

Wordnet

Jejaring kata / Pangkalan data leksikal

David Moeljadi

23 Agustus 2016

Jurusan Linguistik dan Kajian Multibahasa, Fakultas Humaniora dan Ilmu Sosial,
Universitas Teknologi Nanyang

MUSYAWARAH SEKRETARIAT KE-22 MAJELIS BAHASA BRUNEI DARUSSALAM-INDONESIA-MALAYSIA (MABBIM),
MILLENIUM HOTEL SIRIH, JAKARTA PUSAT, INDONESIA

Perkenalan diri

David Moeljadi

2014 - (2018?) S3 Linguistik di Universitas Teknologi Nanyang

1986 lahir - TK - SD - SMP - SMA (Jurusan Bahasa) di Malang, Jawa Timur

2004 - 2005 S1 Sastra Jepang di Univ. Bina Nusantara, Jakarta

2005 - 2006 Pusat Bahasa dan Kebudayaan Jepang, Universitas Kajian Asing Osaka

2006 - 2010 S1 Linguistik di Univ. Tokyo

2010 - 2012 S2 Linguistik di Univ. Tokyo

2012 - 2013 Rakuten Travel, Inc. di Tokyo

2013 - 2014 Lembaga Penelitian Bahasa dan Budaya Asia dan Afrika, Universitas Kajian Asing Tokyo

compling.hss.ntu.edu.sg/who/david/

Proyek saat ini

1. Indonesian Resource Grammar (INDRA)
github.com/davidmoeljadi/INDRA
chimpanzee.ling.washington.edu/demophilin/indra/
delph-in.github.io/delphin-viz/demo/
2. Wordnet Bahasa
3. NTU Multilingual Corpus
compling.hss.ntu.edu.sg/ntumc/
4. Indonesian Loanword Search Engine
david.blogid.me/loanword_searchengine
5. Kamus Pemelajar Kanji Jepang-Indonesia
6. KBBI IV
7. Analisis sentimen teks bahasa Indonesia

Wordnet

1. Wordnet: Apa itu?
2. Wordnet: Apa saja kegunaannya?
3. Berbagai macam Wordnet
4. Wordnet Bahasa
 - Tanya Jawab –

Wordnet: Apa itu?

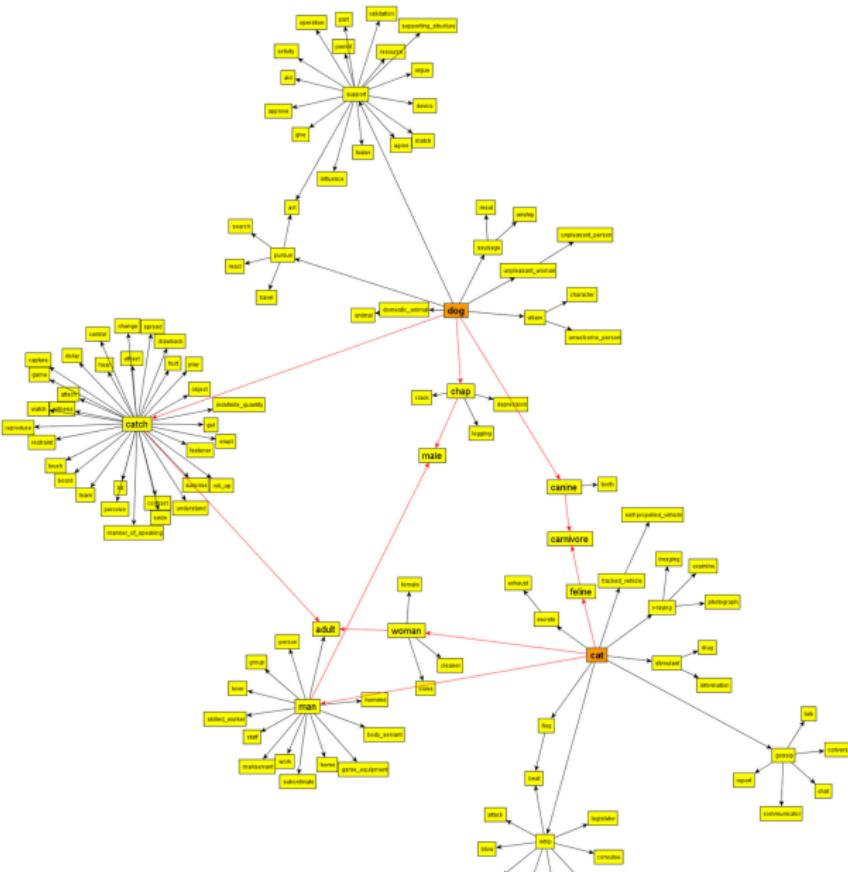
WORDNET: PANGKALAN DATA LEKSIKAL ELEKTRONIK (MILLER, 1995)

