LEMBAR
HASIL PENILAIAN SEJAWAT SEBIDANG ATAU PEER REVIEW
KARYA ILMIAH : JURNAL ILMIAH

Judul artikel : The Morphological Annotation of Reduplication-Circumfix Intersection in Indonesian
Nama penulis : Prihantoro
Jumlah penulis : 1
Status pengusul : Penulis pertama/~~penulis anggota~~/penulis korespondensi
Identitas Prosiding
a. Judul Prosiding : NooJ 2020: Formalizing Natural Languages: Applications to Natural Language Processing and Digital Humanities
b. ISSN : 1865-0937
c. Tahun terbit, tempat pelaks : 2020, Zagreb (Croatia)
d. Penerbit : Springer Nature
f. Alamat repository web : https://link.springer.com/book/10.1007/978-3-030-70629-6
g. Alamat artikel : https://link.springer.com/chapter/10.1007/978-3-030-70629-6_4
h. Terindeks : SCOPUS (Q4) SJR 0.21
Kategori publikasi makalah ☑ Prosiding forum ilmiah internasional (Scimago + Scopus)
□ Prosiding forum ilmiah nasional

Hasil penilaian **peer review 1**

Komponen yang dinilai		Nilai maksimal Prosiding		Nilai akhir yang diperoleh
		Internasional	Nasional	
		[30]	**[ ]**	
a	Kelengkapan unsur isi prosiding (10%)	3.00		10% x 30 = 3
b	Ruang lingkup dan kedalaman pembahasan (30%)	9.00		30% x 30 = 9
c	Kecukupan dan kemutakhiran data / informasi dan metodologi (30%)	9.00		30% x 30= 9
d	Kelengkapan unsur dan kualitas terbitan/jurnal (30%)	9.00		30% x 27 = 8.1
Total 100%		30.00		
Nilai pengusul: 100% x 30 = 30				

Catatan penilaian paper oleh **reviewer 1**

1.Kelengkapan unsur isi prosiding:

Kelengkapan unsur isi artikel prosiding memenuhi kriteria artikel ilmiah

2.Ruang lingkup dan kedalaman pembahasan:

Ruang lingkup dan kedalaman pembahasan sesuai dengan bidang ilmu penulis dan disertai dengan ilustrasi dan figure yang informatif

3.Kecukupan dan kemutakhiran data/informasi dan metodologi:

Memiliki Informasi kebaruan dan kecukupan dan temuan penting melalui metodologi penelitian yang dapat membuktikan bahwa temuan ini bermakna

4.Kelengkapan unsur dan kualitas terbitan:
Kualitas terbitan prosiding internasional bereputasi terindeks SCOPUS (Q4) SJR 0.21

Medan, 4 Mei 2023

Reviewer 1

Tinasinar

Nama	: Prof. T. Silvana Sinar, M.A., Ph.D.
NIP/NIDN	: 195409161980032003
Unit kerja	: Fakultas Ilmu Budaya Universitas Sumatera Utara

LEMBAR
HASIL PENILAIAN SEJAWAT SEBIDANG ATAU PEER REVIEW
KARYA ILMIAH : JURNAL ILMIAH

Judul artikel : The Morphological Annotation of Reduplication-Circumfix Intersection in Indonesian
Nama penulis : Prihantoro
Jumlah penulis : 1
Status pengusul : Penulis pertama/~~penulis anggota~~/penulis korespondensi
Identitas Prosiding
a. Judul Prosiding : NooJ 2020: Formalizing Natural Languages: Applications to Natural Language Processing and Digital Humanities
b. ISSN : 1865-0937
c. Tahun terbit, tempat pelaks : 2020, Zagreb (Croatia)
d. Penerbit : Springer Nature
f. Alamat repository web : https://link.springer.com/book/10.1007/978-3-030-70629-6
g. Alamat artikel : https://link.springer.com/chapter/10.1007/978-3-030-70629-6_4
h. Terindeks : SCOPUS (Q4) SJR 0.21

Kategori publikasi makalah
☑ Prosiding forum ilmiah internasional (Scimago + Scopus)
□ Prosiding forum ilmiah nasional

Hasil penilaian **peer review 2**

Komponen yang dinilai		Nilai maksimal Prosiding		Nilai akhir yang diperoleh
		Internasional	Nasional	
		[30]	**[ ]**	
a	Kelengkapan unsur isi prosiding (10%)	3.00		3.00
b	Ruang lingkup dan kedalaman pembahasan (30%)	9.00		8.50
c	Kecukupan dan kemutakhiran data / informasi dan metodologi (30%)	9.00		9.00
d	Kelengkapan unsur dan kualitas terbitan/jurnal (30%)	9.00		8.00
Total 100%		30.00		28,50
Nilai pengusul: 100% x 30 = 30				

Catatan penilaian paper oleh **reviewer 2**

1. Kelengkapan unsur isi buku

Unsur yang ada dalam nasakah artikel anotasi morphologi cukup lengkap dad menarik, memadai dan lengkap

2.Ruang lingkup dan kedalaman pembahasan

Ruang lingkup terkait anotasi morphology sudah cukup mewakili materi yang penting untuk dikaji

3.Kecukupan dan kemutakhiran data: data dan informasi yang disajikan cukup lengkap dan komprehensif

4.Kelengkapan unsur dan kualitas penerbit:

Badan Penerbit Springer Nature sudah dikenl dalam bidang publikasi pembelajaran linguistik

Bandar Lampung, 4 Mei 2023

Reviewer 2

Nama : Prof Dr. Cucu Sutarsyah, Dip.TESL., MA
NIP/NIDN : 195704061986031002/0006045704
Unit kerja : FKIP Universitas Lampung

LEMBAR PERNYATAAN BEBAS PELANGGARAN KARYA ILMIAH

Yang bertanda tangan di bawah ini

Nama	: Prihantoro
NIP	: 198306292006041002
NIDN	: 3374102906830004
Pangkat (golongan ruang)	: Pembina/IV A
Jabatan Akademik	: Lektor Kepala
Program Studi	: Magister Linguistik
Fakultas/Sekolah	: Fakultas Ilmu Budaya/Universitas Diponegoro

menyatakan bahwa karya ilmiah dengan judul “The Morphological Annotation of Reduplication-Circumfix Intersection in Indonesian” yang dipublikasikan pada (NooJ 2020: Formalizing Natural Languages: Applications to Natural Language Processing and Digital Humanities , vol 1389); di mana saya sebagai (salah satu) penulis, bebas dari atau tidak mengandung pelanggaran kode etik ilmiah.

Demikian surat pernyataan ini kami buat untuk dipergunakan sebagaimana mestinya.

