DEVELOPMENT OF INDONESIAN DICTIONARY AS INDONESIAN NLP'S RESOURCES

Rizki Hilmi Fauzi[1], Ken Kinanti Purnamasari[2]

[1,2] Teknik Informatika – Indonesian Computer Universitas
Dipatiukur Street No. 114-116 Bandung 40132
E-mail : rizkihill@gmail.com[1], kinantiken@gmail.com [2]

ABSTRACT

Dictionary is a book which contains words along with it meanings, usages and translations. The Great Indonesian Dictionary consist of lemmas, dictionary labels and word's meanings. The development of Indonesian dictionary have researched in "Dictionary Development Resource Type Words As NLP Indonesian" has been detecting 38.870 lemmas. Then the research titled "Development of Word's Type Dictionary as Indonesian NLP's Resources" has developed which has been detecting 43.060 lemmas. After that the research by detecting compound word lemmas in research titled "Development of Equipped Compound Words of Word's Type Dictionary as Indonesian NLP's Sources" which has been detecting 51.147 lemmas. The problem from the earlier research was that the dictionary labels other than word's types and word's meanings was not detected yet.
From the result of this research, to detected lemma, dictionary and word meaning, need to do some step, such as preparation of inputted data for fixing the inputted data so that have same entry pattern and step of lemma detection, dictionary labels and word meaning for detecting elements of dictionary.
The result this research has been detected that lemma as dictionary elements, dictionary label, and the words meaning, and reach as much 52.978 lemmas with 100% percentage from all lemmas resulting from counting manualy on KBBI which is used as inputed data. Based on analysis result, increasing of data detector because detection of data using tag mark html strong for lemma and tag mark html em for dictionary label. So can detected all lemmas on inputted data.
Key word : Indonesian dictionary, online dictionary, lemma, Wordnet.

1. PREFACE

Dictionaries are a reliable source of reference in understanding the meaning of the word of a language, because dictionary contains many vocabularies, ideally it is unlimited. Dictionaries are arranged in alphabetical order, accoumpanied by a dictionary label, describing the use of a word and its meaning. Indonesian language or bahasa has several varieties of words that are derivative words, compound words, and repetitive words.

Indonesian dictionary resources are needed for research in NLP (Natural Language Processing) Indonesian language such as early tokenisasi on POS Tag, Wordnet, and question answering. Nowadays the Indonesian dictionary is not only in a hardcopy, but also the development has been done to make Indonesian dictionaries to a digital form such as website, among others; Kbbidaring.info.com, kamusbahasaindonesia.org and etc, but currently digital dictionary development is not an open source. This resulted that the user can only do a word search without being able to download the resources of Indonesian language dictionary that already exist can not be used for research in the field of NLP Indonesian language.

The fourth edition of Indonesian Dictionary published in 2008, contains about 90.049 enteries consists of 41.250 basic words, 24.607 words affixed and 23.536 combined words. In the first study of the Indonesian dictionary entited "Dictionary Development Resource Type Words As NLP Indonesian" by C G Ceppy Efraim Bolly has succesed in detectinng an entry in Indonesian dictionary with a percentage of 73,37% [1]. Then the development has been done by re-analyzing the stages to detect the words and types of words by Arief Adiguna Putra with study entitled "Development of Word's Type Dictionary as Indonesian NLP's Resources" has succesed in detecting the entry with a percentage of 82,28% [2]. Furthetmore, the other development has been done by adding the detection of compound words to increase the number of words detected, by Yoppy Yansyah with a study entitled "Development of Equipped Compound Words of Word's Type Dictionary as Indonesian NLP's Sources" can detected an entry with a percentage of 96,54% [3]. Previous research has focused on word detection and dictionary labels in the form of word types, whereas dictionary labels in the form of word type, and word meaning are not detected. The resulted in an entry that has focused on undetected word type, thus affecting the number of entries that were succesfully

detected. In addition, because it does not detect the meaning of the end result dictionary Indonesian languagecan not be used for research NLP Indonesian language such as early tokenisasi on POS Tags, Wordnet, question answering.

