

Please open your phone and load your Barcode Scanner / QR Code App





kamusi.org

Martin Benjamin
martin@kamusi.org

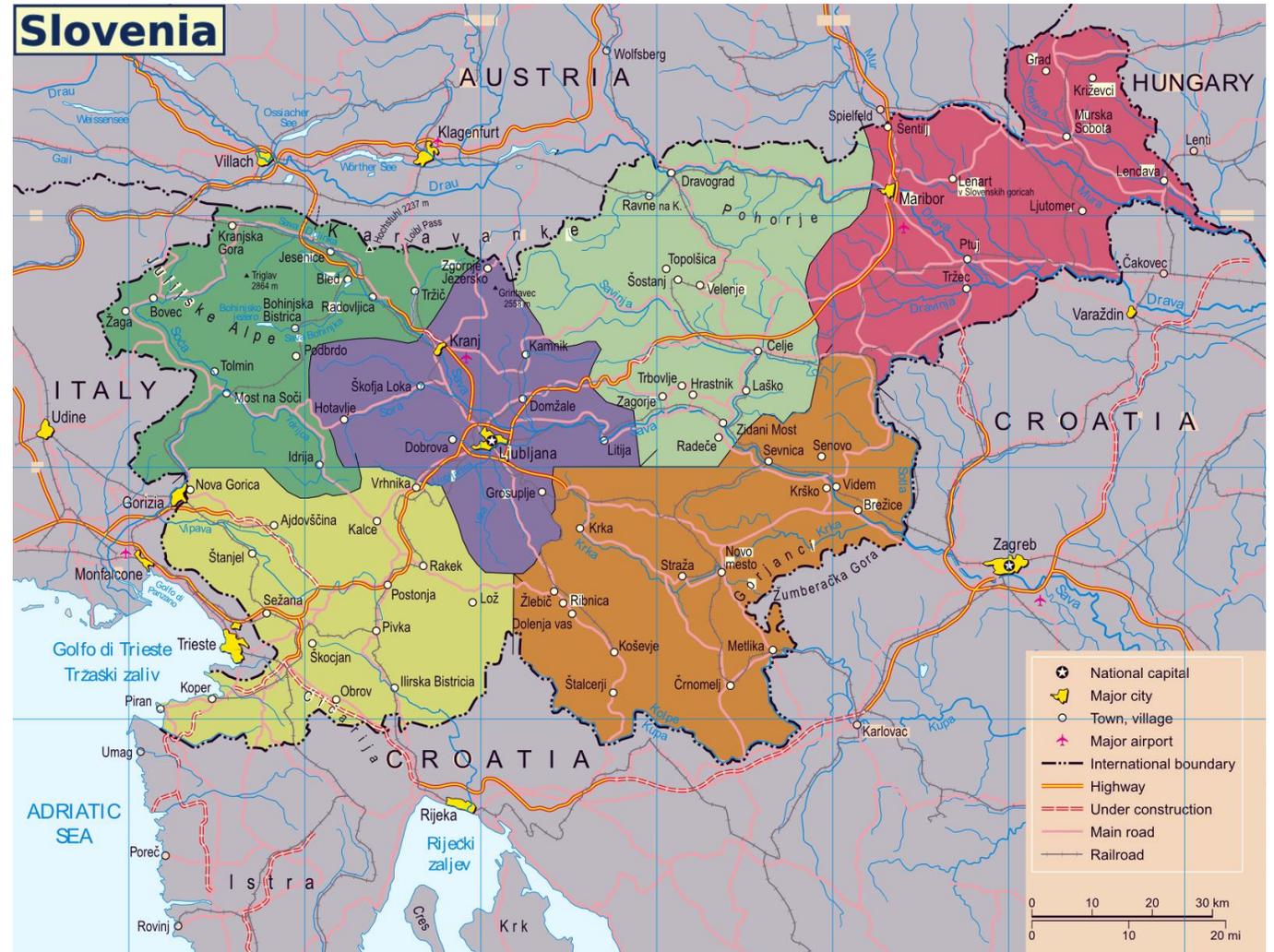
Wordnet as a crowd source for untreated languages, concepts, and data elements

Workshop on Wordnet as Lexicographic Resource
euralex2018

Ljubluana, Slovenia– 16 July, 2018

Wordnet as a crowd source for untreated languages, concepts, and data elements

- Why Wordnet?
- Why *not* Wordnet?
- Wordnet and the Crowd



Why Wordnet?

