of graphemes constituted valid Tongan words. This task measured two key parameters: reaction time, representing the time taken from presentation to determination of word validity, and accuracy of response, indicating whether judgments were correct. Reaction time and accuracy rates were utilized to gauge the efficiency of Tongan lexical processing. To assess the efficacy of this processing, results from Tongan lexical decisions were compared with data from the highest frequency vocabulary in the English (Balota et al. 2007) and the French Lexicon Project (Ferrand et al. 2010).

In an English lexical decision task (Balota et al. 2007), a participant must decide whether a combination of letters forms an English word. For example, when 'cat' is presented, the participant responds YES because it is a real English word. Conversely,

when 'tca' is presented, the participant responds NO as it is not a real English word. The time taken from the presentation of the stimulus word to the participant's correct or incorrect decision is automatically recorded by the computer, and this is referred to as 'reaction time'. The result of the correct or incorrect judgment is also recorded by the computer. In studies of lexical decisions, errored items are excluded when averaging reaction times. Therefore, a high accuracy rate is crucial for obtaining correct reaction time measurements. The present study applied the same method for the lexical decision task to collect data on reaction times and accuracies as conducted by the English (Balota et al. 2007) and the French Lexicon Project (Ferrand et al. 2010). Using these comparative data, the study aimed to determine whether the 4,013 selected Tongan words were processed as efficiently (both accurately and speed) as their English and French counterparts. The findings were then used to ascertain whether these chosen Tongan words could be classified as fundamental core vocabulary in Tongan.

3.1 Participants

A total of 168 participants (109 females and 59 males) residing on the island of Tongatapu took part in the second study. Their ages ranged from 16 years and 11 months to 33 years and 4 months, with an average age of 25 years and 11 months and a standard deviation of 4 years and 8 months. All participants were high school graduates, including some students at the Tongan National University. Many were employed in various roles, such as in retail shops, restaurants (as waitstaff), hospitality, as receptionists, and as government officials. While the participants for Study 2 were drawn from a broader segment of the Tongan-speaking population, they still represent a more educated group. This selection was intentional, as we sought participants with a strong knowledge of Tongan vocabulary. All participants received monetary compensation for their involvement and provided written informed consent. Notably, participants from the first study did not take part in the second study. Informed consent was obtained from all participants, and the experiment was conducted in accordance with the Declaration of Helsinki.

3.2 Words and nonwords

A total of 4,256 commonly used Tongan words were selected from the *Students' English-Tongan and Tongan-English Dictionary* (Thompson and Thompson 1992). Additionally, 608 nonwords were generated for the study. These words were organized into seven lists, each comprising 608 real words (resulting in 4,256 real

words in total across all lists) and 608 nonwords (with identical nonwords included in each list). Every participant (totaling 168 individuals) completed the lexical decision task for 608 words and 608 nonwords from one of the seven lists. Each list had a minimum of 22 participants assigned to it. Furthermore, all 168 participants evaluated lexical decisions for all nonwords.

3.3 Procedure and apparatus

The experiment on Tongan lexical decisions followed the procedure outlined in Figure 1. Initially, participants were presented with an eye fixation marker ('+') at the center of the computer screen for 500 ms, followed by the appearance of a target word. Participants were instructed to determine whether the presented word existed in Tongan by pressing the YES key for "exists" or the NO key for "does not exist". For example, as depicted in Figure 1, the word *fale*, a noun in Tongan meaning 'house' or 'building', would prompt a YES response from the participant. Following the participant's response, the next trial commenced after a 200 ms interval. Participants were instructed to perform this lexical decision task with both speed and accuracy.

All stimulus words were presented in a random order for each participant. Each participant completed 16 blocks of lexical decisions, with each block containing 38 words and 38 nonwords, totaling 76 items per block. Therefore, across all 16 blocks, participants encountered a total of 1,216 items (16 blocks × 76 items). Following the completion of each block, participants were advised to take a minimum rest period of 30 s. Prior to the commencement of the actual experiment, participants completed a practice session consisting of 12 practice items. The experiment was conducted individually in a quiet room, with a native Tongan experimenter administering

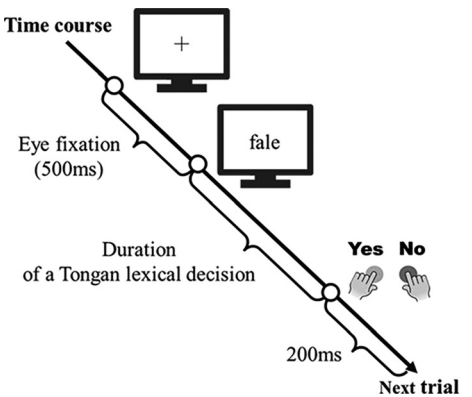


Figure 1: A single trial of Tongan lexical decisions.

the tasks face-to-face using the Tongan language. A computer connected to the experimental program Pavlovia (https://pavlovia.org/) was used to administer the experiment.

3.4 Results for lexical decision times and accuracy for the core Tongan vocabulary

Reaction times were calculated after excluding incorrectly responded items. If the error rate was too high, it compromised the accuracy of reaction time measurement. Therefore, a threshold of 70 % correct response rate was established. Following this criterion, 243 Tongan words with correct response rates below 70 % were excluded from the analysis, resulting in a final set of 4,013 words. The words with low accuracy rates included terms unfamiliar to the younger participants in their teens to early 30s, such as *hakisoa* ('hacksaw'), *sela* ('jailer'), and *haeane* ('clothes-iron'). Additionally, English loanwords expressed in Tongan, like *saoa* ('shower'), *sanuisi* ('sandwich'), and *hone* ('honey' or 'bee'), also had low accuracy rates. Since English is frequently used, native Tongan speakers often encounter these words in their original English spelling, leading to their misidentification when presented in Tongan spelling.

These remaining 4,013 words were considered as the candidate set for the core Tongan vocabulary. Similarly, 49 nonwords with correct response rates below 70 % were removed, leaving 559 nonwords for analysis. The average accuracy of the 168 participants for the final list of 4,013 Tongan words was 95.40 %, with a standard deviation of 6.33 %. Reaction times for lexical decisions regarding these 4,013 words ranged from 421 to 949 ms, with an average reaction time of 575 ms and a standard deviation of 68 ms. For the 559 nonwords, the average accuracy rate was 92.62 %, with a standard deviation of 6.49 %. Reaction times for nonwords ranged from 527 to 801 ms, with an average reaction time of 617 ms and a standard deviation of 47 ms.

To mitigate skewness in the distribution, reaction times were subjected to two transformations: natural logarithm (loge) and reciprocals (–1,000/rt, where rt represents reaction time). Quantile-quantile (QQ) plots illustrating these transformations are depicted in Figure 2. In a QQ plot, linearity of points indicates a normal distribution (Wilk and Gnanadesikan 1968). The raw reaction time data deviates from the identity line between sample quantiles and theoretical quantiles. Although data transformed to natural logarithm still displays slight deviation from the identity line, reciprocal transformation appears to exhibit better linearity, making it the preferred method for analyzing the reaction times of Tongan lexical decisions, as observed in Figure 2.

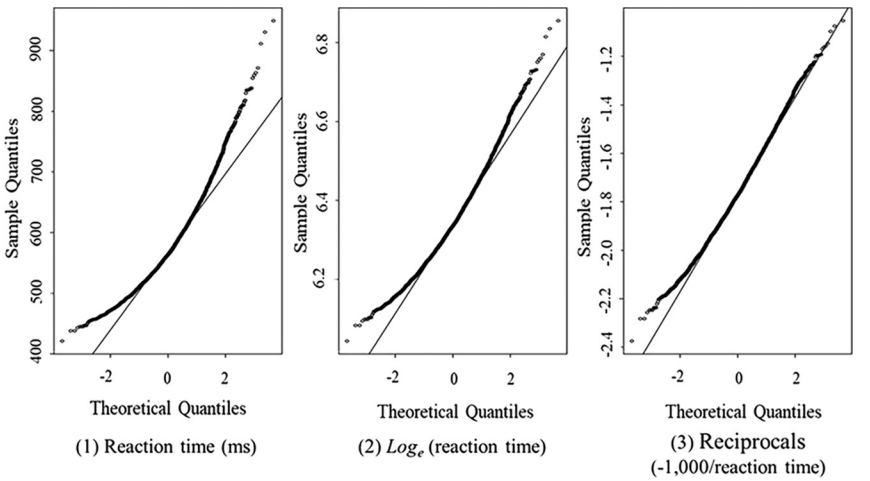


Figure 2: Quantile-quantile (QQ) plot of reaction times for 4,013 Tongan words.

The accuracy data exhibit significant skewness due to the large number of Tongan words correctly identified by 100 % (1.00) of the participants. The distribution of the accuracy data is negatively skewed (skewness = –1.83) and displays a steeper peak (kurtosis = 3.09). Therefore, the percentage of correct responses serves as a crucial criterion for comprehension in selecting stimuli from the core 4,013 Tongan words.