- Laboratorium Ilmu **Kognitif** Universitas Princeton
- Princeton WordNet (PWN):
wordnetweb.princeton.edu/perl/webwn
- Jaringan semantik leksikal untuk bahasa Inggris (Amerika)
- Entri leksikal disusun berdasarkan **makna**

Bagi pengguna: cara baru menggunakan kamus untuk mencari dan mempelajari kata

Bagi komputer: sumber untuk melakukan analisis semantik pada teks bahasa alami (analisis teks), aplikasi kecerdasan buatan

WORDNET: JEJARING KATA (KONSEP)



SYNSET “SYNONYM SET” / HIMPUNAN SINONIM

- Satuan dasar dalam Wordnet
- Tiap synset mewakili sebuah **konsep** (bdk. “buku” dan “kitab” di KBBI IV)
- Makna kata diwakili oleh:
 - **sinonim**
 - **kalimat definisi**
 - **hubungannya dengan synset lainnya**
- Anggota synset adalah kata-kata berkonsep sama
- **Wordnet = Kamus + Tesaurus**
- Kata yang bermakna lebih dari satu menjadi anggota lebih dari satu **synset**
- 5 kategori synset:
 - n** (nomina), **v** (verba), **a** (adjektiva), **r** (adverbia),
x (kata penggolong, kata seru, gelar/sebutan)

CONTOH

Contoh kata bermakna banyak dan memiliki synset lebih dari satu:

<nomina.tubuh> **kepala**, hulu – (bagian tubuh yang di atas leher (pada manusia dan beberapa jenis hewan merupakan tempat otak, pusat jaringan saraf, dan beberapa pusat indra))

<nomina.persona> **kepala**, ketua, pemimpin, penghulu – (orang yang mengepalai atau memimpin (rapat, dewan, perkumpulan, dsb); “**kepala** pasukan”)

<nomina.kognisi> **kepala**, kepintaran, kecerdikan, akal – (kemampuan mental; “jika kita masih punya **kepala**, kita harus diberi jatah satu kilogram gula”)

...

HUBUNGAN ANTAR-SYNSET NOMINA

hipernim istana → tempat tinggal, kediaman, rumah

holonim jari → tangan

 tepung → roti, adonan, kue

 juri → mahkamah, pengadilan, tribunal

contoh Mozart, Wolfgang Amadeus Mozart → contoh komponis, komposer, ...

juga ragam bahasa, bidang ilmu (domain) dsb.
(bdk. label-label di kamus)

Wordnet: Apa saja kegunaannya?

GOOGLE TERJEMAHAN

translate.google.com/about/intl/id_ALL/license.html

Mengelompokkan kata-kata terjemahan alternatif

Contoh:

translate.google.co.id/#en/id/head

KEUNTUNGAN MENGGUNAKAN WORDNET

- Untuk pengguna awam:
 - Gratis (Daring)(+ Luring)
 - Multibahasa
- Untuk peneliti bahasa dan pekamus:
 - Jaringan semantik leksikal dengan informasi semantik, sintaks, definisi, contoh, domain, tautan derivasi kata
 - Pembuatan kamus (cetak) multilingual dengan \LaTeX
[www.overleaf.com/latex/examples/
dictionary-template/pdztbwjxrpmz#.V7a16nV96kA](http://www.overleaf.com/latex/examples/dictionary-template/pdztbwjxrpmz#.V7a16nV96kA)
- Untuk peneliti linguistik komputasional:
 - Pemrosesan bahasa alami (NLP), analisis teks, disambiguasi makna, kecerdasan buatan, misalnya melalui NLTK (Natural Language Toolkit/Perangkat Bahasa Alami) untuk Python:
www.nltk.org

Berbagai macam Wordnet

BERBAGAI PROYEK WORDNET DI DUNIA

- Banyak Wordnet untuk berbagai bahasa telah dikembangkan!
- Proyek Wordnet awal yang penting:
EuroWordNet (Vossen, 2004) untuk bahasa-bahasa Eropa Barat
BalkaNet (Tufiş, Cristea & Stamou, 2004) untuk
bahasa-bahasa Eropa Timur
- Berbagai proyek Wordnet yang lain, dengan lisensi yang berbeda
<http://globalwordnet.org/wordnets-in-the-world/>