So in this research, we will be do the element detection in the form of entries, dictionary labels and meaning of words to increase the number of succesfully detected entries. In addition, we wil analyze the processes taken in the previous research to find a solution to the problems that accurred in the previous research.

2. RESEARCH CONTENT

Based on the background that has been presented, that in prevoius research not detecting dictionary elements in the form of a dictionary label other than word type and dictionary elements the form of meaning of word, so on this research will be added by detecting the dictionary with lemma, dictionary label, and meaning of word.

2.1 Dictionary

Etymologically, dictionary word comes from arabic language, that is qamus (plural from qawamus). Arabic language absorb dictionary words from ancient greek, okeanos which means the ocean.

Kridalaksana said that if the dictionary is a reference book that lists the word combined word with descriptions of various meanings and its use in the language, dictionaries are usually arranged alphabetically. [4] Based on the dictionary sense expressed by the experts, it can be summed up as follows.

1. The dictionary includes a reference book containing the words or word combination of language.
2. The words are arranged alphabetically.
3. The words are given a description of meaning and usage.
4. The words is in addition to being given a description of its meaning, also given information on how to saying, spelling, and others.
5. The description of that meaning is given also in other languages; in this research dictionary used input data source to obtain word and type of word.

In this research dictionary used as input sources to obtain the word, dictionary label and meaning of words.

2.2 Problem Analysis

problem in previous studies that effected the number of succesfully detected lemma did previous research only detect lemma and word type, so lemma has dictionary labels other than word type, for example a dictionary label ark (arkais) not detected. Other than that, lemma which means more than one word (homonym) it only detect one lemma because there is no word which means distinguishing lemma from another one, for example lemma ‘A’ the first alphabet in Indonesian with ‘a’ being ampere in unit of measure of electric current. Futher analyzed in previous research, so it can solve problems that occurred in previous research.

Table 1 The Process of The Previous Research Process

No	**Research Ceppy Bolly [1]**	**Research Arief Adiguna Putra [3]**	**Research Yoppy Yansyah [4]**
1	Penghilangan Digit Diawal Kalimat	Penghilangan Kalimat Dengan Delimeter Digit	-
2	-	-	Penghapusan Nomor Makna
3	Penghilangan Spasi Kosong Diawal Kalimat	Penghilangan Spasi Kosong Diawal Kalimat	-
4	Penghilangan Spasi Ganda	Penghilangan Spasi Ganda	-
5	-	-	Penghapusan Karakter Entitas *Html*
6	Penghilangan *Blank Line*	Penghilangan Baris Kosong	Penghapusan Baris Kosong
7	Penghilangan Baris Di awali Simbol	Penghilangan Baris Di awali Simbol	-
8	-	-	Penyamaan Simbol Kata Majemuk
9	-	-	Pengisian Kata dan Jenis Kata Untuk Kata Majemuk
10	-	-	Penyatuan Kata yang Terpisah
11	-	-	Pemisahan Kata Majemuk
12	Penghilangan Simbol Kecuali	Penghilangan Simbol Kecuali	Pembersihan

	Strip	Strip	
13	Pengambilan Kata dan Jenis Kata	Pengambilan Kata dan Jenis Kata	Pengambilan Kata dan Jenis Kata
14	Penghilangan Baris yang kurang dari dua kata	Penghilangan Baris yang kurang dari dua kata	-
15	Pengecekan Kata Perbaris	Pengecekan Kata Perbaris	Pemisahan Kata Dasar Dengan Kata Turunannya
16	Pengurutan	Pengurutan	-
17	-	Pengecekan Kata Duplikasi	-
18	-	-	Penghapusan Label Kecuali Label Jenis Kata
19	-	-	Pengkategorian

2.3 Solution Analysis

solution analysis is the stage of understanding of the system or application created by identifying existing problems and the needs needed to make system better.

On lemma detection process research, the dictionary label and meaning of word will use method role-based, by determining the steps needed to get the lemma detection, dictionary label and meaning of word use function from regular expression or often called regex by searching and replace the world base on the pattern of its appearance in the dictionary.

In this study will input source html format because it can detect the writing bold and italics. In Indonesian dictionary bold writing to write lemma detection process, dictionary label and meaning of the words so the result are better.

