ISSN: 1803-3814 (Printed), 2571-3701 (Online)
https://doi.org/10.13164/mendel.2020.2.023

MENDEL Soft Computing Journal

THE CLASSIFICATION OF DOCUMENTS IN MALAY AND INDONESIAN USING THE NAIVE BAYESIAN METHOD USES WORDS AND PHRASES AS A TRAINING SET

Marvin Chandra Wijaya✉

Maranatha Christian University, Computer Engineering Department, Indonesia

marvinchw@gmail.com✉

Abstract

Malay Language and Indonesian Language are two closely related languages, sharing a lot in common in the meanings of words and grammar. Classifying the two languages automatically using a tool is a challenge because the two languages are very similar. The classification method that is widely used today is the Naive Bayesian method. This method needs to be implemented in a particular way to increase the level of classification accuracy. In this study, a new method was used, by using a training set in the form of words and phrases instead of just using a training set in the form of words only. With this method, the level of classification accuracy of the two languages is increased.

Keywords: *Malay, Indonesia, Language, Naïve Bayesian, Classification.*

Received: 09 December 2020
Accepted: 21 December 2020
Published: 21 December 2020

1 Introduction

Indonesian is the official language of the Republic of Indonesia in Southeast Asia. Indonesia is an archipelago located on the equator with a population of 267 million people. Malay is the official language of the Malaysian Federation which is also in Southeast Asia. Malaysia is located adjacent to Indonesia and has a population of around 27 million people. The speakers of these two languages add up to nearly 300 million people [14].

Indonesian and Malay are languages that have similarities between them, but differ at the same time. The two languages are generally understandable, but have differences in vocabulary, pronunciation, grammar, and spelling. Moreover, the same word in the two languages sometimes has different meanings.

The Malay Language is part of the Austronesian language family. The Malay language is spoken in several countries in Southeast Asia such as Malaysia, Indonesia, Singapore, and Brunei. Malay Language as the official language, is called the National Language. In Singapore and Brunei, the language is called Bahasa Melayu, while in Malaysia it is called Bahasa Malaysia. In Indonesia, the language is called the Indonesian Language. The Indonesian Language is the lingua franca (Permersatu) or a unifying language (Language of Unity) in the Republic of Indonesia.

Recognition of the Malay language is different in Malaysia and Indonesia. Malay is the national language in Malaysia, while in Indonesia it is only a regional language for several residents in the western part of the island of Kalimantan and the east coast of the island of Sumatra.

In Indonesia, the Malay Language is at the same level as other regional languages, such as Javanese, Sundanese, Balinese, Batak, Madurese, Bugis, Minangkabau, Betawi, Acehnese, and other regional languages [6].

Classification is processed by looking at the similarities of the features that exist in each language. A good classification must meet non-arbitrator, exhaustive and unique requirements. Non-arbitrator means that the classification criteria must only have one criterion, then the results will be exhaustive. The classification results must also be unique, which means that if a language has been classified into one group, it cannot be included in another group. If the classification falls into two or more groups, it means that the classification results are not unique. Typological classification is carried out based on the similarity of types found in several languages. This type is a certain element that can occur repeatedly in a language. This typology classification can be done at all language levels. Sociolinguistic classification is carried out based on the relationship between language and the factors prevailing in society, to be precise based on the function, assessment, and status that society gives to that language.

This classification is based on four characteristics or criteria. The first criterion is historicity concerning the history of language development or the history of the use of that language. The second criterion is standardization concerning its status as standard or non-standard language or its status in formal or informal usage. The third criterion is vitality regarding whether the language has speakers who use it in their daily activities actively or not. The fourth criterion is homogeneity concerning whether the lexicons and grammar of the language are derived.

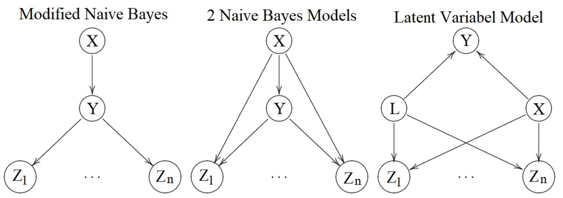

Figure 1: Three Naïve Bayesian Approaches.

2 Related Work

Naive Bayesian has been widely used in the classification of various things since decades ago. Various modifications to the classification method have been carried out to improve the accuracy of the classification.

