Towards genuine stemming and lemmatization in Malay/Indonesian

野元　裕樹（東京外国語大学）

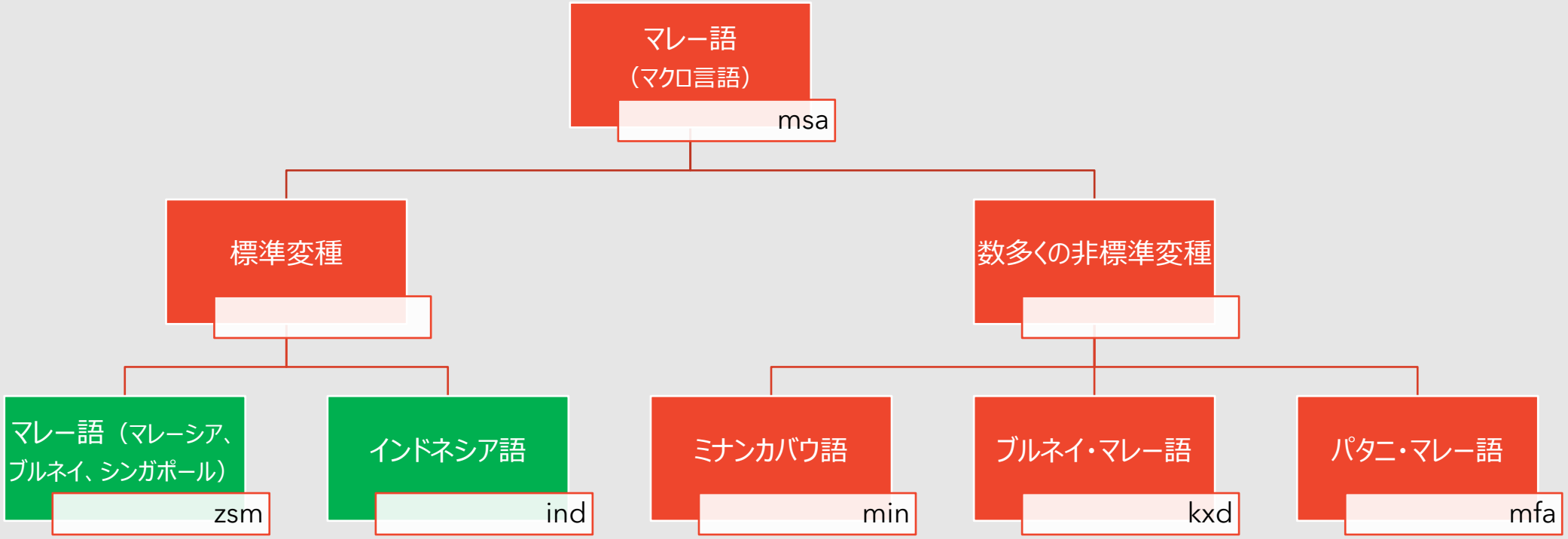
言語処理学会第26回年次大会、2020年3月18日

東京外国語大学
Tokyo University of Foreign Studies

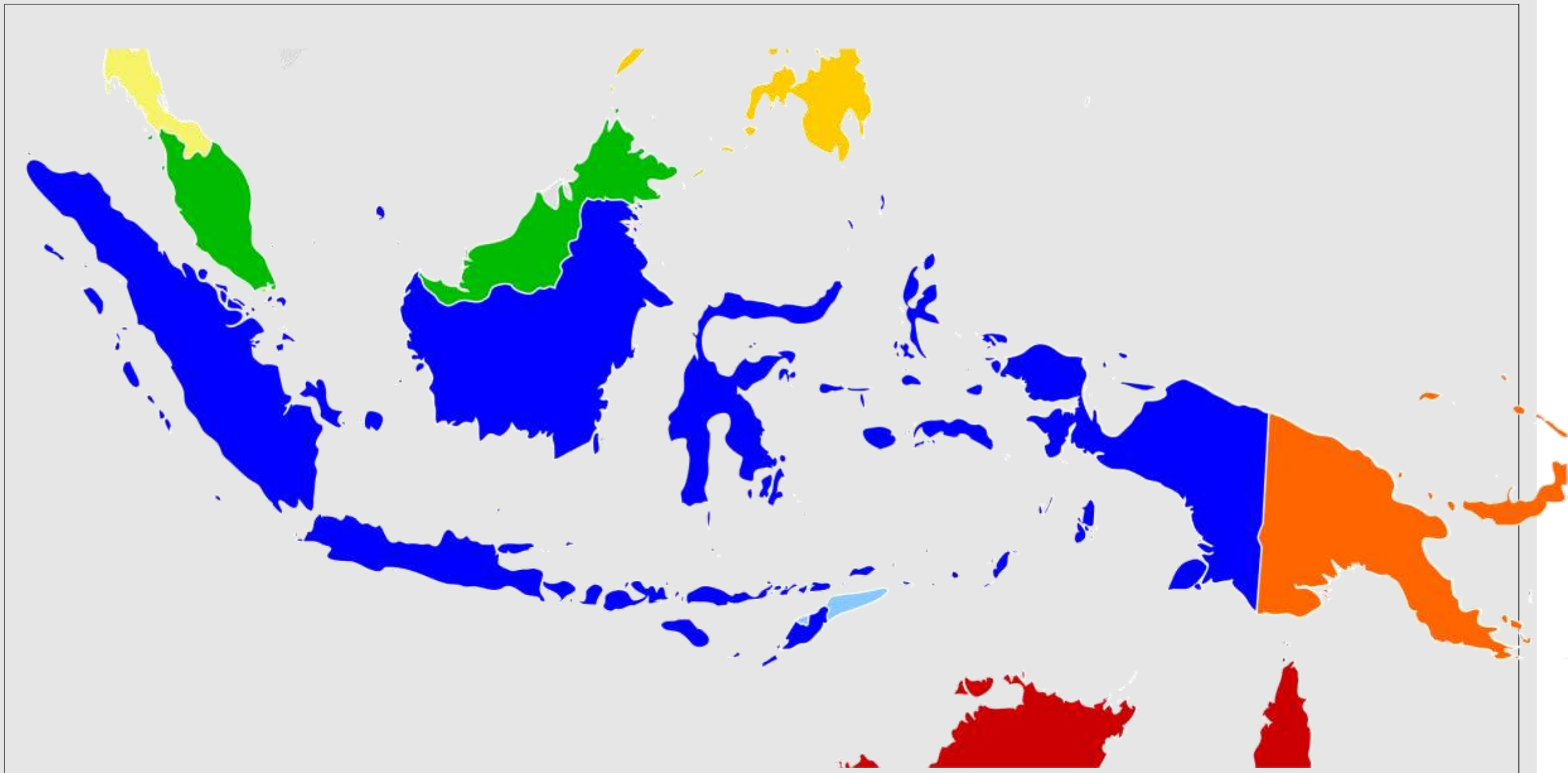
本発表の目的

- マレー・インドネシア語のステミング、レンマ化の現状について紹介する
- 発表者が開発するマレー・インドネシア語形態情報辞書MALINDO Morph (Nomoto et al. 2018)を紹介する
