Download Print Save to PDF Add to List Create bibliography

Lexicography • Volume 9, Issue 2, Pages 175 - 193 • 2022

SANTI-morf dictionaries

Prihantoro ✉
Save all to author list

[a] Universitas Diponegoro, Semarang, Indonesia

Document type
Article

Source type
Journal

ISSN
21974292

DOI
10.1558/lexi.23569

View more

Full text options Export

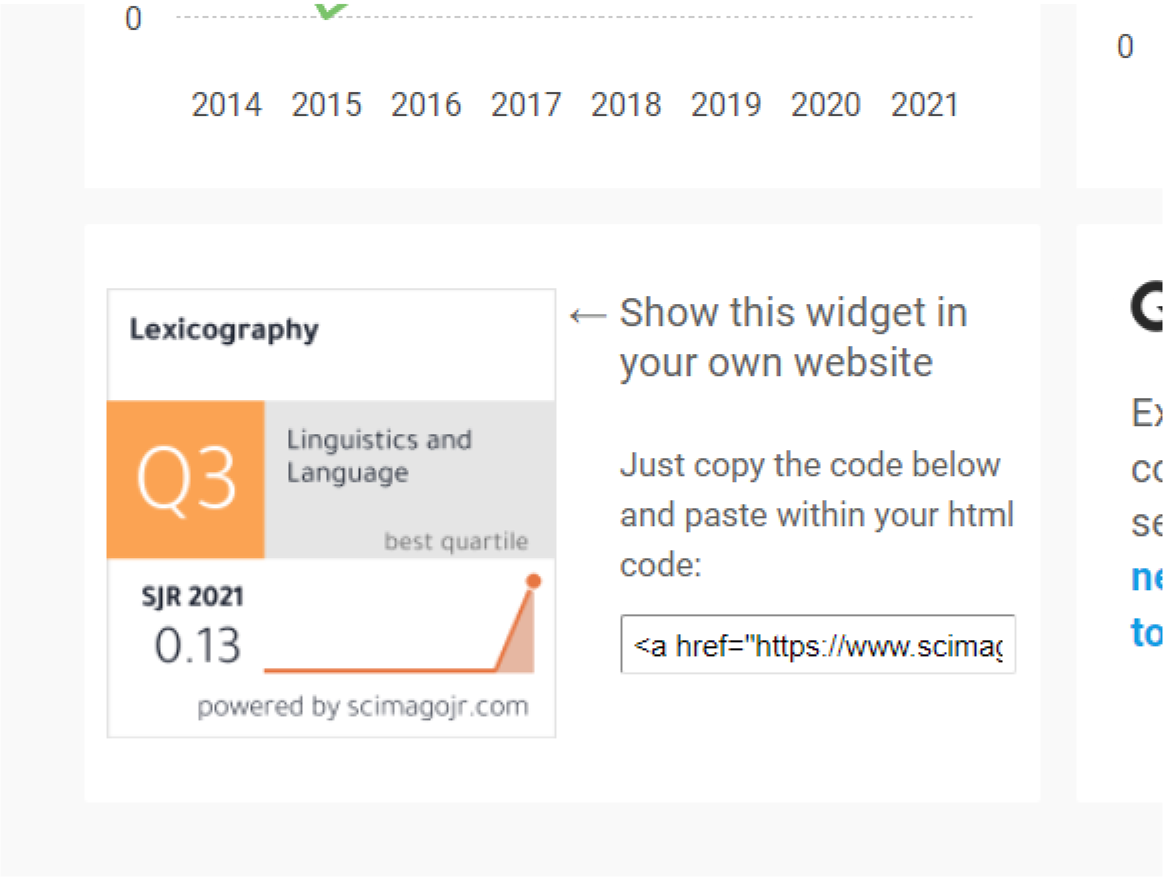
Abstract

Author keywords

SciVal Topics

Abstract

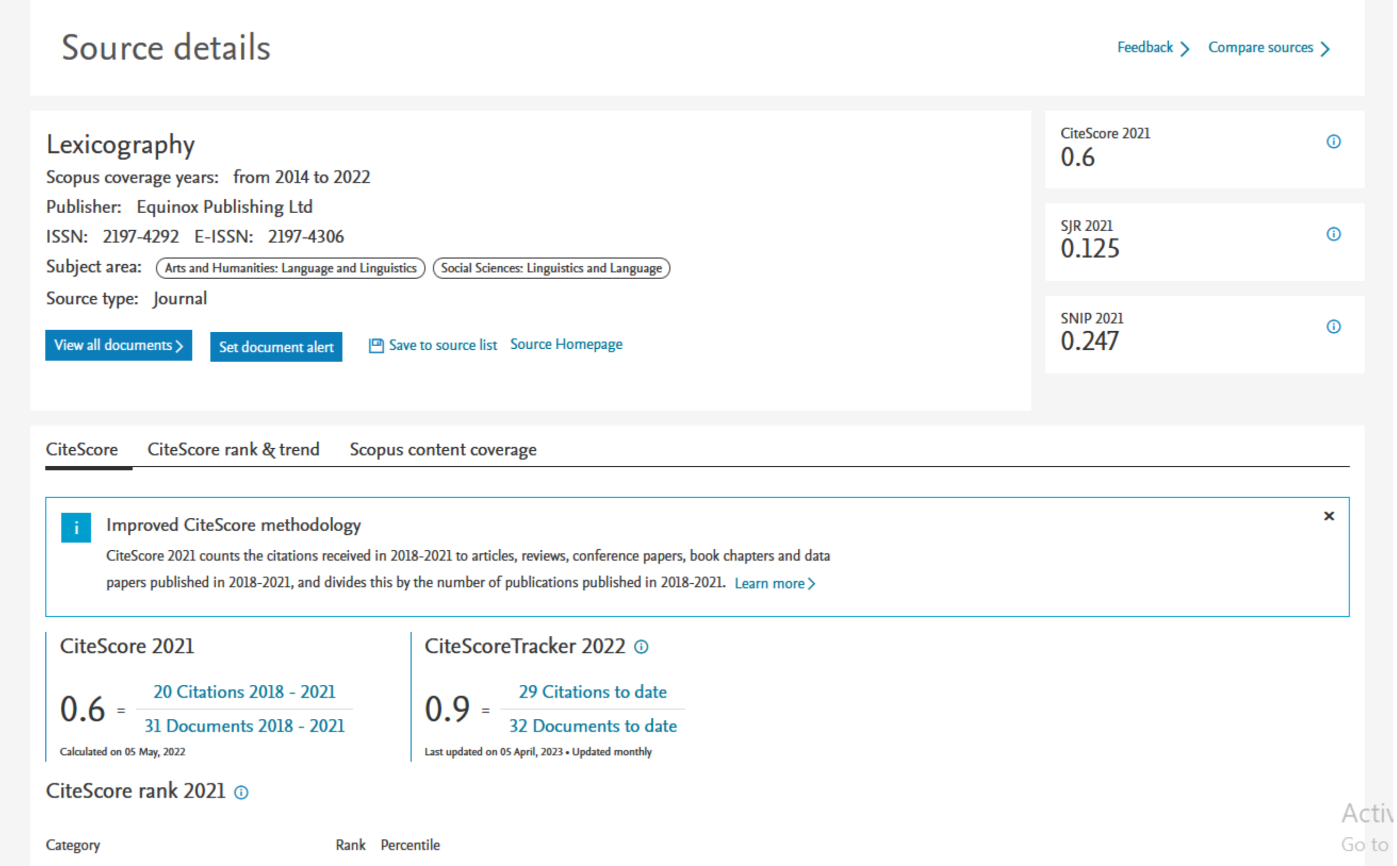
This article highlights the structure of dictionaries used in SANTI-morf (Sistem Analisis Teks Indonesia – morfologi), a multi-module pipeline system that performs annotations for an Indonesian corpus at the morpheme level and built using NooJ (Silberztein, 2003, 2016). SANTI-