All 4,013 words were categorized into major parts of speech, with detailed numbers provided in Table 2. Some words were classified as multiple parts of speech. Tongan verbs have a relatively simple system of inflection compared to languages with more complex verbal morphology, such as European languages. Tongan verbs can be modified to express various grammatical categories, including tense, aspect, mood, and voice. For example, the present tense is often unmarked or marked with the particle *ʻoku*, as in *ʻoku ou lea* 'I am speaking'. The past tense is marked with the particle *naʻe*, as in *naʻe lea* 'spoke'. The future tense is marked with the particle *te*, as in *te u lea* 'I will speak'. The imperative mood is typically formed by using the verb stem without a subject pronoun, as in *lea!* 'Speak!'.

Reaction times and accuracies for the primary words limited to a single part of speech were as follows:

The mean processing time for 1,472 nouns was 577 ms (*SD* = 68 ms) and the mean accuracy was 95.02 % (*SD* = 6.52 %).

The mean processing time for 1,390 verbs was 575 ms (*SD* = 69 ms) and the mean accuracy was 95.82 % (*SD* = 6.10 %).

The mean processing time for 124 adjectives was 568 ms (*SD* = 76 ms) and the mean accuracy was 95.96 % (*SD* = 5.59 %).

Table 2: Numbers of parts of speech among 4,013 Tongan words.

Part of speech	Number
adjective	124
adjective/adverb	108
adjective/adverb/verb	2
adjective/noun	42
adjective/pronoun	1
adjective/verb	438
adverb	142
adverb/interrogative	1
adverb/noun	1
adverb/pronoun	2
adverb/verb	6
article	1
conjunction	14
conjunction/adverb	1
conjunction/pronoun	1
expression	16
interjection	18
interrogative	4
interrogative/adverb	1
noun	1,472
noun/verb	28
numeral	28
plural marker	1
prefix	2
preposition	12
preposition/demonstrative	3
pronoun	146
pronoun/adverb	1
quantifier	3
tense marker	4
verb	1,390
Total	4,013

The mean processing time for 142 adverbs was 568 ms (*SD* = 66 ms) and the mean accuracy was 96.00 % *(SD* = 6.19 %).

Nouns and verbs demonstrated slightly slower decision times compared to adjectives and adverbs, while accuracies remained consistent at approximately 95–96 %. Overall, there appeared to be no discernible difference in lexical decisions based on the part of speech when considering the means of reaction times and accuracies.

3.5 Comparing 4,013 Tongan words to their counterpart English words

Lexicon projects have been established for major languages, providing extensive data for investigating the cognitive mechanisms of lexical processing. The English Lexicon Project (Balota et al. 2007) initially comprised data for naming latencies and lexical decision times for 40,481 English words, alongside an equal number of nonwords. Subsequently, the English Lexicon Project (last updated on September 28, 2019) expanded to encompass data for 79,672 words (downloaded on November 30, 2021). Given the variation in word frequencies influenced by socio-cultural factors, it becomes challenging to select semantically matched words between English and Tongan. Therefore, instead of matching lexical meanings, we selected 4,013 English words, corresponding in number to the Tongan words, based on their highest frequencies.

Distributions of lexical decision times for 4,013 Tongan and English words were visualized using a density plot (refer to Figure 3). Density was computed using kernel density estimation (Sheather and Jones 1991). Reaction times for the 4,013 English words ranged from 489 to 921 ms, with an average of 636 ms (*SD* = 60 ms). In comparison, the average reaction time for the 4,013 Tongan words was 575 ms (*SD* = 68 ms), indicating that Tongan words were processed 61 ms faster than their English counterparts.

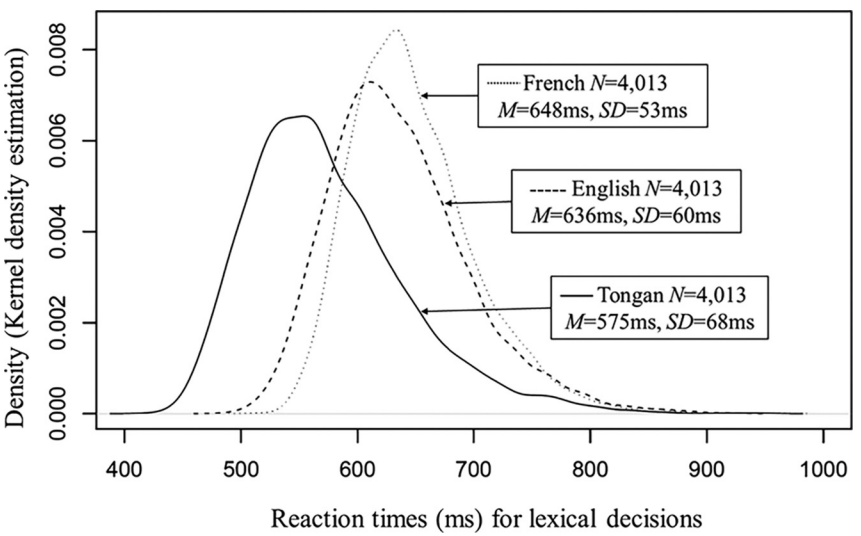


Figure 3: Density plots of lexical decision times.

The present study encompassed nearly all words (i.e., 4,256 words) found in the fundamental dictionary utilized by students up to Grade 12 (Thompson and Thompson 1992). Out of these 4,256 words, 243 exhibited an accuracy rate of less than 70 %, leading to their exclusion from the database of core Tongan words. These 243 words could be regarded as challenging 'distractors' for YES (correct) responses. However, the reduced number of challenging distractors in the Tongan experiment may have accelerated lexical decisions compared to the English Lexicon Projects (Balota et al. 2007), which encompass a larger number of words, including rare ones.

Additionally, unlike English, the Tongan vocabulary demonstrates regularity and consistent pronunciation and spelling. The adoption of the alphabet to phonetically represent Tongan is closely tied to the arrival of European missionaries in the early 19th century. Before European contact, Tongan was exclusively an oral language with no indigenous writing system. European missionaries, primarily from the United Kingdom, saw the need to create a written form of Tongan to translate the Bible. The first Tongan Bible, translated by missionaries, was published in 1862, setting the use of the alphabet for the Tongan language. The Tongan alphabet is simple and phonetic, with each letter corresponding to a specific sound. It consists of 16 letters: five vowels (a, e, i, o, u) and 11 consonants (f, h, k, l, m, n, p, s, t, v, ʻ). The letter or symbol ʻ, known as the *fakauʻa*, represents the glottal stop, an essential phoneme in Tongan. This orthography was designed to be easily accessible to native speakers and straightforward to learn. Today, the Tongan alphabet is firmly established as the standard writing system for the language.

Tongan has highly regular letter-to-sound correspondences, making it relatively easy to read and write once the basic rules are understood. For example, the word *talanoa*, meaning 'to talk,' consists of seven letters, each corresponding to a distinct phoneme. The word is pronounced /talanoa/, with each letter representing a clear and consistent sound. Additionally, the word can be broken down into four moras: /ta la no a/. This regularity is a key feature of Tongan, making it straightforward for learners to pronounce words correctly based on their spelling. This characteristic may have facilitated lexical decisions for Tongan, contributing to the efficiency of lexical processing.

To achieve a smoother graph, accuracies were rounded up to the nearest whole number. Subsequently, distributions of lexical decision accuracies for the 4,013 Tongan, English, and French words were depicted on the density plot shown in Figure 4. Accuracies for the English word counterparts ranged from 70.00 to 100 %, with an average accuracy of 97.38 % (*SD* = 3.54 %). In contrast, the average accuracy for the Tongan words was 95.40 % (*SD* = 6.33 %), indicating that English words were processed 1.98 % more accurately than Tongan words. Nonetheless, both Tongan and English words exhibited accuracies exceeding 95 %. Based on the comparison of

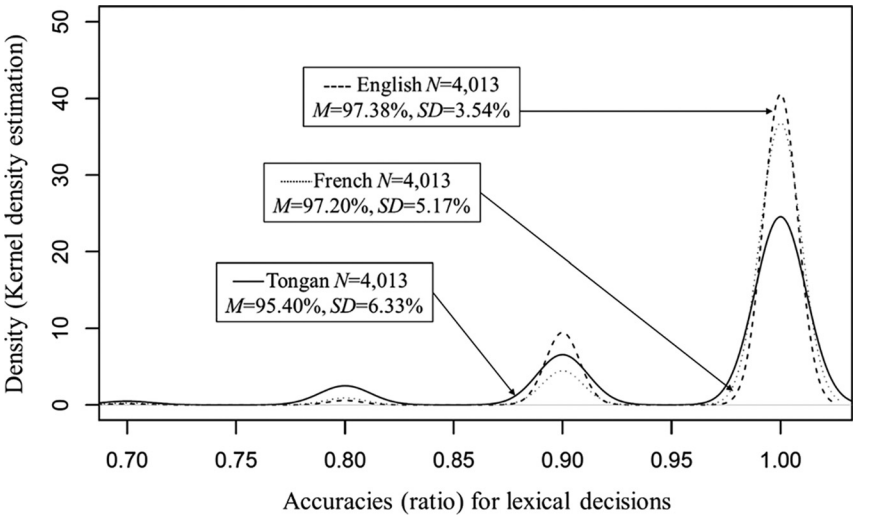


Figure 4: Density plots of lexical decision accuracies.

reaction times and accuracies with their English counterparts, we assert that the 4,013 Tongan words can be considered a core Tongan vocabulary.