- MALINDO Morphへの語幹（stem）と見出し語（lemma）情報追加について報告する（ガイドライン）
  - 前提となる言語分析
  - データの具体的な表記法
  - 問題点・注意すべき点
- 今後の開発の方向性について、ご意見をいただく

マレー・インドネシア語について



文法はほぼ同じ
語彙は約90%共通
話者人口約2億5千万人



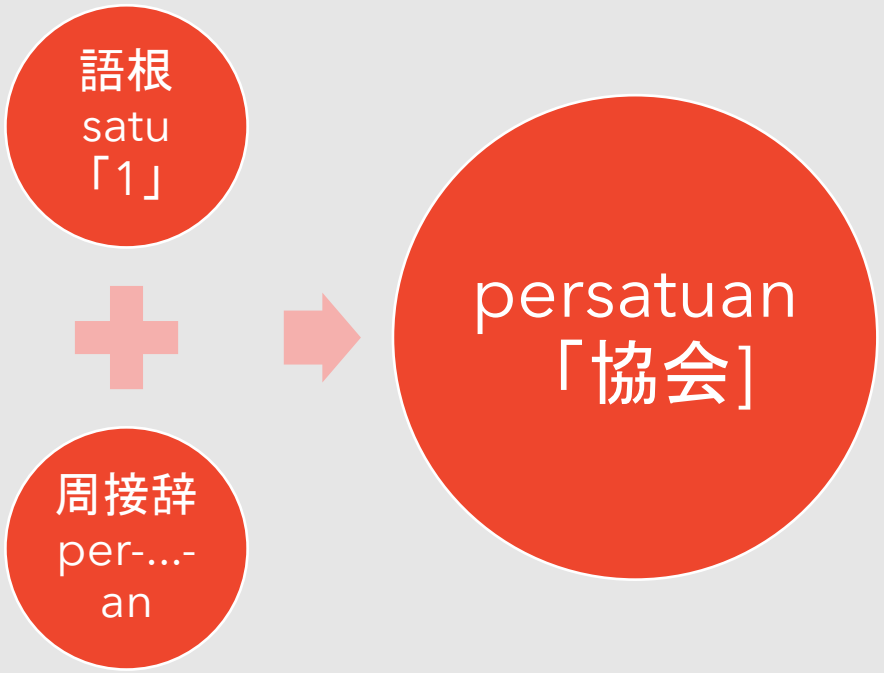
https://commons.wikimedia.org/wiki/File:Maritime_South_East_Asia.svg

