Tokenization and Lemmatization on German Learning Textbook Level A1 of CEFR Standard

M. Kharis
Universitas Negeri Surabaya
Universitas Negeri Malang

Kisyani Laksono*
Universitas Negeri Surabaya

Suhartono
Universitas Negeri Surabaya

Agus Ridwan
Universitas Negeri Surabaya

Mintowati
Universitas Negeri Surabaya

Yuniseffendri
Universitas Negeri Surabaya

This study aims to compare the number of vocabulary types and tokens contained in the German language learning textbooks Themen Neu, Studio D, and Netzwerk A1 level of CEFR standards and to describe the use of German Lemmatizer to identify the lemmas in the textbook. The SpaCy lemmatizes and parses the words. Both discussions with expert judgment and German language experts validate the data. Based on the analysis results, the number of vocabulary types and tokens in the three books in each chapter is always rises changes to always rises at the beginning, but increases and decreases unsteadily from the middle to the last chapter. In addition, the SpaCy lemmatizer is able to lemmatize and parse the form of the words and classify the word classes in German, although there are still errors in its analysis. Therefore SpaCy still has to improve the system in its dataset, to make a better analyze the form of lemmas and classify the word classes.

Keywords: SpaCy, lemmatizer, vocabulary type, token, word count

INTRODUCTION

Research on the number of lemma and vocabulary in languages has been done substantially. However, there are minor research reports and publications on vocabulary related to lemma mastery in tiered textbooks. Moreover, research on the measurement of lemma mastery of German learning textbooks in Indonesia at levels A1, A2, and B1 of the Common European Framework of Reference for Languages (in short CEFR) has never been carried out. In context of foreign language learning, research on measuring lemma and vocabulary is used to measure the level of language skills. This study is important because the breadth and depth of vocabulary have a significant correlation with language learning outcomes (Chen, 2020). Based on an initial search through the search engine Google Scholar by using the keyword 'lema bahasa Jerman' (German lemma in Bahasa Indonesian version) as the title, it was not found any documents related to the measurement of lemma and vocabulary. Meanwhile, when searching with the keyword "German lemma (in English version)" only found three documents and six citations.

Wikipedia explains that a lemma is an input word or phrase in a dictionary outside the definition given in the entry. Thus, lemma can be in the form of a root word, affixed word, repeated word, compound word, phrase, or acronym. For example, in English, the words *table* and *tables* are in different types and tokens, but these types are called lemma. Thus, a lemma is the head of the word, the form of the inflection and the form of reduction of the word (Nation, 2013). Knowledge of lemma is used to map vocabulary to lexical sources and establish relationships between their inflected forms. This is especially necessary for morphologically rich languages and for overcoming the distribution of variations in unstated forms (Müller et al., 2015).

In the context of foreign language skills, it takes at least 800 lemmas that are used in everyday life to be able to communicate well. However, more lemmas are required if one aims at producing more complex sentences. For instance, to understand dialogue in film or television a learner must acquire around 3,000 lemmas and to understand very well novels, newspapers, written articles, a learner needs to acquire 8,000 to 9,000 lemmas (Hutami, 2018). As the findings of the research of Ibrahim et al. Karami and Schmitt, an individual requires at least 8,000 to 9,000 lemmas to be able to read a variety of authentic reading texts well (Ibrahim et al., 2016; Karami et al., 2019; Milton, 2010; Schmitt et al., 2020), e.g. newspapers, magazines, and/or manuals. However, for an individual oral communication, it only takes about 5,000 to 7,000 lemmas (Schmitt et al., 2020). By having around 8,000 to 9,000 lemmas, it can be regarded that an individual acquires a good level of vocabulary mastery. It is a fact that the more vocabulary a person acquires, the better his/her language skills are.