In the analysis of system solutions to be run has 2 stages that is stages of compilation of input data and lemma detection, dictionary label and meaning of word, then put in the database, then the data can be converted into several formats and output. Explanation of each stage can be seen in the following figure.

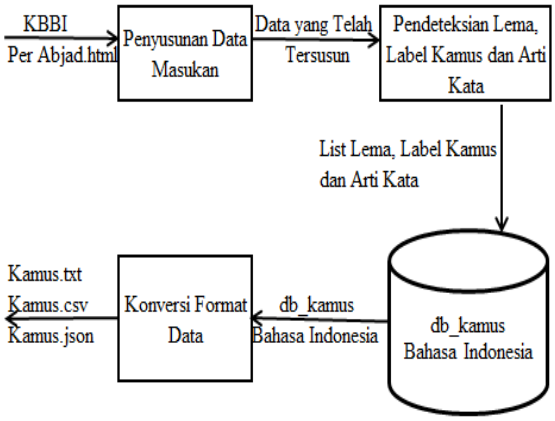


Figure 1. General Stages of The System

In the picture above the description of general stages of system will bw described as follows.

1. In the early stages including the data sources that is Indonesian dictionary that has been divided alphabetically from A to Z with html format into the system.
2. In the compilation stage of input data will be processed removal of html entity characters, unified compound words, deletion dropcap, deletion tika, the unification of the meaning of word and the removal of punctuation. This stage aims to set the pattern of input data previously decapitated by space and enter, so the detection process gets better results.
3. Next, phase of word detection, dictionary label and meaning of word synonym process is moved, replacement of compound symbols, meaning of the word number replacement, entry elements detection. The end result of this stage is the world list, dictionary label and meaning of words.
4. Then the word lists, dictionary label and meaning of word stored in the database. In this study the data base used ismySQL. The final step is to convertthe data format into several format namely .txt, .csv, .json.

2.4 Analysis of Input Data

Analysis of input data on the Indonesian dictionary development system is describe the data input process in the form of Indonesian dictionary data with html format to be processed by the system. Data Indonesian dictionary input with pdf format will be converted to html format with 2 stages among others :

a. Division File

The initial input data is a file that is divided in to alphabetically from A to Z using the Nitro Pro 10 program. This process aims to avoid technical problems, when the system is running because of the amount of data being processed simultaneously. For more details described in the following block diagram form.

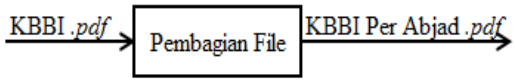


Figure 2 Block Diagram Division File

The following is an example of initial input data before the file sharing process development of files that can be seen in the following figure.

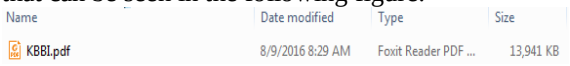


Figure 3 Indonesian Dictionary With Pdf Format

The following is an example of input data after the Indonesian dictionary alphabetically file sharing process can be seen in the picture below.

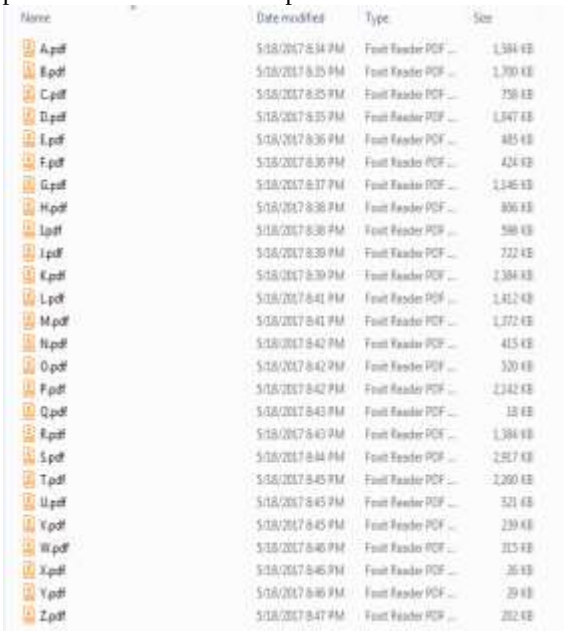


Figure 4 Indonesian Dictionary Alphabetically With Pdf Format

b. Data Conversion

The input data conversion process is done 2 times, first pdf format input data in conversion to doc format using the nitro pro 10 program. This process is done for the convertion processto be html format to get better result. For more details described in the following block diagram form.

