Predictive Text System for Bahasa with Frequency, n-gram, Probability Table and Syntactic using Grammar

Derwin Suhartono, Garry Wong, Polim Kusuma and Silviana Saputra
Computer Science Department, Bina Nusantara University, K H. Syahdan 9, Jakarta, Indonesia

Keywords: Predictive Text, Word Prediction, n-gram, Prediction, KSPC, Keystrokes Saving.

Abstract: Predictive text system is an alternative way to improve human communication, especially in matter of typing. Originally, predictive text system was intended for people who have flaws in verbal and motor. This system is aimed to all people who demands speed and accuracy in typing a document. There were many similar researches which develop this system that had their own strengths and weaknesses. This research attempts to develop the algorithm for predictive text system by combining four methods from previous researches and focus only in Bahasa (Indonesian language). The four methods consist of frequency, n-gram, probability table, and syntactic using grammar. Frequency method is used to rank words based on how many times the words were typed. Probability table is a table designed for storing data such as predefined phrases and trained data. N-gram is used to train data so that it is able to predict the next word based on previous word. And syntactic using grammar will predict the next word based on syntactic relationship between previous word and next word. By using this combination, user can reduce the keystroke up to 59% in which the average keystrokes saving is about 50%.

1 INTRODUCTION

Conventional process of typing documents using a typewriter has become obsolete due to technological advances. This is clear as computer can help people considerably in many daily activities. This progress can also be felt when a computer help to predict which words are going to be typed by user. The ability to predict the word that is going to be typed by user is often referred to predictive text system. It is also often called as the word prediction system.

Predictive text is a part of the research in the field of artificial intelligence especially on natural language processing (NLP). NLP is a field of computer science that focused on getting computers to perform useful tasks involving human language, tasks like enabling human-machine communication, improving human to human communication, or simply doing useful processing of text or speech (Jurafsky and Martin, 2008), and predictive text itself is a technique that helps the input process that was used by people with disabilities or not on a desktop system, handheld devices, and augmentative communication systems (MacKenzie and Ishii, 2007). Originally, predictive text system was intended for people who have flaws (defects) in verbal and motor (Vitoria and Abascal, 2006). Over time, the usage of this system began to change and now is aimed to all people who demands speed and accuracy in typing a document.

In the previous research, there are several prediction methods in predictive text, such as prediction using frequencies, prediction using word probability tables, syntactic prediction using probability tables, syntactic prediction using grammar, and semantic prediction (Vitoria and Abascal, 2006). N-gram is also introduced by other researchers as another prediction method in a predictive text system (Verberne, et.al, 2012).

However, if only one of the methods is used, it will not efficient and effective enough for a predictive text system. It is because the weakness of the chosen method cannot be supported by another method so when the system is being evaluated by considering the value of keystrokes per character (KSPC) and keystrokes saving, the result is not satisfactory. Therefore, there are many other researches that try to combine the existing methods to achieve maximum result.

This research focused on predictive text in Bahasa (Indonesian language) by combining some prediction methods that are expected to be a more

Suhartono D., Wong G., Kusuma P. and Saputra S..
Predictive Text System for Bahasa with Frequency, n-gram, Probability Table and Syntactic using Grammar.
DOI: 10.5220/0004756603050311
In *Proceedings of the 6th International Conference on Agents and Artificial Intelligence* (ICAART-2014), pages 305-311
ISBN: 978-989-758-015-4

optimal solution than the previous researches. Bahasa was selected as the focus of this research because many existing researches used foreign languages such as English, Swedish, etc., as its main focus. Lack of knowledge of the Indonesian people about good and proper grammar is the other reason. It caused slang or colloquial language which is often unconsciously used in the official documents that should have used the proper language.

This research is expected to help Indonesian people in the process of typing a document more quickly and precisely based on the proper Indonesian language.

2 RELATED WORKS

Previous research that discuss about measuring performance of predictive text system was a research about keystrokes per character (KSPC) and keystroke saving (MacKenzie, 2002). The result of the research said that the smaller value of KSPC will give a better performance for the system. A year later, there was another research to evaluate the accurate measurement of predictive system in case of typing errors caused by users (Soukoreff and Mackenzie, 2003). This evaluation was done by minimum string distance (MSD) error rate and KSPC. The research results were a new equation for MSD error rate and development of KSPC formula. Using this development, the used bandwidth which represents useful information that was transferred will be determined. Besides, it could determine the wasted bandwidth and the total error rate.