OPEN MULTILINGUAL WORDNET (BOND & PAIK, 2012)

- compling.hss.ntu.edu.sg/omw/
- Kumpulan berbagai Wordnet dengan lisensi sumber terbuka

Wordnet	Lang	Synsets	Words	Senses	Core	Licence	Data	Citation
Albanet	als	4,676	5,990	9,602	31%	CC BY 3.0	als.zip (+xml)	cite:als; (.bib)
Arabic WordNet (AWN)	arb	10,165	14,595	21,751	48%	CC BY SA 3.0	arb.zip (+xml)	cite:arb; (.bib)
BulTreeBank Wordnet (BTB-WN)	bul	4,999	6,783	9,056	100%	CC BY 3.0	bul.zip (+xml)	cite:bul; (.bib)
Chinese Open Wordnet	cmn	42,312	61,533	79,809	100%	wordnet	cmn.zip (+xml)	cite:cmn; (.bib)
Chinese Wordnet (Taiwan)	qcn	4,913	3,206	8,069	28%	wordnet	qcn.zip (+xml)	cite:qcn; (.bib)
DanNet	dan	4,476	4,468	5,859	81%	wordnet	dan.zip (+xml)	cite:dan; (.bib)
Greek Wordnet	ell	18,049	18,227	24,106	57%	Apache 2.0	ell.zip (+xml)	cite:ell; (.bib)
Princeton WordNet	eng	117,659	148,730	206,978	100%	wordnet	eng.zip (+xml)	cite:eng; (.bib)
Persian Wordnet	fas	17,759	17,560	30,461	41%	Free to use	fas.zip (+xml)	cite:fas; (.bib)
FinnWordNet	fin	116,763	129,839	189,227	100%	CC BY 3.0	fin.zip (+xml)	cite:fin; (.bib)
WOLF (Wordnet Libre du Français)	fra	59,091	55,373	102,671	92%	CeCILL-C	fra.zip (+xml)	cite:fra; (.bib)
Hebrew Wordnet	heb	5,448	5,325	6,872	27%	wordnet	heb.zip (+xml)	cite:heb; (.bib)
MultiWordNet	ita	34,728	40,343	61,558	83%	CC BY 3.0	ita.zip (+xml)	cite:ita; (.bib)
Japanese Wordnet	jpn	57,184	91,964	158,069	95%	wordnet	jpn.zip (+xml)	cite:jpn; (.bib)
Multilingual Central Repository	cat	45,826	46,531	70,622	81%	CC BY 3.0	cat.zip (+xml)	cite:cat; (.bib)
Multilingual Central Repository	eus	29,413	26,240	48,934	71%	CC BY 3.0	eus.zip (+xml)	cite:eus; (.bib)

Wordnet Bahasa

WORDNET BAHASA (BOND, LIM, TANG & RIZA, 2014)

- `wn-msa.sourceforge.net`
- Sumber terbuka
- Menggabungkan tiga wordnet:
 - Wordnet bahasa Melayu (Lim & Hussein, 2006)
 - Wordnet bahasa Indonesia (Riza, Budiono & Hakim, 2010)
 - Wordnet Bahasa Terbuka (Nurril Hirfana, Suerya & Bond, 2011)
- Menandai lema yang digunakan dalam:
 - bahasa Malaysia (`zsm`)
 - bahasa Indonesia (`ind`)
 - keduanya (`msa`)
- Anotasi korpus NTU-MC

ANGGOTA WORDNET BAHASA SAATINI

- Dr Francis Bond (Universitas Teknologi Nanyang, Singapura)
- David Moeljadi (Universitas Teknologi Nanyang, Singapura)
- Dr Hammam Riza (Badan Pengkajian dan Penerapan Teknologi, Indonesia)
- Dr Tang Enya Kong (Linton University College, Malaysia)
- Dr Lim Lian Tze (Malaysia)

KONEKSI DENGAN PRINCETON WORDNET DAN TIM WORDNET GLOBAL



Konferensi Wordnet Global ke-8 di Rumania (gwc2016.racai.ro/)

- Prof Christiane Fellbaum (Universitas Princeton)
- Prof Piek Vossen (VU Amsterdam)
- ...