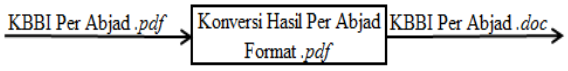


Figure 5 Block Diagram Conversion Input Data With Doc Format

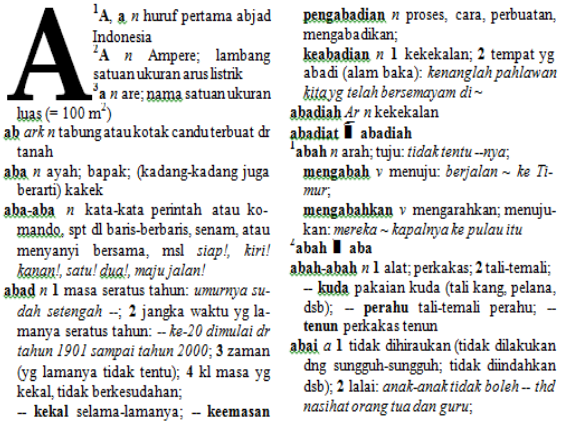
The following is an example of input data before the data conversion process can be seen in the following figure.



A, **a** *n* huruf pertama abjad Indonesia
A *n* Ampere; lambang satuan ukuran arus listrik
a *n* are; nama satuan ukuran luas (= 100 m²)
ab *ark n* tabung atau kotak candu terbuat dr tanah
aba *n* ayah; bapak; (kadang-kadang juga berarti) kakek
aba-aba *n* kata-kata perintah atau komando, spt dl baris-berbaris, senam, atau menyanyi bersama, msl *siap!, kiri! kanan!, satu! dua!, maju jalan!*
abad *n* **1** masa seratus tahun: *umurnya sudah setengah* --; **2** jangka waktu yg lamanya seratus tahun: -- *ke-20 dimulai dr tahun 1901 sampai tahun 2000*; **3** zaman (yg lamanya tidak tentu); **4** kl masa yg kekal, tidak berkesudahan;
-- **kekal** selama-lamanya; -- **keemasan** **pengabadian** *n* proses, cara, perbuatan, mengabadikan;
keabadian *n* **1** kekekalan; **2** tempat yg abadi (alam baka): *kenanglah pahlawan kita yg telah bersemayam di* --
abadiah *Ar n* kekekalan
abadiat → **abadiah**
abah *n* arah; tuju: *tidak tentu --nya*;
mengabah *v* menuju: *berjalan ~ ke Timur*;
mengabahkan *v* mengarahkan; menujukan: *mereka ~ kapalnya ke pulau itu*
abah → **aba**
abah-abah *n* **1** alat; perkakas; **2** tali-temali; -- **kuda** pakaian kuda (tali kang, pelana, dsb); -- **perahu** tali-temali perahu; -- **tenun** perkakas tenun
abai *a* **1** tidak dihiraukan (tidak dilakukan dng sungguh-sungguh; tidak diindahkan dsb); **2** lalai: *anak-anak tidak boleh -- thd nasihat orang tua dan guru*;

Figure 6 Input Data With Pdf Format

The following conversion result after conversion data into doc format using the Nitro Pro 10 program. Can be seen in the following figure.