The next research was a survey that revealed several factors associated with the predictive text system (Vitoria and Abascal, 2006). The research stated that there are eight important factors that affect a predictive text system. They were size of the text block, dictionary structure, prediction method, effect of the language used, the system interface, system adaptability, system usability, and other special features. The result stated that there are five prediction methods that can be used in predictive text. They are prediction using frequencies, prediction using word probability tables, syntactic prediction using probability tables, syntactic prediction using grammar, and semantic prediction. The survey also concluded that the result of a predictive text system was expressed in terms of keystroke saving and hit ratio or predictive accuracy of a system can be considered as another measure tools of predictive text system.

In 2008, there was a research that found a standard of keystroke saving in evaluating a word prediction system (Trnka and McCoy, 2008). The result of this research stated that there are two limits or boundaries that can become a standard evaluation of a word prediction system. The two limits are theoretical keystroke saving limit and vocabulary limit.

Furthermore, there was a new development of the predictive text system by incorporating some prediction methods, such as using the rules of English grammar to help text prediction and by adapting to the amount of word usage frequency (Nalavade, Mahule and Ketkar, 2008). The research result declares the incorporation of these methods can reduce KSPC by 26.91% compared to the T9 predictive text system.

The combination of semantic methods, frequency, and part-of-speech model on keypads was used in the next research (Gong, Tarasewich, and MacKenzie, 2008). The result showed that it can improve the text entry speed by 10% and reduce errors as much as 20% depending on the keypads. A year later, subsequent research did a combination of syntactic and semantic method (Ganslandt, Jorwall, and Nugues, 2009). The result declared that it can reduce KSPC error in the Sweden corpus as much as 12.4%. In addition, when the combination of syntactic and semantic coupled with the bigram method, it can reduce the error up to 29.4%.

The next research was about a predictive text system based on n-gram method (Verberne, et.al, 2012). N-gram was known as buffer and there are two forms of buffer types (n-gram) which are 'current prefix of the word' and 'buffer15'. The 'buffer15' gave a better result than 'prefix of the current word'. The summary of the combination of predictive methods can be seen in Table 1.

3 PROPOSED ALGORITHM

The purpose of this research is to develop predictive text system by combining some prediction methods that hopefully can give smaller KSPC value than previous researches. Methods that are used in this research are:

3.1 Frequency

Frequency method is used to rank words in the word table. It is based on how many times the word were typed by the user. This method works by adding the value of used word incrementally. By using this method, predictive text system will offer words that

Table 1: Predictive Methods from Previous Research.

Researchers	Year	Prediction Methods	Result
Nalavade, Mahule, and Ketkar	2008	Frequency and rules of English grammar	Decrease KSPC by 26.91%
Gong, Tarasewich, and MacKenzie	2008	Semantic methods, frequency and part-of-speech model	Improve text entry speed by 10% and decrease error as much as 20%
Ganslandt, Jorwall, and Nugues	2009	Syntactic and semantic method	Decrease KSPC error as much as 12.4%
Ganslandt, Jorwall, and Nugues	2009	Syntactic, semantic method and bigram	Decrease error up to 29.4%

are frequently used by user.

3.2 Probability Tables

Probability table is a table that is designed to store data. The data are predefined phrases from Indonesian dictionary and corpus that has been trained. Phrases are stored as static so user can select faster on the prediction.

3.3 n-gram

N-gram is as a buffer that can be trained to predict the next word based on previous word. In this research, the used n-gram is bigram as the differences between bigram and trigram do not produce a significant difference and trigram makes computing more complex. Therefore, bigram is the most appropriate choice for this research. Training result from bigram will be stored into probability table.

3.4 Syntactic using Grammar

In this method, the system predicts the next word based on syntactic relationship between previous word and next word that has a greater frequency. This relationship can be determined from data training by n-gram. When data training is finished, it will show the best probability of syntactic relationship that can be used.

Database structure of this predictive text system will contain three tables: word table, probability table, and syntactic relationship table. Word table contains all proper words that exist in Indonesian dictionary: *Kamus Besar Bahasa Indonesia* (KBBI), 3rd edition.