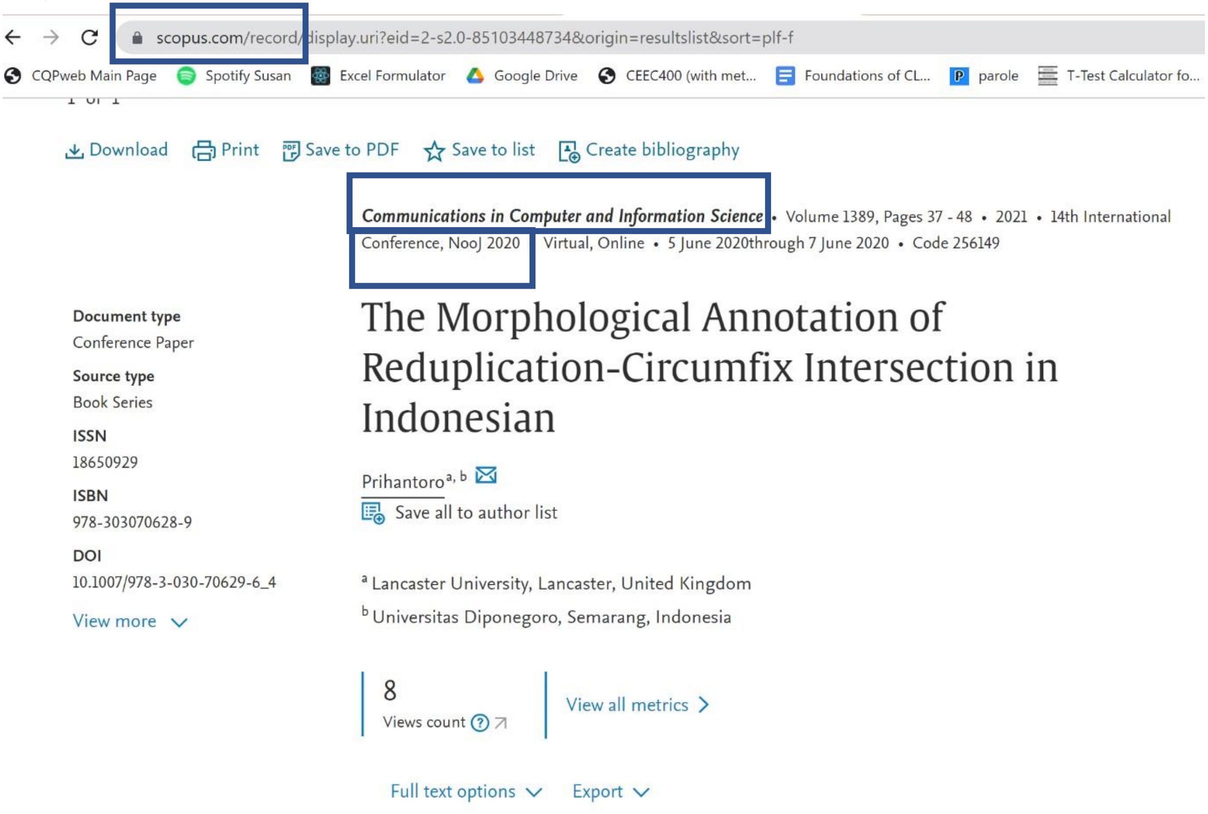
Semarang, 04 Maret 2021

Yang Menyatakan

10000
METERAI
TEMPEL
5A545AJX017204510

Prihantoro
NIP. 198306292006041002

Scopus



Communications in Computer and Information Science • Volume 1389, Pages 37 - 48 • 2021 • 14th International Conference, NooJ 2020 Virtual, Online • 5 June 2020through 7 June 2020 • Code 256149

Document type
Conference Paper

Source type
Book Series

ISSN
18650929

ISBN
978-303070628-9

DOI
10.1007/978-3-030-70629-6_4

View more

The Morphological Annotation of Reduplication-Circumfix Intersection in Indonesian

Prihantoro[a, b]

Save all to author list

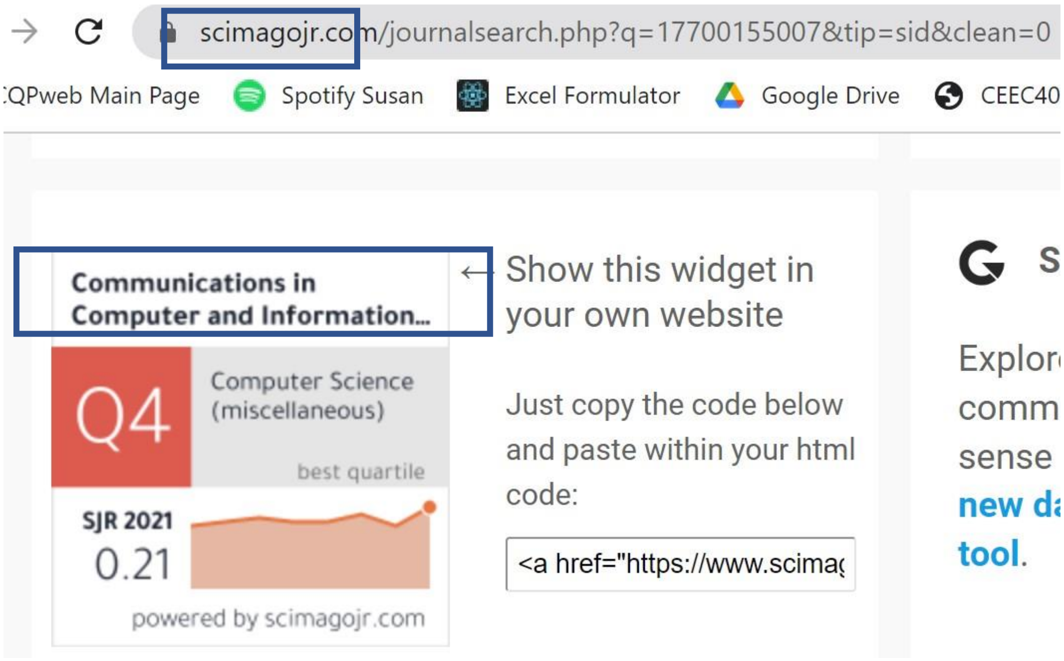
[a] Lancaster University, Lancaster, United Kingdom

[b] Universitas Diponegoro, Semarang, Indonesia

8
Views count

View all metrics

Scimagojr Q4



Communications in Computer and Information...

Q4 Computer Science (miscellaneous)

best quartile

SJR 2021
0.21

powered by scimagojr.com

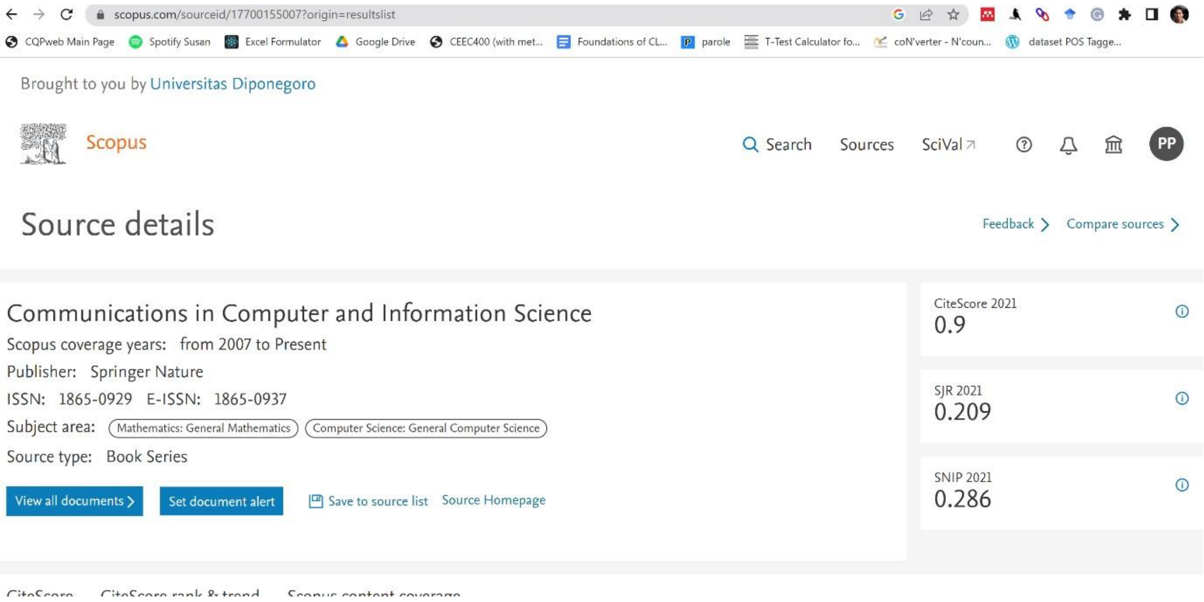
Show this widget in your own website

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

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

Explor comm sense new da tool.

Impact Factor 0.9



Brought to you by Universitas Diponegoro

Source details

Communications in Computer and Information Science

Scopus coverage years: from 2007 to Present

Publisher: Springer Nature

ISSN: 1865-0929 E-ISSN: 1865-0937

Subject area: Mathematics: General Mathematics Computer Science: General Computer Science