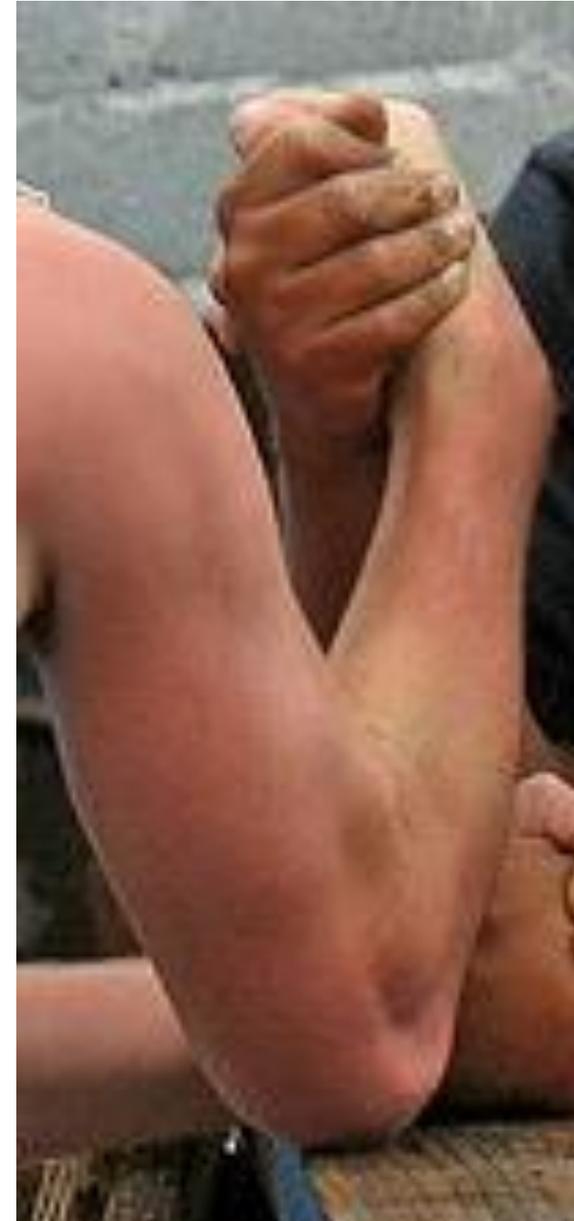
- Low hanging fruit
- Many languages
- Concept linked
- Open data
- Frequently the result of human review
- Linked to many other projects

Why *not* Wordnet?

- English as the center of gravity
- Many languages are computationally inferred
- Inconsistent composition of “Wordnets”
- First approximations vs. considered lexicography
- Poor English definitions
- Frozen data
- Canonical forms only, and definitions occasionally
- Extraneous terms (especially names)
- Missing terms from concept set

Why *not* Wordnet?

- English as the center of gravity
- Many languages are computationally inferred
- Inconsistent composition of “Wordnets”
- First approximations vs. considered lexicography
- Poor English definitions
- Frozen data
- Canonical forms only, and definitions occasionally
- Extraneous terms (especially names)
- Missing terms from concept set



Why *not* Wordnet?

- English as the center of gravity
- Many languages are computationally inferred
- Inconsistent composition of “Wordnets”
- First approximations vs. considered lexicography
- Poor English definitions
- Frozen data
- Canonical forms only, and definitions occasionally
- Extraneous terms (especially non-English)
- Missing terms from concept sets

03485997	-n	'the appendage to an object that is designed to be held in order to use or move it';	Search WN	
Albanian		<i>dorezë</i>		
Arabic		مقبض		
Catalan		<i>aferrall , nansa , aferrador , agafall , aga</i>		
Chinese (simplified)		柄, 把柄, 把手, 手柄		
Greek		λαβή		
English		<i>handle₄ (▶) , grip₁ (▶) , hold , handgrip</i>		
Basque		<i>eskutoki , eskuleku , euskarri , orakarri ,</i>		
Farsi		قبضه شمشیر		
Finnish		<i>sanka , kahva , kädensija , ripa</i>		
French		<i>poignée , prise , manche</i>		
Galician		<i>mango , asa , agarra , agarradoira , pica</i>		
Croatian		<i>ručka , drška , ručica</i>		
Indonesian		<i>telinga , batang , gagang , pemegang , t</i>		
Italian		<i>maniglia , presa , impugnatura , manopo</i>		
Japanese		把手, 柄, 取り所, ハンドル, つ分, 掴み, 手持部分, 手		
Dutch		<i>oor , kruk , handel</i>		
Polish		<i>rażka</i>		
Portuguese		<i>alça , cabo , trinco , taramela</i>		
Chinese (traditional)		把		
Romanian		<i>coadă , mâner</i>		
Slovak		<i>ručka , držadlo</i>		
Slovene		<i>ročica , roč , držaj , prijem , ročaj , držal</i>		
Spanish		<i>agarradera , asa , asidero</i>		
Thai		มือจับ , ด้าม , ที่จับ , ด้ามจับ , คั่น , หูหิ้ว		
Malaysian		<i>batang , gagang , hendal , tangkai , pega</i>		

01804414	-v	'show and train'; V2;		
English		<i>handle (▶▶)</i>		
Finnish		<i>pitää käsissä , pidellä</i>		
French		<i>gérer , traïter</i>		
Indonesian		<i>menekel , membehandel</i>		
Italian		<i>maneggiare</i>		
Japanese		操る		
Romanian		<i>conduce</i>		
Thai		ฝึก		
Malaysian		<i>memperlakukan</i>		

Why *not* Wordnet?