The level of mastery of German and other languages in Europe is not categorized on how many lemmas and vocabulary a learner acquires, yet language mastery of a language learner is explained and described in the already mentioned CEFR, as a framework and guidelines for European language standards that apply internationally. There are six levels of mastery of German, specifically A1 and A2 (basic user), B1 and B2 (independent user), and C1 and C2 (profient). Each of these levels and the needed receptive and productive abilities are (Glaboniat et al., 2013). In Indonesia, textbooks that have been used for learning German up to level B1 in higher education include *Themen Neu* (Aufderstraße, 2002), *Studio D* (Funk et al., 2013). Currently, the book used in the university educational setting is *Netzwerk* (Dengler et al., 2017).

In this study, vocabulary learning refers to learning German vocabulary. Learning German vocabulary is integrated into the learning of four language skills which are usually thematic and tiered based on the level of difficulty, both word choice and grammar study. This research is important because the degree and comprehensiveness of vocabulary have a significant correlation with language learning outcomes (Chen, 2020).

This study aims at examining German language learning textbooks at the A1 level of CEFR, specifically *Themen Neu, Studio D*, and *Netzwerk*, which used by student as the German learner. The research focuses on comparing the number of types of vocabulary and tokens and the form of lemmas contained in the three books by using a software known as lemmatizer. There are several lemmatizers that can be used for lemmatizing the German language. The researchers used SpaCy to assist the lemmatization process. This is in line with Ferreira who stated that SpaCy is more user friendly and allows to process text through the

entire pipeline with one function, allows better synchronization of each pipeline stage, is personal, and can make editing more clear (Ferreira et al., 2019).

This study intends to compare the performance of the three textbooks mentioned above. Research results are used for the development and or selection of textbooks and/or textbooks used in subsequent German language learning.

METHODOLOGY

Research Goal

The research was carried out by collecting data, data categorization, data analysis, and drawing conclusions. Before the research began, observations, data collection and data sorting related to the research focus were performed. The data were obtained from the A1 level German learning textbook. The data collected was classified based on the focus of the study. To provide more focused discussion, thus it formulated two major research goal as follows:

1. To describe the types of vocabulary, tokens, and lemma are contained in the textbooks of *Themen Neu, Studio D,* and *Netzwerk A1.*
2. To describe how can SpaCy-Lemmatizer identify entry forms in data sources.

Sample and Data Collection

The data of this research were words in the book *Themen, Studio D* and *Netzwerk*. The three books served as a corpus or research database. Data collection was carried out by a combination of several software, specifically the Simple Concordance Program (SCP), Visual Studio Code, Python, and SpaCy Lemmatizer. These software assisted the researchers to generate corpus data from data sources.

Analyzing of Data

The data collected was obtained by using purposive sampling technique, specifically sorting the words in sequence until the data were obtained with as many variations as possible to determine the number of lemmas and the existing vocabulary. The data were analyzed using qualitative descriptive techniques by describing all important elements found in the data based on the facts. The validity of the data was carried out by means of repeated careful and detailed observations of the data found. In addition, the validity of the data was carried out through interpretation, and discussion with expert judgment, German language experts, and Focus Group Discussion (FGD) with colleagues.

RESULT

The process of obtaining data was carried out by calculating all vocabulary contained in the data source. Certain vocabulary was ignored, for instance foreword, table of contents, pictures and numbers which were not included as learning material because they were not related to learning material. The researchers carried out the process of cleaning the sections per this chapter manually. After being declared clean from unnecessary data, the file was processed with the Simple Concordance Program (SCP) software to determine the number and frequency of vocabulary words.