世界人権宣言前文 https://www.ohchr.org/EN/UDHR/Pages/UDHRIndex.aspx

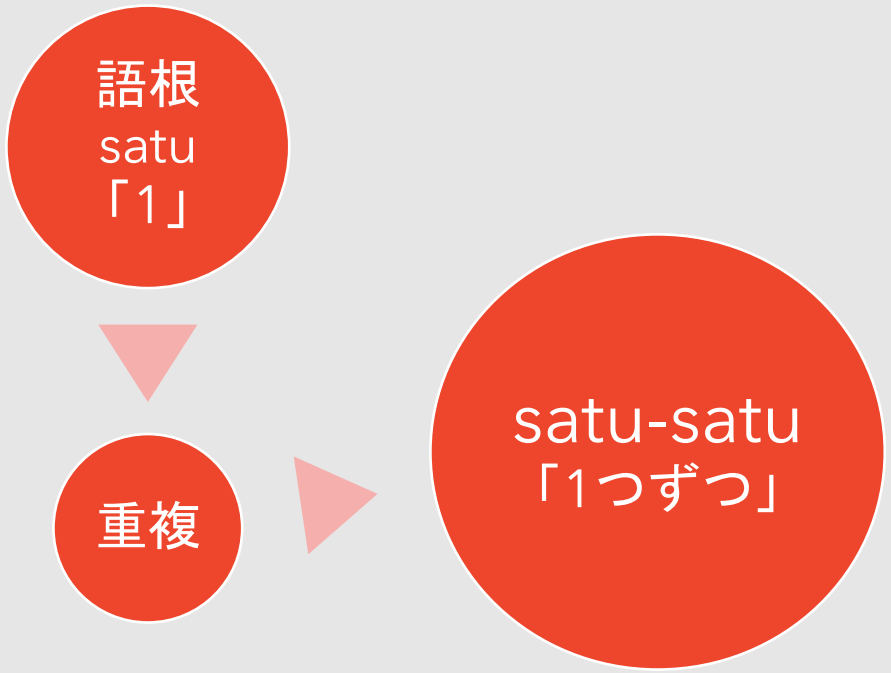
- **[zsm]** Bahawasanya pengiktirafan keutuhan kemuliaan dan hak samarata serta asasi yang tak terpisah bagi seluruh umat manusia adalah asas kebebasan, keadilan dan kedamaian dunia.
- **[ind]** Menimbang bahwa pengakuan atas martabat alamiah dan hak-hak yang sama dan mutlak dari semua anggota keluarga manusia adalah dasar kemerdekaan, keadilan dan perdamaian di dunia,
- **[jpn]**人類社会のすべての構成員の固有の尊厳と平等で譲ることのできない権利とを承認することは、世界における自由、正義及び平和の基礎であるので、
- ローマ字、語境界はホワイトスペース
- 多くの規則的派生語（緑の語）

派生語の作り方

接辞付加



重複



Pranala (*link*): https://kbbi.web.id/satu

satu */sa·tu/* *num* 1 bilangan yang dilambangkan dengan angka 1 (Arab) atau I (Romawi); 2 nama bagi lambang bilangan asli 1 (angka Arab) atau I (angka Romawi); 2 urutan pertama sebelum ke-2; 3 bilangan asli terkecil sesudah 0;
-- **bahasa** *ki* satu anggapan (pikiran, pandangan, dan sebagainya); sependapat;
-- **rayu** peras`an pilu sayu; rawan hati;

辞書は語根の下に派生語が並ぶ
（＝派生語は語根に戻して引く）

satu-satu */sa·tu-sa·tu/* *adv* 1 *cak* satu per satu; satu demi satu; 2 masing-masing; 3 tiap-tiap kali satu;

bersatu */ber·sa·tu/* *v* 1 berkumpul atau bergabung menjadi satu; menjadi satu: *bangsa-bangsa Asia Tenggara - dalam ASEAN;* 2 sepakat; seia sekata: *- kita teguh bercerai kita runtuh;* **- hati** sepakat; seia sekata;

menyatu */me·nya·tu/* *v* menjadi satu; berpadu; manunggal;

menyatukan */me·nya·tu·kan/* *v* 1 menjadikan satu; mengumpulkan (menggabungkan dan sebagainya) menjadi satu; 2 memusatkan (mengarahkan) kepada satu tujuan;