3.6 Comparing the 4,013 Tongan words to counterpart French words

The French Lexicon Project (Ferrand et al. 2010) encompasses data for lexical decision times for 38,840 real words. Similar to the approach used for English words, 4,013 French words were selected based on their highest frequencies. Figure 3 illustrates the distribution of lexical decision times for the French, Tongan, and English words. Reaction times for these 4,013 French words ranged from 516 to 963 ms, with an average reaction time of 648 ms (*SD* = 53 ms). Given that the average reaction time in Tongan was 575 ms (*SD* = 68 ms), the Tongan words were processed 73 ms faster than their French counterparts. The distribution of lexical decision accuracies for the three languages is depicted in Figure 4. Accuracies for the French words ranged from 40.00 to 100 %, with an average of 97.20 % (*SD* = 5.17 %). In comparison, the average accuracy in Tongan was 95.40 %, indicating that the French words were processed 1.80 % more accurately than the Tongan words. Therefore, by comparing reaction times and accuracies with their French counterparts, the 4,013 Tongan words are considered to be the core vocabulary.

3.7 Comparing the number of letters to English and French words

Tongan is classified as a mora language, where each mora typically consists of a consonant and a vowel (i.e., a consonant-vowel, or CV, phonological structure) without a final consonant. In contrast, some languages, like Japanese, distinguish between moras and syllables, with the mora being a smaller unit. For example, the Japanese word for 'newspaper,' pronounced /sinbun/, is a two-syllable word divided into /shin/ and /bun/. However, the nasal sound /n/ is treated as a separate mora. Therefore, the word can be divided into four moras: /si/, /n/, /bu/, and /n/, and it consists of six phonemes: /s/, /i/, /n/, /b/, /u/, and /n/. As illustrated, syllables, moras, and phonemes represent distinct phonological units. In Tongan, the phonological structure is simpler, with each mora aligning closely with the CV structure. This makes the unit of mora in Tongan potentially smaller and more regular than the syllables found in languages like English and French.

Unlike Japanese, Tongan does not clearly differentiate between moras and syllables. Instead, Tongan primarily adheres to a consistent CV (consonant-vowel) structure, with no consonants appearing at the end of words (Garellek and Tabain 2020). This distinction is crucial for understanding phonological processing in Tongan, as it may influence how words are segmented and processed during lexical tasks (e.g., Kureta et al. 2006; O'Séaghdha et al. 2010; Tamaoka et al. 2022; Verdonschot et al. 2022). When searching for phonological similarities between English, French, and Tongan, it's important to note that the basic units differ: syllables in English and French versus moras in Tongan. As such, using graphemes – units that typically represent phonemes – provides a more appropriate basis for comparison. Therefore, comparing the number of letters (graphemes) used to spell words across Tongan, French, and English is a more standardized method for making such comparisons.

As depicted in Figure 5, the average number of letters (N = 4,013) was: 7.59 (SD = 2.96) in Tongan, 6.02 (SD = 1.90) in English, and 6.44 (SD = 1.95) in French. Given

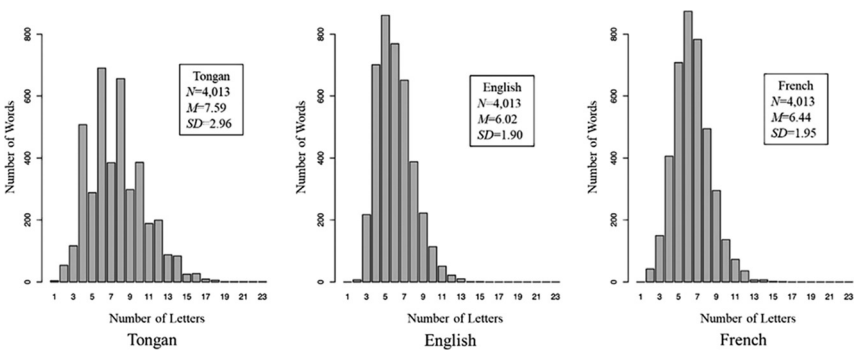


Figure 5: Distribution of letters spelling 4,013 Tongan, English and French words.

the large sample size, a one-way ANOVA yielded a significant main effect for the variable of languages (Tongan, French, and English) [$F(2, 12{,}036) = 495.00$, $p < 0.001$]. Even the small difference of 0.42 (6.44–6.02) between French and English was statistically significant [$t(8{,}024) = 9.88$, $p < 0.001$]. In summary, Tongan words comprised more letters than both English and French words, and French words also had more letters than English words. Additionally, Tongan words with an even number of letters were more frequent than those with an odd number of letters. This trend can be attributed to Tongan's fundamental CV mora structure, where a single mora is often represented by two letters. Consequently, the CV mora structure contributes to the higher frequency of even-numbered letters in Tongan words.

The introduction of new concepts from English that were previously nonexistent in Tongan may have necessitated translations using multiple or compound words. Consequently, what could be expressed with a single word in English or French might be conveyed in Tongan through the use of multiple or compound words, resulting in longer letter lengths. For instance, while the Tongan word *fale* represents a noun meaning 'house', the compound word *fale faiva* refers to a 'theatre'. In this compound, *faiva* signifies 'drama', 'concert', or 'film', expanding the meaning of *fale faiva* to 'drama/concert/film house'. However, it is noteworthy that Tongan lexical items incorporating such longer-length words were processed more rapidly for lexical decisions compared to their English or French counterparts. Therefore, considering differences in word length when comparing the efficiency of lexical processing between Tongan and English or French appears reasonable.

The study conducted an analysis using 4,013 words across three languages – Tongan, English, and French – predicting lexical decision times based on the length of the words using a linear regression analysis. The *R* command used was *lm*(–1,000/rt ~ language*letters, data), with Tongan set as the reference for the language variable. The results are reported in Table 3. As shown in Table 3, the linear regression analysis revealed significant interactions between language and letter

Table 3: The result of the linear regression model for reaction times.

Variables	Estimate	SE	*t* value	Pr(>|t|)	*p*
(Intercept)	−1.793	0.007	−273.357	2.00E−16	***
Langage English	−0.004	0.010	−0.409	0.682	
Langage French	0.175	0.011	16.629	2.00E−16	***
Letters	0.004	0.001	4.760	1.96E−06	***
Language English*Letters	0.031	0.001	20.921	2.00E−16	***
Language French*Letters	0.006	0.001	4.178	2.96E−05	***

***$p < 0.001$. SE, standard errors. Item total = 4,013. The regression model used was *lm*(–1,000/rt ~ language*letters, data). Tongan was set as the reference for the language variable.

length: specifically, interactions between Tongan and English [t = 20.921, p < 0.001] and between Tongan and French [t = 4.178, p < 0.001].

Figure 6 was created to visualize the interaction results. As shown in Figure 6, even when the number of letters increased, Tongan lexical decisions were made with only a slight delay. In contrast, for English lexical decisions, the increase in the number of letters caused a significant delay. For French, lexical decisions were slower than for Tongan at a consistent rate, but the change in delay was not as pronounced as in English. Tongan is a mora language with highly regular spelling and pronunciation. In contrast, English has many irregular words in both spelling and pronunciation. As the number of letters in English words increases, phonological information may contribute less to lexical decisions. Longer spellings in English may cause a significant delay in processing time because visual segments of longer words would take longer and accumulate delay for lexical processing. French, while more regular than English, still has less regularity between spelling and pronunciation than Tongan, leading to a slight increase in delay for lexical decisions compared to Tongan.

The Tongan lexical decision experiment adhered to the experimental methodology employed in the English (Balota et al. 2007) and the French Lexicon Project (Ferrand et al. 2010), ensuring consistency across the studies. However, we also recognize that differences may still arise due to variations in the experimental setup, the methods used to calculate reaction times, or the criteria for removing outliers. Despite these potential differences, the significant delay observed in English lexical decisions – particularly as word length increases – could suggest that this delay is

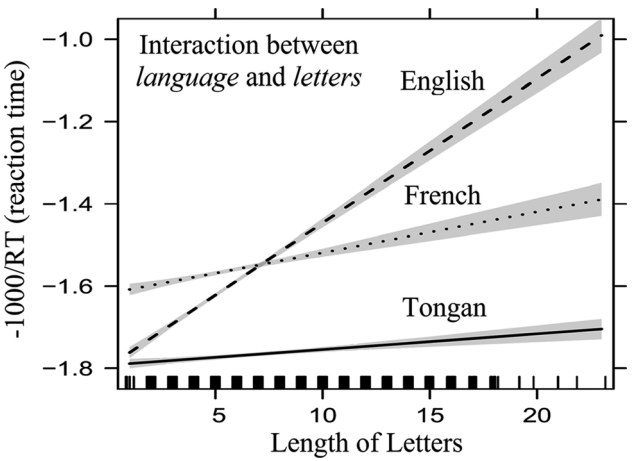


Figure 6: Regression lines by length of letters per language.

primarily attributable to the increased number of letters in English words, rather than to methodological discrepancies.