← Show this widget in your own website

Just copy the code below and paste within your html code:

```
<a href="https://www.scimag
```



Source details

Feedback > Compare sources >

Lexicography

Scopus coverage years: from 2014 to 2022

Publisher: Equinox Publishing Ltd

ISSN: 2197-4292 E-ISSN: 2197-4306

Subject area: Arts and Humanities: Language and Linguistics | Social Sciences: Linguistics and Language

Source type: Journal

View all documents > | Set document alert | Save to source list | Source Homepage

CiteScore 2021
0.6

SJR 2021
0.125

SNIP 2021
0.247

CiteScore | CiteScore rank & trend | Scopus content coverage

Improved CiteScore methodology

CiteScore 2021 counts the citations received in 2018-2021 to articles, reviews, conference papers, book chapters and data papers published in 2018-2021, and divides this by the number of publications published in 2018-2021. Learn more >

CiteScore 2021

$0.6 = \frac{\text{20 Citations 2018 - 2021}}{\text{31 Documents 2018 - 2021}}$

Calculated on 05 May, 2022

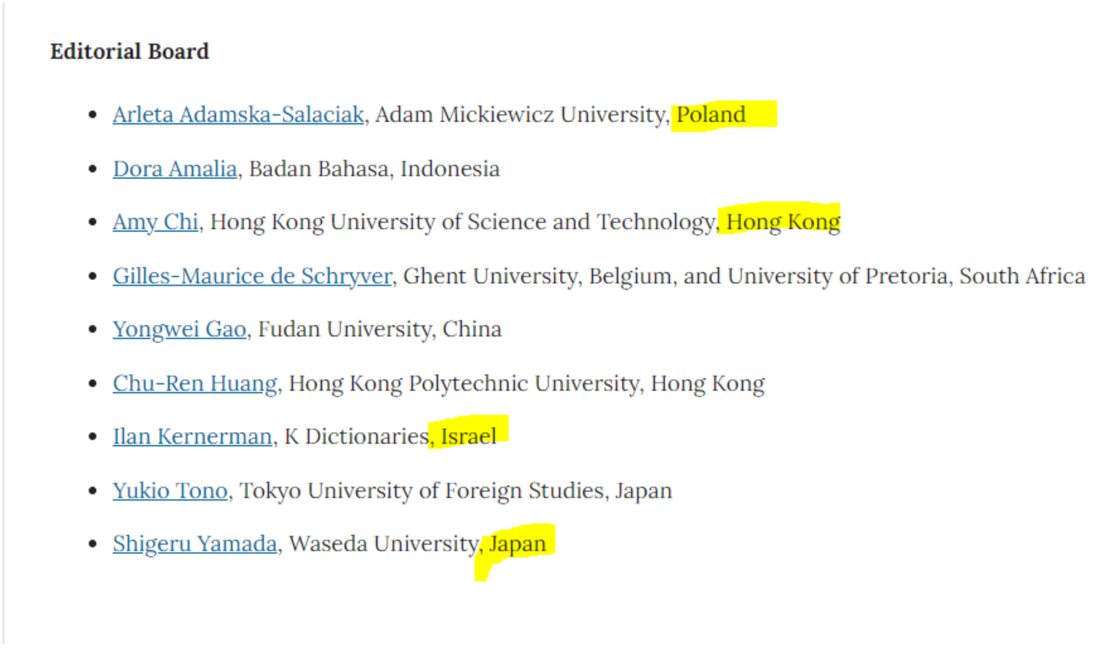
CiteScoreTracker 2022

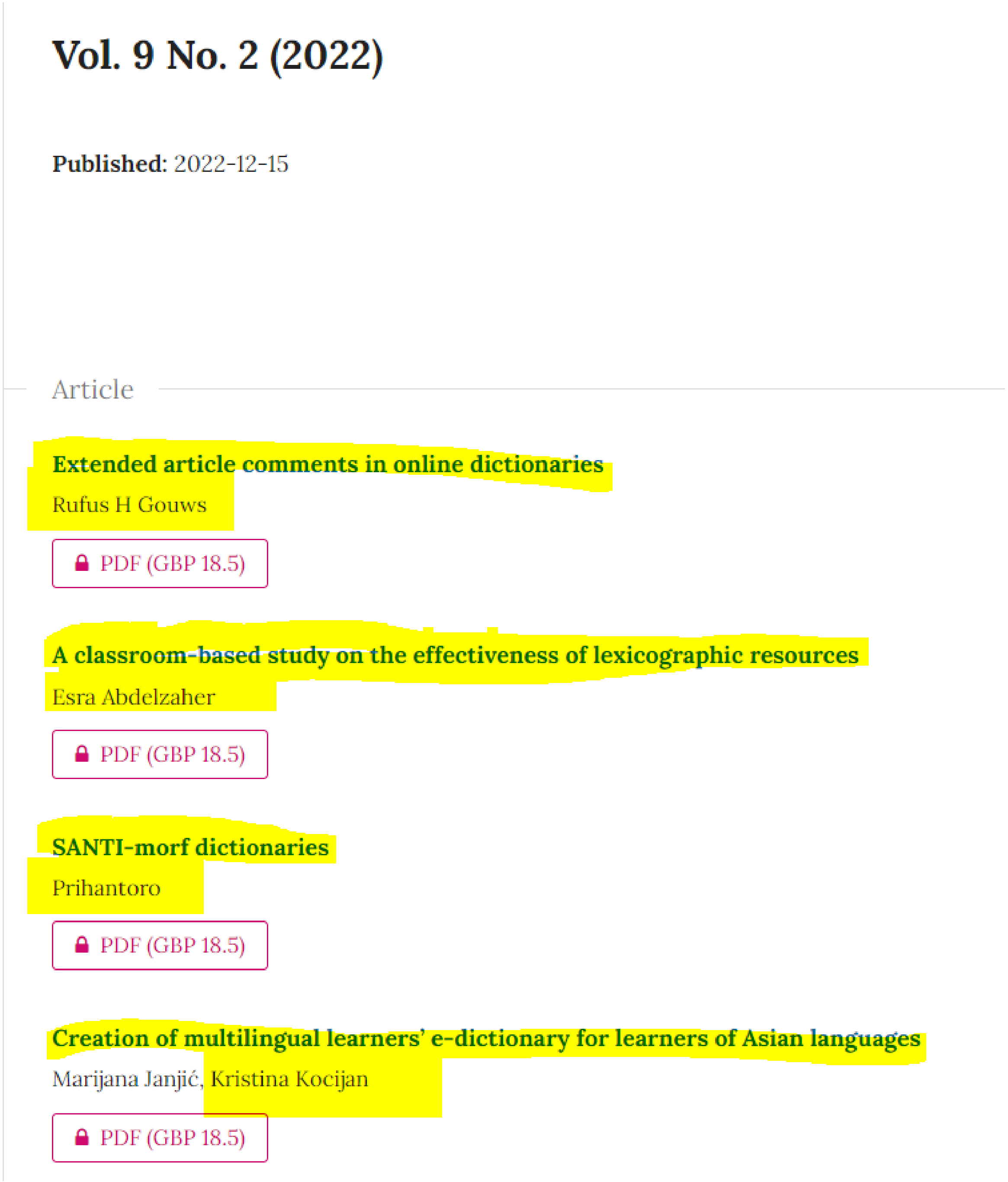
$0.9 = \frac{\text{29 Citations to date}}{\text{32 Documents to date}}$

