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

ANDROID APP ON Google Play

Download on the App Store

Martin Benjamin
martin@kamusi.org

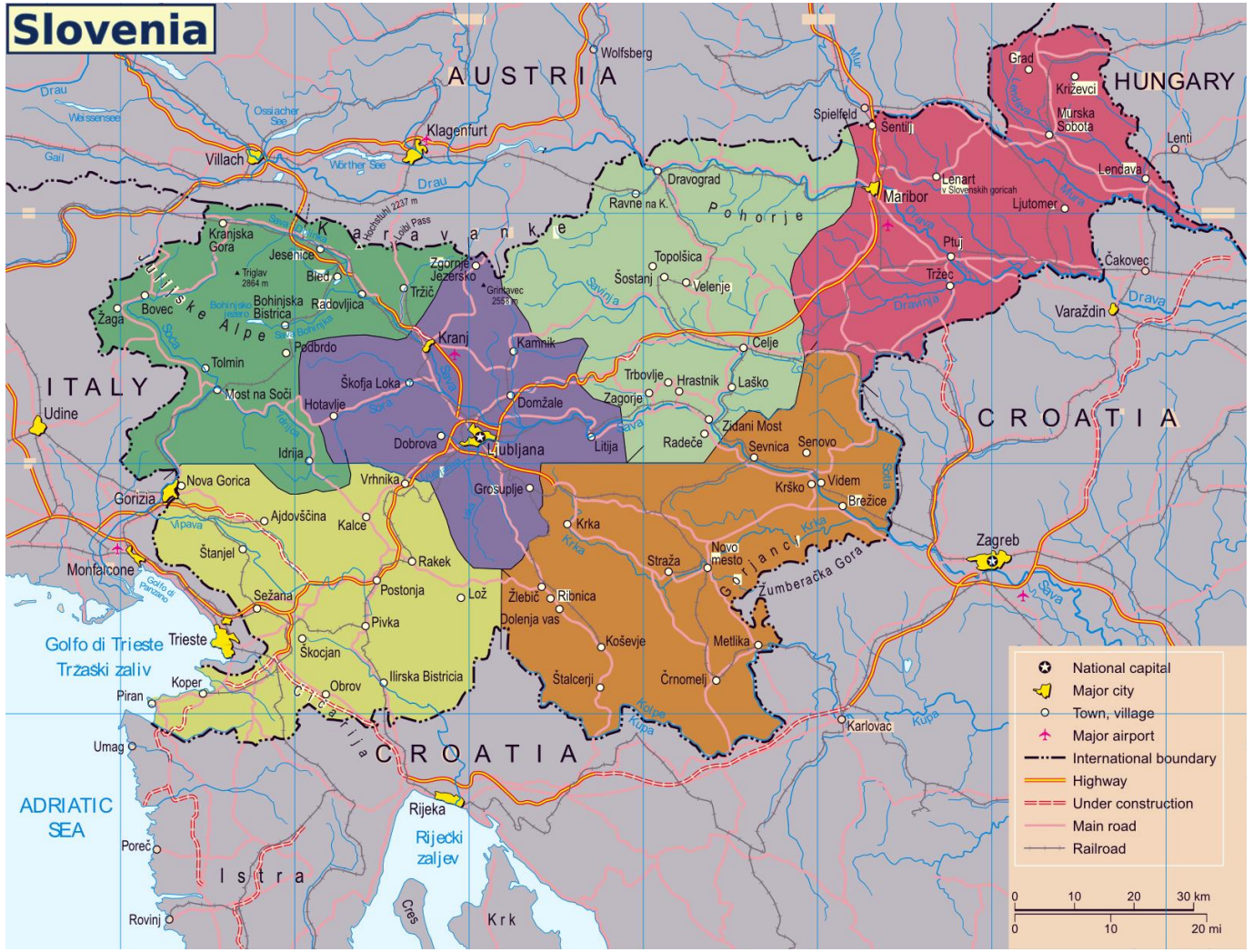
Wordnet as a crowd source for untreated languages, concepts, and data elements

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?