3.8 Correlations of letters, phonemes and moras among the 4,013 core Tongan words and prediction of reaction times

The phonological units of Tongan are detailed in Columns 4 and 5 of Section 4.1. Tongan's fundamental rhythmic unit comprises moras, which consist of regular combinations of consonants and vowels. Notably, the number of letters and phonemes in Tongan words are nearly identical. Specifically, the average number of letters was 7.59 (*SD* = 2.96), while the average number of phonemes was 7.63 (*SD* = 2.92), reflecting a negligible difference of only 0.04. Furthermore, the average number of moras was 4.17 (*SD* = 1.58), slightly smaller than the average number of

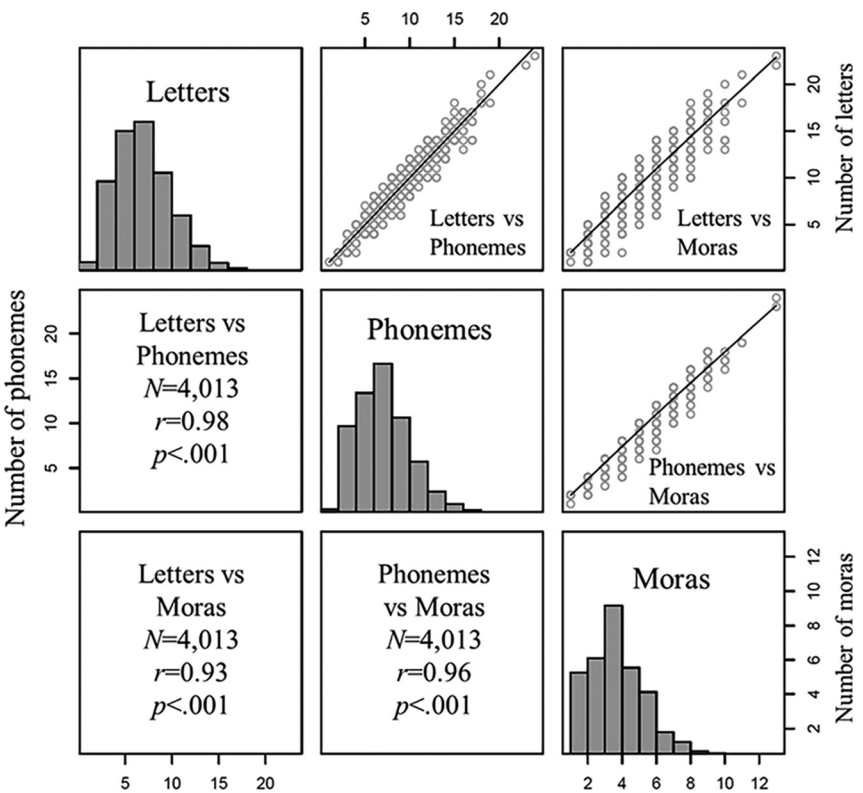


Figure 7: Distributions and correlations of letters, phonemes and moras.

letters and phonemes. Illustrated in Figure 7, the number of letters, phonemes, and moras among the 4,013 Tongan words exhibited exceptionally high correlations, with $r = 0.98$ ($p < 0.001$) between letters and phonemes, $r = 0.93$ ($p < 0.001$) between letters and moras, and $r = 0.96$ ($p < 0.001$) between phonemes and moras. These strong correlations indicate the consistent relationship between spelling and sounds.

Quadratic functions are particularly significant because they capture non-linear relationships that linear functions cannot. When the number of letters, phonemes, or moras is small, the distinction between words is less pronounced, leading to longer processing times. Figure 8 illustrates the predictions of reciprocal reaction times (–1,000/RT) for lexical decisions, showing that quadratic functions provide a better fit than linear functions for these variables. Words with fewer letters, phonemes, or moras exhibit longer reaction times compared to typical 6–8 phoneme words. Specifically, words with only 1–2 letters require more time for lexical processing than those with 6–7 letters, as fewer letters make it harder to distinguish between words. Processing speed initially increases as the number of letters rises from 1 to 7, but then reverses, with reaction times lengthening as the sequence extends to around 22 letters. Similar trends are observed for phonemes and moras. This non-linear pattern, which is crucial for understanding lexical decision speeds in Tongan, is effectively captured by quadratic functions.

In a regression analysis predicting reaction times using the number of letters (calculated by R statistical language, *lm* linear models), the quadratic function ($t = 3.951$, $p < 0.001$) was determined as follows: reaction times = $-1{,}000/(-1.731 - 0.013 \times \text{letters} + 0.001 \times \text{letters}^2 + e)$. For example, when the number of letters was 15, the reaction time was predicted to be 588 ms using the formula $-1{,}000/(-1.731 - 0.013 \times 15 + 0.001 \times 15^2)$. Similarly, in a regression analysis predicting

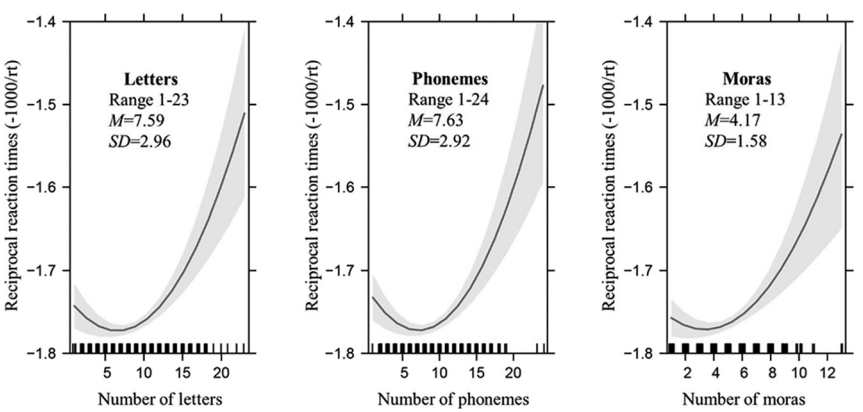


Figure 8: Reaction times predicted by the numbers of letters, phonemes and moras.

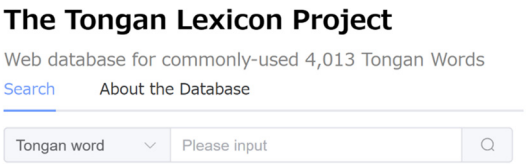
reaction times based on the number of phonemes, a quadratic function was determined (t = 4.176, p < 0.001), yielding the formula: reaction times = –1,000/(–1.719 – 0.015 × phonemes + 0.001 × phonemes^2 + e). Additionally, in a regression analysis predicting reaction times based on the number of moras, another quadratic function was determined (t = 4.176, p < 0.001), leading to the formula: reaction times = –1,000/(–1.743 – 0.017 × moras + 0.003 × phonemes^2 + e). Essentially, predictions of reaction times based on the number of letters, phonemes, and moras exhibited a similar trend regardless of the unit used.

3.9 Discussion

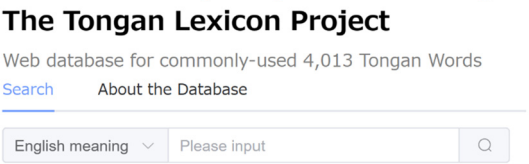
Due to the absence of a large-scale corpus for the Tongan language to calculate lexical frequencies for identifying a Tongan core vocabulary, this study utilized vocabulary selected from the *Students' English-Tongan and Tongan-English Dictionary* (Thompson and Thompson 1992), which is commonly used in Tongan schools. This dictionary compiles words used in both Tongan and English in everyday life. Using these words, the present study identified the basic core vocabulary by examining processing speed and accuracy rates in lexical decision tasks. Additionally, an equal number of 4,013 Tongan words were selected for comparison with their English and French counterparts. The results revealed that Tongan words were processed more quickly than the English and French words. However, the accuracy for the Tongan words was, on average, about 2 % lower than for the English and French words. Despite this, the average accuracy for the Tongan words remained above 95 %, indicating a high overall comprehension level. Given the faster processing speeds observed for the Tongan words, this study suggests that these words represent a core vocabulary in Tongan. Furthermore, this study has compiled a database of various characteristics of these 4,013 words, making it accessible online. The details of this database will be introduced in the following section.

4 Study 3 – Development of a web-accessible search engine for the core 4,013 Tongan words

Creating a web-accessible search engine (https://tonganlex.vercel.app/) to serve as a repository for the 4,013 core Tongan words, along with their parts of speech, meanings, and psycholinguistic data, provides a valuable resource for ongoing research endeavors. To achieve this objective, the third study was dedicated to developing a freely accessible search engine. This platform offers two search



(1) Search for related multiple Tongan words from a Tongan word



(2) Search for multiple Tongan words from an English meaning

Figure 9: Web-accessible search engine for 4,013 commonly-used Tongan words.

methods: users can input either a Tongan word or an English meaning. By utilizing the "Tongan word" search tab (as depicted in Figure 9), users can explore information related to identical words, as well as associated and compound words that share the same spelling. For instance, upon searching for *fale*, users can discover 47 lexical items such as *fale faiva* (theatre), *fakafaletolo* (veranda), and *fale motokā* (garage). All relevant items can be conveniently accessed by scrolling through the right sidebar.

It is also possible to search for an English meaning using the "English meaning" search tab. For example, entering 'house' in English yields 21 lexical items, including the search target word. Some examples of 'house' include *fale* 'house', *fale kai* 'restaurant', *mata fale* 'house front', and *fale paaki* 'printing-house'. Once again, users can view all items by utilizing the right side scroll bar. This dual search function proves beneficial for identifying intended Tongan words for use as stimulus items in new experiments focusing on word, phrase, and sentence processing.