In 2009, Chen Jing-Nian, Huang Hou-Kuan, Tian Sheng-Feng, and QuYou-li conducted a study on text classification. To improve the accuracy of the classification results, feature selection is carried out. In this study, two feature evaluations were applied to text classification. The first feature evaluation is the Multi-Class Odd Ratio which is called MOR. The second feature evaluation is the Class Discrimination Measure called CDM. The use of these two feature evaluations improves the classification results in text documents [2].

In 2010, Toon Calders and Sisco Verwer made modifications to the Naive Bayesian method for classification. The modification used is to use three naïve bayesian to perform discrimination-free classification [1]. The first approach of the Naïve Bayesian modification used is the modification of the probability of the classification result to be positive. The second approach of the Naive Bayesian modification used is to train each model for each attribute value that is considered important. The third approach of the Naive Bayesian modification used is to add a latent variable that is free from bias and perform Bayesian modeling which results in maximizing the classification results (see Figure 1). The formula for joint distribution via class X, sensitive Y, and all attributes $Z_1, \ldots, Z_n$ is:

$$P(X, Y, Z_1, \ldots, Z_n) = P(Y)P(X|Y)P(Z_1|X) \cdots P(Z_N|X) \quad (1)$$

In 2012 Wan Chin-Heng, Lee Lam-Hong, Rajkumar Rajprasad, and Isa Dino conducted a study for text classification by integrating SVM (Support Vector Machine) and KNN (K-Nearest Neighbor) [13]. The two approaches are integrated into SVM-NN (Support Vector Machine Nearest Neighbor) (see Figure 2). The results of the text document classification are calculated based on the closest average distance of support vectors and data point testing. In the experiments conducted in this study, the results obtained increased accuracy by adding training sets and testing sets.

In 2016, Jiang Liang-Xiao, Wang Sha-Sha, Li Chao-Qun, and Zhang Lung-An conducted a study on the multinomial naive Bayes [4]. Modifications are made by extending the structure on multinomial naive Bayes. This modification is called SEMNB (Structure Extended Multinomial Naive Bayes). SEMNB which has a simple but effective algorithm (see Figure 3) has succeeded in increasing classification accuracy compared to MNB. This study resulted in a significant increase in the results of text data sets.

In 2017, Krzystof Krawiec conducted a study to search for drivers using genetic programming [5]. In this study, multiple tests were carried out to find suitable drivers, and according to the desired classification. In this study, there is a search for object clustering using Bayesian Information Criterion to increase the success rate in finding suitable programs. In the same year, Ivars Namatevs and Ludmila Aleksejeva conducted a study on donor-recipient matching. In that study, a decision algorithm was made using the greedy algorithm. Also, BBN (Bayesian Belief Network) is used to determine the right results [7].

In 2018, Pavel Skrabanek and Sule Yildirim Yayilgan conducted a study on performance dependency classification [10]. The method used is WECIA (Weighting Coefficients Impact Assessment) graph. In the same year, Radek Hrebik and Jaromir Kukal conducted a study on the classification of Context Out [3]. The method used in this study is a hidden class system. The result of this study is to get the best sensitivity in the output classes.

In 2019, Eslam Amer and Ivan Zelinka conducted a study on classification to detect malware [15]. In this study, the minimum feature set method was used. The experiments on these studies yielded significant results. The accuracy of this study produces accuracy that competes with other methods (XGB, Estra Tree Classifier, LDA, Ada Boost, Random Forest, Decision Tree, MLP, SVM, and K-Nearest Neighbor). In the same year, Mucahid Mustafa Saritas and Ali Yasar conducted a study to classify data [9]. This study used the Naive Bayes algorithm and ANN (Artificial Neural Network). This method is successful in classifying medical data.

The process of classifying a problem can be done using the methods discussed in this section, but can also use tools such as using the software. For example Matlab has a toolbox that helps the classification process using the Classification Learner app.

3 Method

3.1 Semantics

The meaning in a word needs to be analyzed, using the study of morphemes [16] (free morphemes [12] and bound morphemes) [11]. Therefore, there are theories related to the problems in this study, namely morpheme theory, homonym theory, connotation theory, and the theory of language derivation or branching [8]. The differences between the two languages are spelling, writing, punctuation, pronunciation, and vocabulary.