Last updated on 05 April, 2023 • Updated monthly

CiteScore rank 2021

Category	Rank	Percentile





Vol. 9 No. 2 (2022)

Published: 2022-12-15

Article

Lexicography:
Journal of
ASIALEX

LEXI (ONLINE) ISSN 2197–4306

Article

SANTI-morf dictionaries

Prihantoro

Abstract

*This article highlights the structure of dictionaries used in SANTI-morf (*Sistem Analisis Teks Indonesia – morfologi*), a multi-module pipeline system that performs annotations for an Indonesian corpus at the morpheme level and built using NooJ (Silberztein, 2003, 2016). SANTI-morf dictionaries, together with other SANTI-morf components, enable the system to tokenize each word in an Indonesian corpus into morphemes (e.g., cliticized and non-cliticized roots, affixes, reduplications) and associate these morphemes with their corresponding tags. Each entry in the SANTI-morf dictionary is encoded with a tag composed of morphological analysis (MA) labels. In most cases, these labels are combined with system implementation (SI) labels. Morphological analysis labels consist of formal and functional morphological criteria labels and are typically used for searching the annotated corpus (e.g., root part of speech (POS) labels). System implementation labels are used for system implementation and are mostly of interest to developers rather than end users. They include morphotactic and morphophonemic constraint labels, which are processed when the monomorphemic entries in dictionaries work together with SANTI-morf grammars (rules).*

KEYWORDS: SANTI-MORF; CORPUS; DICTIONARY; INDONESIAN; MORPHOLOGY

Affiliation

Universitas Diponegoro, Semarang, Indonesia
email: prihantoro@live.undip.ac.id

Received: 17 July 2022 Accepted after revision: 2 October 2022

LEXI VOL 9.2 2022 175–193 doi: https://doi.org/10.1558/lexi.23569

1 SANTI-morf

Malay is genetically affiliated with the Austronesian language family. Over time, it has developed into different language varieties throughout Southeast Asia. Some of these varieties are standardized, and they serve as the official languages of a number of countries in this region (Indonesia, Malaysia, Brunei, and Singapore).

Indonesian is one of the standardized Malay varieties used as the national as well as the official language of the Republic of Indonesia. Lewis and co-workers (2009) note that Indonesian is spoken by almost 200 million speakers. According to the most recent 2020 Indonesia national census, the population is now 270.2 million (https://www.bps.go.id/press-release/2021/01/21/1854/hasil-sensus-penduduk-2020.html; last accessed February 9th, 2022). Thus, the number of Indonesian speakers is likely to increase. This makes Indonesian the most widely used standardized Malay variety among other varieties spoken in Southeast Asia.

Let us now discuss the structure of Indonesian, specifically morphology, an area of linguistics relevant to the focus of this study. Indonesian polymorphemic words can be formed using a variety of morphological processes such as affixation, compounding, reduplication, cliticization, or a combination thereof, among many others. Such processes can be analyzed using automatic computational morphology tools, whose resources are specifically designed to handle Indonesian morphology.

Pisceldo and colleagues (2008) created a two-level morphological analyzer for Indonesian. Later, Larasati and associates (2011) built MorphInd, presented as an advancement of Pisceldo and co-workers' tool. I reviewed MorphInd's morphological annotation scheme and suggested some improvements (see more details in Prihantoro, 2021b). In order to implement the scheme, I created SANTI-morf. SANTI-morf is an acronym for *Sistem Analisis Teks Indonesia – morfologi*, or in English, "Indonesian text analysis system – morphology." It is a new morphological analysis system for Indonesian text, whose evaluation is fully explained in Prihantoro (2021a). The system itself is already available for use (http://www.nooj4nlp.org/resources.html).

SANTI-morf is a rule-based morphological annotation system for Indonesian which fully tokenizes and annotates Indonesian words at the morpheme level, not at the word level. SANTI-morf adopts the morphological annotation scheme devised by Prihantoro (2019). Dictionaries and grammars are two core components of SANTI-morf. These resources are grouped into four modules: Annotator, Guesser, Improver, and Disambiguator (see Prihantoro, 2021a). SANTI-morf is implemented using NooJ (http://www.nooj4nlp.org) (Silberztein 2003, 2016), a finite-state rule-based text analyzer program.

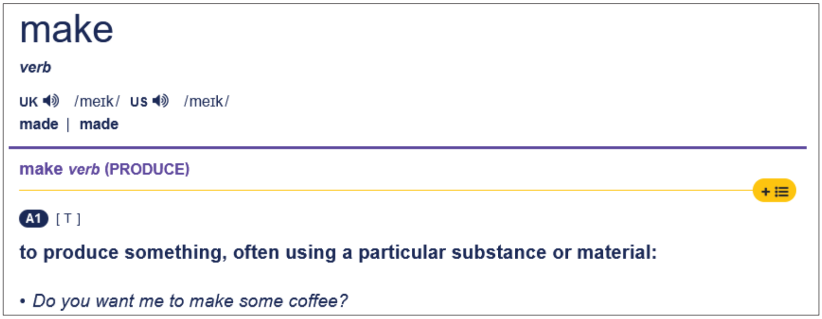
Once a text is annotated using SANTI-morf, a user can search a morpheme (or a combination of morphemes) based on several morphological aspects: the morpheme(s), formal and functional morphological categories, or combinations thereof. SANTI-morf can contribute to applications in different fields such as informatics, corpus linguistics, or lexicography. The application of SANTI-morf to support lexicographic work is demonstrated in Section 4 of this article. There is a wide range of aspects of SANTI-morf to discuss, but in this study I will focus on describing the architecture of SANTI-morf dictionaries.

2 Dictionary

2.1 Dictionary entry

One typically consults a dictionary by observing its entries. Consider Figure 1. It shows how *make* is structured as a dictionary entry in the online version of the *Cambridge Dictionary* (https://dictionary.cambridge.org/dictionary/english/make). The entry consists of a number of lexicographic components such as head(word), phonetic transcription, part of speech label, and definition, among many others.

This is a typical structure of an entry found in a human-readable dictionary (HRD), a dictionary that targets human readers, such as students, researchers, lexicographers, etc. Another type of dictionary targets computer programs; such a dictionary is a machine-readable dictionary (MRD). Such dictionaries are also often called lexicons. They are used for various natural language processing (NLP) applications such as topic modeling, a dialogue system, text summarization, or automatic annotation. For automatic annotation software, such as TreeTagger (Schmid, 1994), Unitex



make

verb

UK 🔊 /meɪk/ US 🔊 /meɪk/

made | made

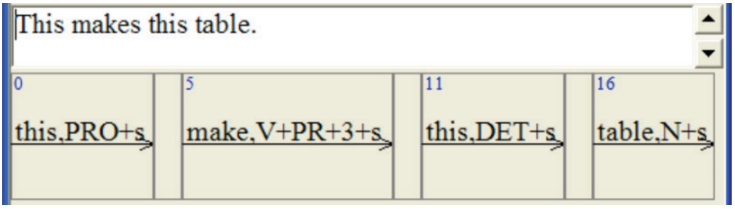
make *verb* (PRODUCE)

A1 [T]

to produce something, often using a particular substance or material:

- *Do you want me to make some coffee?*

Figure 1: *make* in the *Cambridge Dictionary* (online)



This makes this table.