Within the "About the Database" tab of the search engine (located next to "Search"), users can access detailed descriptions of the core Tongan words stored in the database. Additionally, the entire spreadsheet file containing these words can be obtained in the form of an online spreadsheet under "Downloads" at the top section of the "About the Database" page. This spreadsheet constitutes the original data of the web-accessible search engine. While users cannot modify the data of the 4,013 Tongan words in Google Sheets, the entire database can be downloaded as an Excel file, allowing users to manipulate the data as needed for their own purposes.

4.1 Data for the 4,013 core Tongan words

All lexical items presented in the web-accessible search engine and the Excel database of 4,013 words include the following information, organized into 10 columns.

Column 1: Tongan Words (Spelling): The first column displays Tongan words alphabetically.

Column 2: Dictionary Page: The words are sourced from *The Students' English-Tongan and Tongan-English Dictionary* by Richard H. Thompson and Thompson (1992), available at https://friendlyisles.press/books/. Each word's corresponding page number in the dictionary is recorded in this column. For instance, *afi* meaning 'fire' is listed on Page 71. English words are situated in the first section of the dictionary, while Tongan words are arranged in alphabetical order in the latter section. Consequently, words like *afi* appear on relatively later pages.

Column 3: Number of Letters: This column denotes the number of letters in each Tongan word, ranging from 1 to 23. The mean length is 7.59, with a standard deviation of 2.96. Tongan words are written in the Latin script, where each letter typically represents a phonemic unit. However, there are nuances to consider, particularly with long vowels, the spelling of *ng*, and the glottal plosive. First, long vowels, indicated by a macron (e.g., *ā* and *ō*), are written with a single letter but counted as two phonemes. For example, the English loanword *kāloti* ('carrot') is written with six letters but counts as seven phonemes (/k/, /a/, /a/, /l/, /o/, /t/, and /i/). Second, the spelling of ng is counted as two letters, although it represents a single consonant, the voiced velar nasal [ŋ]. For instance, the word *engeenga* ('yellow') is written with eight letters but contains six phonemes (/e/, /ng/, /e/, /e/, /ng/, and /a/) and four moras (/e/, /nge/, /e/, and /nga/). Third, the glottal plosive, indicated by a reversed apostrophe (ʻ) before a vowel, is considered a single consonant and counted as one letter. For example, the word *foʻi ʻakau* ('fruit') is counted as 9 letters and 9 phonemes (/f/, /o/, /ʻ/, /i/, /ʻ/, /a/, /k/, /a/, and /u/). Additionally, hyphens are not counted as letters. For instance, the word *anga-faingata'a* ('not easy-going' or 'difficult to approach or relate with') is hyphenated, but the hyphen itself is not considered a letter.

Column 4: Number of Phonemes: Tongan features a restricted set of phonemes, encompassing twelve consonants (p, m, f, v, t, s, n, k, l, ŋ, ʔ, and h) and five vowels (i, e, a, u, o; see Garellek and White 2015). The phonetic ʔ represents a glottal plosive (known as *fakauʻa* in Tongan) and is written as a reversed apostrophe (ʻ). The count of phonemes within a single Tongan word varies from 1 to 24, with a mean of 7.63 and a standard deviation of 2.92. Similar to the considerations in the "Number of Letters" section, the presence of long vowels, the voiced velar nasal *ng*, and the glottal plosive (ʻ) were taken into account when determining the number of phonemes in a word.

Each vowel and the glottal stop or *fakauʻa* (represented as /ʔV/) followed by a vowel are separately searchable in the search engine. These vowel-initial words would be particularly important for phonetic studies. To clarify this, we have added Table 4, which shows the numbers and percentages of vowel-initial words among the 4,013 words, including long vowels marked by a macron (i.e., ā, ē, ī, ō, and ū) and words beginning with a glottal stop or *fakauʻa* followed by a vowel. As illustrated in Table 4, the /ʻ+vowel/sequence has the greatest number of occurrences, with 198 words (4.93 %). In general, the /ʻ+vowel/pattern appears more frequently than words starting with just a vowel. Specifically, the number of /u/-initiated words is also higher, with 48 words (0.87 %) and the number of /ʻu/-initiated words is also higher, with 76 words (1.25 %).

In addition, we presented the top five most frequently-used initial letters (graphemes) in Table 5. The number of /f/-initiated words is notably high, totaling 1,036 words (25.82 %), accounting for about a quarter of the 4,013 words. The second most frequent is /t/-initiated words, with 703 words (17.52 %). The third is /m/-initiated words, with 405 words (10.09 %). The fourth is /h/-initiated words, with 253 words (6.30 %), and the fifth is /k/-initiated words, with 209 words (5.21 %). Collectively, these five letters account for 2,606 words, encompassing 64.94 % of the total 4,013 words. This indicates that many Tongan words begin with a relatively limited set of initial sounds.

Column 5: Number of Moras: Tongan rhythm is structured around moras, with each mora consisting of a strong initial beat followed by a weaker second beat (Hayes 1995). Tongan adheres to a strict syllabic pattern featuring only (C)V syllables

Table 4: Numbers of vowel-initial words and their percentages among 4,013 words.

Vowel	Frequency	ʻ+Vowel	Frequency	Long vowel	Frequency	ʻ+Long vowel	Frequency
a	35	ʻa	198	ā	0	ʻā	26
e	5	ʻe	50	ē	1	ʻē	3
i	16	ʻi	49	ī	1	ʻī	0
o	21	ʻo	84	ō	1	ʻō	4
u	48	ʻu	76	ū	1	ʻū	4
Vowel	Percentage	ʻ+Vowel	Percentage	Long vowel	Percentage	ʻ+Long vowel	Percentage
a	0.87 %	ʻa	4.93 %	ā	0.00 %	ʻā	0.65 %
e	0.12 %	ʻe	1.25 %	ē	0.02 %	ʻē	0.07 %
i	0.40 %	ʻi	1.22 %	ī	0.02 %	ʻī	0.00 %
o	0.52 %	ʻo	2.09 %	ō	0.02 %	ʻō	0.10 %
u	1.20 %	ʻu	1.89 %	ū	0.02 %	ʻū	0.10 %

Table 5: Top five frequently-used initial letters.

Letter (grapheme)	Frequency	Percentage
f	1,036	25.82 %
t	703	17.52 %
m	405	10.09 %
h	253	6.30 %
k	209	5.21 %

(Anderson and Otsuka 2006; Tamaoka et al. 2022). Monomoraic words in Tongan, such as *ki* meaning 'to/towards' and *pe* meaning 'or', typically serve grammatical functions. Words conveying meaning, like *kolo* meaning 'town', generally comprise more than one mora. A straightforward method to ascertain the number of moras in a Tongan word is to count the number of vowels. For instance, *feke*, signifying 'octopus', contains two vowels, thus equating to two moras (/fe/ and /ke/). It is worth noting that the long vowel in mā (/ma/ + /a/) effectively represents a double vowel /a/ and is counted as two moras. The count of moras ranges from 1 to 13, with a mean of 4.17 and a standard deviation of 1.58.

In response to the observation about vowel sequences, Tables 6 and 7 provide detailed information on words containing consecutive vowels. Table 6 shows the numbers and percentages of words with the VV sequential pattern, while Table 7 focuses on the V'V sequential pattern. In Table 6, numbers on the diagonal represent long vowels marked by a macron (/ā/, /ē/, /ī/, /ō/, /ū/), and in Table 7, they indicate patterns with the ' glottal stop like /a'a/, /e'e/, /i'i/, /o'o/, and /u'u/. Notably, the long vowel /ā/ appears frequently, in 398 words (9.92 %), while the diphthongs /au/ and /ai/ are also common, found in 362 words (9.02 %) and 258 words (6.43 %), respectively.

Table 6: Word Frequencies Containing VV (Vowel + Vowel) sequential pattern.

VV type	Frequencies of occurances out of 4,013 words						Percentages out of 4,013 Words				
	a	e	i	o	u	Total	a	e	i	o	u
a	398	39	258	63	362	1120	9.92 %	0.97 %	6.43 %	1.57 %	9.02 %
e	68	103	61	5	26	263	1.69 %	2.57 %	1.52 %	0.12 %	0.65 %
i	139	46	35	40	34	294	3.46 %	1.15 %	0.87 %	1.00 %	0.85 %
o	112	9	18	122	89	350	2.79 %	0.22 %	0.45 %	3.04 %	2.22 %
u	194	34	65	27	69	389	4.83 %	0.85 %	1.62 %	0.67 %	1.72 %
Total	911	231	437	257	580		22.70 %	5.76 %	10.89 %	6.40 %	14.45 %

Table 7: Word Frequencies Containing V‘V (Vowel + ‘ glottal stop) sequence type.

Frequencies of occurrences out of 4,013 words						Percentages out of 4,013 words				
a	**e**	**i**	**o**	**u**	**Total**	**a**	**e**	**i**	**o**	**u**
128	25	70	19	29	271	3.19 %	0.62 %	1.74 %	0.47 %	0.72 %
45	25	21	3	6	100	1.12 %	0.62 %	0.52 %	0.07 %	0.15 %
37	15	46	13	4	115	0.92 %	0.37 %	1.15 %	0.32 %	0.10 %
14	0	24	21	8	67	0.35 %	0.00 %	0.60 %	0.52 %	0.20 %
50	0	31	3	57	141	1.25 %	0.00 %	0.77 %	0.07 %	1.42 %
274	65	192	59	104		6.83 %	1.62 %	4.78 %	1.47 %	2.59 %

These three patterns combined account for 1,018 words (25.37 %). In the V‘V pattern, /a‘a/ is the most frequent, appearing in 128 words (3.19 %). The vowel /a/ is the most frequently used among all vowel sequences. We believe this information will be valuable for phonetic studies.