Source type: Book Series

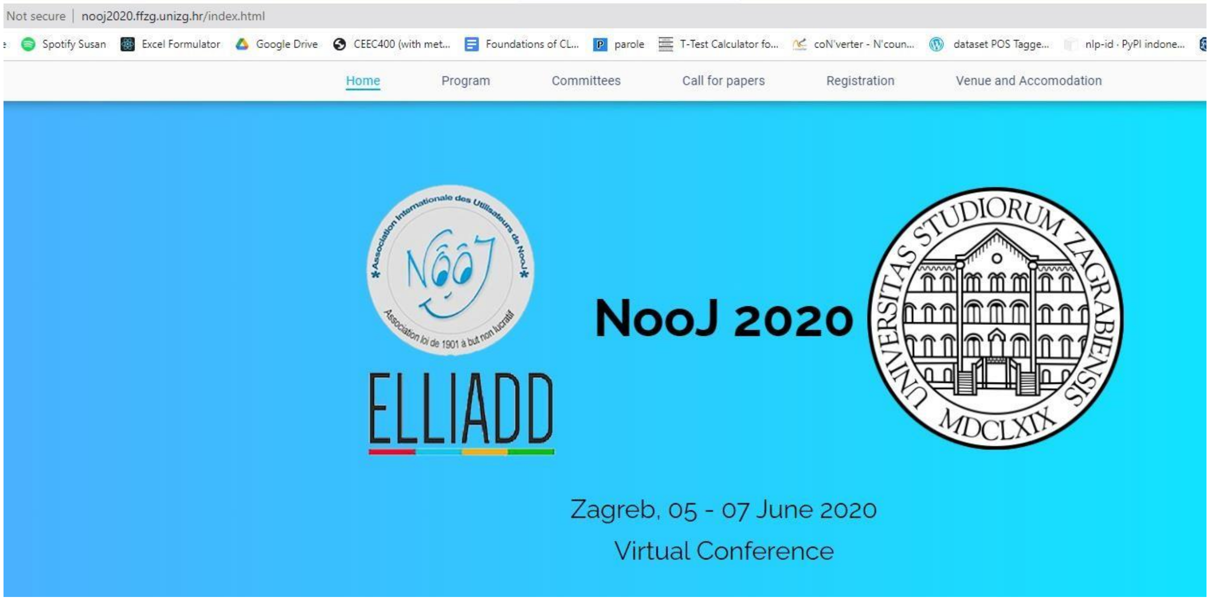
CiteScore 2021
0.9

SJR 2021
0.209

SNIP 2021
0.286

CiteScore CiteScore rank & trend Scopus content coverage

Conference web: http://nooj2020.ffzg.unizg.hr/index.html

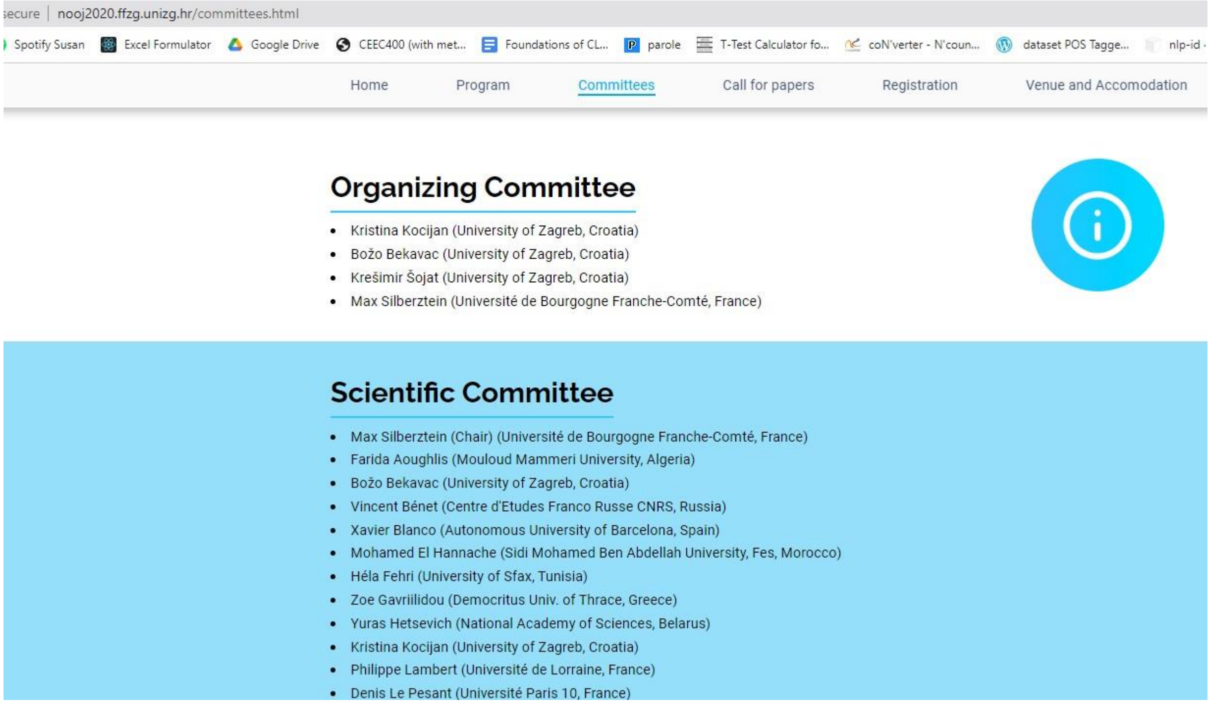


NooJ 2020

Zagreb, 05 - 07 June 2020

Virtual Conference

Committee : http://nooj2020.ffzg.unizg.hr/committees.html



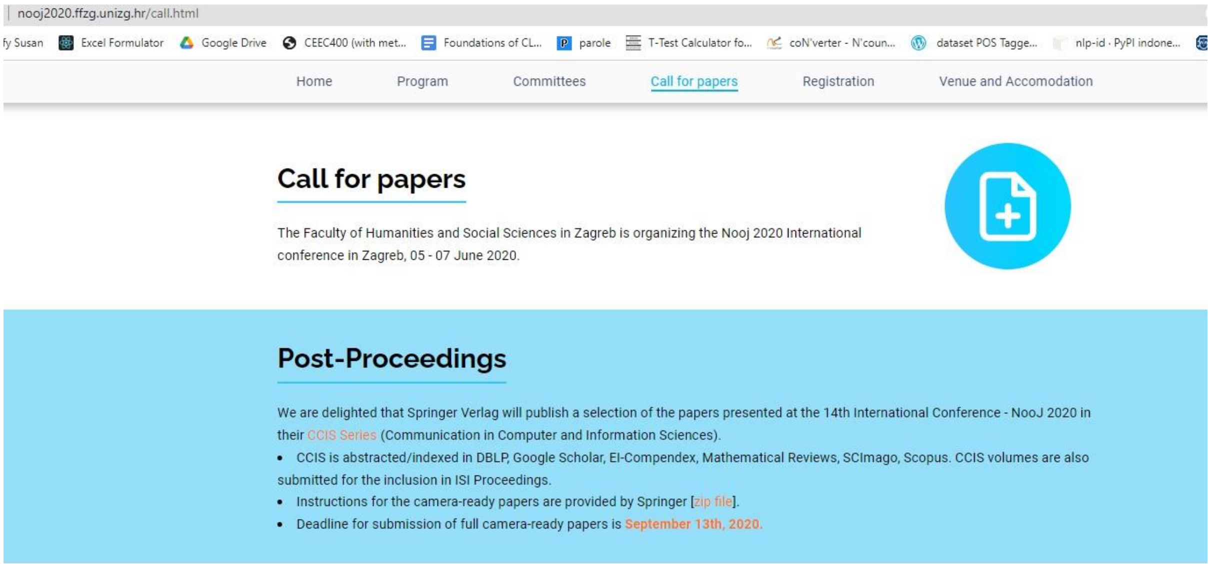
Organizing Committee

- Kristina Kocijan (University of Zagreb, Croatia)
- Božo Bekavac (University of Zagreb, Croatia)
- Krešimir Šojat (University of Zagreb, Croatia)
- Max Silberztein (Université de Bourgogne Franche-Comté, France)

Scientific Committee

- Max Silberztein (Chair) (Université de Bourgogne Franche-Comté, France)
- Farida Aoughlis (Mouloud Mammeri University, Algeria)
- Božo Bekavac (University of Zagreb, Croatia)
- Vincent Bénet (Centre d'Etudes Franco Russe CNRS, Russia)
- Xavier Blanco (Autonomous University of Barcelona, Spain)
- Mohamed El Hannache (Sidi Mohamed Ben Abdellah University, Fes, Morocco)
- Héla Fehri (University of Sfax, Tunisia)
- Zoe Gavriilidou (Democritus Univ. of Thrace, Greece)
- Yuras Hetsevich (National Academy of Sciences, Belarus)
- Kristina Kocijan (University of Zagreb, Croatia)
- Philippe Lambert (Université de Lorraine, France)
- Denis Le Pesant (Université Paris 10, France)

Call for papers: http://nooj2020.ffzg.unizg.hr/call.html



Call for papers

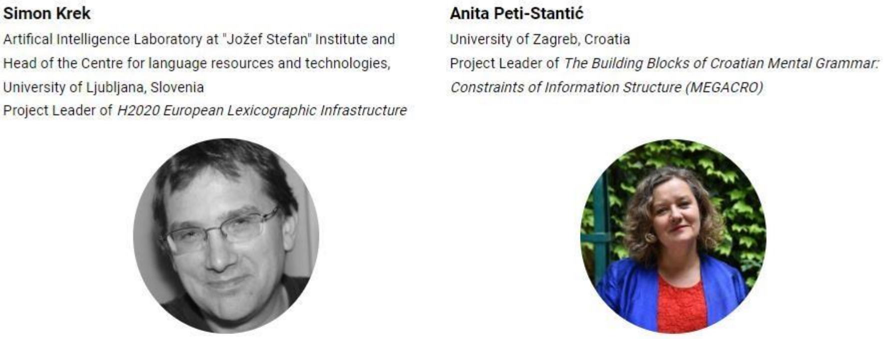
The Faculty of Humanities and Social Sciences in Zagreb is organizing the Nooj 2020 International conference in Zagreb, 05 - 07 June 2020.

Post-Proceedings

We are delighted that Springer Verlag will publish a selection of the papers presented at the 14th International Conference - NooJ 2020 in their CCIS Series (Communication in Computer and Information Sciences).