satuan */sa·tu·an/* *n* 1 bilangan bulat positif terkecil dari bilangan seluruhnya (bilangan satu): *bilangan 235 -nya 5, puluhannya 3, dan ratusannya 2;* 2 standar atau dasar ukuran (takaran, sukatan, uang, dan sebagainya): *meter ialah - ukuran panjang, sedangkan gram - berat;* 3 sekelompok orang (tentara, alat-alat, dan sebagainya) yang merupakan keutuhan: *- pasukan bermotor;* 4 perangkat; unit; **- kerja** kelompok orang yang melakukan suatu kegiatan yang sama; **- tempur** sekelompok orang (tentara, prajurit) yang melakukan pertempuran; **- tugas** sekelompok orang yang mempunyai kegiatan atau tugas yang sama;

penyatu */pe·nya·tu/* *n* 1 orang yang menyatukan; 2 alat yang menyatukan;

persatuan */per·sa·tu·an/* *n* 1 gabungan (ikatan, kumpulan, dan sebagainya) beberapa bagian yang sudah bersatu: *bahasa Indonesia adalah bahasa - bangsa Indonesia;* 2 perserikatan; serikat;

mempersatukan */mem·per·sa·tu·kan/* *v* menjadikan bersatu; menyatukan;

pemersatu */pe·mer·sa·tu/* *n* 1 orang yang mempersatukan; 2 alat untuk mempersatukan;

pemersatuan */pe·mer·sa·tu·an/* *n* proses, cara, perbuatan mempersatukan**penyatuan** */pe·nya·tu·an/* *n* proses, cara, perbuatan menyatukan;

【問題点】従来のステミング・レンマ化＝語根化

- Sastrawi stemmer https://github.com/sastrawi/sastrawi
  - 名前はステマーだが、語根を返す
  - persatuan「協会」のステミング結果が、satu「1」☹
- Malaya (Husein 2018) https://github.com/huseinzol05/malaya
  - Sastrawiの問題をそのまま継承

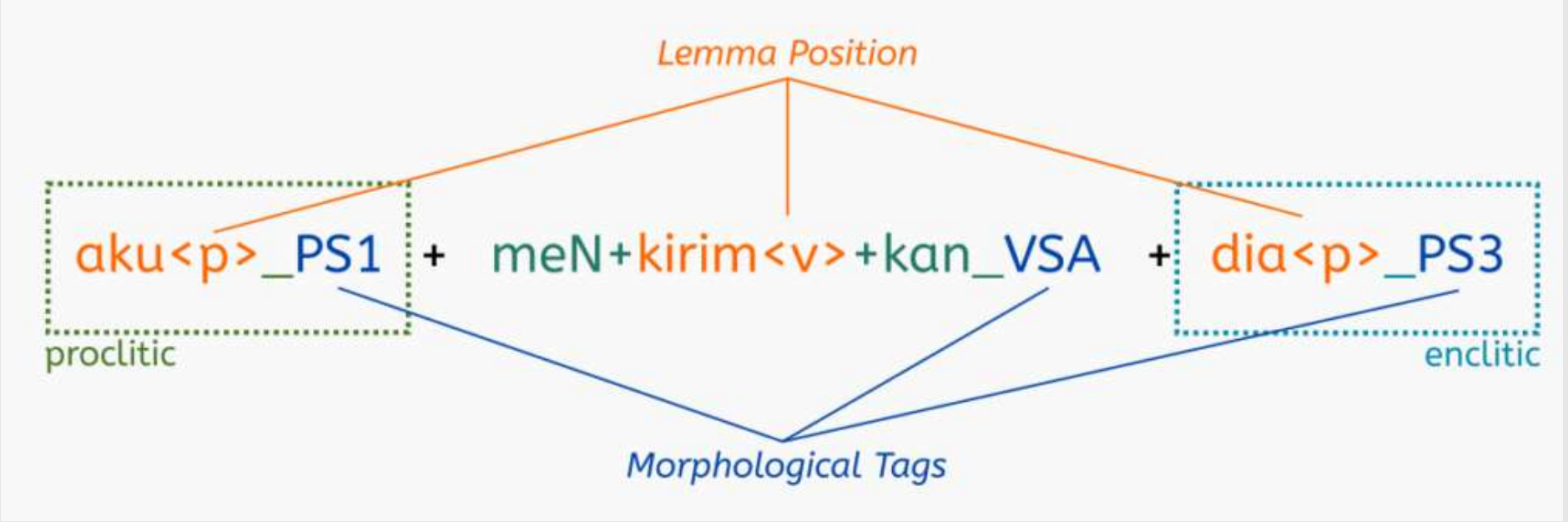
```
In [8]: stemmer = malaya.stem.deep_model('lstm')
        stemmer.stem('saya sangat sukakan awak tetapi awak sangatlah sakai')
Out[8]: 'saya sangat suka awak tetapi awak sangat sakai'
```