Column 6: Parts of Speech: The parts of speech were categorized into 16 distinct groups: adjective, adverb, article, conjunction, (idiomatic) expression, interjection, interrogative, noun, numeral, plural marker, prefix, preposition, pronoun, quantifier, tense marker, and verb. Since many words serve multiple parts of speech, these are described in the column with a forward slash, such as adjective/verb, adjective/noun, conjunction/adverb, interrogative/adverb, noun/verb, etc.

Column 7: Meaning in English: The meanings of the 4,013 core Tongan words are exclusively sourced from *The Students’ English-Tongan and Tongan-English Dictionary* (Thompson and Thompson 1992). An illustrative format for this column entry is exemplified by *fo‘ou* ‘(adj/v) new, fresh; strange, unfamiliar’, where the definition is preceded by its part(s) of speech. Meanings were cross-referenced with multiple dictionaries, including *A Simplified Dictionary of Modern Tongan* (Tu’Inukuafe and Holding 1992).

Column 8: Number of Participants: This column indicates the total number of participants who responded correctly to each target word. For instance, if 24 native Tongan speakers responded correctly to the word *fale ta‘o-mā* ‘bakery’, the entry in this column would be ‘24’. The number of participants for each word ranged from 22 to 32, with responses controlled to ensure a minimum of 22 participants for all Tongan words. The mean number of participants was 26.04, with a standard deviation of 2.94.

Column 9: Reaction Times: This column records the reaction times for Tongan lexical decisions associated with each word. Detailed characteristics of reaction times are discussed in the second study.

Column 10: Accuracy: The average accuracies of the 4,013 words are recorded in this column. The detailed characteristics of accuracies are discussed in the second study.

4.2 Nonwords used for Tongan lexical decisions

The data for 559 nonwords judged by all 168 participants are recorded in an Excel file, which can be downloaded from the 'About the Database' section at https://tonganlex.vercel.app/. This dataset comprises four columns: (1) nonword spelling, (2) number of letters, (3) reaction times, and (4) accuracies. While there is not a web-accessible search engine specifically for nonwords, users have the option to download the entire spreadsheet as an Excel file from Google Sheets. Further details regarding reaction times and accuracies are provided in Section 3.4.

5 General discussion

In 2020, UNESCO classified the Tongan language as 'potentially vulnerable,' emphasizing concerns about its preservation and continuity. Despite the prevalence of English in educational and business contexts in Tonga, this study aimed to explore the usage patterns and proficiency levels of Tongan and English among the population. Through a questionnaire-based approach, the study assessed the frequency of Tongan usage in daily life and the proficiency levels in both English and Tongan. Results indicated that Tongan is significantly more frequently used than English in various aspects of daily communication. Furthermore, the questionnaire revealed that respondents exhibited relatively high levels of comprehension in both languages, with Tongan being more comprehensible across all four language skills: listening, speaking, reading, and writing. These findings suggest that native Tongan individuals residing on the main island of Tongatapu continue to prioritize and maintain their proficiency in the Tongan language, despite the growing influence of English.

The present study further administered a lexical decision task involving 4,013 Tongan words to 168 native Tongan speakers. Impressively, participants exhibited high accuracies in their lexical decisions, with a mean accuracy of 95.40 % and a standard deviation of 6.33 %. The average reaction time for these decisions was 575 ms, with a standard deviation of 68 ms. To provide comparative insights, an equal

number of words were selected based on frequency rankings in English (Balota et al. 2007) and French (Ferrand et al. 2010). Through this comparison, it was observed that Tongan words were processed more swiftly for lexical decisions in comparison to their English and French counterparts. Furthermore, the mean accuracies across all three languages – Tongan, English, and French – surpassed 95 %. Given the efficiency and accuracy with which these 4,013 Tongan words were processed, they were deemed representative of the core vocabulary within the Tongan language. This underscores the significance and prevalence of these words in Tongan linguistic communication.

Following the identification of the core 4,013 Tongan words, the present study developed a web-accessible search engine dedicated to these words (https://tonganlex.vercel.app/). This innovative tool enables users to seamlessly search for the English meanings of Tongan words based on their Tongan spelling, and conversely, to search for Tongan expressions using English meanings. This database of Tongan core vocabulary encompasses essential linguistic attributes such as numbers of letters, phonemes, and moras, alongside detailed information on parts of speech and meanings. Additionally, it provides psycholinguistic data, including lexical decision times and accuracies. The implementation of this search engine marks a significant advancement in Tongan language research, offering researchers a robust platform for generating experimental stimuli and facilitating further investigations into the intricacies of the Tongan language.

The present study furthermore revealed strong correlations among the number of letters, phonemes, and moras in the 4,013 commonly-used Tongan words, with all correlations exceeding $r = 0.93$. Additionally, predictive models for reaction times in lexical decision tasks were developed based on these linguistic units. The predictive analysis demonstrated that words with fewer letters, phonemes, and moras tended to exhibit longer reaction times, contrasting with those containing approximately 6–8 letters and phonemes or 3 to 4 moras, which showed shorter reaction times. However, reaction times increased again after reaching around 8 letters and phonemes. Notably, quadratic functions provided a better fit for predicting reaction times across all three linguistic units. Given Tongan's status as a moraic language, characterized by the regular combination of consonants and vowels, it is unsurprising that the trends observed in reaction time prediction were consistent across letters, phonemes, and moras. This highlights the robustness of the predictive models and underscores the unique linguistic characteristics of the Tongan language.

Indeed, the bilingual nature of Tongans, proficient in both Tongan and English, presents an intriguing avenue for studying the inter-language relations of words in the mental lexicon. This phenomenon is conducive to exploring various aspects of bilingual lexical access, as evidenced by prior research (Cristoffanini et al. 1986;

Davis et al. 2010; De Groot and Nas 1991; Dijkstra et al. 1999, 2010; Duyck et al. 2007, 2008; García et al. 2014; Lemhöfer and Dijkstra 2004; Nakayama 2002). While Tongan shares characteristics with other moraic languages, such as Japanese, the alphabetic script used in Tongan presents an intriguing contrast. Investigating how this script difference influences phonological processing is a compelling area of study, with potential implications for understanding bilingual phonological processing (Nakayama et al. 2016; O'Séaghdha et al. 2010; Verdonschot et al. 2013, 2015; Wang et al. 2018; You et al. 2012; Zhang and Damian 2019). Furthermore, the VSO (Verb-Subject-Object) word order of Tongan, in contrast to the SVO (Subject-Verb-Object) order of English, offers opportunities to explore verb processing in bilinguals within sentence contexts using the lexical pool of the 4,013 words (van Assche et al. 2013). These investigations could provide valuable insights into language processing mechanisms and the influence of language structure on bilingual cognition.

In conclusion, this study sheds light on various aspects of Tongan language processing, highlighting its unique characteristics and its role in the bilingual context of Tongans proficient in both Tongan and English. The examination of the 4,013 core Tongan words, coupled with the development of a comprehensive web-accessible search engine, opens possibilities for further research and applications in psycholinguistics and bilingualism. Moreover, the findings contribute to the broader understanding of language processing mechanisms and the interplay between linguistic structure, script, and cognitive processing in bilingual individuals. As Tongan language and culture flourish, exploring its linguistic complexities holds the key to deepening our comprehension of human language cognition and bilingualism across diverse cultural contexts.

Acknowledgments: The authors greatly appreciate the contribution of native Tongan speakers who participated in this research.

Research ethics: The present study involving human participants were reviewed and approved by the Research Ethics Committee of Humanities Department at Tohoku University, Japan. The participants provided their written informed consent to participate in this study.

Author contributions: KT (first and corresponding author) developed the theoretical framework, designed the experiments, and wrote the initial draft of the manuscript. SY created the experimental program and developed the web-accessible search engine. JZ organized the data into an analyzable format and coordinated with the native Tongan experimenter. KM conducted data analysis. MK, as the principal investigator, provided overall guidance, suggestions, and financial support.

Research funding: This study was supported by the Grant-in-Aid for Japan Society for the Promotion of Science (#19H05589, PI: Masatoshi Koizumi) and by the Grant-in-Aid for Japan Society for the Promotion of Science (#23K00628; PI: Emeritus

Professor Katsuo Tamaoka at Graduate School of Humanities, Nagoya University, Nagoya, Japan).

Data availability: The Tongan real words and nonwords data that supported the findings of this present study can be downloaded as Excel files from tonganlex.-vercel.app/.