00672433-v (50) V1, V2	estimate, guess, judge, gauge, approximate	judge tentatively or form an estimate of (quantities or time)
	ثَمَّن, حكم على, قارب, ثمن, كان رأيا, حكم قضائيا, قِيم, قدر, تبار, فصل, خمن, خَمَّن, حزر, قوم, استنتج, قاس, عين سعة شيء ما, ظن, قدر, حاكم	

inconsistent composition of
"Wordnets"

- First approximations vs. considered lexicography
- Poor English definitions

00672433-v (50) V1, V2	estimate, guess, judge, gauge, approximate	judge tentatively or form an estimate of (quantities or time)
	見立てる, 見積る, 予算+する, 目算, 積もる, 目算+する, 見積 もる, 予算, 積る, 推算, 推算 +する	

- Missing terms from concept set

1. ثَمَّن, evaluated; 2. حكم على, judged; 3. قارب, compared; 4. ثمن, price; 5. كان رأيا, had an idea about; 6. حكم قضائيا, verdict; 7. قِيم, evaluated; 8. قدر, considered; 9. تبار, focused; 10. فصل, separated; 11. خمن, guessed; 12. خَمَّن, quantified; 13. حزر, guessed; 14. قوم, measured; 15. استنتج, concluded; 16. قاس, measured; 17. عين سعة شيء ما, set capacity of; 18. ظن, doubted; 19. قدر, evaluated; 20. حاكم, put to trial

1. 見立てる to judge or diagnose [kanji for see and stand up] (make a visual estimation such as a physical exam, or take measurements for clothing); 2. 見積る, 3. 見積もる to estimate [kanji for see and stack] (predict price and time for a job); 4. 予算+する, 5. 予算 to estimate or budget [kanji for calculate and beforehand] (calculate anticipated expenses); 6. 目算, 7. 目算+する to estimate [kanji for calculate and look] (an inexact number such as ml in a cup or remaining moves in Go); 8. 積もる, 9. 積る to estimate [kanji for stack] (uncountable things such as snow or emotions); 10. 推算, 11. 推算+する estimation [kanji for calculate and guess] (less-knowable or unknowable things such as a coin flip, the size of a crowd, or evaluation of a crime scene)

Why *not* Wordnet?

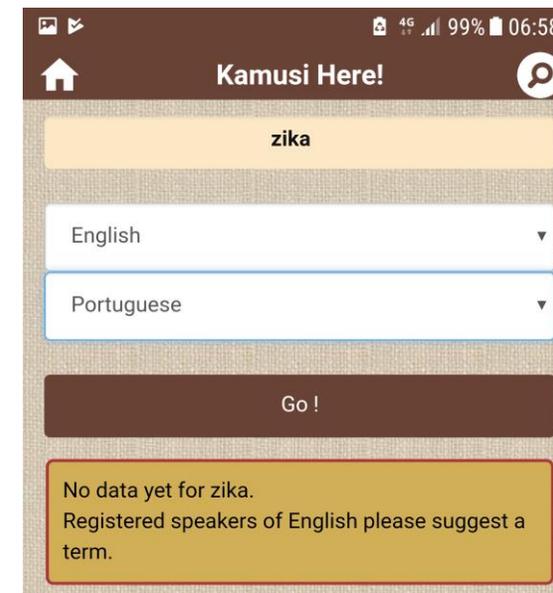
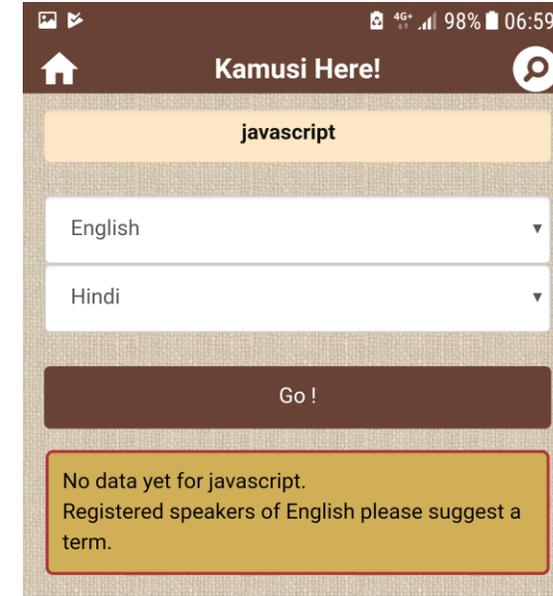
- English as the center of gravity
- Many languages are computationally inferred
- Inconsistent composition of “Wordnets”
- First approximations vs. considered lexicography
- **Poor English definitions**
- Frozen data
- Canonical forms only, and definitions occasionally
- Extraneous terms (especially names)
- Missing terms from concept set

Language	Term	Description
English	policewoman (n) police matron (n)	Definition: a woman policeman
Romanian	polițistă (n)	Definition: Femeie polițist

Language	Term	Description
English	law practice (n)	Definition: the practice of law
Dutch	gericht (n) judicatuur (n) rechtspraak (n) berechting (n) justitie (n) recht (n)	Definition in Dutch not available from original source. Member-provided definitions coming soon.

Why *not* Wordnet?