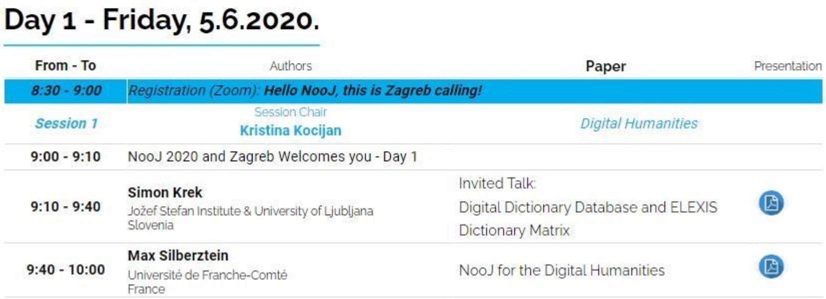
- CCIS is abstracted/indexed in DBLP, Google Scholar, EI-Compendex, Mathematical Reviews, SCImago, Scopus. CCIS volumes are also submitted for the inclusion in ISI Proceedings.
- Instructions for the camera-ready papers are provided by Springer [zip file].
- Deadline for submission of full camera-ready papers is **September 13th, 2020.**

Jadwal: http://nooj2020.ffzg.unizg.hr/program.html



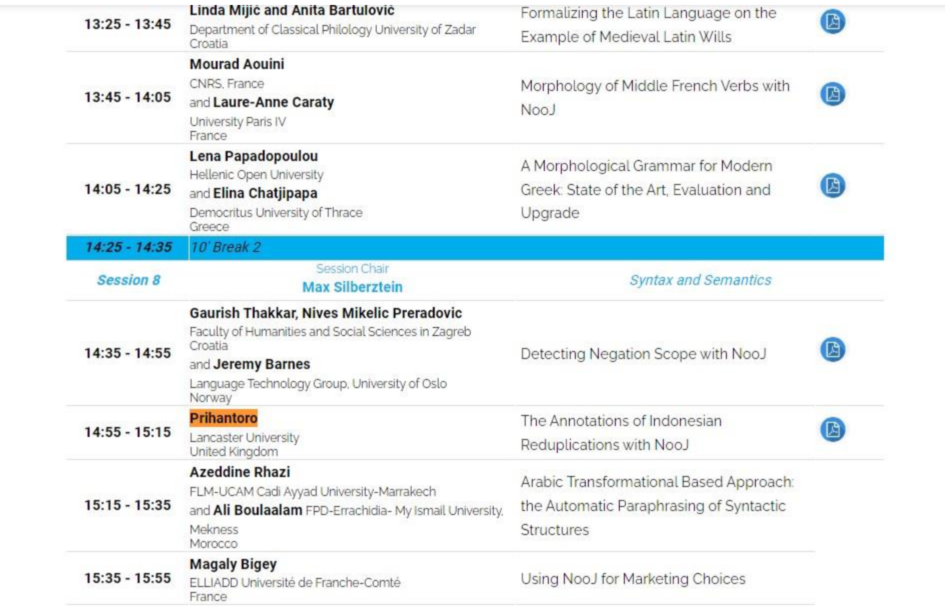
Simon Krek
Artifical Intelligence Laboratory at "Jožef Stefan" Institute and Head of the Centre for language resources and technologies, University of Ljubljana, Slovenia
Project Leader of *H2020 European Lexicographic Infrastructure*

Anita Peti-Stantić
University of Zagreb, Croatia
Project Leader of *The Building Blocks of Croatian Mental Grammar: Constraints of Information Structure (MEGACRO)*



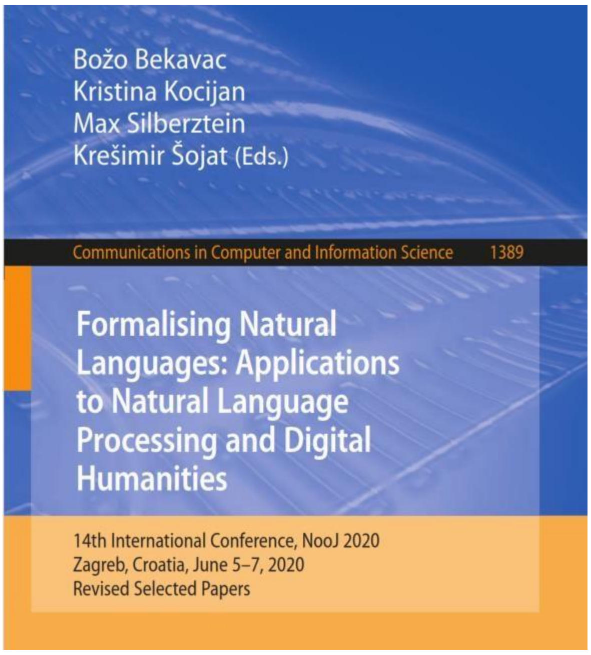
Day 1 - Friday, 5.6.2020.

From - To	Authors	Paper	Presentation
8:30 - 9:00	*Registration (Zoom): **Hello NooJ, this is Zagreb calling!***		
Session 1	Session Chair **Kristina Kocijan**	*Digital Humanities*	
9:00 - 9:10	NooJ 2020 and Zagreb Welcomes you - Day 1		
9:10 - 9:40	**Simon Krek** Jožef Stefan Institute & University of Ljubljana Slovenia	Invited Talk: Digital Dictionary Database and ELEXIS Dictionary Matrix	🗎
9:40 - 10:00	**Max Silberztein** Université de Franche-Comté France	NooJ for the Digital Humanities	🗎



13:25 - 13:45	**Linda Mijić and Anita Bartulović** Department of Classical Philology University of Zadar Croatia	Formalizing the Latin Language on the Example of Medieval Latin Wills
13:45 - 14:05	**Mourad Aouini** CNRS, France and **Laure-Anne Caraty** University Paris IV France	Morphology of Middle French Verbs with NooJ
14:05 - 14:25	**Lena Papadopoulou** Hellenic Open University and **Elina Chatjipapa** Democritus University of Thrace Greece	A Morphological Grammar for Modern Greek: State of the Art, Evaluation and Upgrade
14:25 - 14:35	*10' Break 2*	
Session 8	Session Chair **Max Silberztein**	*Syntax and Semantics*
14:35 - 14:55	**Gaurish Thakkar, Nives Mikelic Preradovic** Faculty of Humanities and Social Sciences in Zagreb Croatia and **Jeremy Barnes** Language Technology Group, University of Oslo Norway	Detecting Negation Scope with NooJ
14:55 - 15:15	**Prihantoro** Lancaster University United Kingdom	The Annotations of Indonesian Reduplications with NooJ
15:15 - 15:35	**Azeddine Rhazi** FLM-UCAM Cadi Ayyad University-Marrakech and **Ali Boulaalam** FPD-Errachidia- My Ismail University, Mekness Morocco	Arabic Transformational Based Approach: the Automatic Paraphrasing of Syntactic Structures
15:35 - 15:55	**Magaly Bigey** ELLIADD Université de Franche-Comté France	Using NooJ for Marketing Choices