↑語根化としては完璧。だが、ステマーとしては問題あり。

【注意】この言語が分からなくても、簡単に（問題のある）結果が得られてしまう…

MorphInd (Larasati et al. 2011)

- 最も広く用いられている形態解析器
- https://septinalarasati.com/morphind/
- 語根をレンマとみなしている



語幹（stem）、見出し語（lemma）、語根（root）の関係

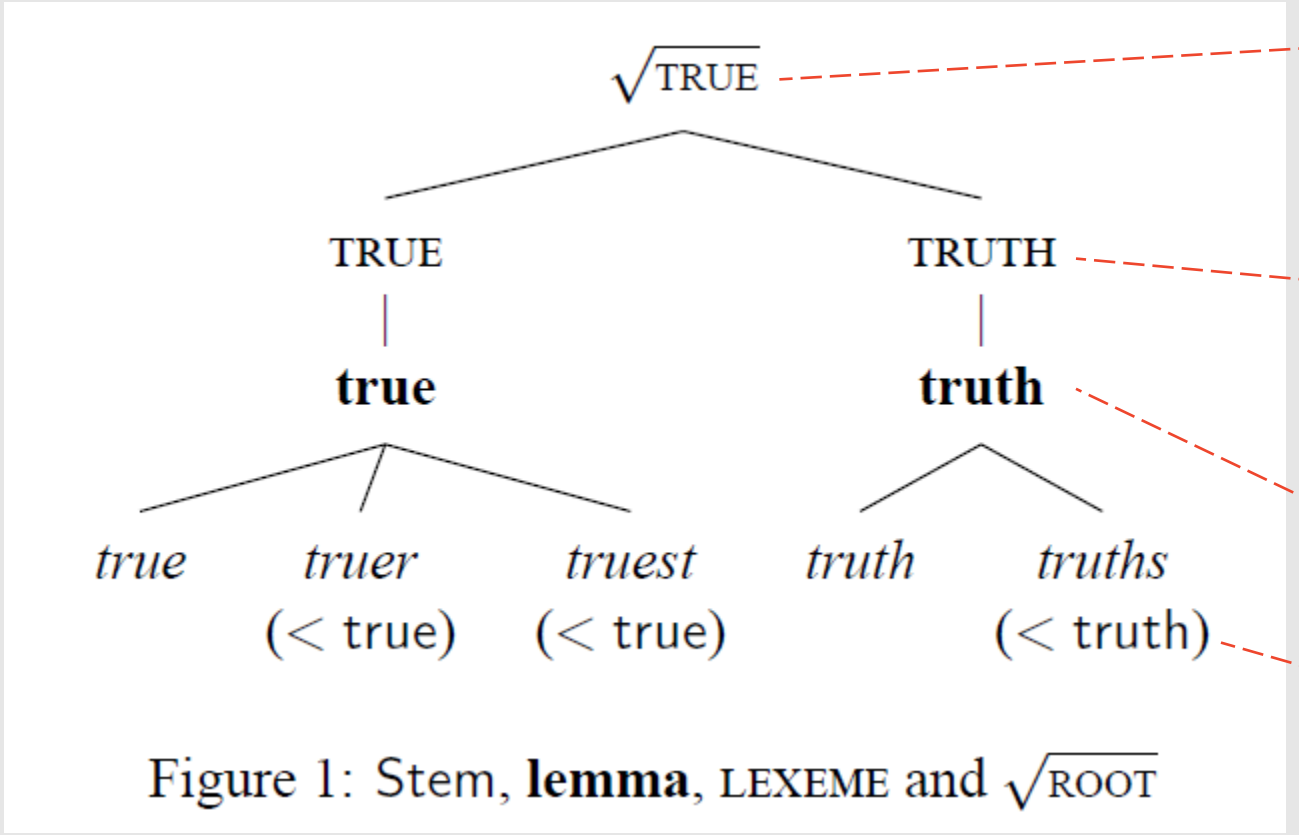
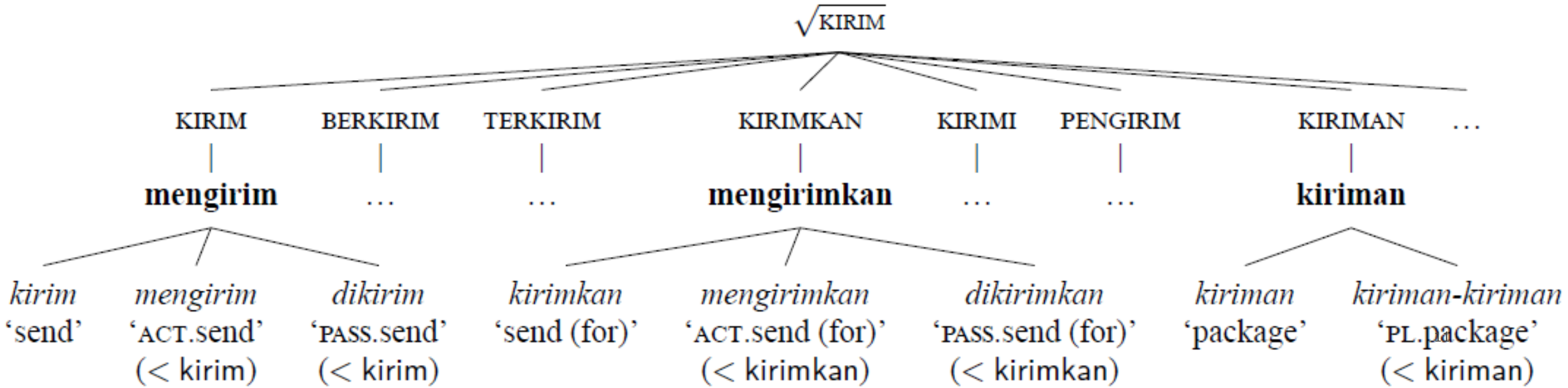