BENGKEL WORDNET BAHASA PERTAMA DAN KEDUA



Universitas Teknologi Nanyang, Singapura, 26–27 Oktober 2014
compling.hss.ntu.edu.sg/events/2014-ws-wn-bahasa



Universitas Teknologi Nanyang, Singapura, 15–16 Januari 2016
compling.hss.ntu.edu.sg/events/2016-ws-wn-bahasa/

BENGKEL WORDNET BAHASA KETIGA



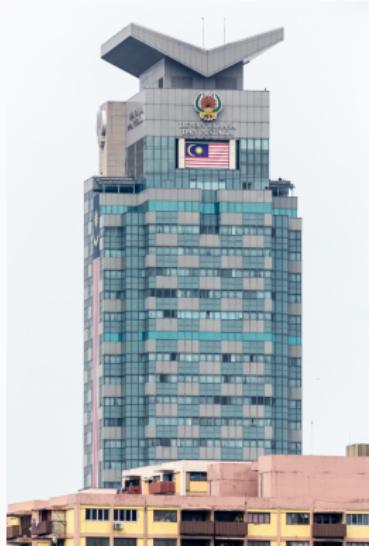
Badan Bahasa



Universitas Indonesia

Tahun 2017?

BENGKEL WORDNET BAHASA KEEMPAT



Malaysia



Brunei Darussalam

Tahun 2018?

WORDNET BAHASA DALAM ANGKA (BOND ET AL., 2014)

Wordnet	Bhs	Synset	Kata	Makna
Wordnet B.Indonesia	ind	27 506	30 358	57 560
Wordnet B.Malaysia	zsm	23 953	23 833	48 996
Wordnet Bahasa	ind	19 316	19 522	48 111
	zsm	19 347	19 572	48 181
Gabungan	ind	48 689	58 541	133 005
	zsm	38 736	45 664	114 025

- Menambah dan memperbaiki data menggunakan:
 - Kamus Besar Bahasa Indonesia (Indonesia)
 - Penerjemahan judul artikel Wikipedia dan Wikidata (Malaysia)
Flora dan fauna
 - Data bahasa Melayu Brunei (Brunei Darussalam?)

BEBERAPA PERMASALAHAN

- Beberapa konsep yang biasa/sering dipakai masih belum ada
- Sedikit kalimat definisi
- Kekeliruan (kata-kata dalam bahasa Malaysia dan bahasa Indonesia)
- Kesalahan ketik dan konsep yang tidak ada padanannya

AKHIR KATA

“Dengan Wordnet, mari kita bersama memajukan leksikografi dan penelitian bahasa di Brunei Darussalam, Indonesia, dan Malaysia”

Pertanyaan?

Ucapan terima kasih

- Terima kasih kepada Badan Bahasa (Ibu Dora Amalia) atas kesempatan yang diberikan di MABBIM
- Terima kasih kepada Lim Lian Tze untuk slides presentasi tentang Wordnet
- Terima kasih kepada Francis Bond atas dukungan dan bantuannya untuk Wordnet

Daftar Pustaka I

- Bond, F., Lim, L. T., Tang, E. K. and Riza, H. (2014). The combined Wordnet Bahasa. NUSA: Linguistic studies of languages in and around Indonesia, 57, 83–100. Retrieved from <http://hdl.handle.net/10108/79286>
- Bond, F. and Paik, K. (2012). A survey of wordnets and their licenses. In Proceedings of the 6th Global WordNet Conference (GWC 2012) (pp. 64–71). Matsue, Japan.
- Lim, L. T. and Hussein, N. (2006). Fast prototyping of a Malay WordNet system. In Proceedings of the Language, Artificial Intelligence and Computer Science for Natural Language Processing (LAICS-NLP) Summer School Workshop (pp. 13–16). Bangkok, Thailand.
- Miller, G. A. (1995). WordNet: a lexical database for English. Communications of the ACM, 38(11), 39–41.

Daftar Pustaka II

- Nurril Hirfana, M. N., Suerya, S. and Bond, F. (2011). Creating the Open Wordnet Bahasa. In Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation (PACLIC 25) (pp. 255–264). Singapore.
- Riza, H., Budiono and Hakim, C. (2010). Collaborative work on Indonesian Wordnet through Asian Wordnet (AWN). In Proceedings of the 8th Workshop on Asian language resources (pp. 9–13). Beijing.
- Tufiş, D., Cristea, D. and Stamou, S. (2004). BalkaNet: aims, methods, results and perspectives – a general overview. Romanian Journal of Information Science and Technology Special Issue, 7(1), 9–43.
- Vossen, P. (2004). EuroWordNet: a multilingual database of autonomous and language-specific wordnets connected via an Inter-Lingual-Index. Special Issue on Multilingual Databases, International Journal of Linguistics, 17(2).