The SCP is an open-source software that can be downloaded at http://www.textworld.com/scp/ for free and can run on Windows and MacOS systems. SCP helps users to create word lists and search natural language text files for words, phrases and patterns. Thus, users can generate concordances and search for words or phrases through the KWIC concordance feature. The SCP presents a degree of statistical data related to the entered text. However, in this study, not all data analyzed in the SCP were presented in this section. Thus, only data related to the number of vocabulary words and lemmas were taken. After analyzing the three books, the following data were obtained:

TABLE 1
THE NUMBER OF VOCABULARY TYPES AND WORD COUNT

Lesson (L)	Vocabulary Types			Word count (Tokens)		
	Themen Neu	Studio D	Netzwerk	Themen Neu	Studio D	Netzwerk
	0	348*	0	0	897*	0
1	361	559	321	1029	1938	1273
2	298	644	498	975	2145	1579
3	576	386	549	1948	2570	1921
4	229	515	539	548	2527	2051
5	570	679	499	1859	2827	2145
6	553	718	679	1880	2206	2427
7	560	736	755	1576	2483	2674
8	619	606	499	1931	2441	2106
9	678	742	704	2034	2613	2575
10	725	680	624	1820	2627	2174
11	685	683	666	2163	2665	2412
12	765	713	617	2407	2334	2380
13	1000	n.a.	n.a.	2718	n.a.	n.a.
14	608	n.a.	n.a.	1465	n.a.	n.a.
15	829	n.a.	n.a.	2801	n.a.	n.a.
Total all chapters	9,056	8,009	6,950	27,154	30,273	25,717

Note: * additional chapter only by Studio D.
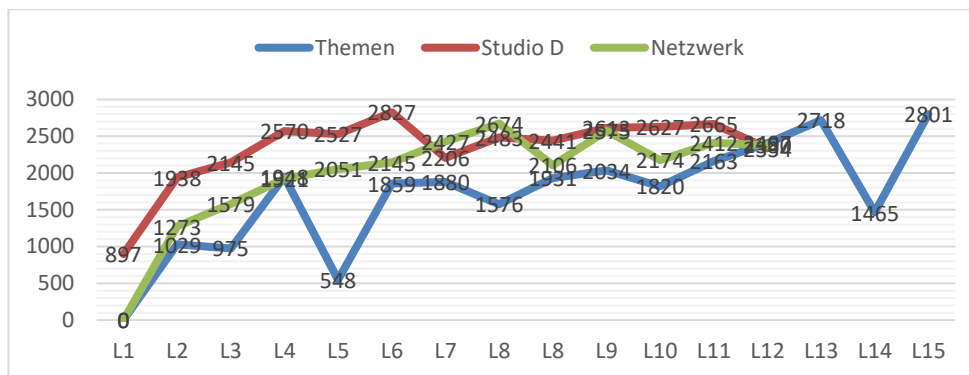n.a. stands for not applicable.

In diagrammatic form, the number of vocabulary types contained in the three books is described as follows:

FIGURE 3
THE NUMBER OF VOCABULARY TYPES
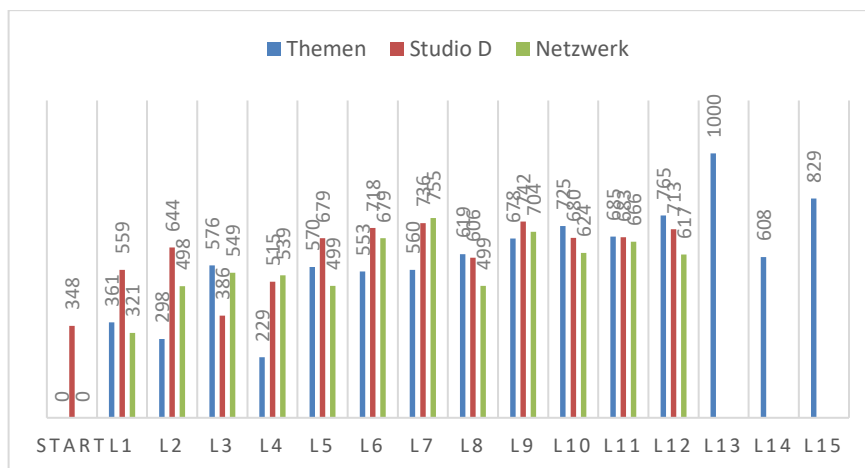
In diagrammatic form, the number of wordcount (tokens) can be described as follows:

FIGURE 4
THE NUMBER OF WORD COUNT (TOKENS)

Vocabulary type is the number of different words based on the form of writing. This means that words that have the same form are counted as one vocabulary. Meanwhile, wordcount (token) is the entire vocabulary, both the same and different forms. The diagram above shows that both the number of vocabulary types and wordcount (tokens) are relatively stable at the beginning, but rise and fall unstable in the middle of the textbook to the last chapter.