- **Spelling and writing**: The impact of British colonization in Malaysia and the Netherlands in

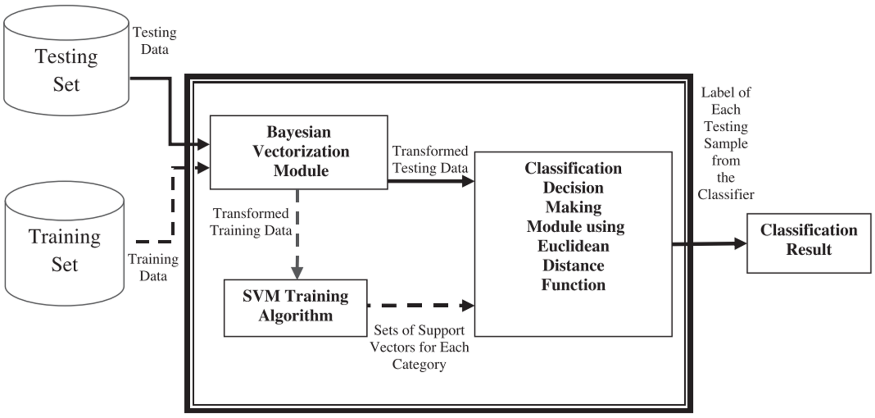
Testing Set
Testing Data
Training Set
Training Data
Bayesian Vectorization Module
Transformed Testing Data
Transformed Training Data
SVM Training Algorithm
Sets of Support Vectors for Each Category
Classification Decision Making Module using Euclidean Distance Function
Label of Each Testing Sample from the Classifier
Classification Result

Figure 2: SVM-NN [13].

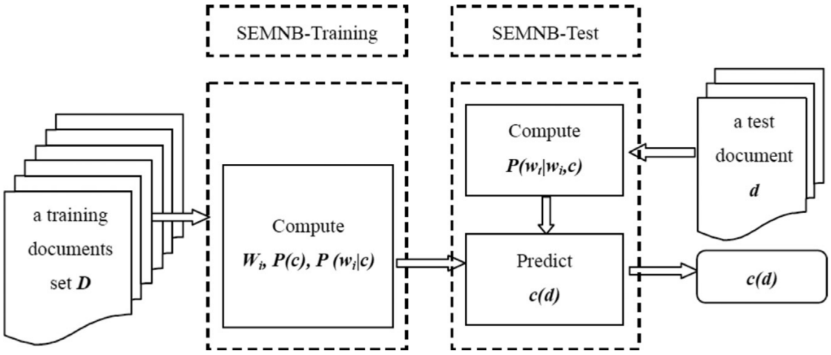

Figure 3: SEMNB algorithm framework [4].

Indonesia greatly influenced language. Today, the representations of the speech sounds in the two languages are remarkably identical, but some minor spelling differences apply for historical reasons.

- **Punctuation**: Malay language and Indonesian language differ in the use of punctuation marks, including the decimal sign. Standard Malay language, influenced by English, uses a decimal point. In the Indonesian language, the decimal point is used which is influenced by the Dutch system.
- **Pronunciation**: The pronunciation of the two languages is sometimes different. But because this study did not involve pronunciation, this difference was ignored in this study.
- **Vocabulary**: There is a marked difference in the vocabulary of the two languages in loan words. The English word "Christmas" in Indonesia uses the word "Christmas", but in the Malay language, the word "Krismas" is derived from English. The English word "College" in Malay uses the word "Kolej" while in Indonesian it is "Sekolah Tinggi".

Table 1 shows an example of the differences in words in Malay and Indonesian.

Phrases are one of the terms in linguistic studies. Phrases are linguistic units that are larger than words and smaller than clauses and sentences. The meaning of the phrase is not different from the meaning of the word which is the head/essence of the phrase. Phrases have no new meaning. The meaning in a phrase will not be far from the meaning of the word that forms it, but the meaning of a phrase can differentiate between Malay Language and Indonesian Language (Table 2).