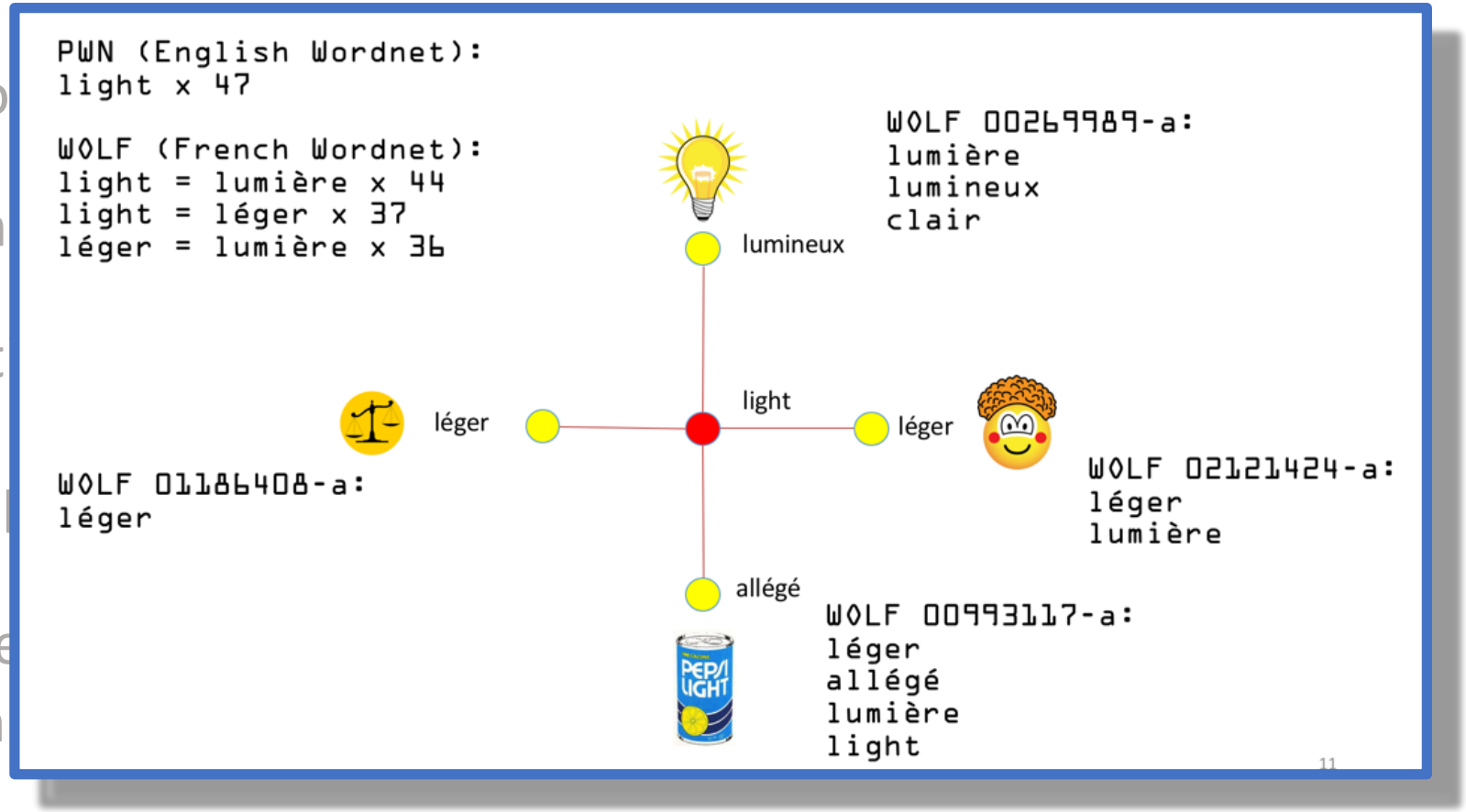
- English as the center of gravity
- Many languages are computationally inferred
- Inconsistent composition of "Wordnets"
- First approximations vs. considered lexicography
- Poor English definitions
- Frozen data
- Canonical forms only, and definitions occasionally
- Extraneous terms (especially names)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



Why *not* Wordnet?