0	5	11	16
this,PRO+s	make,V+PR+3+s	this,DET+s	table,N+s

Figure 2: Annotation of an English sentence using NooJ (Silberztein, 2003, p. 154)

(Paumier, 2014), or NooJ (Silberztein, 2003), dictionaries are essential resources which determine the quality of an annotation system.

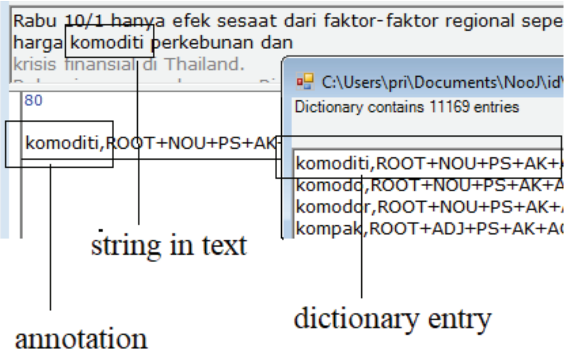
The structure of an entry for MRDs differs from HRDs. Consider the entry line <makes,make,V+PR+3+s> from an MRD of English, used in the NooJ software for part-of-speech tagging/annotation of English. The first part *makes* is the orthographic form of the target word, while the second part *make* is its corresponding lemma. The subsequent parts *V+PR+3+s* are a collection of annotation labels (in this case morphosyntactic), which inform that *makes* in this context is a present-tense third-person singular verb.

The process of using MRDs for linguistic annotation can be described as follows. Typically, the software performs a dictionary look-up for each word (or compound word) in a text. If a word (or compound word) in a text matches a lexical entry in the dictionary, then the software will label the word with the entry's corresponding annotation labels (Figure 2).

The word *makes* in the text is annotated on two levels: lemma and part of speech (POS). The lemma annotation label is *make,* and the POS annotation labels are *V+PR+3+s*. This annotation is obtained from the previous entry line, *makes,make,V+PR+3+s*. Note that in NooJ, when the orthographic and lemma or citation forms are identical, only one form is present in the dictionary entry. For instance, the entry line *make,V+PR+3+p* is used to annotate *make* as a present tense verb when it agrees with a third-person plural subject

2.2 SANTI-morf dictionary

SANTI-morf dictionaries are MRDs used for morphological annotation purposes. When the SANTI-morf system detects a string of characters in a text, it will always first perform a cross-examination with SANTI-morf dictionary entries before checking other types of resources (i.e., grammars). When matches are found in the dictionaries, SANTI-morf will annotate



Rabu 10/1 hanya efek sesaat dari faktor-faktor regional sepe
harga komoditi perkebunan dan
krisis finansial di Thailand.

80

komoditi,ROOT+NOU+PS+AK

C:\Users\pri\Documents\NooJ\id\
Dictionary contains 11169 entries

komoditi,ROOT+NOU+PS+AK+
komodo,ROOT+NOU+PS+AK+A
komodor,ROOT+NOU+PS+AK+
kompak,ROOT+ADJ+PS+AK+A

Figure 3: Annotation based on a match found in one of the SANTI-morf dictionaries

the string based on the labels encoded in the corresponding dictionary entries (Figure 3).

A dictionary file in SANTI-morf can be described as a file containing a collection of entry lines. Each entry line contains a head and its corresponding tag (one or more annotation labels), delineated by a comma. At this point, let us focus solely on the head. In all the examples in this section, I replace the tags with an arbitrary code TAG for conciseness. For example, entry line (1) below includes a head *ikan* "fish," whose actual tag is replaced by TAG.

(1) ikan,TAG

In terms of the number of morphemes, the entries can be divided into two categories: root and full form (polymorphemic). For instance, *getar* "(generic) vibrate" is a root, but *gemetar* "(body part) tremble" is a full form. Note that full form entries are reserved for words that are created using non-productive morphemes such as infix *-em-*. Words produced by productive morphemes such as *-an* in *getaran* "vibration" or *ber-* are analyzed using a combination of dictionaries and grammars, another annotation resource. SANTI-morf dictionaries and grammars are the core components of SANTI-morf, but in this article we will only focus on discussing the architecture of SANTI-morf dictionaries, thus grammars are not discussed here.

(2) getar,TAG (monomorphemic entry line)

(3) gemetar,TAG (polymorphemic entry line)

A head may consist only of letters, non-letter symbols, or a combination thereof. This might seem trivial from a linguistic standpoint, but they are widely present in authentic texts and thus must be dealt with. For example, in a chemistry text, chemical compounds may be written by combining letters and numbers (e.g., h2so4).

(4) h2so4,TAG

In NooJ, some symbols are used for computational purposes, such as for inflection, derivation, or transformation, among many other processes. Thus, when a head consists of one (or more) of these symbols, the symbol(s) must be preceded by a backslash, as shown in example (5).

(5) \:),TAG

The head overall refers to a smiley emoticon :) but must be written as \:) because a colon is a special symbol in NooJ for performing a derivation. Unlike the colon, a question mark can be used independently without having to be escaped from using backslash; this is because there is no computational operation in NooJ that is specified by this symbol (see example (6) below).

(6) ?,TAG

Some non-letter characters may be incorporated into the head on purpose, in order to deal with orthographic variations. For example, the equals sign in the entry *kura=kura* allows SANTI-morf to recognize both *kura-kura* "turtle" and *kura kura* "turtle." This is a useful recognition feature, as in a running text of Indonesian, the hyphen in *kura-kura* is often replaced by a space. By fully incorporating *kura=kura* as the head of an entry line, both forms (with or without a hyphen) can be recognized. Note that *kura-kura* is monomorphemic, even though it looks like reduplication.

(7) kura=kura,TAG

Another source of orthographic variation is the use of optional space characters For instance, *saputangan* "handkerchief" is sometimes written with an extra space as *sapu tangan.* To allow both forms to be analyzed as single morphemes, instead of a combination of *sapu* "broom" and *tangan* "hand," the following entry line must be created.

(8) sapu_tangan,TAG

We can see that the head of this entry line is written as *sapu_tangan.* The underscore that delineates *sapu* "broom" and *tangan* "hand" means that the entry line can recognize two orthographic variations in which the two elements may be written cohesively or with a space. Once incorporated, a dictionary with this entry will allow the system to recognize both *saputangan* and *sapu tangan,* tag them as single morphemes, and assign the corresponding tags. If the system does not find any match in the dictionary, it will analyze the target string as a combination of two morphemes. For instance, *seekor* is analyzed as a combination of two morphemes, namely, *satu* "one (numeral)" and *ekor* "animal classifier" (literally means "tail"), as *seekor* is not present as a dictionary entry.

Another aspect of a head is case sensitivity. If a head is written in full lowercase, it will be used case-insensitively. A full lowercase head such as *bagian* (see example (9) below) matches *bagian, Bagian*, or *BAGIAN*.

(9) bagian,TAG

However, if one or more uppercase letters are present in the head, the matching will be case-sensitive. For instance, the entry line whose head is *Bandung* (see example (10) below) is used to annotate the name of a city in Indonesia, which always begins with an uppercase letter. Therefore, the head also begins with an uppercase letter. For this reason, this entry line will always case-sensitively match *Bandung*, which begins with an uppercase letter in the text, not *bandung* which is written in full lowercase, or *BANDUNG* which is written in full uppercase.

(10) Bandung,TAG

3 Dictionary tag

As previously discussed, an entry line in a SANTI-morf dictionary is composed of a head and a tag. In the previous section, I replaced the tag with an arbitrary label TAG, as we were focusing on discussing the head structure. In subsequent sections, we will discuss the format of the SANTI-morf tag in more detail. A SANTI-morf tag can be defined as a label or a sequence of labels connected using a plus symbol (+). The labels can be further classified into two groups: analytic and system implementation labels. The ordering of labels is technically free, but is presented in a fixed order in this study (analytic first, then implementation) for ease of reading.