- English as the center of gravity
- Many languages are computationally inferred
- Inconsistent composition of “Wordnets”
- First approximations vs. considered lexicography
- Poor English definitions
- Frozen data
- Canonical forms only, and definitions occasionally
- Extraneous terms (especially names)
- Missing terms from concept set



Why *not* Wordnet?

- English as the center of gravity
- Many languages are computationally inferred
- Inconsistent composition of “Wordnets”
- First approximations vs. considered lexicography
- Poor English definitions
- Frozen data
- **Canonical forms only, and definitions occasionally**
- Extraneous terms (especially names)
- Missing terms from concept set

English	kid (n) child (n)	Definition: a human offspring (son or daughter) of any age Example: they had three children. they were able to send their kids to college.
Romanian	copil (n)	Definition: Un individ al speciei umane (indiferent de vârstă) văzut prin prisma faptului ca este fiul sau fiica unor părinți

		Articulated
Masculine singular	copil	copilul
Masculine plural	copii	copiii
Feminine singular	copilă	copila
Feminine plural	copile	copilele

Why *not* Wordnet?

- English as the center of gravity
- Many languages are computationally inferred
- Inconsistent composition of “Wordnets”
- First approximations vs. considered lexicography
- Poor English definitions
- Frozen data
- Canonical forms only, and definitions occasionally
- Extraneous terms (especially names)
- Missing terms from concept set

Language	Term	Description
English	cooperstown (n)	Definition: a small town in east central New York
Punjabi	No data yet	Registered speakers of Punjabi please input your suggestion. (Future Feature)

10908313-n 'English navigator who claimed the east coast of Australia for Britain and discovered several Pacific islands (1728–1779)';

Catalan	<i>James Cook</i>
English	<i>Cook , James Cook , Captain Cook , Captain James Cook</i>
Finnish	<i>Cook , kapteeni James Cook , kapteeni Cook , James Cook</i>
French	<i>cuisinier , chef , Philippe Lefebvre , cook , cuisinière , James Cook</i>
Galician	<i>James Cook</i>
Indonesian	<i>Cook , Captain James Cook , Captain Cook , James Cook</i>
Japanese	<i>ジェームズ・クック</i>
Portuguese	<i>James cook , James Cook</i>
Slovene	<i>kuhar , cook , kuharica , James Cook</i>
Spanish	<i>James Cook</i>
Thai	<i>เจมส์ คูก , กัปตันเจมส์ คูก , กัปตันคูก</i>
Malaysian	<i>Captain James Cook , Captain Cook , James Cook , Cook</i>

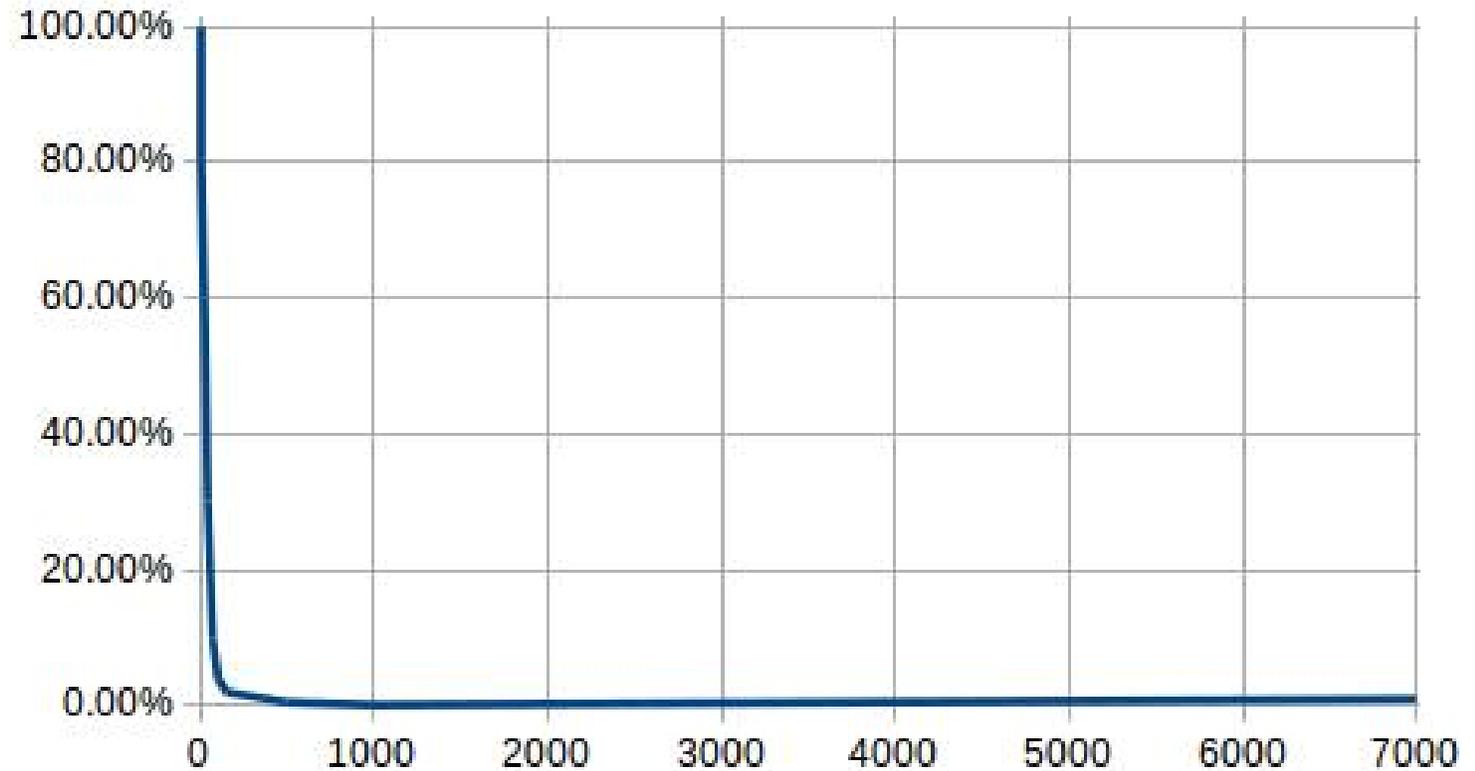
Why *not* Wordnet?