Sampul Depan



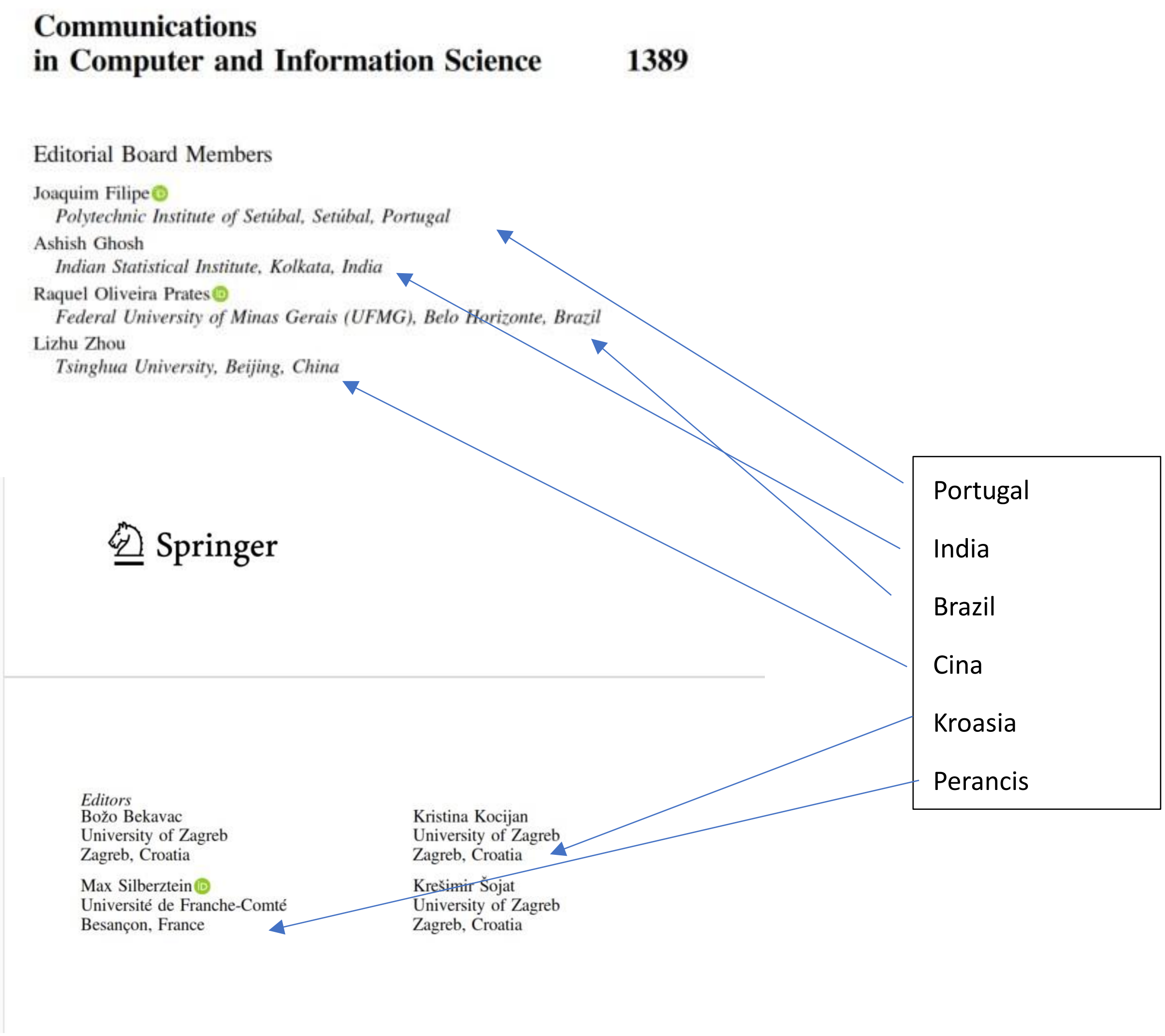
Božo Bekavac
Kristina Kocijan
Max Silberztein
Krešimir Šojat (Eds.)

Communications in Computer and Information Science 1389

Formalising Natural Languages: Applications to Natural Language Processing and Digital Humanities

14th International Conference, NooJ 2020
Zagreb, Croatia, June 5–7, 2020
Revised Selected Papers

Dewan Editor minimal 2 negara berbeda



Communications
in Computer and Information Science 1389

Editorial Board Members

Joaquim Filipe
Polytechnic Institute of Setúbal, Setúbal, Portugal

Ashish Ghosh
Indian Statistical Institute, Kolkata, India

Raquel Oliveira Prates
Federal University of Minas Gerais (UFMG), Belo Horizonte, Brazil

Lizhu Zhou
Tsinghua University, Beijing, China

Springer

Editors

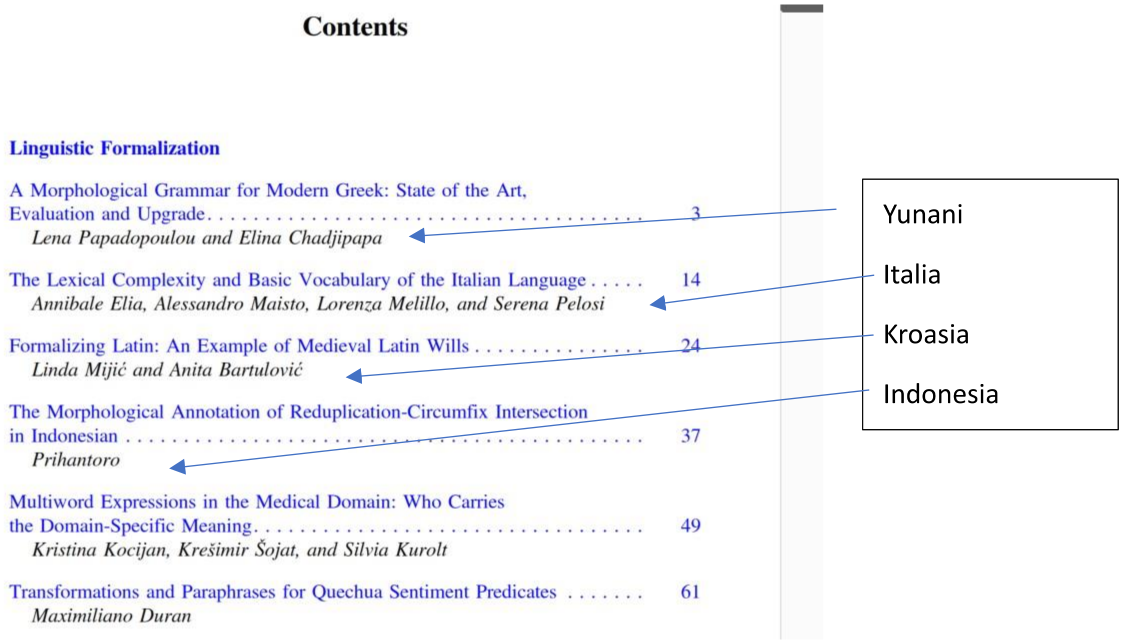
Božo Bekavac
University of Zagreb
Zagreb, Croatia

Kristina Kocijan
University of Zagreb
Zagreb, Croatia

Max Silberztein
Université de Franche-Comté
Besançon, France

Krešimir Šojat
University of Zagreb
Zagreb, Croatia

Penulis dalam terbitan minimal 2 negara berbeda



Contents

Linguistic Formalization

The Morphological Annotation of Reduplication-Circumfix Intersection in Indonesian

Prihantoro[1,2](B)

[1] Lancaster University, Lancaster, UK prihantoro2001@yahoo.com

[2] Universitas Diponegoro, Semarang, Indonesia

Abstract. In this paper, I report on the implementation of a morpheme-level annotation scheme for Indonesian [1], particularly the annotation of reduplications. Utilizing the NooJ program [2] and a set of novel linguistic resources, the majority of reduplications formed according to a number of distinct patterns were successfully annotated. However, one reduplication-circumfix intersection pattern could not be annotated. This is because the current NooJ morphological grammar is not designed to read the hyphen symbol, an orthographical cue that connects a root to its copy in Indonesian reduplication. Failure to read this symbol disconnects the opening and closing elements of the circumfix that surrounds its reduplication base. To overcome this problem, I introduced additional circumfix rules into the current morphological grammar without using any hyphen symbols in the rule definitions. The circumfix elements in the rules are linked as a single dependent unit using syntactic grammar whose rules NooJ allows to contain the hyphen symbol. However, this method promotes undesirable ambiguities. To overcome this side effect, I have modified the existing syntactic grammar to eliminate these ambiguities.

Keywords: NooJ · Annotation · Reduplication · Circumfix · Indonesian

1 Introduction

In this paper, I present some experimental results on the morphological annotation of Indonesian, focusing on the circumfix-reduplication interface. Indonesian is the official as well as the national language of the Republic of Indonesia, spoken by over 200 million people [3]. It is one of the standard varieties of Malay, found throughout Southeast Asia, and genetically affiliated to Austronesian languages [4].

In Indonesian, reduplication is one of the most productive word-formation operations. It is thus clearly important that reduplications should be accurately annotated. Other productive word-formation operations are affixation and compounding [4].

How are reduplications annotated by the currently available Natural Language Processing (NLP) tools built for Indonesian? Wicaksono and Purwarianti [5] constructed a POS tagger for Indonesian (IPOS). This NLP application invariably tokenizes both monomorphemic and polymorphemic words as single-word tokens, as shown in Table 1. Input 1 is monomorphemic. Input 2 is polymorphemic, formed by fully reduplicating the root. Despite differing in shape, both are functionally tagged as NN (noun). The tool supplies no tags for categories of morphological form, such as whether a word is formed by reduplication or not. Therefore, for users who rely solely on this tool's output, it is impossible to run a query to find reduplications.

B. Bekavac et al. (Eds.): NooJ 2020, CCIS 1389, pp. 37–48, 2021.
https://doi.org/10.1007/978-3-030-70629-6_4

Table 1. IPOS tagger input and output samples

	Input	Word formation	Output
1	*buku* 'book'	Root word	buku/NN
2	*buku-buku* 'books'	Reduplication	buku-buku/NN

Let us now turn to MorphInd [6], an automatic morphological analyser for Indonesian.MorphIndsuppliestwotypesoftags:wordtagsandroottags.Larasatietal.[6]asse rt thatMorphIndperformsmorphemicsegmentation.However,wecannotobservethissegmen tation for reduplicated words in MorphInd's output, as shown in Table 2. The tokenizationofthereduplication*buku-buku*'books'isidenticaltothemonomorphemic/nonreduplicated word *buku* 'book'. They are distinguished only by the word tag (NSD = singular noun versus NPD = plural noun).