3.1 Analytic label

Analytic labels, in this context morphological analysis (MA) labels, reflect the formal and functional analytic categories users are likely to be interested in for searching. These labels are designed based on users' anticipated needs.

For example, the monomorphemic head *pohon* "tree" has two analytic categories. The first label is ROOT, which signifies its formal category as a root morpheme. The second label is NOU, which corresponds to a noun (the root's POS), a functional analytic category. These labels anticipate a user's underspecified query (search all instances of root morphemes) or specified query (search all instances of noun root morphemes). Labels that follow ROOT+NOU are system implementation labels, which will be discussed in the next section. In this section, all system implementation labels are omitted for ease of reading.

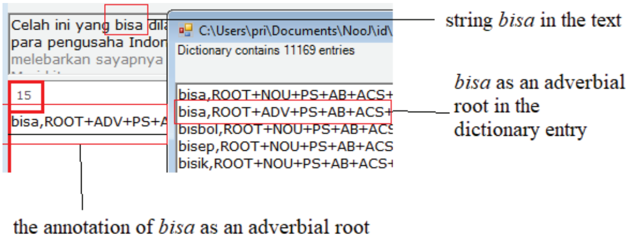
(11) pohon,ROOT+NOU

The functional classification of roots is drawn from the common POS categorization of Indonesian suggested by Alwi and colleagues' (1998) and Sneddon and associates' (2010) reference grammars of Indonesian. For example, *bisa* "can/be able to" is an adverb of modality, and is thus categorized as an adverbial root (Figure 4).

This differs from English, in which its equivalent, *can*, is likely to be classified as a modal verb or just a modal. For instance, in the CLAWS7 tagset (Garside, 1987) the tag for *can* is VM (V = verb, M = modal), in which the modal is under the hierarchy of verbs. Unlike CLAWS, in the Penn Treebank tagset (Marcus et al., 1993), the tag for *can* is MD (modal), which is organized in the same hierarchy of verb tags.

Let us now return to the adverb of modality *bisa* in Indonesian. What analytic category is given to this root in the SANTI-morf dictionary? While *bisa* includes the analysis of modality, only the highest hierarchy (adverb) is documented in the SANTI-morf tagset. Its specification as a modal is not given, thus it is only ROOT+ADV.

(12) bisa,ROOT+ADV



```
Celah ini yang bisa dila
para pengusaha Indon
melebarkan sayapnya

15
bisa,ROOT+ADV+PS+A
```

```
C:\Users\pri\Documents\NooJ\id\
Dictionary contains 11169 entries

bisa,ROOT+NOU+PS+AB+ACS+
bisa,ROOT+ADV+PS+AB+ACS+
bisbol,ROOT+NOU+PS+AB+ACS
bisep,ROOT+NOU+PS+AB+ACS
bisik,ROOT+NOU+PS+AB+ACS+
```

string *bisa* in the text

bisa as an adverbial root in the dictionary entry

the annotation of *bisa* as an adverbial root

Figure 4: Bisa – text, dictionary entry, and annotation

In Indonesian, the monomorphemic word *bisa* is actually ambiguous. It may refer to an adverbial root, as previously suggested, or a noun that means "venom." In an ambiguous case like this, each alternative analysis is also presented as a separate entry line. Thus, in addition to being analyzed as an adverbial root (ROOT+ADV), *bisa* is also analyzed as a noun root (ROOT+NOU). This ambiguity will be resolved later using the Disambiguator module in SANTI-morf.

(13) bisa,ROOT+NOU

SANTI-morf analytic category labels also include a classifier (CLA), a noun categorization morpheme (Sneddon et al., 2010, p. xxi). For instance, *ekor* "tail (literally)" can be used as a classifier for animals. In Indonesian, when used as a classifier, *ekor* is bound to a numeral morpheme, thus is also called a numeral classifier (Aikhenvald, 2001, p. 443). For instance, *ekor* in *dua ekor kucing* "two (animal classifier) cats" is an animal classifier, as its occurrence is preceded by the numeral *dua* "two." The majority of Indonesian classifiers are ambiguous. The morpheme *ekor* can also be used freely as a noun (NOU) when it is not preceded by a numeral, such as *ekor* in *ekor kucing* "cat tail."

(14) ekor,ROOT+CLA

(15) ekor,ROOT+NOU

Some root morphemes in Indonesian cannot occur as monomorphemic words. Thus, their root POS categorization is unclear. The root morpheme *juang* "struggle" can serve to illustrate this. It can only occur within polymorphemic words, such as *ber-juang* "struggle (intr)," or *per-juang-an* "struggle (noun)," among others. If we followed the POS outcome of *perjuangan,* the POS would be a noun root; however, this is problematic, as in another word such as *berjuang* it can be a verb. For this reason, it is essential to establish a unique category for such morphemes. The analytic category label +BOU (bound) is used to specify this kind of root morpheme.

(16) juang,ROOT+BOU

There are 14 POS categories used as analytic category labels in the SANTI-morf tagset (see Table 1). However, only 13 are true POS categories. The remaining category aims to analyze foreign words, that is, non-Indonesian words. Foreign words are analyzed as monomorphemic, even if in the source language they are polymorphemic. For example, "posting" in English is a polymorphemic word. Regardless, SANTI-morf analyzes it as monomorphemic. For example, the word *diposting* (a combination of a passive

Table 1: SANTI-morf root POS

POS	Tag	Examples
Noun	NOU	*nasi* "rice," *jagung* "corn," London "London"
Pronoun	PRO	*aku* "I" (personal), *kenapa* "why" (interrogative), *sini* "this place" (demonstrative)
Numeral	NUM	*satu* "one" (cardinal), *pertama* "first" (ordinal)
Classifier	CLA	*ekor* "animal class," *orang* "human class"
Verb	VER	*pergi* "go," *makan* "eat," lari "run"
Adjective	ADJ	*cantik* "beautiful," *cepat* "quick," *lama* "long"
Adverb	ADV	*selalu* "always," *jarang* "seldom," *hanya* "only"
Preposition	PRE	*di* "at," *ke* "to," *dari* "from"
Conjunction	CON	*dan* "and," *atau* "or," *ketika* "when"
Interjection	INT	*hai* "hi," *aduh* "ouch," *astaga* "oh my god"
Article	ART	*si* "the" (derogatory), *sang* "the" (honorific)
Particle	PAR	*kah*, *lah*, *pun* (all emphasis)
Precategorial	BOU	*juang* "struggle," *nyanyi* "sing"
Foreign	FRG	post, posting (English), *aqua* "water" (Latin), *monggo* "please" (Javanese)

voice prefix *di-* and an English word *posting*) is analyzed as two morpheme tokens in which *posting* is treated as a single root (ROOT+FRG).

Unlike MorphInd (Larasati et al., 2011), there is no "unknown" POS category in the SANTI-morf annotation scheme. When the Annotator module in SANTI-morf fails to perform an analysis, the Guesser (one of the SANTI-morf modules) will offer its best guess rather than just leaving a morpheme unknown. For guessing, no dictionary is needed. The guessing mechanism is fully implemented using grammars, which is not discussed here.

So far, we have discussed different types of entry lines, aiming to describe individual morphemes, mostly root morphemes. However, certain entry lines aim to describe full forms, a combination of morphemes, whose goal is to annotate polymorphemic words generated using unproductive morphological processes or whose meanings are irregular.