Probability table is a table that contains predefined phrase from KBBI and result from training process using n-gram (bigram). Meanwhile, syntactic relationship table is a table that contains data about probability of syntactic relationship from trained words. This table will be used as a reference table for prediction to predict next word from the greatest to the least probability of syntactic relationship.

The sequential steps for predictive text system to produce the desired word prediction are:

1. User types first character of desired word.
2. Predictive text system will trace words from word table which its first character similar with user typed.
3. System will offer collection of words sorted by frequency value from bigger to smaller and the highest syntactic relationship. If the desired word is found, user can choose the word by pressing predefined buttons on the keyboard. In this research, predefined buttons are listed from number one (1) to seven (7). Afterwards, frequency of chosen word will be incremented.
4. If the desired word is not found, user can type the next character and return to second step or word typed until complete.
5. Later on, when user presses space bar button, system will show next prediction from trained word by n-gram method and predefined word as a phrase that are stored in probability table. The word has the biggest probability of syntactic relationship from previous word.
6. When desired word is found, frequency from the selected phrase or trained words by n-gram will also be incremented either in the word table or in the probability table.
7. After pressing space bar button, if the desired word is not found, user can repeat the first step. Then, when desired word is found, user will press space bar button and n-gram (bigram) will learn by catching two words in front of space bar sign.

8. System will look for the syntactic grammar and store it into syntactic relationship table from words that just stored and learned by n-gram (bigram) in probability table. It will be used for the next learning to decide the best probability of syntactic relationship by adding its frequency.
9. The process will be repeated from the beginning to the last step until all words have been completely typed.

In this research, the performance of predictive text system was measured by using KSPC formula without concerning errors or mistakes made by the user and keystroke saving. By this limitation, the used KSPC formula is adopted from MacKenzie and Ishii (2007) as follows:

$$\text{KSPC} = \frac{\sum_{w \in W}(K_w X F_w)}{\sum_{w \in W}(C_w X F_w)}$$

Details of above formula: KSPC is the value of keystrokes per character, *w* is a word in the word model *W, Kw* is the number of keystrokes required to enter *w*, *Fw* is frequency count for *w*, and *Cw* is the number of characters in *w*. The reason of using this KSPC formula is based on previous researches that mostly use KSPC formula to evaluate the performance of predictive text system. KSPC value for QWERTY keyboard is one (1) because each buttons represent a single character. KSPC value must be lower than one (1) or the smallest KSPC value for better performance on predictive text system.

KSPC value represents how many keystrokes are needed to type a document. Meanwhile, keystroke saving represents how many keystrokes that are saved. The used keystroke saving formula is adopted from Trnka and McCoy (2008) which was stated in below formula:

$$KS = \frac{keys_{normal} - keys_{with\ prediction}}{keys_{normal}} \text{X } 100\%$$

Where:

KS = Keystroke saving.
$keys_{normal}$ = The number of keystrokes for every words.
$keys_{withprediction}$ = Number of keystrokes that required to entry a word with predictive text system.

Depicted from above formula, Keystroke Saving (KS) is the amount of how many keystrokes have been saved by the predictive text system.

4 RESULT AND DISCUSSION

To make sure this research goal is achieved, the predictive text system is tested by using the comparison of three prediction method groups. Those method groups are shown in Table 2.

Table 2: Method Groups.

Method Group	Prediction Method Combination				Database
	Frequency Method	N-gram	Probability Table	Syntactic Using Grammar	
Dictionary	-	-	X	-	KBBI
Frequency	X	-	X	-	KBBI
Syntactic	X	X	X	X	KBBI

Where:
x = Used
- = Unused

The testing method uses the best case scenario when user does not make any mistakes in typing articles or documents. The test data or sample was collected manually from *www.liputan6.com*. *Liputan6* is a news program from one of Indonesia's most popular television news channel called *Surya Citra Televisi (SCTV)*. It is known for delivering actual, sharp, and trusted news in Indonesia. The 8 (eight) articles adopted from 4 (four) categories or topics are collected from *Liputan6.com*'s online article on 13th September 2013. The selected topics are business, politic, health, and sport.