Table 2. MorphInd input and output samples

	Input	Word formation	Output
1	*buku* 'book'	Root word	buku<n>_NSD
2	*buku-buku* 'books'	Reduplication	buku<n>_NPD

Only functional tags (like NSD and NPD) are present in MorphInd's tag inventory. The letter P (Plural) in the second position of the word tag can, in practice, be used to distinguish monomorphemic buku (NSD) and the corresponding reduplication bukubuku (NPD). In the MorphInd scheme, all reduplications are analysed as plurals. This analysis, however, can be inaccurate. Other than plurality, reduplication in Indonesian can mark a wide range of grammatical or semantic functions, including manner, distributive, and reciprocal. Thus, the singular versus plural analysis is not fully reliable and cannot fully compensate for the absence of formal morphological category tags. It is also apparent that certain reduplications that intersect with affixes are incorrectly analysed by MorphInd as two unrelated units.

In the example illustrated in Table 3, MorphInd incorrectly segments the input into two separate parts (the break being marked by the string DASH). The tag VSA (= singular verb instead of VPA = plural verb) indicates, inaccurately, that each part is an independent word instead of a single morphological formation, i.e. a reduplication.

The above systems are widely used and are state-of-the-art systems (MorphInd for morphologicalanalyserandIPOStaggerforPOStagger)forIndonesian,althoughothers exist.

Table 3. Inaccurate analysis of reduplication-affix intersection from MorphInd

Input	Word formation	Output
pukul-memukul ‘to hit one and each other’	Reduplication	pukul<v>_VSADASH ˆmeN+pukul<v>_VSA

2 Reduplication and Its Intersection with Affixes

The annotation in this experiment was implemented using NooJ v.5 [7] (June 2020 version). The morphological annotation scheme used in the experiment is that devised and presented in full detail in [1], henceforth abbreviated PM. To implement PM, I constructed NooJ dictionaries and grammars for Indonesian from scratch; to date, no Indonesian language resources are available for NooJ.

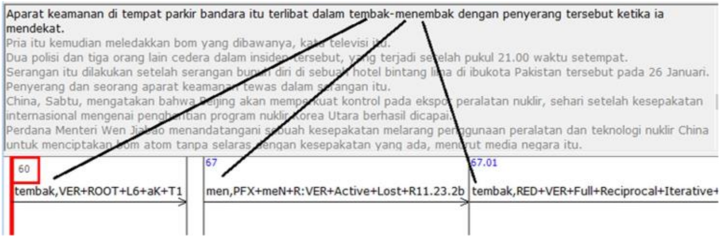
PM dictates that words must be tokenized into morphemes, and each morpheme must be associated with at least one morphological tag. The output format is <token, delimiter (comma), tag>. Thus, the annotation for the verbal root morpheme pukul, ‘to hit’, is as follows: <pukul, VER+ROOT>. The tag is a combination of analytic codes, demarcated by +.

PM follows the view of *morphological reduplication* proposed by Chaer [8]. Central to Chaer’s concept of morphological reduplication is the distinction between a root and its copy. Thus, one of the fundamental principles of the annotation of reduplications in PM is that the annotation of a root and its copy must be clearly distinguished. In the implementation, the first segment of a full reduplication is considered to be the root, and the second its copy.

Each part must be encoded with a distinct analytic tag. The tag for the copy begins with the code RED (reduplication), while the tag for the original root starts with the root’s POS tag (e.g., VER, ADJ, NOM for verbs, adjectives, and nouns, respectively) and is the same as the tag that that root would be assigned in a non-reduplicated context. The NooJ Task Annotation Structure (TAS) in Fig. 1 may serve to illustrate this.

Theannotationof*tembak-menembak* inFig.1isanexampleofaprefix-reduplication intersection, since the copy (but not the root) has the prefix *meN*-. This analysis was obtained in three steps. First, the NooJ dictionaries and morphological grammar were applied. The dictionaries supplied identical POS tags for the root tembak ‘to shoot’ and its copy as verbal roots. The morphological grammar supplied the tag for prefix *meN*-.

This grammar can handle morphophonemic alternation for the allomorphs of *meN-* [9], ofwhich*men*-(asintheaboveTAS)isone.Second,asyntacticgrammarwithanequality constraint [2] was applied to introduce the annotation that indicates a copy (i.e. begins with <RED>) to the copy element. Third, a syntactic grammar with disambiguation rules was applied to the copy in order to remove all annotations except the annotation that begins with <RED>. In concert, these resources can annotate almost all patterns in which reduplications intersect with affixes. However, there was one pattern that these resources failed to annotate. Table 4 enumerates both the analysable and unanalysable patterns.



Aparat keamanan di tempat parkir bandara itu terlibat dalam tembak-menembak dengan penyerang tersebut ketika ia mendekat.
Pria itu kemudian meledakkan bom yang dibawanya, kata televisi itu.
Dua polisi dan tiga orang lain cedera dalam insiden tersebut, yang terjadi setelah pukul 21.00 waktu setempat.
Serangan itu dilakukan setelah serangan bunuh diri di sebuah hotel bintang lima di ibukota Pakistan tersebut pada 26 Januari.
Penyerang dan seorang aparat keamanan tewas dalam serangan itu.
China, Sabtu, mengatakan bahwa Beijing akan memperkuat kontrol pada ekspor peralatan nuklir, sehari setelah kesepakatan internasional mengenai penghentian program nuklir Korea Utara berhasil dicapai.
Perdana Menteri Wen Jiabao menandatangani sebuah kesepakatan melarang penggunaan peralatan dan teknologi nuklir China untuk menciptakan bom atom tanpa selaras dengan kesepakatan yang ada, menurut media negara itu.

tembak,VER+ROOT+L6+aK+T1

men,PFX+meN+R:VER+Active+Lost+R11.23.2b

tembak,RED+VER+Full+Reciprocal+Iterative+

Fig. 1. NooJ TAS for a full reduplication: *tembak-menembak* ‘to shoot one and each other’

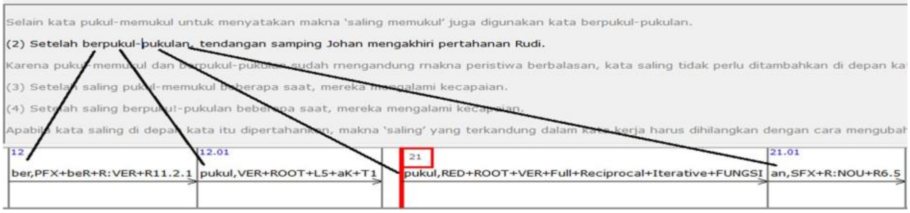
Table 4. An evaluation of Indonesian reduplication patterns tagged by NooJ

	Pattern	Correct tagging
1	<PFX><ROOT> - <PFX><RED>	Yes
2	<ROOT> - <PFX><RED>	Yes
3	<ROOT><SFX> - <RED><SFX>	Yes
4	<ROOT> - <RED><SFX>	Yes
5	<ROOT> - <IFX><RED>	Yes
6	<PFX><ROOT> - <RED><SFX>	Yes
7	<CFX+A><ROOT><CFX+Z> - <CFX+A><RED><CFX+Z>	Yes
8	<CFX+A><ROOT> - <RED><CFX+Z>	No

The NooJ annotation in Fig. 2 exemplifies the eighth pattern. This is unsatisfactory because it analyses *beR-* and *-an* as two independent affixes (prefix + suffix) instead of as a single circumfix.[1] A circumfix is orthographically not distinct from a prefix and suffix combination, as a circumfix is composed of opening and closing elements. But

[1] This is visible from the rule numbers. The rule numbers for *ber-* and *-an* in the TAS are different (R11.2.1 and R6.5, respectively). The numbers should be identical if the circumfix has been correctly analysed.

although there are two elements, we must consider them as one set of affixes instead of as two independent affixes because the circumfix is functionally distinct.



Selain kata pukul-memukul untuk menyatakan makna 'saling memukul' juga digunakan kata berpukul-pukulan.

(2) Setelah berpukul-pukulan, tendangan samping Johan mengakhiri pertahanan Rudi.

Karena pukul-memukul dan berpukul-pukulan sudah mengandung makna peristiwa berbalasan, kata saling tidak perlu ditambahkan di depan ka

(3) Setelah saling pukul-memukul beberapa saat, mereka mengalami kecapaian.

(4) Setelah saling berpukul-pukulan beberapa saat, mereka mengalami kecapaian.