In an entry line which targets such words, the analysis of each morpheme is accumulated in a linear order. The analysis of each morpheme is surrounded by angle brackets. For example, *tersangka* "suspect (noun)" is a polymorphemic word composed of two morphemes: the patientive nominalizer prefix *teR-* and the nominal root morpheme *sangka* "suspect (verb)" (see Table 2).

Table 2: Full form entry structure for *tersangka* "a suspect"

Structure	Head,<1st morpheme entry><2nd morpheme entry>
Head (polymorphemic)	tersangka
First morpheme	<ter,teR,PFX+R_NOU+PTNT+DykaA1>
Second morpheme	<sangka,ROOT+VER+DykaA1>
Full entry line	tersangka,<ter,teR,PFX+R_NOU+PTNT+DykaA1> <sangka,ROOT+VER+DykaA1>+UNAMB

The entry line for the first morpheme (prefix) is <ter,teR,PFX+R_NOU+PTNT+DykaA1>. The head in this entry line has two forms, *ter* and *teR*. In SANTI-morf, this is a format given to a morpheme whose orthographic and citation forms differ. The presence of both orthographic and citation forms in the annotation output is required to anticipate a user's need, in order to carry out both specified and underspecified searches.

For instance, the morpheme *teR-* has two allomorphs, *te-* and *ter-*. When a user wants to retrieve word forms specifically containing either *te* or *ter-*, they will need to specify the search with either *te* or *ter*. However, in some cases, a user might want to retrieve all instances of word forms containing both *te-* and *ter-*, thus, an underspecified query <teR> would suffice. The subsequent labels are analytic, which overall suggests a patientive nominalizer prefix. The meaning of each label subsequent to the head is described in Table 3; a full description of all analytic labels is present in the SANTI-morf documentation.

Note that the number of allomorphs for each morpheme may vary. While *teR-* has two allomorphs, *meN-* and *peN-* have six allomorphs each. Now, let us return to the sequence of morphemes that form *tersangka*. Following the patientive nominalizer prefix *ter-* is a verbal root morpheme *sangka* "to suspect," whose entry line is <sangka,ROOT+VER+DykaA1>. Now that all the required entry lines from the two morphemes are identified, they need to be accumulated in a tag slot as a single entry line, tersangka,<ter,teR, PFX+R_NOU+PTNT+DykaA1><sangka,ROOT+VER +DykaA1>+UNAMB.

Table 3: Description of analytic labels that correspond to the prefix *ter-*

Analytic label	Description
PFX	Prefix
R+NOU	Noun outcome
PTNT	Patientive

This entry line allows SANTI-morf to annotate *tersangka* without support from any grammar. While this may apply to a specific polymorphemic word, *tersangka*, it does not apply to all words, even in the same word family. For example, the polymorphemic word *disangka* "to be suspected" and *menyangka* "to suspect" are not solely annotated using the entry lines in dictionaries, even though they share the same verbal root *sangka*.

To sum up, SANTI-morf dictionaries can be used to analyze polymorphemic words. Entries like this are exceptions, and thus relatively fewer in number than root entries. Today, there are 233 entry lines of polymorphemic word entries, and all of them are manually hard-coded. Conversely, polymorphemic words generated using productive and regular morphological processes are tackled using both dictionaries and grammars. They constitute the majority of SANTI-morf dictionaries.

3.2 System implementation label

System implementation labels are labels used for SANTI-morf implementation purposes, and are typically of interest to developers rather than end users. Let us consider an example. At the very end of each dictionary tag, a system implementation label that marks the name of the dictionary source file is present. There are only three possible labels from three SANTI-morf dictionary files, arbitrarily named as follows: DykaA1, DykaA2, and DykaA3.

DykaA1 consists of entries which are neither a proper name nor a foreign word. It consists of both root and full form entries. For instance, *pohon* "tree" is one of the entries in DykaA1, as it is neither a proper name nor a foreign word. DykaA2 consists of proper name entries such as *Aljazair* "Algeria." DykaA3 consists of non-Indonesian entries such as *response* (from English) (see Table 4).

(17) pohon,ROOT+NOU+**DykaA1**

(18) Aljazair,ROOT+NOU+**DykaA2**

(19) response,ROOT+FRG+**DykaA3**

Table 4: SANTI-morf dictionaries and their corresponding entries

Dictionary	Entry	Total	Examples
DykaA1	Root entries Full form entries	10922 233	*makan* "eat," *dan* "and," *tulang* "bone" *tersangka* "a suspect," *gemetar* "tremble," *pepohonan* "trees"
DykaA2	Proper nouns	60151	*Aljazair, Bandung, Australia*
DykaA3	Foreign	14691	brown, moon, finance

The label of the name of the resource file can be used for debugging purposes. For instance, when an error is detected in an annotation outcome, a developer can quickly retrieve the resource file they suspect to be the source of the error. The developer can then locate the specific entry line and implement the required modification(s).

In addition to a resource file name, system implementation labels also include labels used for rule/grammar constraint purposes. Consider this constraint. One of the affixation rules in Indonesian reference grammars (Alwi et al., 1998, p. 117) dictates that the suffix *-i* cannot attach to bases ending in *i*. SANTI-morf takes this rule into account. The verbal root morpheme *cari* "search," for example, ends in *i*, and thus is marked using a +ZI label. Once this label (+ZI) is detected, the *-i* suffixation rule is blocked for the corresponding root entry (*cari, lari, beri*, and all root morphemes ending in *-i*).

(20) cari,ROOT+VER+PS+AC**+ZI**+ACS+T1+DykaA1

Let us consider another example, this time from a syllable number label. The root entry *bom* "bomb" is a monosyllabic root entry, thus, the entry line consists of a +MS label. This is a useful label for selecting the correct allomorphs. For example, the *meN-* and *peN-* morphemes have six allomorphs each. However, when the corresponding base is monomorphemic, the correct allomorphs are *menge-* and *penge-*. The label ensures that the proper affixation rule is applied.

(21) bom,ROOT+VER+**MS**+AB+ZI+ACS+TX+DykaA1

Another label in the system implementation can be used to suggest transitivity. Each verb root entry is specified for its transitivity, namely: intransitive (+T0), transitive (+T1), or ambitransitive (+T2). Non-verbal root entries are given a +TX label. A transitivity label is actually a grey area label between analytic and system implementation labels.

(22) tabrak,ROOT+VER+PS+AT+ZK+ACS+**T1**+DykaA1

I decided to categorize this label as an implementation label because it is used to set constraints. For example, the reciprocal function for the circumfix *ber—an* is added to the annotation when the verb root is transitive (*tabrak* "hit (trans)" > *ber-tabrak-an* "hit one and each other") (Table 5).

A special label, +UNAMB, shown at the end of Section 3.2, finalizes the tag which corresponds to the word form entry *tersangka* "a suspect." This is one way to resolve ambiguities, which can be illustrated as follows.

Table 5: System implementation labels (…, labels omitted due to space constraints)

System implementation labels	

Dictionary name	DykaA1 = main dictionary
	DykaA2 = proper name dictionary
	DykaA3 = foreign word dictionary

Syllable	MS = monosyllable
	PS = polysyllable

Orthography	AA = begins with letter a AB = begins with letter b … ZA = ends with letter a ZB = ends with letter b … AVW = begins with vowel ACS = begins with consonant

Transitivity	TX= non-verb
	T0 = intransitive
	T1 = transitive
	T2 = ambitransitive

SANTI-morf grammars include a rule for *ter-* affixation, a polysemous prefix. In one context, it can be used to form an accidental passive verb, such as in *tertembak* "to get shot accidentally" or *terbawa* "to be brought." In another context, it can be used as a patientive nominalizer prefix, as in *tersangka* "a suspect."