3.2 Naïve Bayesian

There are five stages in the implementation of Bayesian in document classification. These stages are as follows:

- Identification of prerequisites for training the Naive Bayes classifier
- Document Matrix calculation for each class
- Frequency calculation
- Use of Naive Bayes rules
- Calculation of possible document class

In the Naive Bayesian rule it is necessary to have a set of examples that exist for each category (class) in which part of the text is classified. In this research, there is an intention to classify a document. The document is classified as a document in Malay Language or the document is in the Indonesian Language. There are two requirements used in this study, namely a set of words in Malay and Indonesian. Also, a collection of phrases in Malay Language and Indonesian Language (see Figure 4). In this study, the size of the training

Table 1: Different in words for both languages

English Language	Malay Language	Indonesia Language
March	Mac	Maret
August	Ogos	Agustus
Monday	Isnin	Senin
Challenge	Cabaran	Kecabaran, Tantangan
Speak	Bercakap, bertutur, berbual	Bercakap-cakap, berbicara
Shop	Kedai	Toko
Ticket	Tiket	Tiket, karcis
Because	Kerana	Karena
Hospital	Hospital	Rumah Sakit
Zoo	Zoo	Kebun Binatang
Television	Televisyen	Televisi
University	Universiti	Universitas
Chair	Kerusi	Kursi
Chairman	Pengerusi	Ketua
Orange	Oren	Jeruk
Apple	Epal	Apel
Car	Kereta	Mobil

set (the number of documents known to be in Malay or Indonesian) was made variable to see the level of accuracy.

The document matrix consists of a list of the frequency with which words or phrases appear in the training document. The document matrix is a tenuous rectangular matrix consisting of n words/phrases and m documents as in Table 3.

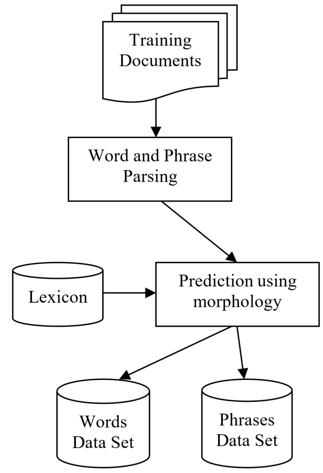

Figure 4: Training Process.

After the matrix document is calculated for each class, the next step is to calculate the frequency and occurrence of each word/phrase as in Table 4.

The formula for naive bayesian used is as follows:

$$P(A|B) = P(B|A)P(A)/P(B) \tag{2}$$

In simple terms $P(A)$ is "prior" and $P(B)$ is "evidence". $P(A)$ and $P(B)$ show the possible predictions of A and B independently of each other. Meanwhile, $P(A|B)$ is "posterior" and $P(B|A)$ is "likelihood". $P(A|B)$ and $P(B|A)$ are the conditional probabilities between A and B. In this study, the formulas of Naive Bayesian are as follows:

$$P(\text{Malay}|x) = P(x|\text{Malay})P(\text{Malay})/P(x) \tag{3}$$

$$P(\text{Indonesian}|x) = P(x|\text{Indonesian})P(\text{Indonesian})/P(x) \tag{4}$$

where x are words and phrases:

$$x = [w_1, w_2, \dots, w_{n_1}] \text{ and } [p_1, p_2, \dots, p_{n_2}]$$

The assumption "Naive" in the Naive Bayes classifier is that the probability of a word or phrase is independent of one another. The result is that "likelihood" is the product of the probability that each word or phrase is present in a collection of Malay or Indonesian documents. The formulas are as follows:

$$P(\text{Malay}|w_1, \dots, w_{n_1}) \propto P(\text{Malay}) \prod_{i=1}^{n_1} P(w_i|\text{Malay}) = P(\text{Malay})P(w_1|\text{Malay}) \cdots P(w_{n_1}|\text{Malay}) \tag{5}$$

$$P(\text{Malay}|p_1, \dots, p_{n_2}) \propto P(\text{Malay}) \prod_{i=1}^{n_2} P(p_i|\text{Malay}) = P(\text{Malay})P(p_1|\text{Malay}) \cdots P(p_{n_2}|\text{Malay}) \tag{6}$$

$$P(\text{Indonesian}|p_1, \dots, p_{n_2}) \propto P(\text{Indonesian}) \prod_{i=1}^{m_1} P(w_i|\text{Indonesian}) = P(\text{Indonesian})P(w_1|\text{Indonesian}) \cdots P(w_{m_1}|\text{Indonesian}) \tag{7}$$

$$P(\text{Indonesian}|p_1, \dots, p_{m_2}) \propto P(\text{Indonesian}) \prod_{i=1}^{m_2} P(p_i|\text{Indonesian}) = P(\text{Indonesian})P(p_1|\text{Indonesian}) \cdots P(p_{m_2}|\text{Indonesian}) \tag{8}$$

In this study, two types of Bayesian naive experiments were carried out. The first experiment was to use a data set of words only (see Figure 5). The second experiment using a data set of words and phrases (see Figure 6).