- English as the center of gravity
- Many languages are computati inferred
- Inconsistent composition of "Wordnets"
- First approximations vs. consid lexicography
- Poor English definitions
- Frozen data
- Canonical forms only, and defir occasionally
- Extraneous terms (especially na
- Missing terms from concept se

03485997–n 'the appendage to an object that is designed to be held in order to use or move it';

Search WN

Albanian	*dorezë*
Arabic	مِقْبض
Catalan	*aferrall , nansa , aferrador , agafall , aga*
Chinese (simplified)	*柄 , 把柄 , 把手 , 手柄*
Greek	*λαβή*
English	*handle$_4$ (▻) , grip$_1$ (▻) , hold , handgrip*
Basque	*eskutoki , eskuleku , euskarri , orakarri ,*
Farsi	قبضه شمشیر
Finnish	*sanka , kahva , kädensija , ripa*
French	*poignée , prise , manche*
Galician	*mango , asa , agarra , agarradoira , pica*
Croatian	*ručka , drška , ručica*
Indonesian	*telinga , batang , gagang , pemegang , t*
Italian	*maniglia , presa , impugnatura , manopo*
Japanese	*把っ手 , 把手 , 柄 , 取り所 , ハンドル , つ 分 , 掴み , 手持部分 , 手*
Dutch	*oor , kruk , handel*
Polish	*rączka*
Portuguese	*alça , cabo , trinco , taramela*
Chinese (traditional)	*把*
Romanian	*coadă , mâner*
Slovak	*rúčka , držadlo*
Slovene	*ročica , roč , držaj , prijem , ročaj , držal*
Spanish	*agarradera , asa , asidero*
Thai	*มือจับ , ด้าม , ที่จับ , ด้ามจับ , คัน , หูหิ้ว*
Malaysian	*batang , gagang , hendal , tangkai , pega*

01804414–v 'show and train'; V2;

English	*handle (▻ ▻)*
Finnish	*pitää käsissä , pidellä*
French	*gérer , traiter*
Indonesian	*menekel , membehandel*
Italian	*maneggiare*
Japanese	*操る*
Romanian	*conduce*
Thai	*ฝึก*
Malaysian	*memperlakukan*

Why *not* Wordnet?

00672433–v (50) V1, V2	estimate, guess, judge, gauge, approximate ثمّن, حكم على, قارب, ثمن, كان رأيا, حكم قضائيا, قيّم, قدر, تبأر, فصل, خمن, خَمَّنَ, حزر, قوم, استنتج, قاس, عين سعة شيء ما, ظن, قدّر, حاكم	judge tentatively or form an estimate of (quantities or time)

- Inconsistent composition of "Wordnets"
- First approximations vs. considered lexicography
- Poor English definitions

00672433–v (50) V1, V2	estimate, guess, judge, gauge, approximate 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	judge tentatively or form an estimate of (quantities or time)