- English as the center of gravity
- Many languages are computationally inferred
- Inconsistent composition of “Wordnets”
- First approximations vs. considered lexicography
- Poor English definitions
- Frozen data
- Canonical forms only, and definitions occasionally
- Extraneous terms (especially names)
- **Missing terms from concept set**



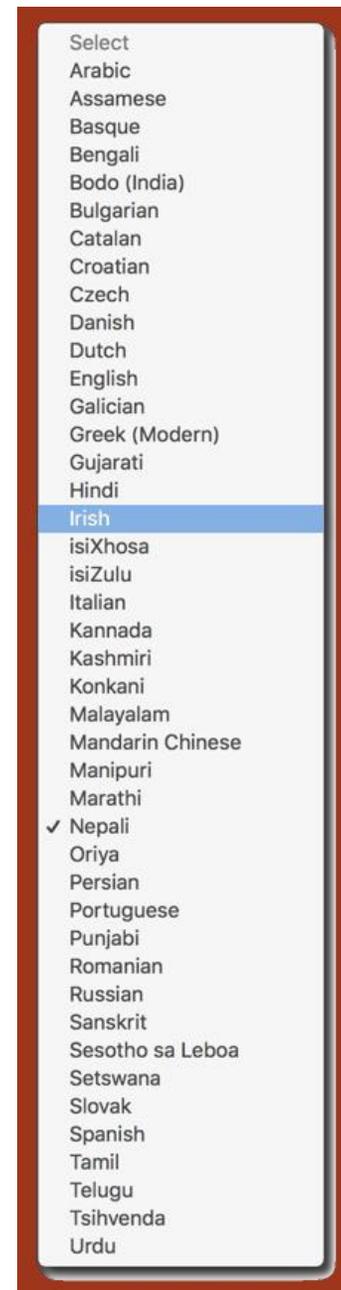
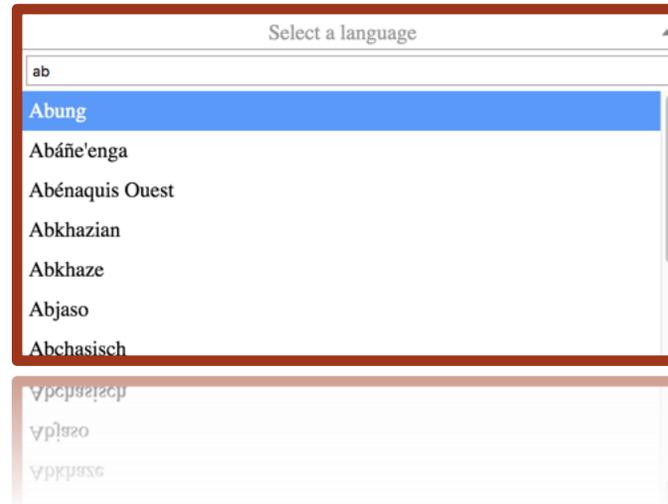
Wordnet and the Crowd

- Untreated languages
- Poorly-treated languages
- Well-treated languages



Wordnet and the Crowd

- Untreated languages
 - Starting from scratch
 - Starting with non-Wordnet datasets (DUCKS)
 - Any language with a sufficient crowd
- Poorly-treated languages
- Well-treated languages



639-3	eng
aaa	Ghotuo
aab	Alumu-Tesu
aac	Ari
aad	Amal
aae	Albanian (Arbëreshë Dialect)
aaf	Aranadan
aag	Ambrak
aah	Abu' Arapesh
aaï	Arifama-Miniafia
aak	Ankave
aal	Afade
aam	Aramanik
aan	Anambé
aaö	Arabic (Algerian Sahara)
aap	Arára, Pará
aaq	Abnaki (Eastern)
aar	Afar
aas	Aasáx
aat	Albanian (Arvanitika)
aaü	Abau
aaw	Solong
aax	Mandobo Atas
(aay)	Aariya
aaz	Amarasi
aba	Abé
abb	Bankon
abc	Ayta, Ambala
abd	Agta, Camarines Norte