After data on the number of vocabulary types and tokens were collected, the data per chapter were analyzed using Lemmatizer SpaCy software to determine the pattern of the lemmas. To run SpaCy, the users must first install Python. This is an object-oriented high-level programming language software. Python is an open-source software, it means that anyone can download and use Python for free (Kedia & Rasu, 2020) and is popular among many programmers. In addition, in recent years, Python can perform sentiment analysis in languages other than English because of its multilingual support (Sharma, 2020). To reduce the complexity of the analysis procedure in Python, the researchers used Visual Studio Code software which runs Python and Lemmatizer SpaCy in one single software. Lemmatization with SpaCy assisted by Visual Studio Code software uses a markup code/ programming language code that looks as follows:

```
import spacy
nlp = spacy.load ("de_core_news_sm")
doc = nlp (u 'Hier sind die Sätze')
for token in doc:
print (token.text, token.lemma_, token.pos_)
```

The text entered in the column was analyzed based on the dataset attained by SpaCy. The sentence in the paragraph was then broken down by word and displayed in the form of the lemmas, as well as the word class category. An example of the results, how SpaCy's lemmatizer analyzes vocabulary and converts them into lemmas is seen in the following table:

TABLE 2
THE ANALYSIS OF LEMMAS IN SPACY

Word	Lemma	Class
Wörter	Wort	NOUN
geht	gehen	VERB
dir	sich	PRON
ist	sein	AUX
und	und	CCONJ
dreizehn	dreizehn	NUM
deutsche	deutsch	ADJ
Ach	Ach	INTJ
Ja	Ja	PART
heißen	heiß	VERB*
die	der	DET*
Ah	Ah	NOUN*

Annotation:

NOUN: noun,
VERB: verb,
PRON: pronoun,
AUX: auxiliary,
CCONJ: coordinating conjunction,

DET: determiner,
NUM: numeral,
INTJ: interjection,
PART: particle,

Another class code word, which SpaCy can recognize is:

ADJ: adjective,
ADP: adposition,
ADV: adverb,
CONJ: conjunction,
PROPN: proper noun,

PUNCT: punctuation,
SCONJ: subordinating conjunction,
SYM: symbol,
X: other,
SPACE: space

Based on the results of lemmatization above, Lemmatizer SpaCy is able to show the pattern of the lemma and determine the word class in German language. However, there was an error in SpaCy's analysis. In the table 2 are the examples of analysis errors (marked *). Based on the analysis results, SpaCy made no mistakes in the PUNCT and SPACE word categories and made only a few errors in the CCONJ, NUM, ADV, ADP analysis. This is because these words are included in words that do not experience changes in form (inflection). However, SpaCy has the potential to make a number of mistakes in the NOUN, PRON, ADJ, VERB, PART and AUX analysis, especially words that are derivative or can change its form. These words would have the potential to mess up the system in the dataset, because are they included in the word flexion category. For example, as is well known, the writing of nouns in German must begin with a capital letter. Meanwhile, imperative sentences always begin with a verb. In addition, verbs that can function as full verbs and auxiliary verbs are also one of the weaknesses of SpaCy analysis, for example the verbs *haben, sein*, and *werden*. Thus, imperative verbs and verbs that can function as full verbs and auxiliary verbs can mess up the SpaCy work system and have the potential for analysis errors. In addition, the DET word class in German consists of three, specifically masculine *der*, feminine *die* and neutral *das*. These three determinants can be changed its form depending on the case, whether nominative, accusative, dative, and genitive. The examples of SpaCy errors in the analysis are shown in the following table:

TABLE 3
THE ERROR IDENTIFICATION FOUND IN SPACY

Vocabulary	Lemma	Identified by SpaCy as	Actual word class
Bilden	Bilden	NOUN	VERB
ERFRAGEN	ERFRAGEN	PROPN	VERB
ein	einen	ADV	DET
das	der	PRON	DET
Danke	Dank	NOUN	AUSRUF

The following is a list of errors made by SpaCy was found in *Netzwerk* which is shown in the table below:

TABLE 4
THE NUMBER OF ERRORS ANALYSIS OF ALL CLASSES IN NETZWERK

Word Class	Errors
VERB	353
NOUN	338
DET	279
ADJ	171
PROPN	87
ADV	53
INTJ	42
AUSRUF	39
ABREVIASI	32
ADP	17
NUM	15
PART	8
PRON	8
CCONJ	3
AUX	2

Interestingly, there are also the same words, yet it generated differences in analysis. For example, it was found that the word Plural was analyzed as a class of words ADV and PROPN. The words Plural, if they start with a capital letter, are included in the NOUN word class category and are in the ADJ word class if they are written in lowercase. This shows that SpaCy is sensitive to the use of capital letters. Another word also occurred in the word sorgen which means to care, and Sorgen has the meaning worries. This also happens to the words bitte and Bitte.

Of all word classes, SpaCy does not give categories to exclamation types of words, in specific words that are utterances, for instance greetings, thanking, and apologizing. This type of data is often found, especially at the beginning of the book, because foreign language textbooks always begin with greetings.

SpaCy is able to identify the types of abbreviated words, although the number of known words is still limited, for example the abbreviation Tel. as Telefon with NOUN word class, usw. as und so weiter with the word class ADV, Okt. as Oktober with the NOUN word class. The abbreviation commonly used in everyday life cannot be recognized, as shown in the following table:

TABLE 5
UNIDENTIFIED ABBREVIATIONS BY SPACY IN NETZWERK TEXTBOOK

Abbreviations	Meaning	Word Class
ca.	circa	ADV
Dr.	Doktor	NOUN
lol	laugh out loud	ADV
mask.	maskulinum	ADJ
neutr.	neutrum	PROPN
P.S.	Postskriptum	NOUN
Zi.	Zimmer	NOUN
qm.	Quadratmeter	NOUN
Do.	Donnerstag	NOUN
Fr.	Freitag	NOUN
Akk.	Akkusativ	NOUN
Nom.	Nominativ	NOUN
z.B.	zum Beispiel	NOUN

Another weakness of SpaCy is its inability to analyze participle form verbs. As is known, verbs in German have three tenses, specifically present, past and participle. From data sources, SpaCy has indeed known numerous participant verbs, including *gegangen, gegessen, geklingelt, gekommen, geöffnet, gelernt, geplant, gemacht, geredet, gesucht, gefallen*, and even separable verbs, for example *angekommen, eingekauft*. However, some of the commonly used participle form verbs are not recognized, for example *gefahren* and *getroffen*.

DISCUSSION

The process of obtaining the lemma from a word is called lemmatization. More specifically, lemmatization is the process of acquiring a basic form, or lemma, a word from its inflected form (Perera & Witte, 2005). Meanwhile, Konrad argues that lemmatization is the process of acquiring the basic (or dictionary) form of a word that might change its lemma (Konrad, 2017).

The process of lemmatization for morphologically complex languages such as German and possibly other languages is not an easy task and this can only be done through particular algorithms. For instance, in German the suffixation changes are *-s, -es, -e, -n, -er,* and *-ern* and vowel changes due to the addition of *Umlaut*. Some examples are presented in table 2 above. From the analysis process in this study, it can be seen that SpaCy can carry out the lemmatization process with several limitations.