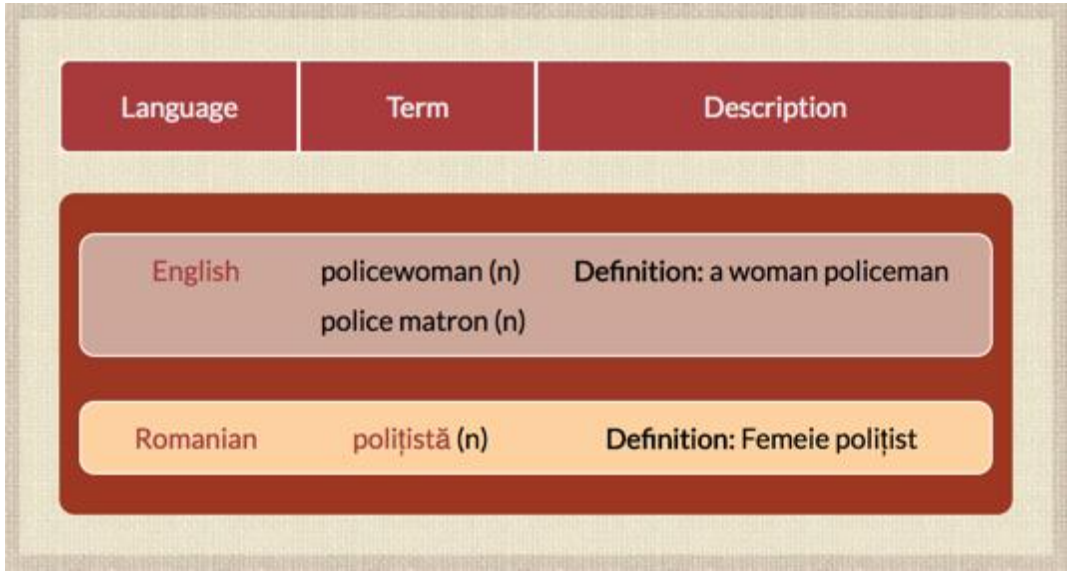
- Missing terms from concept set

1. ثمّن, evaluated; 2. حكم على, judged; 3. قارب, compared; 4. ثمن, price; 5. كان رأيا, had an idea about; 6. حكم قضائيا, verdict; 7. قيّم, evaluated; 8. قدر, considered; 9. تبأر, focused; 10. فصل, separated; 11. خمن, guessed; 12. خَمَّنَ, 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. قدّر, 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)

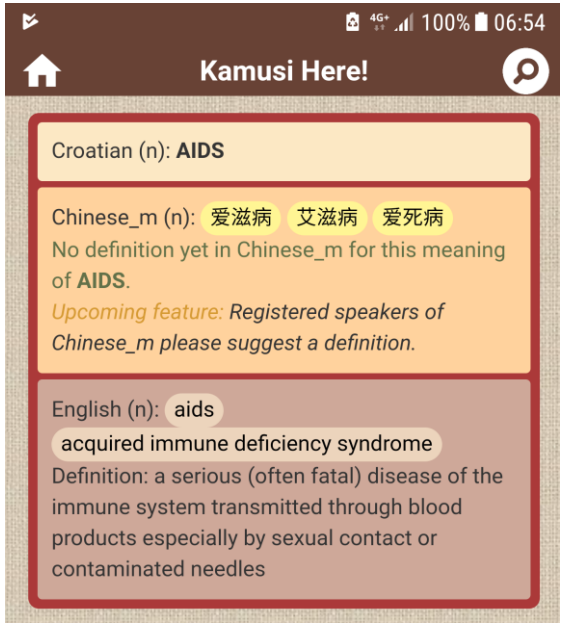
Why *not* Wordnet?

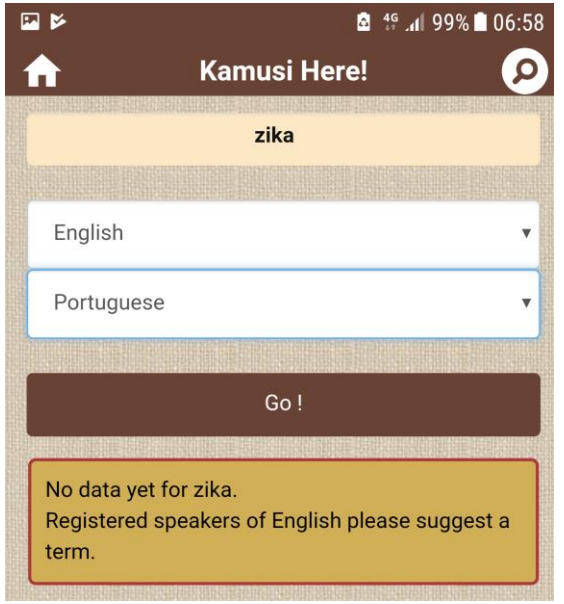
- English as the center of gravity
- Many languages are computationally inferred
- Inconsistent composition of “Wordnets”
- First approximations vs. considered lexicography
- Poor English definitions
- Frozen data
- Canonical forms only, and definitions occasionally
- Extraneous terms (especially names)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



Why *not* Wordnet?