Conflict of interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Appendix A

Savea ki hono mahino'i 'a e lea faka-Tonga

Language Understanding Survey for Tongan Speakers

'Aho 'a e Savea: _____Ta'u _____Mahina _____'Aho
Date of the Questionnaire Survey: _____Year _____Month _____Day

1. 'Aho fa'ele'i: _____Ta'u _____Mahina _____'Aho
   Birthday: _____Year_____ Month _____Day

2. Tangata/Fefine
   Male/Female

3. Ngāhi lea: (Kataki 'o fakafonu mai kotoa e ngaahi lea 'oku ke ngaue'aki):
   Lea faka- Tonga____ Lea faka-Pilitania ____ Ngāahi lea kehe ____
   Language (Please check languages you can use):
   Tongan____ English____ Other languages ____

4. `I he 'aho `e taha, ko e ha e lahi ho'o ngāue'aki e lea faka-Tonga pea mo e lea faka-Pilitania? (eg. 70% Lea faka-Tonga, 30% Lea faka-Pilitania)? Kapau 'oku ke ngāue'aki ha toe lea kehe, kataki 'o hiki hifo heni:
   In one day, how much do you use Tongan and English? (e.g., 70% Speak Tongan, 30% English)? If you are using additional language, please list here:

Lea faka-Tonga_____% Lea faka-Pilitania_____%
Tongan _____% English _____%
Ngāhi lea kehe (kataki 'o fakamahino mai) _____, _____%
Other languages (please specify) _____, _____%

5. Kataki ‘o fakafuofua mai ange ‘a e lahi ho ‘ilo ‘i he lea faka-Tonga ‘aki ha‘o siakale‘i ‘a e noa (0) *ki he* ono (6), ko e 0 ‘okapau ‘oku ‘ikai pe ke ‘iai ha‘o ‘ilo ki he lea faka-Tonga pe ko e ono (6) ki ho mahino‘i kakato e lea faka-Tonga. Please estimate your knowledge of Tongan by circle either 0 to six 6; 0 if you cannot speak any Tongan and 6 if you fully understand Tongan.

(1) Lea/ Talanoa (Speaking)

‘Ikai ke mahino‘i (Not understand at all) Mahino‘i kakato (Perfectly understand)
0 1 2 3 4 5 6

(2) Fanongo (Listening)

‘Ikai ke mahino‘i (Not understand at all) Mahino‘i kakato (Perfectly understand)
0 1 2 3 4 5 6

(3) Lau konga (Reading)

‘Ikai ke mahino‘i (Not understand at all) Mahino‘i kakato (Perfectly understand)
0 1 2 3 4 5 6

(4) Tohi (Writing)

‘Ikai ke mahino‘i (Not understand at all) Mahino‘i kakato (Perfectly understand)
0 1 2 3 4 5 6

6. Kataki ‘o fakafuofua mai ange ‘a e lahi ho ‘ilo ‘i he lea faka-Pilitania ‘aki ha‘o siakale‘i ‘a e noa (0) *ki he* ono (6), ko e 0 ‘okapau ‘oku ‘ikai pe ke ‘iai ha‘o ‘ilo ki he lea faka-Pilitania pe ko e ono (6) ki ho mahino‘i kakato e lea faka-Pilitania. Please estimate your knowledge of English by circle either 0 to six 6; 0 if you cannot speak any English and 6 if you fully understand English.

(1) Lea/ Talanoa (Speaking)

‘Ikai ke mahino‘i (Not understand at all) Mahino‘i kakato (Perfectly understand)
0 1 2 3 4 5 6

(2) Fanongo (Listening)

‘Ikai ke mahino‘i (Not understand at all) Mahino‘i kakato (Perfectly understand)
0 1 2 3 4 5 6

(3) Lau konga (Reading)

‘Ikai ke mahino‘i (Not understand at all) Mahino‘i kakato (Perfectly understand)
0 1 2 3 4 5 6

(4) Tohi (Writing)

‘Ikai ke mahino‘i (Not understand at all) Mahino‘i kakato (Perfectly understand)
0 1 2 3 4 5 6

References

Anderson, Victoria & Yuko Otsuka. 2006. The phonetics and phonology of “definitive accent” in Tongan. *Oceanic Linguistics* 45(1). 21–42.

Balota, David A., Melvin J. Yap, Michael J. Cortese, Keith A. Hutchison, Brett Kessler, Bjorn Loftis, James H. Neely, Douglas L. Nelson, Greg B. Simpson & Rebecca Treiman. 2007. The English lexicon project. *Behavior Research Methods* 39. 445–459.

Barry, Christopher, Katherine W. Hirsh, Robert A. Johnston & Catherine L. Williams. 2001. Age of acquisition, word frequency, and the locus of repetition priming of picture naming. *Journal of Memory and Language* 44. 350–375.

Brezina, Vaclav & Dana Gablasova. 2015. Is there a core general vocabulary? Introducing the "new general service list". *Applied Linguistics* 36(1). 1–22.

Brysbaert, Marc, Matthias Buchmeier, Markus Conrad, Arthur M. Jacobs, Jens Bölte & Andrea Böhl. 2011. The word frequency effect: A review of recent developments and implications for the choice of frequency estimates in German. *Experimental Psychology* 58. 412–424.

Brysbaert, Marc, Michaël Stevens, Paweł Mandera & Emmanuel Keuleers. 2016. The impact of word prevalence on lexical decision times: Evidence from the Dutch Lexicon Project 2. *Journal of Experimental Psychology: Human Perception and Performance* 42. 441–458.

Brysbaert, Marc, Evelyne Lagrou & Michaël Stevens. 2017. Visual word recognition in a second language: A test of the lexical entrenchment hypothesis with lexical decision times. *Bilingualism: Language and Cognition* 20. 530–548.

Brysbaert, Marc, Paweł Mandera & Emmanuel Keuleers. 2018. The word frequency effect in word processing: A review update. *Current Directions in Psychological Science* 27(1). 45–50.

Carter, Ronald. 2012. *Vocabulary applied linguistic perspectives*. London: Routledge.

Churchward, C. Maxwell. 1953. *Tongan grammar*. London and New York: Oxford University.

Coltheart, Max, Kathleen Rastle, Conrad Perry, Robyn Langdon & Johannes C. Ziegler. 2001. DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review* 108. 204–256.

Cop, Uschi, Emmanuel Keuleers, Denis Drieghe & Wouter Duyck. 2015. Frequency effects in monolingual and bilingual natural reading. *Psychonomic Bulletin & Review* 22. 1216–1234.

Cristoffanini, Paula, Kim Kirsner & Dan Milech. 1986. Bilingual lexical representation: The status of Spanish-English cognates. *Quarterly Journal of Experimental Psychology Section* A38. 367–393.

Custis, Tonya. 2004. *Word order variation in Tongan: A syntactic analysis*. Crookston, MN: University of Minnesota dissertation.

Davis, Chris, Rosa Sánchez-Casas, José E. García-Aibea, Marc Guasch, Margarita Molero & Pilar Ferré. 2010. Masked translation priming: Varying language experience and word type with Spanish-English bilinguals. *Bilingualism: Language and Cognition* 13. 137–155.

De Groot, Annette M. B. & Gerard L. J. Nas. 1991. Lexical representation of cognates and noncognates in compound bilinguals. *Journal of Memory and Language* 30. 90–123.

Dijkstra, Ton, Jonathan Grainger & Walter J. B. van Heuven. 1999. Recognition of cognates and interlingual homographs: The neglected role of phonology. *Journal of Memory and Language* 41. 496–518.

Dijkstra, Ton, Koji Miwa, Bianca Brummelhuis, Maya Sappelli & Harald Baayen. 2010. How cross-language similarity and task demands affect cognate recognition. *Journal of Memory and Language* 62. 284–301.

Dixon, Robert M. W. 1979. Ergativity. *Language* 55. 59–138.

Dixon, Robert M. W. 1994. *Ergativity* (Cambridge Studies in Linguistics). New York Cambridge: Cambridge University.

Dryer, Matthew S. & Martin Haspelmath (eds.). 2013. *The world Atlas of Language Structures online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://wals.info.

Duyck, Wouter, Eva van Assche, Denis Drieghe & Robert J. Hartsuiker. 2007. Visual word recognition by bilinguals in a sentence context: Evidence for nonselective lexical access. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 33. 663–679.

Duyck, Wouter, Dieter Vanderelst, Timothy Desmet & Robert J. Hartsuiker. 2008. The frequency effect in second-language visual word recognition. *Psychonomic Bulletin & Review* 15(4). 850–855.

Ferrand, Ludovic, Boris New, Marc Brysbaert, Emmanuel Keuleers, Patrick Bonin, Alain Méot, Maria Augustinova & Christophe Pallier. 2010. The French Lexicon Project: Lexical decision data for 38,840 French words and 38,840 pseudowords. *Behavior Research Methods* 42. 488–496.

Ferrer-i-Cancho, Ramon. 2005. The variation of Zipf's law in human language. *European Physical Journal B: Condensed Matter and Complex Systems* 44(2). 249–257.

García, Adolfo M., Agustín Ibáñez, David Huepe, Alexander L. Houck, Maëva Michon, Carlos G. Lezama, Sumeer Chadha & Alvaro Rivera-Rei. 2014. Word reading and translation in bilinguals: The impact of formal and informal translation expertise. *Frontiers in Psychology* 5. 1302.

Garellek, Marc & Marija Tabain. 2020. Illusions of the IPA: Tongan. *Journal of the International Phonetic Association* 50(3). 406–416.

Garellek, Marc & James White. 2015. Phonetics of Tongan stress. *Journal of the International Phonetic Association* 45(1). 13–34.

Hayes, Bruce. 1995. *Metrical stress theory: Principles and case studies*. Chicago, IL: University of Chicago Press.