A, **a** *n* huruf pertama abjad Indonesia
A *n* Ampere; lambang satuan ukuran arus listrik
a *n* are; nama satuan ukuran luas (= 100 m²)
ab *ark n* tabung atau kotak candu terbuat dr tanah
aba *n* ayah; bapak; (kadang-kadang juga beranti) kakek
aba-aba *n* kata-kata perintah atau komando, spt dl baris-berbaris, senam, atau menyanyi bersama, msl *siap!, kiri! kanan!, satu! dua!, maju jalan!*
abad *n* **1** masa seratus tahun: *umurnya sudah setengah* --; **2** jangka waktu yg lamanya seratus tahun: -- *ke-20 dimulai dr tahun 1901 sampai tahun 2000*; **3** zaman (yg lamanya tidak tentu); **4** kl masa yg kekal, tidak berkesudahan;
-- **kekal** selama-lamanya; -- **keemasan** **pengabadian** *n* proses, cara, perbuatan, mengabadikan;
keabadian *n* **1** kekekalan; **2** tempat yg abadi (alam baka): *kenanglah pahlawan kita yg telah bersemayam di* ~
abadiah *Ar n* kekekalan
abadiat → **abadiah**
abah *n* arah; tuju: *tidak tentu --nya*;
mengabah *v* menuju: *berjalan ~ ke Timur*;
mengabahkan *v* mengarahkan; menujukan: *mereka ~ kapalnya ke pulau itu*
abah → **aba**
abah-abah *n* **1** alat; perkakas; **2** tali-temali; -- **kuda** pakaian kuda (tali kang, pelana, dsb); -- **perahu** tali-temali perahu; -- **tenun** perkakas tenun
abai *a* **1** tidak dihiraukan (tidak dilakukan dng sungguh-sungguh; tidak diindahkan dsb); **2** lalai: *anak-anak tidak boleh -- thd nasihat orang tua dan guru*;

Figure 7 Input Data With Doc Format

Then conversion input data from doc format to html format using html cleaner. For more details described in the nfollowing block diagram form.

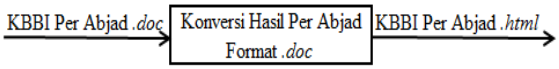


Figure 8 Block diagram conversion input data with html format

The following is an example of data entered after the data conversion process can be seen in the following picture.



Figure 9 Input Data With Html Format

2.5 Proses Process Analysis

Process analysis consist of several stages namely compile input data and word detection, dictionary label and meaning of words.

Stages of compilation of input data consist of 6 processes namely the process of removing html entities, compound word, removing dropcap, removing tika, the union of meaning is separate, and removing of punctuation text. Stages of word detection, dictionary label and meaning of word consists of 4 processes namely synonymous word separation process, symbol replacement, word number replacement, and entry detection. For more details can be seen in the following figure.

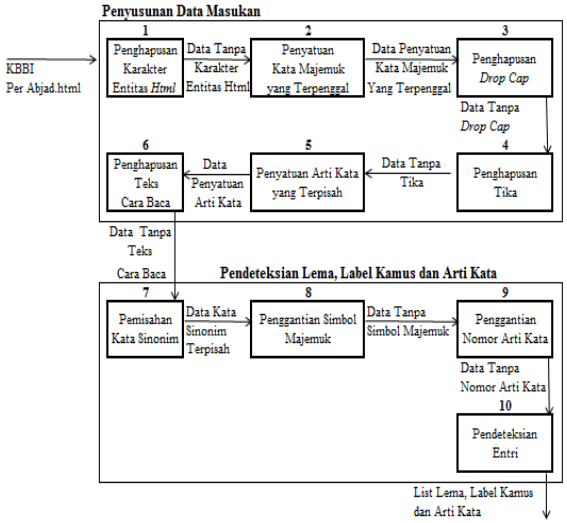


Figure 10. Process Scheme

2.5.1 Stage of Input Data Preparation

Arranging data input stage to correct the input data to have same entry pattern, namely lemma, detection labels and meaning of word. So in detection process obtain better results.

Arranging data input consists of 6 processes namely removing html entities compounding a saparet compound word, removing dropcap, removing tika, the union of meaning is saparet, andremoving of punctuation text.

The following example input data before and after the drafting stage of data input is done.