Wordnet and the Crowd

- Untreated languages
- Poorly-treated languages
 - 4,476 (dan) → 117,659 (PWN) = 3.8%
 - Merge with other datasets (DUCKS)
- Well-treated languages

34 Open Wordnets Merged

Wordnet	Lang	Synsets	Words	Senses	Core
Albanet	als	4,675	5,988	9,599	31%
Arabic WordNet (AWN v2)	arb	9,916	17,785	37,335	47%
BulTreeBank Wordnet (BTB-WN)	bul	4,959	6,720	8,936	99%
Chinese Open Wordnet	cmn	42,312	61,533	79,809	100%
Chinese Wordnet (Taiwan)	qcn	4,913	3,206	8,069	28%
DanNet	dan	4,476	4,468	5,859	81%
Greek Wordnet	ell	18,049	18,227	24,106	57%
Princeton WordNet	eng	117,659	148,730	206,978	100%

Wordnet and the Crowd

- Untreated languages
- Poorly-treated languages
- Well-treated languages
 - Error detection
 - From synset alignment to term alignment
 - Language A → B validation
 - Additional monolingual data

Language	Term	Description
English	suss out (v) look into (v) go over (v) check up on (v) check over (v) check out (v) check into (v) check (v)	Definition: examine so as to determine accuracy, quality, or condition Example: check the brakes. Check out the engine.
Romanian	verifica (v) inspecta (v) controla (v)	Definition: A supune controlului pentru a determina calitatea, acuratețea etc.

What the Crowd can Give

- Warnings
- Terms
- Definitions
- Usage examples
- Inflections
- Alignment
- Grouping and ranking
- Unknown indigenous terms
- Prioritization

What the Crowd can Give

- Warnings
- Terms
- Definitions
- Usage examples
- Inflections
- Alignment
- Grouping and ranking
- Unknown indigenous terms
- Prioritization

The image shows a screenshot of a web interface with two entries for the word "fat". Each entry is contained in a light yellow box with a dark border. The top entry has a "select" button on the left and a "bad definition" button with a close icon on the right. The text for this entry is: "fat {be fat}, adjective", "definition: having a relatively large diameter", "part of speech: adjective", "examples: 'a fat rope'", and "context: fat". The bottom entry has the same "select" and "bad definition" buttons. Its text is: "fat {be fat}, adjective", "definition: having an (over)abundance of flesh", "part of speech: adjective", "examples: 'he hadn't remembered how fat she was'", and "context: fat".

select fat {be fat}, adjective bad definition ×

definition: having a relatively large diameter

part of speech: adjective

examples: "a fat rope"

context: fat

select fat {be fat}, adjective bad definition ×

definition: having an (over)abundance of flesh

part of speech: adjective

examples: "he hadn't remembered how fat she was"

context: fat

What the Crowd can Give

- Warnings
- Terms
- Definitions

Jeu des traductions
Traduisez le mot suivant vers la langue suivante : Français
know *verbe*
Working definition: be familiar or acquainted with a person or an object

connaitre

? Je ne saurais dire... Pas

connaitre

kanasi GAME

English	zap (v) nuke (v) microwave (v) micro-cook (v)	Definition: cook or heat in a microwave oven Example: You can microwave the leftovers.
Tshivenda	No data yet	Registered speakers of Tshivenda please input your suggestion. (Future Feature)
English	microwave (n)	Definition: a short electromagnetic wave (longer than infrared but shorter than radio waves)
Tshivenda	No data yet	Registered speakers of Tshivenda please input your suggestion. (Future Feature)
English	microwave oven (n) microwave (n)	Definition: kitchen appliance that cooks food by passing an electromagnetic wave through it
Tshivenda	No data yet	Registered speakers of Tshivenda please input your suggestion. (Future Feature)

What the Crowd can Give

- Warnings
- Terms
- Definitions
- Usage examples
- Inflections
- Alignment
- Grouping and ranking
- Unknown indigenous terms
- Prioritization



Definition Game ⓘ
Write or vote for a definition in English

go *noun*
Working definition: **a usually brief attempt**

👉 I can write the winning definition for this idea!

An attempt to achieve something with a recognized pos

? I can't say - skip this one...

▶ Keep the working definition. It's spectacular as it is!