Figure 1: Stem, **lemma**, LEXEME and √ROOT

- **語根**: 屈折・派生を問わず、形態的に関連する語をつなぐ抽象的単位
- **語彙素:** 屈折により関係付けられる語の集まりとしての抽象的な語彙的単位
- **見出し語:** 語彙素（lexeme）の代表となる具体的な形式
- **語幹:** 屈折形態論の対象

語根＝見出し語（lemma）でよいか？

- 何を見出し語とするかは、個別言語がそれぞれ決める問題
- マレー・インドネシア語は語根を見出し語とすべきとの意見もある（Knowles & Zuraidah 2006:71）
- きちんとした屈折体系があるのだから、そうすべきではない！（言語を知らずに分析する人は混乱する）



現状のまとめ

- マレー・インドネシア語には、英語のステマーやレンマタイザーに相当するものはない
  （きちんとしたトークナイザーも実はない）
- 「ステマー」、「レンマタイザー」と呼ばれるものは存在するが、それらの実態はすべて「語根検出器」

- 形態情報辞書MALINDO Morph (Nomoto et al. 2018)に語幹と見出し語の情報を追加しよう！

形態情報辞書MALINDO Morph

- マレー・インドネシア語の初めてかつ唯一の形態情報辞書
- 当初版（2018年3月）：232,456行（うち人手チェック済み131,802行）

ID　語根　表層形　接頭辞・後接語　接尾辞・前接語　周接辞　重複のタイプ

- 最新版（2019年9月）：234,274行（うち人手チェック済み134,101行）

ID　語根　表層形　接頭辞・後接語　接尾辞・前接語　周接辞　重複のタイプ　出典　語幹　見出し語

↑
本発表で報告

- プログラムにより機械的に算出
- 結果を5人のインドネシア語話者がチェック

MALINDO Morphデータの例

ID	Root	Surface form	Prefix	Suffix	Circumfix	Redup.	Source	Stem	Lemma
ec-425 93	tarik	Me-nariknya	meN-	-nya	0	0	Leipzig	menariknya @menarik+dia @tarik+dia	menariknya @menarik+dia

3つの可能性

	意味	語幹	見出し語
1.	「興味深いこと、なんと興味深い」	menariknya	menariknya
2.	「彼（女）の／その興味深い～」	menarik + dia	menarik + dia
3.	「彼（女）／それを引く」	tarik + dia	menarik + dia

記号
@: または (atau)
+: トークン境界

＋（トークン境界）の効果

- 現状のホワイトスペース＋αに基づくトークン化では無視されているトークン境界が分かる

	表層形	語幹	意味
1	kupikir	aku+pikir	私は考える
2	kuasaku	kuasa+aku	私の力
3	yakah	ya+kah	そうですか？
4	diatas	di+atas	上で
5	itupun	itu+pun	それも