Table 2 Example stage of compilation of input data

Before
abadiah Ar n kekekalan   abc /abécé/ n 1 abjad Latin: tidak tahu -- , tidak tahu membaca huruf Latin; 2 ki hal- hal pokok yg paling pertama harus diketahui dr suatu keadaan atau perkara: belum tahu -- kehidupan   1acah v, beracah- acah bermain-main; ber- pura-pura;   2acah v, mengacah melanggar: murid yg ~   akan dihukum   asap n gas yg tampak keluar dr sesuatu yg panas, mendidih, atau dibakar;   belum dipanjat -- kemenyan, pb belum kawin; menggantang -, pb melakukan   perbuatan yg sia-sia; berangan-angan yg   hampa;   -- air gas yg keluar dr air yg mendidih; uap; -- api asap dr api;   mengasap v 1 menjadi asap; 2 memasak (mengeringkan, membersihkan, mengha- rumkan, dsb) dng asap;   ~ daging memanggang daging; ~ pakai-   an mengharumkan pakaian dng asap ra- tus dsb; ~ nyamuk mengusir nyamuk dng asap; mengasapi v 1 memberi asap;   2 memasak (mengharumkan dsb) dng asap;
After
abadiah Ar n kekekalan   abc n 1 abjad Latin: tidak tahu --, tidak tahu membaca huruf Latin; 2 ki hal- hal pokok yg paling pertama harus diketahui dr suatu keadaan atau perkara: belum tahu -- kehidupan   acah v, beracah- acah bermain-main; ber- pura-pura;   acah v, mengacah melanggar: murid yg ~ akan dihukum   asap n gas yg tampak keluar dr sesuatu yg panas, mendidih, atau dibakar; belum dipanjat -- kemenyan, pb belum kawin; menggantang --, pb melakukan perbuatan yg sia-sia; berangan-angan yg hampa;   -- air gas yg keluar dr air yg mendidih; uap; -- api asap dr api;   mengasap v 1 menjadi asap; 2 memasak (mengeringkan, membersihkan, mengha- rumkan, dsb) dng asap;   ~ daging memanggang daging; ~ pakaian mengharumkan pakaian dng asap ra- tus dsb ~ nyamuk mengusir nyamuk dng asap; mengasapi v 1 memberi asap;   2 memasak (mengharumkan dsb) dng asap;

2.5.2 Stage Detection Lemma, Dictionary Label and Meaning of Word

Lemma detection stage, dictionary label and meaning of word aims to fill a compound symbol and the number meaning of word, after which it detects entry elements namely lemma, dictionary label and meaning of word.

Lemma detection stage, dictionary label and meaning of words consist of 4 process namely, synonymous word separation process, symbol replacement, word number replacement, and entity detection.

The following are example of data before and after lemma detection stages, dictionary labels and meaning of word.

Table 3 Example Stage Detection Lemma, Dictionary Label And Meaning of Word

Before
abadiah Ar n kekekalan   abc n 1 abjad Latin: tidak tahu --, tidak tahu membaca huruf Latin; 2 ki hal- hal pokok yg paling pertama harus diketahui dr suatu keadaan atau perkara: belum tahu -- kehidupan   acah v, beracah-acah bermain-main; ber- pura-pura;   acah v, mengacah melanggar: murid yg ~ akan dihukum   asap n gas yg tampak keluar dr sesuatu yg panas, mendidih, atau dibakar; belum dipanjat -- kemenyan, pb belum kawin; menggantang --, pb melakukan perbuatan yg sia-sia; berangan-angan yg hampa;   -- air gas yg keluar dr air yg mendidih; uap; -- api asap dr api;   mengasap v 1 menjadi asap; 2 memasak (mengeringkan, membersihkan, mengha- rumkan, dsb) dng asap;   ~ daging memanggang daging; ~ pakaian mengharumkan pakaian dng asap ra- tus dsb ~ nyamuk mengusir nyamuk dng asap; mengasapi v 1 memberi asap;   2 memasak (mengharumkan dsb) dng asap;
After
abadiah Ar n kekekalan   abc n abjad Latin: tidak tahu abc, tidak tahu membaca huruf Latin;   abc n ki hal- hal pokok yg paling pertama harus diketahui dr suatu keadaan atau perkara: belum tahu abc kehidupan   acah vbermain-main; ber- pura-pura;   beracah-acah vbermain-main; ber- pura-pura;   acah v melanggar: murid yg ~ akan dihukum   mengacah v melanggar: murid yg ~ akan dihukum   asap n gas yg tampak keluar dr sesuatu yg panas, mendidih, atau dibakar; belum dipanjat asap kemenyan, pb belum kawin; menggantang asap, pb melakukan perbuatan yg sia-sia; berangan-angan yg hampa;   asap air n gas yg keluar dr air yg mendidih; uap; asap api asap dr api;   mengasap v menjadi asap; asapmemasak (mengeringkan, membersihkan, mengha- rumkan, dsb) dng asap;   mengasap daging v memanggang daging; mengasap pakaian mengharumkan pakaian dng asap ra- tus dsb mengasap nyamuk mengusir nyamuk dng asap; mengasapi v memberi asap;   mengasapi v memasak (mengharumkan dsb) dng asap;