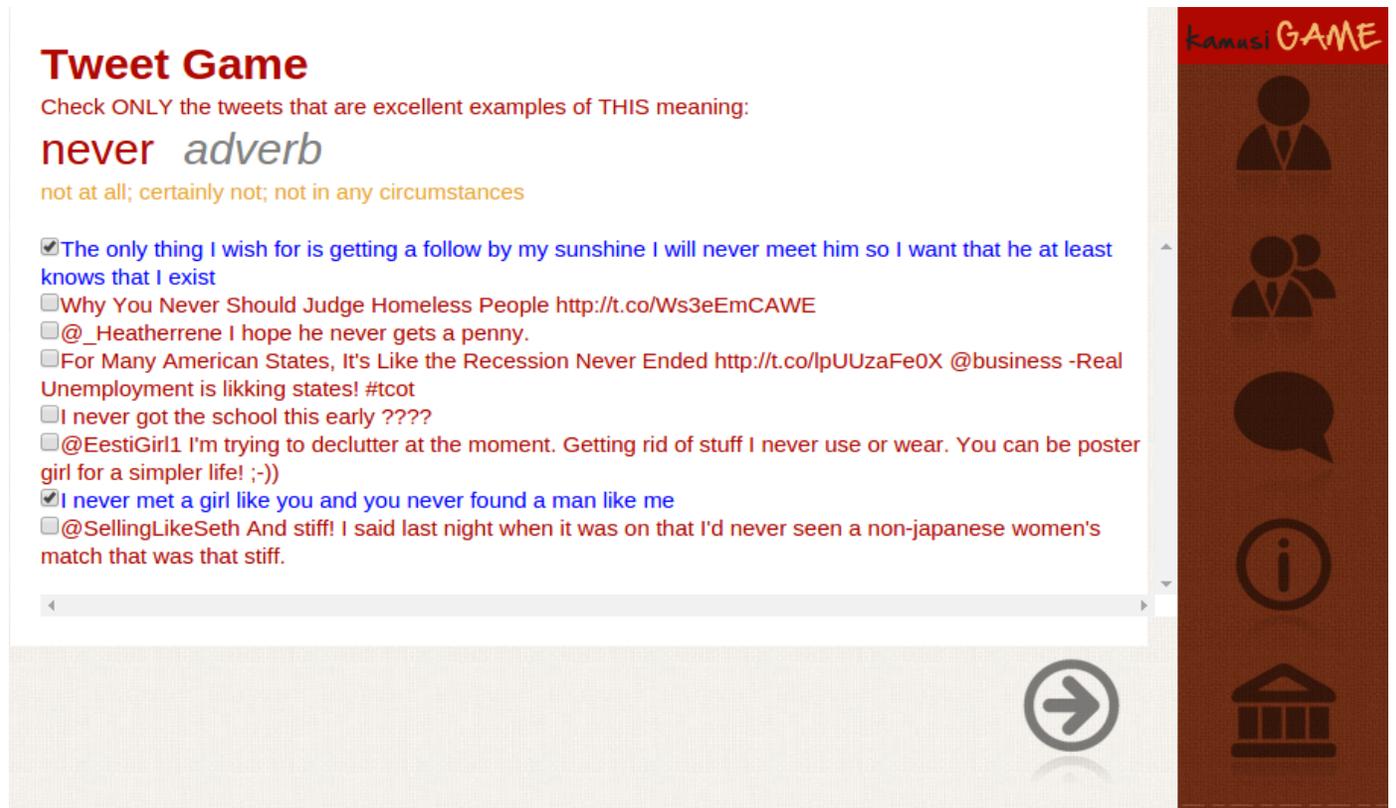
[Wiktionary](#) • [Dictionary.com](#) • [Wordnik](#)

Kamus! GAME
In Play: 90
Banked: 11

Icons: 👤, 👥, 💬, ⓘ, 🏛️

What the Crowd can Give

- Warnings
- Terms
- Definitions
- Usage examples
- Inflections
- Alignment
- Grouping and ranking
- Unknown indigenous terms
- Prioritization



Tweet Game

Check ONLY the tweets that are excellent examples of THIS meaning:

never *adverb*

not at all; certainly not; not in any circumstances

- The only thing I wish for is getting a follow by my sunshine I will never meet him so I want that he at least knows that I exist
- Why You Never Should Judge Homeless People <http://t.co/Ws3eEmCAWE>
- @_Heatherrene I hope he never gets a penny.
- For Many American States, It's Like the Recession Never Ended <http://t.co/lpUUzaFe0X> @business -Real Unemployment is likking states! #tcot
- I never got the school this early ????
- @EestiGirl1 I'm trying to declutter at the moment. Getting rid of stuff I never use or wear. You can be poster girl for a simpler life! ;-))
- I never met a girl like you and you never found a man like me
- @SellingLikeSeth And stiff! I said last night when it was on that I'd never seen a non-japanese women's match that was that stiff.

Navigation icons on the right: Home, Profile, Friends, Messages, Info, and a building icon.

Bottom right navigation icon: A circular arrow pointing right.

What the Crowd can Give

- Warnings
- Terms
- Definitions
- Usage examples
- **Inflections**
- Alignment
- Grouping and ranking
- Unknown indigenous terms
- Prioritization

English	kid (n) child (n)	Definition: a human offspring (son or daughter) of any age Example: they had three children. they were able to send their kids to college.
Romanian	copil (n)	Definition: Un individ al speciei umane (indiferent de vârstă) văzut prin prisma faptului ca este fiul sau fica unor părinți

		Articulated
Masculine singular	copil	copilul
Masculine plural	copii	copiii
Feminine singular	copilă	copila
Feminine plural	copile	copilele

What the Crowd can Give

Language	Term	Description
English	suss out (v)	Definition: examine so as to determine accuracy, quality, or condition Example: check the brakes. Check out the engine. Similar: look into go over check up on check check over check out check into
Romanian	 verifica (v)  inspecta (v)  controla (v)	Definition: A supune controlului pentru a determina calitatea, acuratețea etc.