Steps of testing the predictive text system in this research are:

1. Data will be trained for each prediction method. Prediction method is divided into three groups as shown in Table 2. For the details, those method groups are: Dictionary (prediction is only based on dictionary and probability table without n-gram), Frequency (prediction is based on dictionary with frequency method and probability table without n-gram and syntactic using grammar), and Syntactic (prediction is based on dictionary with frequency, probability, n-gram, and syntactic using grammar).
2. Each method will train 2 (two) articles in one topic sequentially.
3. When the first article has been typed, the KSPC value will be recorded. The process will be repeated to the second article.
4. Those articles will be tested again, and the KSPC value will be recorded.
5. Then, the process will continue to the next topic. Follow the second step until the fourth step with the next topic.

6. Furthermore, KSPC value of each method group will be shown and compared to find the most effective prediction method.

For testing, the algorithm of this research was implemented to a desktop application. It was built in C# programming language and Microsoft Access as the database that contains all of the proper words and phrases. The database contains about 42,000 proper Indonesian words and about 17,000 most used Indonesian phrases. All of the words and phrases were obtained from KBBI, 3rd edition that was published by Indonesia's official language organization.

The application was designed so user can type with the help of predictive text system. User can directly type the article in the application as shown in Figure 1.

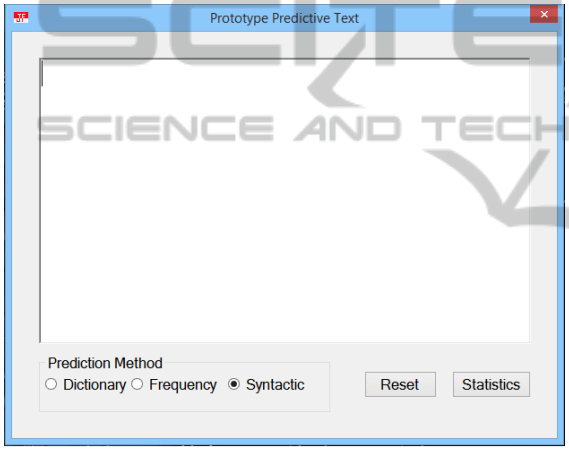

Figure 1: Main Display of Application.

In Figure 1, user can choose the prediction method as explained in first step and there are two buttons, which are "Reset" and "Statistics" buttons. By pressing "Reset" button, it will clear all words from text entry area and by pressing "Statistics" button, it will show a new window as shown in Figure 2.

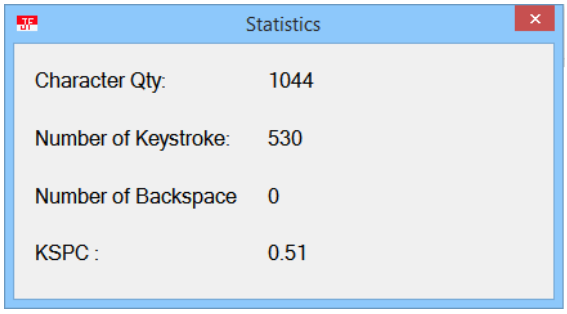

Figure 2: Statistics Window.

In Figure 2, there are 4 (four) statistical values that are displayed. Character Qty represents the length of all words from the article. Number of Keystroke represents how many words that user typed for the article. Number of Backspace represents how many times that backspace button pressed by user. KSPC represents the result from calculation of KSPC formula as stated before.

When user types one character in the application, it will give prediction suggestion. User can choose the desired word by pressing the number that represents the word as shown in Figure 3.

Figure 3: Prediction Suggestion.

When tester does those steps above, it will be shown in Figure 4, Figure 5, and Figure 6.

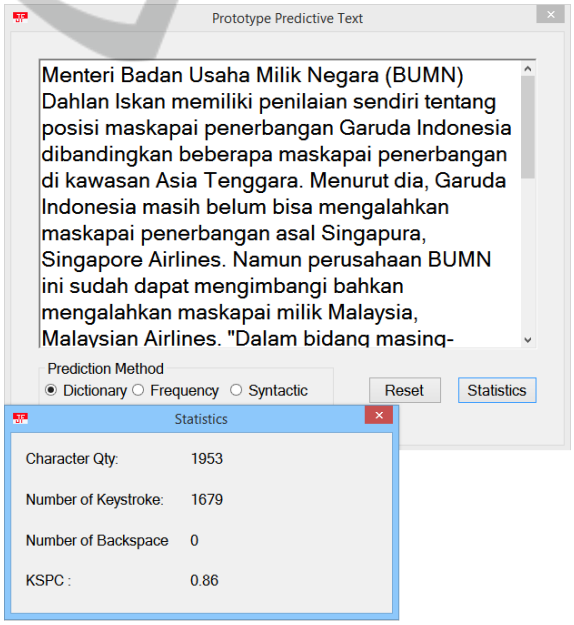

Figure 4: Prediction with Dictionary.