Without a +UNAMB label in the corresponding entry line for *tersangka* in the dictionary, SANTI-morf would generate all possible annotations given by either the dictionary or the rules in the grammar files. This means that there would be multiple annotations on the same token (i.e., ambiguity). However, with the special label +UNAMB given to finalize the corresponding tag for *tersangka* in the dictionary, all the annotations from the rules are blocked. Thus, only the annotation from the lexicon, *ter-* as a patientive nominalizer, is produced. It then overrides the analyses of *ter-* as an accidental passive verbal prefix.

3.3 Residual label

A residual label is used to label non-letter characters. These non-letter characters are grouped into two categories: numerical digits (DGT) and punctuations (PUNC). However, only punctuations are listed as entry lines in the dictionary. In the entry line, every punctuation is identically tagged, using only one label, PUNC, followed by the file name.

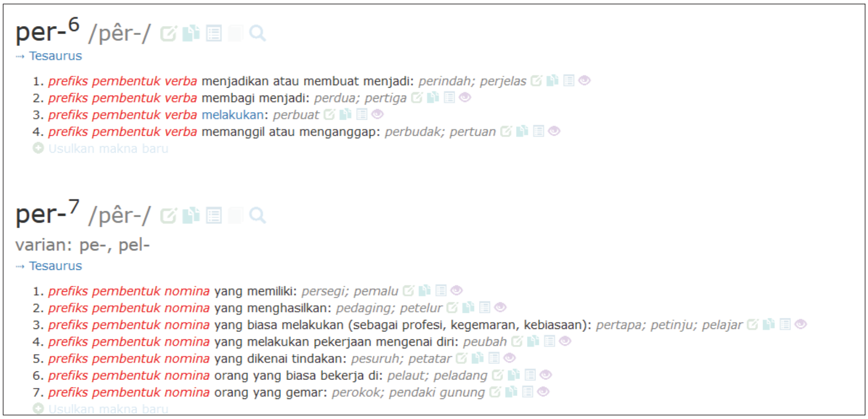
(23) :,PUNC+DykaA1

4 Morpheme list and frequency information

SANTI-morf can be used for a variety of applications. However, let us now focus on using SANTI-morf to support a hypothetical lexicographic project, which focuses on supplying frequency information. Consider the description of a *per-* entry obtained from *Kamus Besar Bahasa Indonesia* (KBBI), or in English, the "Great Dictionary of Indonesian" (available at https://kbbi.kemdikbud.go.id).

A search with a *per-* query returns two entries: *per-* whose outcome POS is a verb (*per-*^{6}) and a noun (*per-*^{7}) (Figure 4). The senses for these two entries vary, but neither have frequency information. In fact, frequency information is a feature that is currently absent from all KBBI entries. Frequency information that corresponds to a morpheme can automatically be derived from a corpus. However, for a bound morpheme such as *per-* the corpus must be annotated at the morpheme level.

The KBBI can potentially benefit from SANTI-morf, as SANTI-morf carries out annotations at the morpheme level. Thus, instead of a word list, the system can produce a morpheme list, which includes frequency information, as shown in Figure 5.



per-^{6} /pêr-/

→ Tesaurus

1. *prefiks pembentuk verba* menjadikan atau membuat menjadi: *perindah; perjelas*
2. *prefiks pembentuk verba* membagi menjadi: *perdua; pertiga*
3. *prefiks pembentuk verba* melakukan: *perbuat*
4. *prefiks pembentuk verba* memanggil atau menganggap: *perbudak; pertuan*

Usulkan makna baru

per-^{7} /pêr-/

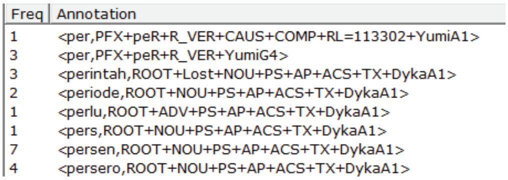
varian: pe-, pel-

→ Tesaurus

1. *prefiks pembentuk nomina* yang memiliki: *persegi; pemalu*
2. *prefiks pembentuk nomina* yang menghasilkan: *pedaging; petelur*
3. *prefiks pembentuk nomina* yang biasa melakukan (sebagai profesi, kegemaran, kebiasaan): *pertapa; petinju; pelajar*
4. *prefiks pembentuk nomina* yang melakukan pekerjaan mengenai diri: *peubah*
5. *prefiks pembentuk nomina* yang dikenai tindakan: *pesuruh; petatar*
6. *prefiks pembentuk nomina* orang yang biasa bekerja di: *pelaut; peladang*
7. *prefiks pembentuk nomina* orang yang gemar: *perokok; pendaki gunung*

Usulkan makna baru

Figure 4: KBBI description for *per-* entry



Freq	Annotation
1	<per,PFX+peR+R_VER+CAUS+COMP+RL=113302+YumiA1>
3	<per,PFX+peR+R_VER+YumiG4>
3	<perintah,ROOT+Lost+NOU+PS+AP+ACS+TX+DykaA1>
2	<periode,ROOT+NOU+PS+AP+ACS+TX+DykaA1>
1	<perlu,ROOT+ADV+PS+AP+ACS+TX+DykaA1>
1	<pers,ROOT+NOU+PS+AP+ACS+TX+DykaA1>
7	<persen,ROOT+NOU+PS+AP+ACS+TX+DykaA1>
4	<persero,ROOT+NOU+PS+AP+ACS+TX+DykaA1>

Figure 5: SANTI-morf morpheme list and frequency information sample

Figure 5 shows two different senses of *per-* in the morpheme list. From the frequency information, the first item has only one instance, while the second item has three instances. While the orthographic forms are identical, the tags for these two types of affixes differ. For this reason, SANTI-morf presents them as two separate items in the morpheme list. Note that both contain the following analytic label, R_VER. This means that the outcome POS for these instances is a verb, which corresponds to the KBBI entry for *per-*^{6}.

The first item, which contains a +CAUS analytic label, corresponds to the first sense of *per-*^{6}, whose frequency is only one. The second item (whose frequency is three) does not contain +CAUS. It corresponds to the remaining senses (2, 3, and 4).

Note that the frequency information in this article is obtained from the BPPT-PAN Localization Corpus (Adriani and Riza, 2008), whose size is relatively small (553,821 words); thus, it may not be fully representative of the Indonesian language. This is also the reason for the low frequency of the two items. With a larger corpus, more representative and reliable frequency information can be obtained. Frequency information can be linked to each KBBI sense, allowing KBBI to produce frequency information automatically. This frequency information can enrich KBBI entries, and KBBI users will find it helpful.

5 Perspectives and recommendations

In this article, I have described the architecture of SANTI-morf dictionaries. These dictionaries work together with other SANTI-morf components, allowing SANTI-morf to automatically annotate Indonesian texts at the morpheme level. I have also demonstrated the application of SANTI-morf, in this case, by supplying frequency information for the *per-* prefix in KBBI. While additional mechanisms are required to port SANTI-morf to KBBI, including creating a corpus from which frequency information can be automatically derived for each entry, this illustration (even though of a hypothetical case) shows how SANTI-morf can potentially be used to support lexicographic work and other areas of study, such as in corpus linguistics, informatics, etc.

While SANTI-morf is specifically designed for Indonesian, the system can also be used to annotate texts from different Malay varieties. Consider the two sentences below, obtained from *Berita Harian* (https://www.beritaharian.sg), whose articles are written in a Malay variety used in Singapore.