2.5.3 Output format conversion

Data from lemma detection process, dictionary label and meaning of word will be saved to the database which results in the SQL format. It will then be converted in to several data formats such as .txt, .csv, .json for more details described in the following block diagram form

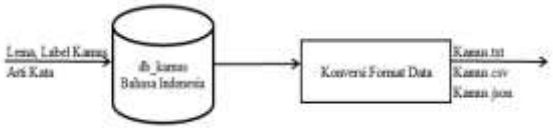


Figure 11 Block Diagram Data Output Conversion Format

2.6 Entity Relationship Diagram (ERD)

System's ERD model can be seen in the figure below.

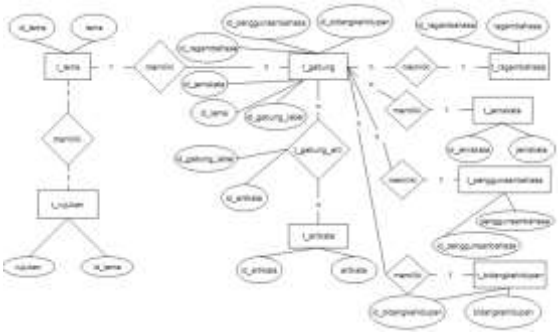


Figure 12 Entity Relation Diagram (ERD)

2.7 Data Flow Diagram (DFD)

System's data flow diagram level 1 model can be seen in the figure below.

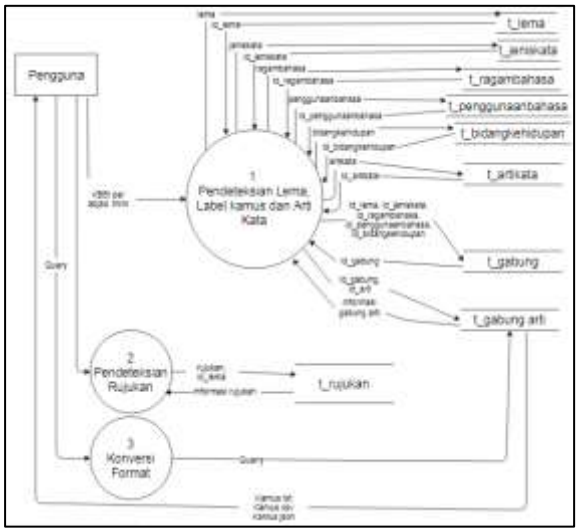


Figure 13 Data Flow Diagram (DFD) Level 1

System's data flow diagram level 2 model can be seen in the figure below.

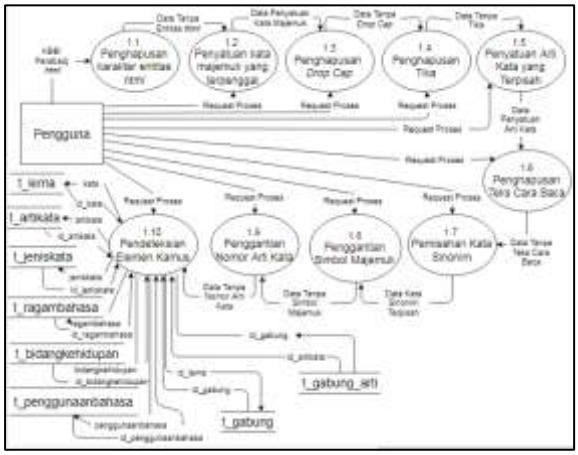


Figure 13 Data Flow Diagram (DFD) Level 2

2.8 System Testing

System testing phase aims to find errors or deferencies in the system being tested. Testing intends to find out weather the program is made in accorclance with the objectives of the study.