Table 2: Different in phrases for both languages

English Language	Malay Language	Indonesia Language
Head office	Ibu Pejabat	Kantor Pusat
Pharmacy	Kedai Ubat	Toko Obat
Restaurant	Kedai Makan	Rumah Makan
Sightseeing	Pusing-Pusing	Jalan-Jalan
Air Force	Tentera Udara	Angkatan Udara
Apartment	Rumah Pangsa	Rumah Susun
Pacific Ocean	Lautan Teduh	Samudera Pasifik
Tap Water	Air Paip	Air Keran
Toilet	Bilik air	Kamar Kecil
United Nations	Pertubuhan Bangsa-Bangsa Bersatu	Perserikatan Bangsa-Bangsa

Table 3: Example Document matrix for each classification

Word	Training 1	Training 2	…	Training m
Buku	1	0		2
University	1	1		0
Tas	0	2		0
Komputer	2	1		0
⋮				
Monitor	2	0		3

Table 4: Example frequency and occurrence calculation

Word	Frequency	Occurrence
Buku	3	0.75
University	2	0.50
Tas	2	0.50
Komputer	3	0.75
⋮		
Monitor	4	1.00

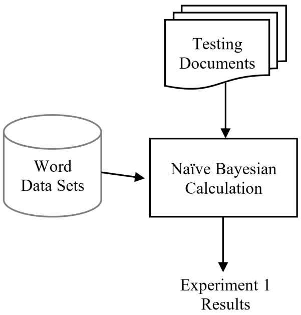

Figure 5: Experiment 1.

The formal decision rules for experiment 1 is:

$$\underset{k\in\{\text{Malay,Indonesian}\}}{\arg\max} P(\text{document}_k)\prod_{i=1}^{n_1} P(w_i|\text{document}_k) \tag{9}$$

The formal decision rules for experiment 2 is:

$$\underset{k\in\{\text{Malay,Indonesian}\}}{\arg\max} \Big\{ P(\text{document}_k)\prod_{i=1}^{n_1} P(w_i|\text{document}_k)\prod_{i=1}^{m_1} P(p_i|\text{document}_k)\Big\} \tag{10}$$

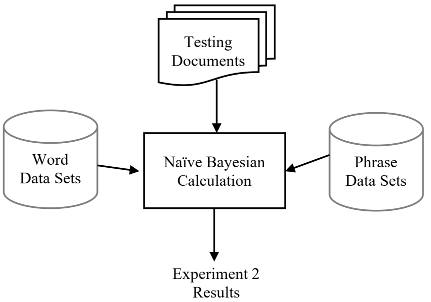

Figure 6: Experiment 2.

Table 5: The level of accuracy in experiment 1 (Words data set)

Training set	Testing Set Document			
Document	20	25	30	Average
30	55%	57%	54%	55%
40	60%	61%	59%	60%
50	67%	66%	67%	67%

4 Results

The experiments conducted in this study looked at the comparison of the classification accuracy results using the word-data set only or with the addition of the phrase-data set as well. Also, experiments were carried out by varying the number of training documents to see their effect on the level of accuracy of document classification. In experiment 1, the experiment was carried out using Naive Bayesian processing using a data set of words. The accuracy testing was carried out using a different number of training document sets. In Table 5, it can be seen that by using 30 training document sets, the average accuracy result is 55%. By increasing the number of training sets to 40 and 50, the accuracy result is also increased to 60% and 67%. In experiment 2, the experiment was carried out using Naive Bayesian processing using a data set of words and phrases. The accuracy in experiment 2 resulted in a higher level of accuracy compared to experiment 1. In the number of training set 30 documents, an accuracy of 67% was obtained. When the number of training set documents is increased to 40 and 50, the accuracy results increase to 74% and 78% (see Table 6).