(24) *Syarikat dengan amalan kerja tidak selamat tidak dibenar ambil pekerja asing baru: MOM*

"Corporates with low safety record are not allowed to recruit new employees: MOM"

(25) *AKSES lebih mudah bagi mendapatkan ganja di negara-negara jiran akan membawa cabaran bagi memastikan Singapura bebas dadah*

"That marijuana being easier to get in neighboring countries is a challenge to ensure a narcotic-free Singapore"

The underlined words in the two example sentences are not typically used in Indonesian. Let us consider *dibenar* "be justified" in the first example. When written in Indonesian, the *-kan* suffix must be added, hence, *dibenarkan*. While this is uncommon in Indonesian, SANTI-morf, in this case, can still correctly analyze the Malay equivalence *dibenar* as a combination of a passive voice prefix and an adjective root. In the subsequent example, *cabaran* "challenge," SANTI-morf can also correctly analyze this word as a combination of a verb root *cabar* "to challenge" and a nominalizer suffix *–an.* In the case of *dadah* "drugs/narcotic" in the same sentence, it is analyzed correctly as a noun root. In Indonesian, the polymorphemic word *cabaran* "challenge" is likely to be morphologically compositional, but instead of *cabar*, *tantang* "to challenge" is typically used in Indonesian to fill out the slot of the verbal root. As for the monomorphemic word *dadah,* its equivalence *narkotika* "narcotic" is used.

While these analyses are relatively acceptable, a thorough evaluation must ideally be carried out over a larger test set, and ultimately should be confirmed by native speakers of the Malay language variety, in order to determine the accuracy of SANTI-morf's analyses for that language variety. Once confirmed, we can devise a plan and take the necessary measures to adapt SANTI-morf resources (grammars, dictionaries, configuration file) to better analyze texts from various Malay varieties.

SANTI-morf is designed to implement full automatic annotations. However, in NooJ, SANTI-morf users can also carry out manual post processing to manually resolve remaining ambiguities. While the level of ambiguity in the testbed corpus was only approximately 1% (see Prihantoro, 2021b, p. 89), SANTI-morf may retain ambiguities when the Disambiguator failed. For instance, in the case of *mengemas,* SANTI-morf's analyses are ambiguous. The word is analyzed into a combination of an active verb prefix *meng-* and a verbal root *kemas* "to pack" (initial consonant deletion applies). However, *mengemas* is also analyzed as a combination of an active verb prefix *meng-* and a nominal root *emas* "a gold" (no deletion; pure concatenation). The reason for the ambiguities is the paucity in the context information embedded in the SANTI-morf resources (dictionaries and grammars). In this case, a user might want to remove the incorrect analysis manually.

However, when matching context information is found in any of the SANTI-morf resources, the system will make a decision to remove the analyses deemed incorrect. For instance, the word form *beruang* can be analyzed into either a monomorphemic word meaning "bear (animal)," or a polymorphemic word *ber-uang* "to have money" composed of the possessive verbalizer prefix *ber-* and nominal root *uang* "money." When encountering this word form, SANTI-morf always chooses one of the analyses using the context information, thus, manual disambiguation is not required. Technical descriptions on this issue are not discussed here, but are available in Prihantoro (2021b, pp. 174–180).

At present, SANTI-POS (Indonesian POS tagger) and SANTI-sense (Indonesian semantic tagger) are being developed. Once completed, they will be integrated with SANTI-morf to create SANTI-network, using which, users can search corpora using complex queries that combine tagsets from different linguistic levels (morphology, morphosyntax, and semantics).

Acknowledgments

I thank Dr. Andrew Hardie (Lancaster University) and the anonymous reviewers for their valuable suggestions and comments on the manuscript. I would also like to thank the editorial team of *Lexicography: Journal of ASIALEX* for their professional service in processing this manuscript from the stage of submission to publication.

About the author

Prihantoro is an associate professor of corpus linguistics in the department of Linguistics, Universitas Diponegoro, Indonesia. He earned his Ph.D from Lancaster University, and he manages some corpora in CQPweb Lancaster (https://cqpweb.lancs.ac.uk/). He is the author of SANTI-morf (a morphological annotation system for Indonesian) and *Buku Referensi Pengantar Linguistik Korpus* (Introduction to corpus linguistics reference book, written in Indonesian). He can be reached via prihantoro@live.undip.ac.id, or his website http://prihantoro.rf.gd/

References

Adriani, M., and Riza, H. (2008). *Research report phase 2.1: Final design report on statistical machine translation network*. Jakarta: Badan Pengkajian dan Penerapan

Teknologi (BPPT).

Aikhenvald, A. Y. (2001). *A typology of noun categorization device*. Oxford: Oxford University Press.

Alwi, H., Dardjowidjojo, S., Lapoliwa, H., and Moeliono, M. (1998). *Tata Bahasa Baku Bahasa Indonesia* (3rd ed.). Jakarta: Balai Pustaka.

Garside, R. (1987). The CLAWS word-tagging system. In R. Garside, G. Leech, and G. Sampson (Eds.), *The computational analysis of English: A corpus-based approach* (pp. 31–41). London: Longman.

Larasati, S.-D., Kuboň, V., and Zeman, D. (2011). Indonesian morphology tool (MorphInd): Towards an Indonesian corpus. In C. Mahlow and M. Piotrowski (Eds.), *Systems and frameworks for computational morphology* (pp. 119–129). Berlin and Heidelberg: Springer. https://doi.org/10.1007/978-3-642-23138-4_8

Leclère, C. (2005). The lexicon-grammar of French verbs: A syntactic database. In Y. Kawaguchi, S. Zaima, T. Takagaki, K. Shibano, and M. Usami (Eds.), *Linguistic informatics – state of the art and the future* (pp. 29–45). Amsterdam: John Benjamins Publishing. https://doi.org/10.1075/ubli.1.05lec

Lewis, M.-P., Simons, G.-F., and Fennig, C.-D. (2009). *Ethnologue: Languages of the world* (vol. 16). Dallas: SIL International.

Marcus, M.-P., Marcinkiewicz, M.-A., and Santorini, B. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, *19*(2), 313–330. https://doi.org/10.21236/ADA273556

Paumier, S. (2014). *Unitex manual version 3.1*. Paris: Université Paris-Est Marne-la-Vallée and LADL.

Pisceldo, F., Mahendra, R., Manurung, R., and Arka, I.-W. (2008). *A two level morphological analyser for the Indonesian language* (pp. 142–150). Tasmania: Australasian Language Technology Association Workshop.

Prihantoro, P. (2019). *A new tagset for morphological analysis of Indonesian* (pp. 176–181). International Corpus Linguistics Conference, Cardiff.

Prihantoro, P. (2021a). An evaluation of MorphInd's morphological annotation scheme for Indonesian. *Corpora*, *16*(2), 287–299. https://doi.org/10.3366/cor.2021.0221

Prihantoro, P. (2021b). *An automatic morphological analysis system for Indonesian*. PhD thesis. Lancaster: Lancaster University Press.

Schmid, H. (1994). *Probabilistic part-of-speech tagging using decision trees*. Proceedings of the International Conference on New Methods in Language Processing, Manchester.

Silberztein, M. (2003). *NooJ Manual*. www.nooj4nlp.net

Silberztein, M. (2016). *Formalizing natural languages: Nooj approach*. London: Wiley. https://doi.org/10.1002/9781119264125

Sneddon, J.-N., Adelaar, A., Djenar, D.-N., and Ewing, M.-C. (2010). *Indonesian reference grammar* (2nd ed.). New South Wales: Allen & Unwin.