@（または）は難しいので、後ほど

何を語幹・見出し語とするか？

- 語幹は、最も基本的な形を選ぶ、言語学的分析で自動的に決まる

  …のだが、既存の文法記述ではほとんど言及がない

  →ステマー開発者がこぞって「語幹＝語根」と誤解

- 一方、見出し語は、基本的に何にしてもよい

  →マレーシア、インドネシアの辞書・文法書での慣習に従う

屈折① 他動詞

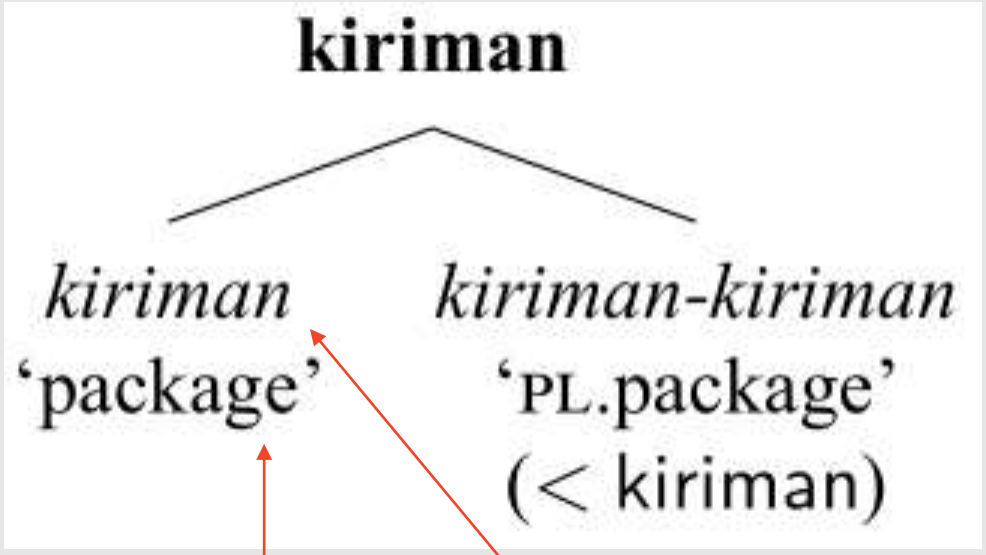
- すべての他動詞がこのパラダイムではない
- meN-形の自動詞・形容詞もある

a.	Mereka sudah *baca* buku itu. 3PL already read book that	能動1
b.	Buku itu sudah mereka *baca*. book that already 3PL read	受動1
c.	Mereka sudah *mem-baca* buku itu. 3PL already ACT-read book that	能動2
d.	Buku itu sudah *di-baca* oleh mereka. book that already PASS-read by 3PL 'They already read the book./ The book was already read by them.'	受動2