In German, changes in word form and class occur in many word classes. The morphology of word formation in German can also be characterized by a number of suffixation and prefixation processes (Schmid et al., 2004). In this case SpaCy is very sensitive to the writing of capital letters and this actually has the potential for errors in its analysis. For example, in the following analysis:

Inoffiziell → *inoffiziell* → ADV*
Inoffiziell → *Inoffiziell* → VERB*

The writing of capital letters *inoffiziell* results in differences in the classification of the word. Differences in analysis such as in the example above can also be experienced by other word classes. A number of lemmatization algorithms have been developed to recognize changes in a word. Normally, the way lemmatizers work is to cut off the suffix and mark other changes taking into account morphological features. For example, in the word *Ziele*, lemmatizer works by removing the suffix *e* in the word *Ziele*, thus, the lemma presented is *Ziel*. In word *Kinder*, lemmatizer works by removing the suffix *-ern* thus, the word

Kind is displayed. In *Häusern* words the process of lemmatization is more complicated, because in addition to removing the suffix *-ern*, it also removes the *Umlaut* sign *ä* thus, it generates *Haus*.

In general, existing software often cannot determine the shape of the lemma with 100 percent accuracy. Until now, we found no Lemmatizer which is able to detect and recognize the perfect derivation of all vocabulary words. In this study, SpaCy also has not been able to do lemmatization with a tolerance of 0%, as shown in table 2 above. This is motivated by the language system having complexity in it, including language style. For example, SpaCy cannot recognize spoken words that fall into the colloquial language and *Denglisch* (a mixture of German and English) categories e. g. *hab', find', geht's, lol, cool, babysitten*. This is because SpaCy only has a dataset for formal languages and cannot recognize foreign words. This inability is one of the weaknesses of the existing lemmatizers.

The diversity of vocabulary in the group cannot be given in one learning process, but in learning per lesson in stages both in terms of the number of words, the complexity of the word phonotactic pattern, and the number of words in a sentence. These criteria are in accordance with those set by Room to Read, specifically the complexity of words and sentences, the average number of sentences per page, the average number of words per sentence, topics, and illustrations (Nugroho, 2017). Ideally, through establishing these criteria the learner's language skills increase along with the material being studied (Chabibah et al., 2018). The data in figure 3 and 4 above show that the vocabulary types and tokens contained in the three books are not entirely ideal, because there are relatively unstable up and down charts. The number of types and tokens in each chapter should be stable and continue to rise, therefore, it is in accordance with the Room to Read guidelines.

From the explanation above, it can be concluded that the lemma can be used to measure an individual's vocabulary acquisition. In addition, knowledge of lemmas on a test may reflect knowledge of other lemma members. This is in accordance with Kremmel's statement which states that lemma measurement activities are proven to have advantages over the word field in terms of measuring lexical diversity (Kremmel, 2016). The words (root words and their inflections) can have word groupings to some degree, and at the same time can reflect the knowledge of the other words in the group. This is because the members of the lemma differ only in grammatical form.

CONCLUSION

The textbooks *Themen Neu, Studio D* and *Netzwerk* teach vocabulary type and tokens by introducing these vocabularies in stages both in terms of number of words, complexity of phonotactic patterns, and number of words in sentences and in paragraphs. However, the number of types and tokens in the three books are less than ideal because the numbers are not stable up, but up and down.

Based on the results of the lemmatization process, the SpaCy lemmatizer is able to show the form of a lemma as well as the word class in German, although there are errors in the analysis. Thus, lemma analysis still requires validation from human being.

REFERENCES

Aufderstraße, H. (Ed.). (2002). *Themen neu: Lehrwerk für Deutsch als Fremdsprache. 1. Kursbuch: ...* (3. Aufl). Hueber.

Chabibah, S., Kisyani, D., & Hum, M. (2018). *Perkembangan Kosakata dan Fonotaktik pada Buku Ajar BIPA Tingkat A1—C2 Terbitan Kemendikbud*, *1*, 9.