Apabila kata saling di depan kata itu dipertahankan, makna 'saling' yang terkandung dalam kata kerja harus dihilangkan dengan cara mengubah

12	12.01	21	21.01
ber,PFX+beR+R:VER+R11.2.1	pukul,VER+ROOT+L5+aK+T1	pukul,RED+ROOT+VER+Full+Reciprocal+Iterative+FUNGSI	an,SFX+R:NOU+R6.5

Fig. 2. TAS for the annotation of reduplication-circumfix intersection: *berpukul-pukulan*

Silberztein [2] shows how morphological grammar can be used to annotate reduplications in Quechua, in whose orthography the two elements of a reduplication are agglutinated without any space or demarcating symbol. But the same approach is insufficient for the circumfix-reduplication example in Fig. 5. This is because the root and copy in the pattern in question are separated by a hyphen, which is not recognized by NooJ's morphological grammar parser. As a result, the two parts of the circumfix cannot be linked (and therefore, incorrect results are produced). Any NooJ morphological grammar that targets the above pattern would thus necessarily fail to analyse the interface.

3 Possible Solutions and the Current Experiment

3.1 Possible Solutions

I have identified several potential solutions (see Table 5), all of which maintain PM's morpheme-level annotation. Option 1 is to build a kind of "unified grammar," which can function as both a morphological and syntactic grammar.[2] This would be an elegant and simple approach. Reduplications which are handled by using a series of rule applications could instead be handled by just one simplified rule. While NooJ seems to be moving in this direction, at present, the engine is not ready, and thus this solution is not presently feasible (Silberztein, personal communication). Option 2, making the morphological grammar read hyphens, would also require internal modification of the NooJ engine, but might not be as time-consuming as option 1 to implement.

Option 3 is expected to compensate for the morphological grammar's restriction on accepting a hyphen. However, in the current version of NooJ, encoding a hyphen as a dictionary entry cannot override this restriction. This solution is thus not feasible. Option 4, manually coding all full morphological analyses involved into the dictionary, would work, but has the disadvantage of treating productive morphological operations like non-productive morphological operations. This stands in contrast with the purpose of automatic annotation, which is to minimize manual work.

[2] In Nooj v.5, morphological and syntactic grammars are two separate types of grammar (.nom and .nog, respectively).

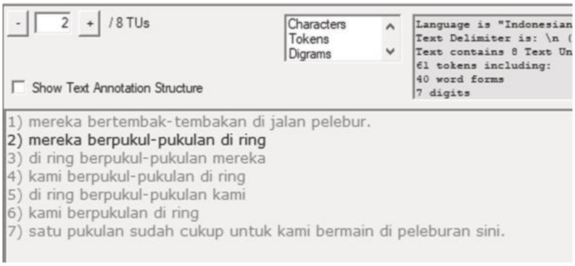
Table 5. Possible solutions to annotate the reduplication-circumfix intersection.

	Option	Automaticity	Complexity	NooJ engine modification
1	Unify morphological and syntactic grammar	High	Low	Required
2	Allow morphological grammar to accept a hyphen	High	Low	Required
3	Encode hyphen as a dictionary entry	High	Low	Required
4	Manually list all full-form reduplications in the dictionary and manually incorporate the corresponding analyses	Low	Low	None
5	Incorporate additional rules that treat a circumfix like a prefix and suffix combination into the current morphological grammar	Low	High	None

In this paper, I explore option 5. Three major advantages of this approach are that (1) it offers a greater degree of automation (no new dictionary entries need be introduced), (2) it is feasible without modifying the current NooJ engine, and (3) the side-effects it causescanbepredictedandanticipated.Theintroductionofrulesthatanalyseacircumfix as a combination of a prefix and a suffix causes ambiguities to greatly increase as a sideeffect. To eliminate this side-effect, the current disambiguation module is revised, as shown later in Sect. 3.2.

3.2 Current Experiment

In the experiment, all resources were simplified to target only the problematic sequence. Thesizeofthetextforthisexperimentislessthan50wordsbutcontainssamplesrelevant to the pattern in question; the reduced dictionary contains only those roots present in the experimental corpus (Fig. 3 and Table 6).



```
Language is "Indonesian
Text Delimiter is: \n (
Text contains 8 Text Un
61 tokens including:
40 word forms
7 digits
```

1) mereka bertembak-tembakan di jalan pelebur.
2) mereka berpukul-pukulan di ring
3) di ring berpukul-pukulan mereka
4) kami berpukul-pukulan di ring
5) di ring berpukul-pukulan kami
6) kami berpukulan di ring
7) satu pukulan sudah cukup untuk kami bermain di peleburan sini.

Fig. 3. Text for the experiment

The morphological grammar rules for this experiment were also adapted to contain only rules relevant to this experiment. In this grammar, the opening element <CFX+A> and closing element <CFX+Z> of the *ber–an* circumfix are introduced in conjunction with the prefix and suffix (on lines 3 and 5, respectively) (Table 7).

Table 6. Sample root entry dictionary for the experiment

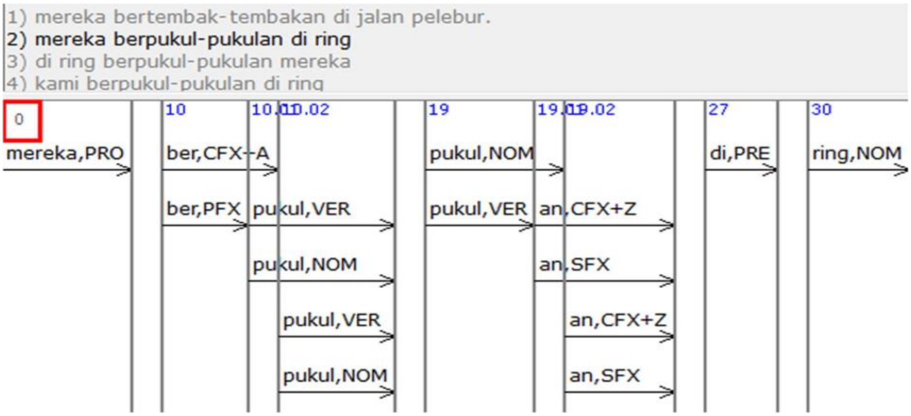
Entry	
pukul,	VER
pukul,	
NOM	
tembak,	
VER main,	
VER	
cukup, ADJ	

Table 7. Morphological grammar rules for the experiment

	Entry
1	Main = :ber | :an | :pe | :ber-an| :pe-an;
2	root = $(X <L>* $) (<E>/<$X=:ALU>) <E>/<$1L, $1C>;
3	ber = (ber/<ber, CFX+A> | ber/<ber, PFX>) :root;
4	…
5	an = :root (an/<an, CFX+Z>| an/<an, SFX>);
6	ber-an = ber/<ber, CFX+A> :root an/<an,CFX+Z>;
7	…

Let us target the most challenging sequence in this corpus, *berpukul-pukulan*, ‘to hit each other repeatedly’, whose root is *pukul*, ‘hit’. This sequence is the most challenging for two reasons. First, the root is ambiguously analysed as both a verb (*pukul* ‘hit’) and a noun (*pukul* ‘o’clock’), where the correct interpretation in this context is a verb. Therefore, the noun analytic code <NOM> has to be removed. Second, we need to remove the incorrect annotations (<PFX> and <SFX>) and replace them with correct analyses (<CFX+A> and <CFX+Z>). Overall, this experiment aims to convert the ambiguous annotations in Fig. 4 into the unambiguous annotations in Fig. 5.

The first step is to insert RED into the copy of the root using the syntactic grammar containingtherulesshowninTable8.Atthispoint,thereisnoattemptatdisambiguation: the analyses of both root and copy remain ambiguous, as shown in Fig. 6.



1) mereka bertembak-tembakan di jalan pelebur.
2) mereka berpukul-pukulan di ring
3) di ring berpukul-pukulan mereka
4) kami berpukul-pukulan di ring

Fig. 4. Ambiguous annotations for *berpukul-pukulan* caused by the morphological grammar rules

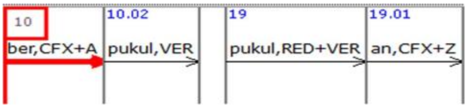


Fig. 5. Expected result for *berpukul-pukulan* after the application of syntactic grammar disambiguation rules

Table 8. The syntactic grammar rules to insert <RED> annotation

Rule
<CFX> $(A <NOM> $) - <E>/<$A_, **RED+NOM** $(B <VER> $) <E>/<$A_=$B_><E>/> <CFX>;
<CFX> $(A <VER> $) - <E>/<$A_,**RED+VER** $(B <VER> $) <E>/<$A_=$B_><E>/> <CFX>

The second step involves a syntactic grammar with the following tasks: (1) remove incorrect analyses of the root <NOM>, (2) remove the incorrect copy <RED+NOM>, (3) remove the incorrect analyses of the affixes (<PFX> and <SFX>). These three tasks are completed successfully via the rule shown in Table 9.