Table 6: The level of accuracy in experiment 2 (Words and phrases data set)

Training set Document	Testing Set Document 20	25	30	Average
30	67%	69%	65%	67%
40	75%	74%	73%	74%
50	78%	78%	79%	78%

From the two experiments, it was found that the accuracy of the document classification results would be greater if the combined words and phrases data set were combined as shown in Figure 7.

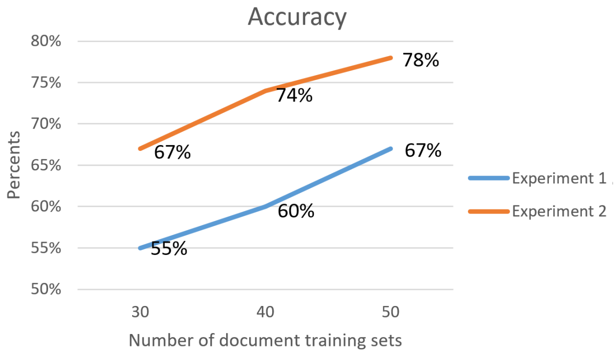

Figure 7: Comparison of the accuracy of experiment 1 and experiment 2.

5 Conclusion

In the study of document classification in Malay Language and Indonesian Language, several conclusions were obtained. With the increase in the number of training set documents, the level of accuracy will increase. To increase the level of accuracy, you can use a combined data set in the form of words and phrases for each language. But the difference between the level of accuracy in experiment 1 and two will decrease with increasing the number of training sets. The level of accuracy of these two languages is not too high because the two languages are very similar so that for future work, other methods can be searched to improve the accuracy.

References

[1] Calders, T., and Verwer, S. Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery 21* (2010), 277–292.

[2] Chen, J., Huang, H., Tian, S., and Qu, Y. Feature selection for text classification with naïve bayes. *Expert Systems with Applications 36* (2009), 5432–5435.

[3] Hrebik, R., and Kukal, J. Context out classifier. *MENDEL 24* (2018), 101–106.

[4] Jiang, L., Wang, S., Li, C., and Zhang, L. Structure extended multinomial naive bayes. *Information Sciences 329* (2016), 346–356.

[5] Krawiec, K. Opening the black box: Alternative search drivers for genetic programming and test-based problems. *MENDEL 23* (2017), 1–6.

[6] Nababan, P. Language in education: The case of indonesia. *International Review of Education 37* (1991), 115–131.

[7] Namatevs, I., and Aleksejeva, L. Decision algorithm for heuristic donor-recipient matching. *MENDEL 23* (2017), 33–40.

[8] Ortmann, A. Connecting the typology and semantics of nominal possession: alienability splits and the morphology–semantics interface. *Morphology 28* (2018), 99–144.

[9] Saritas, M., and Yasar, A. Performance analysis of ann and naive bayes classification algorithm for data classification. *International Journal of Intelligent Systems and Applications in Engineering 73* (2019), 88–91.

[10] Skrabanek, P., and Yayilgan, S. WECIA Graph: Visualization of classification performance dependency on grayscale conversion setting. *MENDEL 24* (2018), 41–48.

[11] Soh, H., and Nomoto, H. The malay verbal prefix men- and the unergative/unaccusative distinction. *Journal of East Asian Linguistics 20* (2011), 77–106.

[12] Sosial, J., and Vol, B. Perbedaan semantik antara bahasa indonesia dan bahasa malaysia: Satu kajian awal upaya mengelak kesalahpahaman dan perbedaan budaya antara bangsa serumpun di asia tenggara fakultas tarbiyah dan keguruan , uin sultan syarif kasim riau. *Jurnal Sosial Budaya 9* (2012), 261–282.

[13] Wan, C., Lee, L., Rajkumar, R., and Isa, D. A hybrid text classification approach with low dependency on parameter by integrating k-nearest neighbor and support vector machine. *Expert Systems with Applications 39* (2012), 11880–11888.

[14] Yap, M., Liow, S. R., Jalil, S., and Faizal, S. The malay lexicon project: A database of lexical statistics for 9,592 words. *Behavior Research Methods 42* (2010), 992–1003.

[15] Zelinka, I., and Amer, E. An ensemble-based malware detection model using minimum feature set. *MENDEL 25* (2019), 1–10.

[16] Zhang, D., Koda, K., and Leong, C. Morphological awareness and bilingual word learning: a longitudinal structural equation modeling study. *Reading and Writing 29* (2016), 383–407.