Hino, Yasushi & Stephen J. Lupker. 1998. The effects of word frequency for Japanese kana and kanji words in naming and lexical decision: Can the dual-route model save the lexical-selection account? *Journal of Experimental Psychology: Human Perception and Performance* 24. 1431–1453.

Hunt, Alan & David Beglar. 2005. A framework for developing EFL reading vocabulary. *Reading in a Foreign Language* 17. 23–59.

Kiyama, Sachiko, Katsuo Tamaoka, Jungho Kim & Masatoshi Koizumi. 2013. Effect of animacy on word order processing in kaqchikel maya. *Open Journal of Modern Linguistics* 3. 203–207.

Kiyama, Sachiko, Meng Sun, Jungho Kim, Katsuo Tamaoka & Masatoshi Koizumi. 2016. Interference of Context and bilinguality with the word order preference in Kaqchikel reversible sentences. *Tohoku Psychologica Folia* 75. 22–34.

Koizumi, Masatoshi, Yoshiho Yasugi, Katsuo Tamaoka, Sachiko Kiyama, Jungho Kim, Juan Esteban Ajsivinac Sian & Lolmay Pedro Oscar García Mátzar. 2014. On the (non) universality of the preference for subject-object word order in sentence comprehension: A sentence-processing study in kaqchikel maya. *Language* 90(3). 722–736.

Koizumi, Masatoshi, Yasuhiro Takeshima, Ryo Tachibana, Riku Asaoka, Godai Saito, Keiyu Niikuni & Jiro Gyoba. 2020. Cognitive loads and time courses related to word order preference in kaqchikel sentence production: An NIRS and eye-tracking study. *Language, Cognition and Neuroscience* 35(2). 137–150.

Kongaika, Isileli T. & J. T. Fisher. 2019. *Tongan dictionary: Tongan to English/English to Tongan*. Tonga: Create Out Loud.

Kuo, Grace & Chad Vicenik. 2012. The intonation of Tongan. *UCLA Working Papers in Phonetics* 111. 63–91.

Kuperman, Victor & Julie A. van Dyke. 2013. Reassessing word frequency as a determinant of word recognition for skilled and unskilled readers. *Journal of Experimental Psychology: Human Perception and Performance* 39. 802–823.

Kureta, Yoichi, Takao Fushimi & Itaru I. Tatsumi. 2006. The functional unit in phonological encoding: Evidence for moraic representation in native Japanese speakers. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 32(5). 1102–1119.

Lemhöfer, Kristin & Ton Dijkstra. 2004. Recognizing cognates and interlingual homographs: Effects of code similarity in language-specific and generalized lexical decision. *Memory & Cognition* 32. 533–550.

Linders, Guido M. & Max M. Louwerse. 2023. Zipf's law revisited: Spoken dialog, linguistic units, parameters, and the principle of least effort. *Psychonomic Bulletin & Review* 30. 77–10.

Monaghan, Padraic, Ya-Ning Chang, Stephen Welbourne & Marc Brysbaert. 2017. Exploring the relations between word frequency, language exposure, and bilingualism in a computational model of reading. *Journal of Memory and Language* 93. 1–21.

Monsell, Sylvia, Michael C. Doyle & Patrick N. Haggard. 1989. Effects of frequency on visual word recognition tasks: Where are they? *Journal of Experimental Psychology: General* 118. 43–71.

Nakayama, Michiko. 2002. The cognate status effect in lexical processing by Chinese-Japanese bilinguals. *Psychologia* 45. 184–192.

Nakayama, Mariko, Sachiko Kinoshita & Rinus G. Verdonschot. 2016. The emergence of a phoneme-sized unit of speech planning in Japanese-English bilinguals. *Frontiers in Psychology* 7. 175.

Nation, I. S. Paul. 2001. *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Nation, I. S. Paul. & Robert Waring. 1997. Vocabulary size, text coverage and word lists. In Norbert Schmitt & Michael McCarthy (eds.), *Vocabulary: Description, acquisition, and pedagogy*, 6–19. Cambridge: Cambridge University Press.

O'Seaghdha, Padraig G., Jenn-Yeu Chen & Train-Min Chen. 2010. Proximate units in word production: Phonological encoding begins with syllables in Mandarin Chinese but with segments in English. *Cognition* 115(2). 282–302.

Otsuka, Yuko. 2000. *Ergativity in Tongan*. Oxford: University of Oxford dissertation.

Otsuka, Yuko. 2005a. Two derivations of VSO: A comparative study of niuean and Tongan. In Andrew Carnie, Heidi Harley & Sheila Ann Dooley (eds.), *Verb first: On the syntax of verb-initial languages*, 281–302. Amsterdam: John Benjamins.

Otsuka, Yuko. 2005b. Scrambling and information focus: VSO-VOS alternation in Tongan. In Joachim Sabel & Mamoru Saito (eds.). *The free word order phenomenon: Its syntactic sources and diversity*, 243–279. Berlin: Mouton de Gruyter.

Otsuka, Yuko. 2007. Making a case for Tongan as an endangered language. *The Contemporary Pacific* 19(2). 446–473.

Piantadosi, Steven T. 2014. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review* 21(5). 1112–1130.

Sheather, Simon J. & M. Chris Jones. 1991. A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society: Series B* 53(3). 683–690.

Taft, Marcus. 1979. Recognition of affixed words and the word frequency effect. *Memory & Cognition* 7. 263–272.

Tamaoka, Katsuo, Jingyi Zhang, Masatoshi Koizumi & Rinus G. Verdonschot. 2022. Phonological encoding in Tongan: An experimental investigation. *Quarterly Journal of Experimental Psychology* 76(10). 2226–2231.

Tamaoka, Katsuo, Shaoyun Yu, Jingyi Zhang, Yuko Otsuka, Hyunjung Lim, Masatoshi Koizumi & Rinus G. Verdonschot. 2024. Syntactic structures in motion: Investigating word order variations in verb-final (Korean) and verb-initial (Tongan) languages. *Frontiers in Psychology* 15. 1360191.

Taumoefolau, Melenaite L. 1998. *Problems in Tongan lexicography*. Auckland: University of Auckland Ph.D. dissertation.

Taumoefolau, Melenaite L. 2002. *Stress in Tongan* (MIT working Papers in linguistics 44). Cambridge, MA: MIT.

Thompson, Richard H. & ‘Ofa Thompson. 1992. *The students’ English-Tongan and Tongan-English dictionary*. Tonga: Faletohi ‘Otumotu Angl’ofa’.

Tu’Inukuafe, Edgar & Robert G. Holding. 1992. *A simplified dictionary of modern Tongan*. Tonga: Polynesian Press.

UNESCO. 2020. *World atlas of languages*, 3rd edn. United Nations Educational, Scientific and Cultural Organization. Available at: https://en.wal.unesco.org.

Van Assche, Eva, Wouter Duyck & Marc Brysbaert. 2013. Verb processing by bilinguals in sentence contexts: The effect of cognate status and verb tense. *Studies in Second Language Acquisition* 35(2). 237–259.

Verdonschot, Rinus G., Mariko Nakayama, Qingfang Zhang, Katsuo Tamaoka & Niels O. Schiller. 2013. The proximate phonological unit of Chinese-English bilinguals: Proficiency matters. *PLoS One* 8(4). e61454.

Verdonschot, Rinus G., Lai Jun, Feng Chen, Katsuo Tamaoka & Niels O. Schiller. 2015. Constructing initial phonology in Mandarin Chinese: Syllabic or sub-syllabic? A masked priming investigation. *Japanese Psychological Research* 57. 61–68.

Verdonschot, Rinus G., Hoàng Thị Lan Phương & Katsuo Tamaoka. 2022. Phonological encoding in Vietnamese: An experimental investigation. *Quarterly Journal of Experimental Psychology* 75(7). 1355–1366.

Wang, Jie, Andus Wing-Kuen Wong & Hsuan-Chih Chen. 2018. Time course of syllabic and sub-syllabic processing in Mandarin word production: Evidence from the picture-word interference paradigm. *Psychonomic Bulletin & Review* 25(3). 1147–1152.

Wilk, M. B. & Ramanathan Gnanadesikan. 1968. Probability plotting methods for the analysis for the analysis of data. *Biometrika* 55(1). 1–17.

Yano, Masataka, Daichi Yasunaga & Masatoshi Koizumi. 2017. Event-related brain indices of gap-filling processing in Kaqchikel. In Samuel R. Harris (ed.), *Event-related potential (ERP): Methods, outcomes and research insights*. Hauppauge, NY: Nova Science Publishers.

Yonelinas, Andrew P. 2002. The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language* 46. 441–517.

You, Wenping, Qingfang Zhang & Rinus G. Verdonschot. 2012. Masked syllable priming effects in word and picture naming in Chinese. *PLoS One* 7(10). e46595.

Zhang, Qingfang & Markus F. Damian. 2019. Syllables constitute proximate units for Mandarin speakers: Electrophysiological evidence from a masked priming task. *Psychophysiology* 56(4). e13317.

Zipf, George K. 1932. *Selected studies on the principle of relative frequency in language*. Harvard, MA: Harvard University Press.

Zipf, George K. 1935. *The psychobiology of language: An introduction to dynamic philology*. Boston, MA: Houghton Mifflin.