Chen, C.L. (2020). The role of vocabulary breadth and depth in IELTS academic reading tests. *Reading in a Foreign Language*, *32*(1), 1–27. Retrieved from http://Nflrc.Hawaii.Edu/Rfl

Dengler, S., Rusch, P., Schmitz, H., & Sieber, T. (2017). *Netzwerk: Deutsch als Fremdsprache. A1, Kursbuch: Kursbuch A1: mit DVD und Audio-CDs* (1. Auflage) [Medienkombination]. Ernst Klett Sprachen.

Ferreira, J., Oliveira, H.G., & Rodrigues, R. (2019). Improving NLTK for Processing Portuguese. In R. Rodrigues, J. Janousek, L. Ferreira, L. Coheur, F. Batista, & H.G. Oliveira (Eds.), *8th Symposium*

on Languages, Applications and Technologies (SLATE 2019) (Vol. 74, pp. 18:1–18:9). Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik. https://doi.org/10.4230/OASIcs.SLATE.2019.18

Glaboniat, M., Müller, M., Rusch, P., Schmitz, H., & Wertenschlag, L. (2013). *Profile deutsch: Lernzielbestimmungen, Kannbeschreibungen und kommunikative Mittel für die Niveaustufen A1, A2, B1, B2, C1 und C2 des "Gemeinsamen Referenzrahmens für Sprachen".* Klett Sprachen.

Hutami, T. (2018). *Berapa Banyak Kata yang Harus Dikuasai untuk Bisa Berbicara Bahasa Asing?* Retrieved from https://womantalk.com/lifehack/articles/berapa-banyak-kata-yang-harus-dikuasai-untuk-bisa-berbicara-bahasa-asing-xem7R

Ibrahim, E.H.E., Sarudin, I., & Muhamad, A.J. (2016). The Relationship between Vocabulary Size and Reading Comprehension of ESL Learners. *English Language Teaching*, *9*(2), 116–123.

Karami, A., Howlett, K.M., & Bowles, F.A. (2019). Implementing Dynamic Assessment In Assessing Vocabulary Development: What The Literature Says. *I-Manager's Journal on English Language Teaching*, *9*(21). https://doi.org/DOI: 10.26634/jelt.9.2.15518

Kedia, A., & Rasu, M. (2020). *Hands-On Python Natural Language Processing: Explore tools and techniques to analyze and process text with a view to building real-world NLP applications* (1st Edition). Packt Publishing.

Konrad, M. (2017, May 19). Lemmatization of German language text. *WZB Data Science Blog*. Retrieved from https://datascience.blog.wzb.eu/2017/05/19/lemmatization-of-german-language-text/

Kremmel, B. (2016). Word Families and Frequency Bands in Vocabulary Tests: Challenging Conventions. *TESOL Quarterly*, *50*(4), 976–987. https://doi.org/10.1002/tesq.329

Milton, J. (2010). The development of vocabulary breadth across the CEFR levels. In *Communicative proficiency and linguistic development: Intersections between SLA and language testing research* (pp. 211–232). First published by Eurosla.

Müller, T., Cotterell, R., Fraser, A., & Schütze, H. (2015). Joint Lemmatization and Morphological Tagging with Lemming. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 2268–2274. https://doi.org/10.18653/v1/D15-1272

Nation, I.S.P. (2013). *Learning vocabulary in another language* (Second Edition). Cambridge University Press.

Nugroho, D.A. (2017). *Best Practice Book Leveling Classification by Room To Read in SDN Sukorame Gresik* (p.8).

Perera, P., & Witte, R. (2005, October). A Self-Learning Context-Aware Lemmatizer for German. *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*, (pp. 636–643). Vancouver. Association for Computational Linguistics.

Schmitt, N., Nation, P., & Kremmel, B. (2020). Moving the field of vocabulary assessment forward: The need for more rigorous test development and validation. *Language Teaching*, *53*(1), 109–120. https://doi.org/10.1017/S0261444819000326

Sharma, M. (2020). Polarity Detection in a Cross-Lingual Sentiment Analysis using spaCy. *2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, pp. 490–496. https://doi.org/10.1109/ICRITO48877.2020.9197829