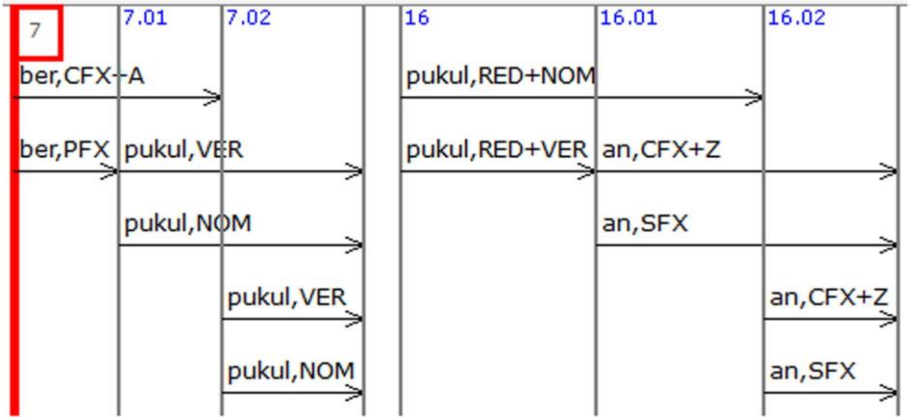


Fig. 6. The result of applying the syntactic grammar rules in Table 8.

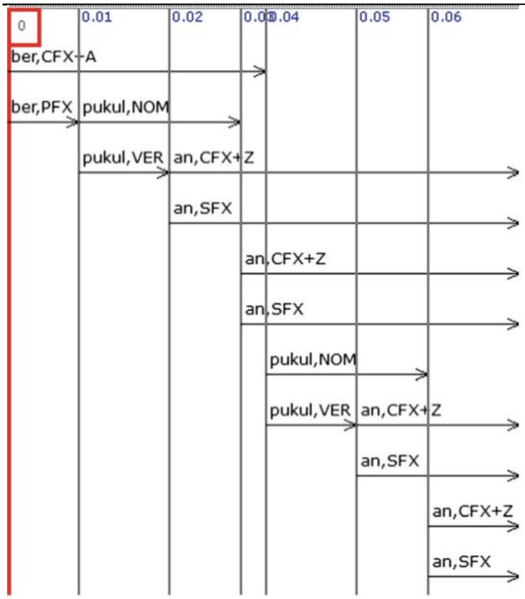
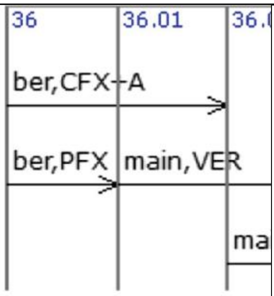
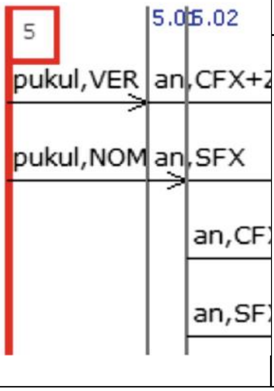
Table 9. Syntactic grammar rules to remove ambiguities in circumfix annotations

Rule
Main = <CFX>/<CFX> $(A <VER>/<VER> $) - $(B <RED+VER>/<RED+VER> $) <E>/<$A_=$B_> <CFX>/<CFX> ;

Introducing new rules to target the problem has proven successful. The expected resultinFig.5wasobtained.However,asmentionedpreviouslyinSect.3.1,thisapproach has a side-effect. Some morphemes that used to be unambiguously analysed have now become unexpectedly ambiguous. Table 10 shows examples of ambiguously annotated circumfix (reciprocal circumfix *ber–an* in *berpukulan*), prefix (intransitive verb marker *ber-* in *bermain*), and suffix (nominalizer suffix *-an* in *pukulan*), respectively. The roots in examples 1 and 3 are also ambiguous, adding yet more ambiguities to be resolved.

Table 10. Ambiguous annotations for *berpukulan, bermain,* and *pukulan* due to the introduction of the morphological grammar rules in

berpukulan ‘to hit each other’	*bermain* ‘to play’

 0 0.01 0.02 0.03 0.04 0.05 0.06 ber,CFX+A ber,PFX pukul,NOM pukul,VER an,CFX+Z an,SFX an,CFX+Z an,SFX pukul,NOM pukul,VER an,CFX+Z an,SFX an,CFX+Z an,SFX	 36 36.01 36. ber,CFX+A ber,PFX main,VER ma	
	 5 5.01 5.02 pukul,VER an,CFX+Z pukul,NOM an,SFX an,CF an,SF	*pukulan* ‘hit’

To disambiguate these annotations, the disambiguation rules in Table 11 must be applied in order. The rule in point 1 targets the annotations of the circumfix *ber–an,* while the rules in point 2 target the annotations of the prefix *ber*- and the suffix -*an*.

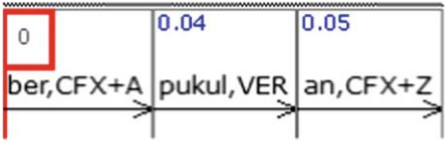
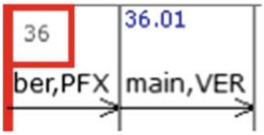
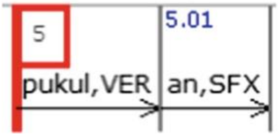
Table 11. Rules to disambiguate noisy annotations for *berpukulan, bermain* and *pukulan*

	Rules
1	Main = <CFX>/<CFX><VER>/<VER><CFX>/<CFX>;
2	Main = <PFX>/<PFX><VER><!-> | <!CFX+A><VER>/<VER><CFX+Z>/<SFX>;

The resulting disambiguated sequences are shown in Table 12.

Table 12. Unambiguous annotation of *berpukulan, bermain* and *pukulan*

Word	TAS

berpukulan	 0 0.04 0.05 ber,CFX+A pukul,VER an,CFX+Z
bermain	 36 36.01 ber,PFX main,VER
pukulan	 5 5.01 pukul,VER an,SFX

4 Conclusion

Despite the experimental nature of the work presented here, the application of the modified morphological analysis resources and disambiguation module has proven successful in overcoming the reduplication-circumfix intersection problem I observed in Sect. 2. The approach has also successfully eliminated the side-effect identified in Sect. 3.1. These results suggest that this is a promising approach worth extending to a full-scale trial.

Further research could explore the effectiveness of this approach when implemented toimprovetheresourcespresentlyusedtoperformmorphologicalannotationsonIndonesiantexts.Aprecisemeasurementoftheeffectivenessofthisapproachwouldbeobtained byevaluatingtheannotationsofafullcorpustowhichtheimprovedresourcesareapplied.

Acknowledgments. This paper has been written during my PhD candidacy at Lancaster University, Lancaster, United Kingdom. I would like to extend my deepest gratitude to the Indonesia Endowment Fund for Education[3] (*Lembaga Pengelola Dana Pendidikan*, or LPDP) for its full sponsorship of my PhD study at Lancaster University. I am also highly indebted to Andrew Hardie, my PhD supervisor, for his useful feedback on this manuscript. All errors are mine.

References

1. Prihantoro: A new tagset for morphological analysis of Indonesian. In: International corpuslinguistics conference, Cardiff (2019)
2. Silberztein, M.: Formalizing Natural Languages NooJ Approach. Wiley, London (2016)
3. Badan Pusat Statistik, I.: Buku 7: Pedomen Kode Provinsi Kabupaten Kota Negara Suku BangsaKewarganegaraan Bahasa dan Lapangan Usaha Sensus 2010. BPS, Jakarta (2010)
4. Tadmor, U.: Malay-Indonesian. In: Major World Languages, pp. 791–818. Routledge, NewYork (2004)

[3] https://www.lpdp.kemenkeu.go.id/.

5. Wicaksono, A.-F., Purwarianti, A.: HMM based part-of-speech tagger for Bahasa Indonesia.In: Proceeding of the Fourth International MALINDO Workshop (MALINDO2010), Jakarta (2010)
6. Larasati, S.-D., Kubonˇ, V., Zeman, D.: Indonesian morphology tool (MorphInd): towardsan indonesian corpus. In: Systems and Frameworks for Computational Morphology, Zurich, Switzerland (2011)
7. Silberztein, M.: NooJ manual (2003). www.nooj4nlp.net
8. Chaer, A.: Morfologi Bahasa Indonesia. Rhinneka Cipta, Bandung (2008)
9. Sneddon, J.-N., Adelaar, A., Djenar, D.-N., Ewing, M.-C.: Indonesian Reference Grammar,2nd edn. Allen & Unwin, New South Wales (2010)