In Figure 4, predictive text system only makes a prediction based on dictionary and probability table without n-gram. And in Figure 5, the system is based on dictionary with frequency method and probability table without n-gram and syntactic using grammar.

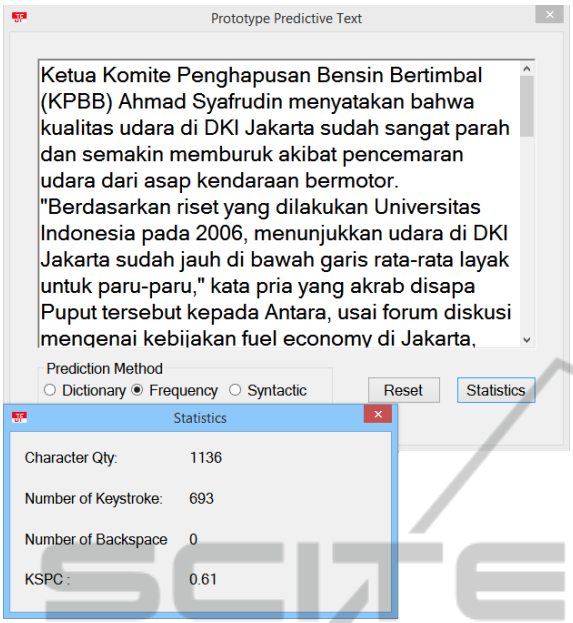

Figure 5: Prediction with Frequency.

In Figure 6, the system makes a prediction based on dictionary with frequency method, probability table, n-gram, and syntactic using grammar.

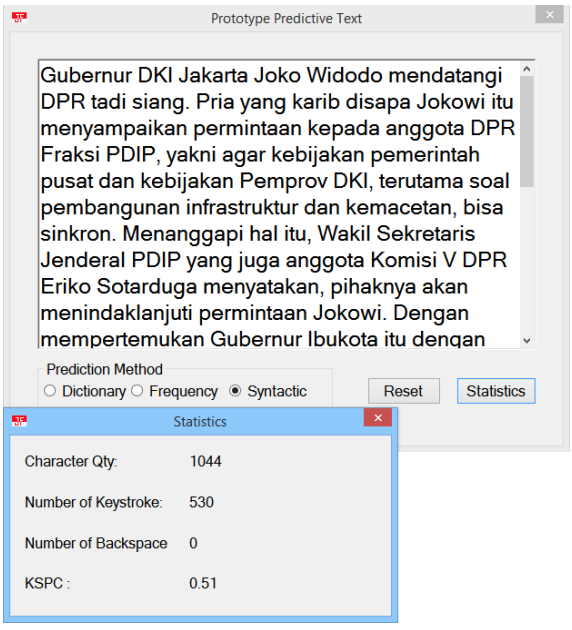

Figure 6: Prediction with Syntactic.

After testing in accordance to the steps above, the result are shown in Table 3, Table 4, and Table 5.

In Table 3, all of the words are typed which included the punctuation marks, foreign language, and people or organization name. In this table, there

Table 3: Test Results for Training the Articles.

Topic	Article	KSPC		
		Dictionary	Frequency	Syntactic
Business	1	0.85	0.74	0.76
	2	0.82	0.75	0.75
Politic	3	0.83	0.72	0.75
	4	0.88	0.79	0.83
Health	5	0.86	0.74	0.79
	6	0.88	0.70	0.68
Sport	7	0.92	0.81	0.82
	8	0.91	0.79	0.84
Average		0.87	0.76	0.78

Where:
Topic = Selected topics from source.
Article = Article's sequence number.
KSPC = Keystroke per character from each method groups.
Average = KSPC average from each method group.

are no significant numbers as it is the first time that system learns (for Frequency and Syntactic). The result shows that the KSPC value is still high.

After training the data, the articles are tested again included the punctuation marks, foreign language, and people or organization name. And the result is shown in Table 4.

Table 4: Test Results for Testing the Articles.