- Alignment
- Grouping and ranking
- Unknown indigenous terms
- Prioritization

Merge Game

How well do the following two words correspond?

English
abacus
noun
a calculator that performs arithmetic functions by manually sliding counters on rods or in grooves

English - Vietnamese
abacus - bàn tính
danh từ
Bàn tính là một công cụ tính toán được sử dụng chủ yếu để thực hiện các phép toán

 Exactly

 Nearly

 Not at all

 I don't know

Kamus GAME

In Play: 3
Banked: 0



What the Crowd can Give

- Warnings
- Terms
- Definitions
- Usage examples
- Inflections
- Alignment
- Grouping and ranking
- Unknown indigenous terms
- Prioritization

The screenshot shows the 'KAMUSI GROUPING TOOL | ZANA JUMUISHI KU'ORODHESHA KAMUSI' interface. On the left, there's a sidebar for 'THE KAMUSI PROJECT' with a 'Translate' section where 'tafuta' is entered. Below that are navigation links for 'DICTIONARIES', 'DISCUSSION', 'LEARNING GUIDE', 'AFRICA GUIDE', 'BE AN EDITOR', 'PARTICIPANTS', 'CONTACT US', 'HOW TO HELP', and 'QUESTIONS'. The main area displays a list of entries for '-tafuta' with various English definitions and a 'BREAK—KITENGO—' separator. On the right, there are buttons for 'Move To Top | Juu Kabisa', 'Move Up | Pandisha', 'Add Break | Tenga', 'Move Down | Telemsha', and 'Move To Bottom | Chini Kabisa'. At the bottom, there are instructions for the tool and a 'Submit Ordering | Peleka' button.

The screenshot shows the 'Kamusi Here!' mobile application interface. At the top, there's a search bar with the Swahili word '𐞏'. Below it are dropdown menus for 'Mandarin Chinese' and 'Greek (Modern)'. A 'Go!' button is present. The main content area displays a list of entries for 'Chinese_m (a): 𐞏'. Each entry includes the source language, the word, and its definition in the target language. For example, the first entry shows 'Greek (Modern) (a): στεγνός, ἕληρος' with the definition 'αὐτός που δεν είναι βρεγμένος, διαμοισμένος από κάποιο υγρό, ἴδιως από νερό'. Other entries include 'English (a): dry', 'English (a): dried-up', and 'English (n): trunk, tree trunk, bole'. At the bottom, there are buttons for 'Sources' and 'Mandarin Chinese' and 'Greek (Modern)'. The status bar at the top shows the time as 15:28 and battery level at 80%.

What the Crowd can Give

- Warnings
- Terms
- Definitions
- Usage examples
- Inflections
- Alignment
- Grouping and ranking
- **Unknown indigenous terms**
- Prioritization



- from log files (frequent null searches)
- from SlowBrew (user translation texts)
- from Visual Dictionary
- from corpus (tagging games)
- from DUCKS (non-Wordnet datasets)

What the Crowd can Give

- Warnings
- Terms
- Definitions
- Usage examples
- Inflections
- Alignment
- Grouping and ranking
- Unknown indigenous terms
- Prioritization

00128638-n 'a hit that flies up in the air';

English	<i>fly₁</i> (≡), <i>fly ball</i>
Finnish	<i>korkea pallo</i>
French	<i>mouche</i> , <i>voler</i>
Japanese	飛球, フライ
Spanish	<i>pelota elevada</i> , <i>elevado</i>
Thai	ฟลายบอล

02190166-n ★ 'two-winged insects characterized by active flight'

Arabic	نُبابَة
Bulgarian	<i>myxa</i>
Catalan	<i>mosca</i>
Chinese (simplified)	蝇, 苍蝇
Danish	<i>flue</i>
Greek	μύγα
English	<i>fly₆</i> (≡)
Basque	<i>euli</i>
Finnish	<i>kärpänen</i>
French	<i>mouche</i> , <i>voler</i>
Galician	<i>mosca</i>
Croatian	<i>muha</i>
Indonesian	<i>lalat</i>
Icelandic	<i>fluga</i>
Italian	<i>Mosca</i> , <i>mosca</i>
Japanese	蝇, フライ, ハエ
Dutch	<i>vlieg</i>
Nynorsk	<i>fluge</i>
Bokmål	<i>flue</i>
Polish	<i>mucha</i>
Portuguese	<i>Moscas</i> , <i>mosca</i> , <i>Diptera</i>
Romanian	<i>muscă</i>
Slovene	<i>muha</i>
Spanish	<i>mosca</i>
Swedish	<i>fluga</i>
Thai	แมลงวัน, แมลงวัน
Malaysian	<i>lalat</i>



kamusi.org

Martin Benjamin
martin@kamusi.org

Wordnet as a crowd source for untreated languages, concepts, and data elements

Workshop on Wordnet as Lexicographic Resource
euralex2018

Ljubluana, Slovenia– 16 July, 2018