語幹
最も基本的な形なので
（偶然、語根と同形）

見出し語
現地の辞書で見出しにされるので

屈折②　可算名詞



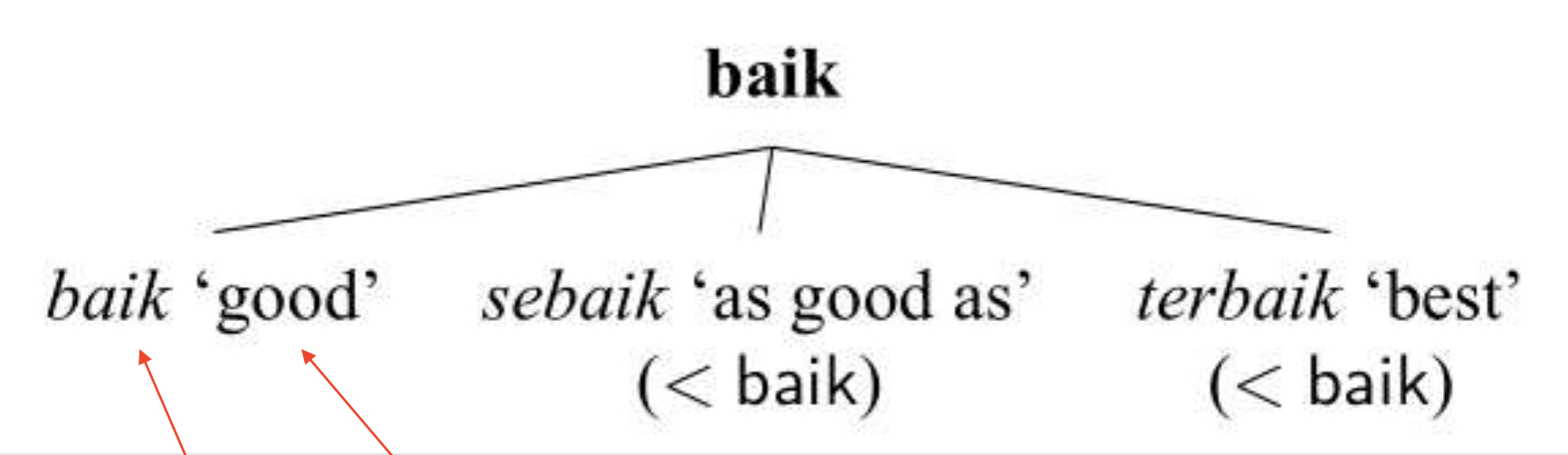
- 単数形が重複形の名詞もある

  例：kanak-kanak「子供」
- 重複形は名詞以外の品詞にもある

語幹
最も基本的な形なので
（語根とは違う形）

見出し語
現地の辞書で見出しにされるので

屈折③　形容詞



語幹
最も基本的な形なので
（語根と同形）

見出し語
現地の辞書で見出しにされるので

形容詞の屈折でなく、派生に関わる接頭辞
se-、ter-もある

@（または）の例

接頭辞　se-

1. 数詞satu「1」の接頭辞形　sebuah「1つ」
   - トークン化する　→ satu + buah
   - 語幹・見出し語はsatu
2. 「～じゅう、全～」（派生）
   - トークン化しない　sekampung「村じゅう」
   - 語幹・見出し語はse-の付いた形
3. 同等比較「同じくらい～」（屈折）　sebaik
   - トークン化しない　「同じくらいよい」
   - 語幹・見出し語はse-を外した形　→ baik

接尾辞・前接語　-nya

1. 3人称代名詞dia「彼(女)、それ」の接語形
   - トークン化する
   - 語幹・見出し語はdia
   tinggi**nya**「彼(女)の身長」
   → tinggi + dia
2. 感嘆文形成、名詞化、副詞形成
   - トークン化しない
   - 語幹・見出し語は-nyaの付いた形
   tinggi**nya**「背が高いなぁ／こと」
   → tinggi**nya**

@（または）を含む比率

語幹	@を含む行	全体	比率
コア・チェック済み（cc）	143	84,415	0.17%
拡張・チェック済み（ec）	138	49,686	0.28%
拡張・未チェック（ex）	13,085	100,173	13.1%

見出し語	@を含む行	全体	比率
コア・チェック済み（cc）	2,635	84,415	3.12%
拡張・チェック済み（ec）	293	49,686	0.59%
拡張・未チェック（ex）	21,579	100,173	21.5%

まとめ

- マレー・インドネシア語の既存のステミング・レンマ化は「語根化」（root-ing）だった
  - 言語学的に不適格
  - 言語が分からないユーザーは（英語式の）語幹・見出し語が得られたと思い込んでしまう
- 形態情報辞書MALINDO Morphに語幹と見出し語の情報を追加した
  - 各種の曖昧性が関与し、かなり難しい

活用法の例

- トークン境界情報（+）により、ホワイトスペース＋αを超えたトークン化が可能に
- 語根でなく、本物の語幹・見出し語を使うことで、情報検索の精度が向上
  例：persatuan「協会」の語幹・見出し語による検索
  satu「１」で検索 → persatuan「協会」で検索
- Wordnet Bahasa (Bond et al. 2014)の見出し語の整理

01437888-v 'send via the postal service'; V2, V3;

Inggeris	*mail*$_{12}$ (▷▷▷▷▷▷▷▷) , *get off*
Bahasa Indonesia	*kirim* , *mengepos* , *mengirim*
Bahasa Malaysia	*kirim* , *mengepos* , *mengirim*

謝辞

◦ 本研究はJSPS国際的な活躍が期待できる研究者の育成事業および科研費 JP18K00568の助成を受けたものです。

参考文献

- Bond, Francis, Lian Tze Lim, Enya Kong Tang, and Hammam Riza. 2014. The combined Wordnet Bahasa. *NUSA* 57: 83–100.
- Husein Zolkepli. 2018. Malaya. GitHub repository. URL https://github.com/huseinzol05/malaya.
- Knowles, Gerald O., and Zuraidah Mohd Don. 2006. *Word Class in Malay: A Corpus-Based Approach*. Kuala Lumpur: Dewan Bahasa dan Pustaka.
- Larasati, Septina Dian, Vladislav Kuboň, and Daniel Zeman. 2011. Indonesian morphology tool (MorphInd): Towards an Indonesian corpus. In *Systems and Frameworks for Computational Morphology*, ed. Cerstin Mahlow and Michael Piotrowski, 119–129. Verlag: Springer.
- Nomoto, Hiroki, Hannah Choi, David Moeljadi, and Francis Bond. 2018. MALINDO Morph: Morphological dictionary and analyser for Malay/Indonesian. In *Proceedings of the LREC 2018 Workshop "The 13th Workshop on Asian Language Resources"*, ed. Kiyoaki Shirai, 36–43.