Topic	Article	KSPC		
		Dictionary	Frequency	Syntactic
Business	1	0.85	0.61	0.53
	2	0.82	0.60	0.51
Politic	3	0.83	0.61	0.51
	4	0.88	0.66	0.56
Health	5	0.86	0.63	0.52
	6	0.88	0.63	0.54
Sport	7	0.92	0.72	0.62
	8	0.91	0.71	0.63
Average		0.87	0.65	0.55

In Table 4, there are many differences that occur from the result, especially in Frequency and Syntactic method group. Both of them have smaller KSPC value than Table 3. In Dictionary method group, there is no difference from the previous experiment because the prediction is only based on dictionary (KBBI) and probability table without n-gram. But the result of Syntactic method group is not satisfactory because of the limitation of dictionary and based on previous research which stated keystrokes saving in practice can achieve 50 until 60% (Trnka and McCoy, 2008).

Based on the previous experiment, the articles are tested again and focused solely on Bahasa (without foreign language and people or

organization name). The result is shown in Table 5.

Table 5: Test Results for Testing the Filtered Articles.

Topic	Article	KSPC		
		Dictionary	Frequency	Syntactic
Business	1	0.81	0.50	0.41
	2	0.81	0.58	0.48
Politic	3	0.83	0.60	0.48
	4	0.87	0.64	0.53
Health	5	0.85	0.59	0.47
	6	0.85	0.60	0.50
Sport	7	0.90	0.67	0.56
	8	0.90	0.67	0.59
Average		0.85	0.61	0.50

In Table 5, the result is much better than before. It shows that Syntactic method group is the most effective combination for predictive text system and can help people to save the keystroke about 50%.

5 CONCLUSIONS

Based on the test results, it can be concluded that the most effective method is Syntactic method group for Bahasa (prediction is based on dictionary with frequency, probability table, n-gram, and syntactic using grammar methods). It can save the keystrokes until 50% (average) from each article with the best saving is 59% and the lowest is 41%. In this research, there are still many limitations for this predictive text system, caused by vocabulary limit. This research cannot find the newest edition of dictionary (*Kamus Besar Bahasa Indonesia*, 4thedition) because it is not released as a digital data yet, as so many articles contain special name or acronym that is not supported by the system. With a better and complete Bahasa database, the predictive text system should be able to improve the keystroke saving up to 60% focused solely on Bahasa.

REFERENCES

Ganslandt, S., Jorwall, J., Nugues. P., 2009. Predictive Text Entry using Syntax and Semantics. *Proceedings of the 11th International Conference on Parsing Technologies (IWPT)*. Association for Computational Linguistics.

Gong, J., Tarasewich, P., MacKenzie, I. S., 2008. Improved Word List Ordering for Text Entry on Ambiguous Keypads. *Proceedings of the Fifth Nordic Conference on Human-Computer Interaction - NordiCHI 2008*. ACM.

Jurafsky, D., Martin, J. H., 2008. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Pearson, Prentice Hall. New Jersey, 2nd edition.

MacKenzie, I. S., 2002. KSPC (keystrokes per character) as a characteristic of text entry techniques. *Proceedings of the Fourth International Symposium on Human–Computer Interaction with Mobile Devices*. Springer-Verlag.

MacKenzie, I. S., Ishii, K. T., 2007. *Text entry systems: mobility, accessibility, universality*. Morgan Kaufmann Publishers. San Francisco.

Nalavade, D., Mahule, T., Ketkar, H., 2008. PreText: A Predictive Text Entry System for Mobile Phones. *Proceedings of the World Congress on Engineering 2008 Vol III*. International Association of Engineers.

Soukoreff, R. W., MacKenzie, I. S., 2003. Metrics for text entry research: An evaluation of MSD and KSPC, and a new unified error metric. *Proceedings of the ACM Conference on Human Factors in Computing Systems - CHI 2003*. ACM.

Trnka, K., McCoy, K. F., 2008. Evaluating Word Prediction: Framing Keystroke Savings. *HLT-Short '08 Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers Pages 261-264.* Association for Computational Linguistics.

Verberne, S., Bosch, A. V. D., Strik, H., Boves, L., 2012. The effect of domain and text type on text prediction quality. *EACL '12 Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

Vitoria, N. G., Abascal, J., 2006. Text prediction system: a survey. *Universal Access in the Information Society*. Springer-Verlag.