2.8.1 Testing of Detection Results

On the accuracy of the test will be calculated the accurate value of the amount of lemma was detected with lemma in the input data. Previously already done the calculation manually against the lemma of the input data Kamus Besar Bahasa Indonesia (KBBI) by way inserting lemma into Microsoft excel then do the deletion of lemma duplication, so as get result of lemma amounts as much as 52.978 lemma. Then the result of calculation accuracy as follows.

- Quantities of manual couting lemma : 52.978
- The number of successful lema detected : 52.978
- Accouracy = (the number of successful lema detected / Quantities of manual couting lemma)
  = (52.978 / 52.978) * 100%
  = 100%

Then in this study yields an accurate value of 100%

The amount of lemma that was detected in this study has increased from previous studies, comparison the results of the detection in the previous study can be seen in the following graph.

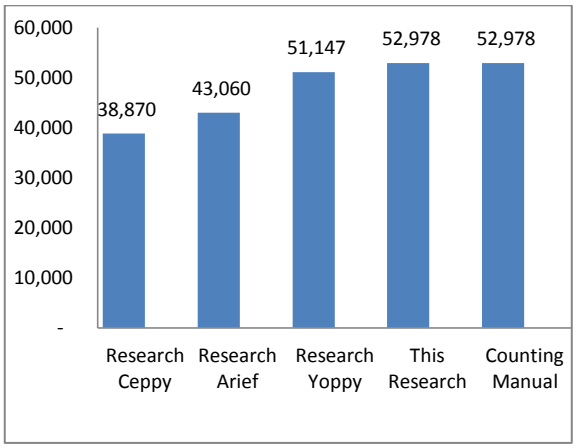


Figure 15 Charts Comparison of Lemma Detection Results

In research conducted by ceppy only detected lemma in the form of basic word and dictionary labels in the form of word type. Then in research conducted by Arief only detected lemma a basic word and dictionary label a type of words and there is duplication of lemma because how to read is also detected as lemma, then in research conducted by Yoppy just detecting lemma in the form of a basic word and compound word and detecting lemma in the form of a word and there are errors consisting of two words, because it is detected in to two different lemma.

This research has successfully overcame the problems that occurred in previous research and has sucecesfully detected the dictionary elements namly dictionary label and meaning of word that were not detected in previous research number of data entry consisting of lemma, dictionary label and meaning of word as much 76.551 entry data this number includes lemma homonyms which have more than one meaning.

3. CLOSING

In this section will explain the conclusion and suggestions of that research.

3.1 Conclusion

In the construction of a Indonesian dictionary as an Natural Language Processing (NLP) Indonesian has got 52.978 lemma results. With an accuracy rate of 100% based on the results of manual count source in the from of Indonesian dictionary (KBBI). This research can detect the lemma contained in the input source because it converts the input source format to html it aims to detect the writing bold. Used to write lemma on Indonesian dictionary. The detection results do not match the amount of lemma released the fourth edition of Indonesian dictionary there are 90.049 lemmas. Because the source of Indonesian dictionary is still in editing stage not a fourth edition of Indonesian dictionary in the development stage (beta version).

3.2 Suggestion

Suggestion for further research are using input data of Indonesian dictionary new edition which contains many lemmas.

BIBLIOGRAPHY

[1] C G Ceppy Efraim Bolly, *Pembangunan Kamus Jenis Kata Sebagai Sumber Daya NLP Bahasa Indonesia*, vol. I, Agustus 2016.
[2] Arief Adiguna Putra, *Pengembangan Pendeteksian Kamus Jenis Kata Sebagai Sumber Daya NLP Bahasa Indonesia*, vol. I, April 2017.
[3] Yoppy Yansyah, *Pengembangan Kamus Jenis Kata yang Dilengkapi Kata Majemuk Sebagai Sumber Daya NLP Bahasa Indonesia*, vol. I, April 2017.
[4] Badan Pengembangan dan Pembinaan Bahasa [Online] Badan Bahasa. [Online]. http://badanbahasa.kemdikbud.go.id/