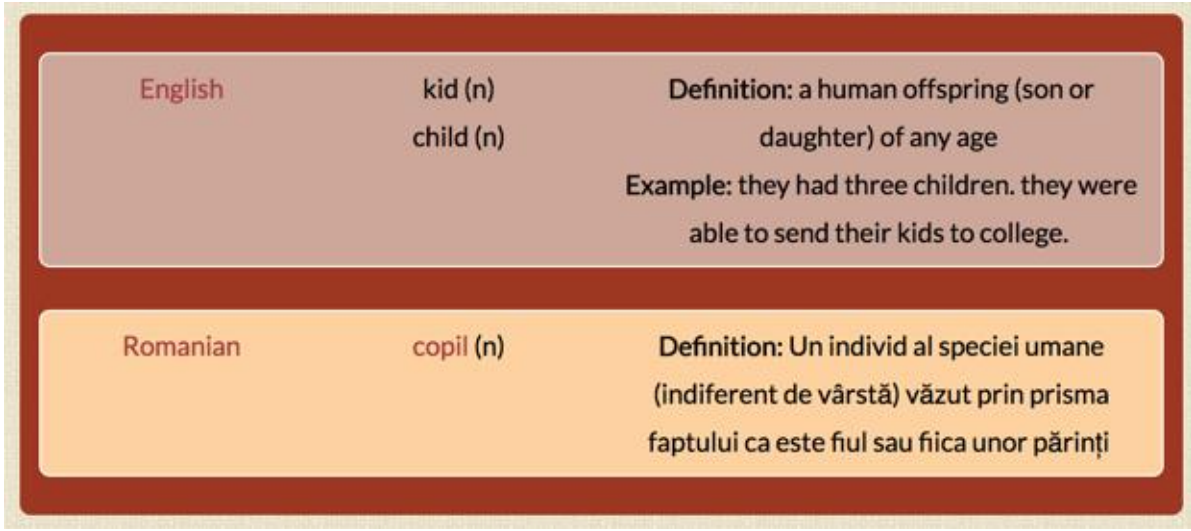
- English as the center of gravity
- Many languages are computationally inferred
- Inconsistent composition of “Wordnets”
- First approximations vs. considered lexicography
- Poor English definitions
- Frozen data
- Canonical forms only, and definitions occasionally
- Extraneous terms (especially names)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





Why *not* Wordnet?

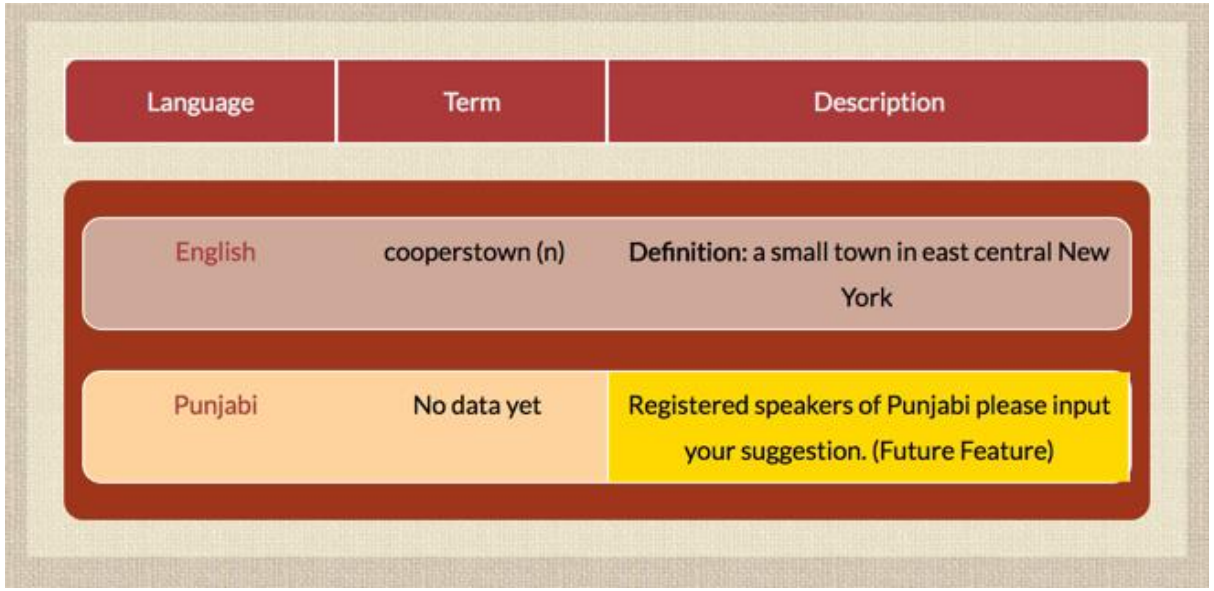
- English as the center of gravity
- Many languages are computationally inferred
- Inconsistent composition of "Wordnets"
- First approximations vs. considered lexicography
- Poor English definitions
- Frozen data
- Canonical forms only, and definitions occasionally
- Extraneous terms (especially names)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



		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)';

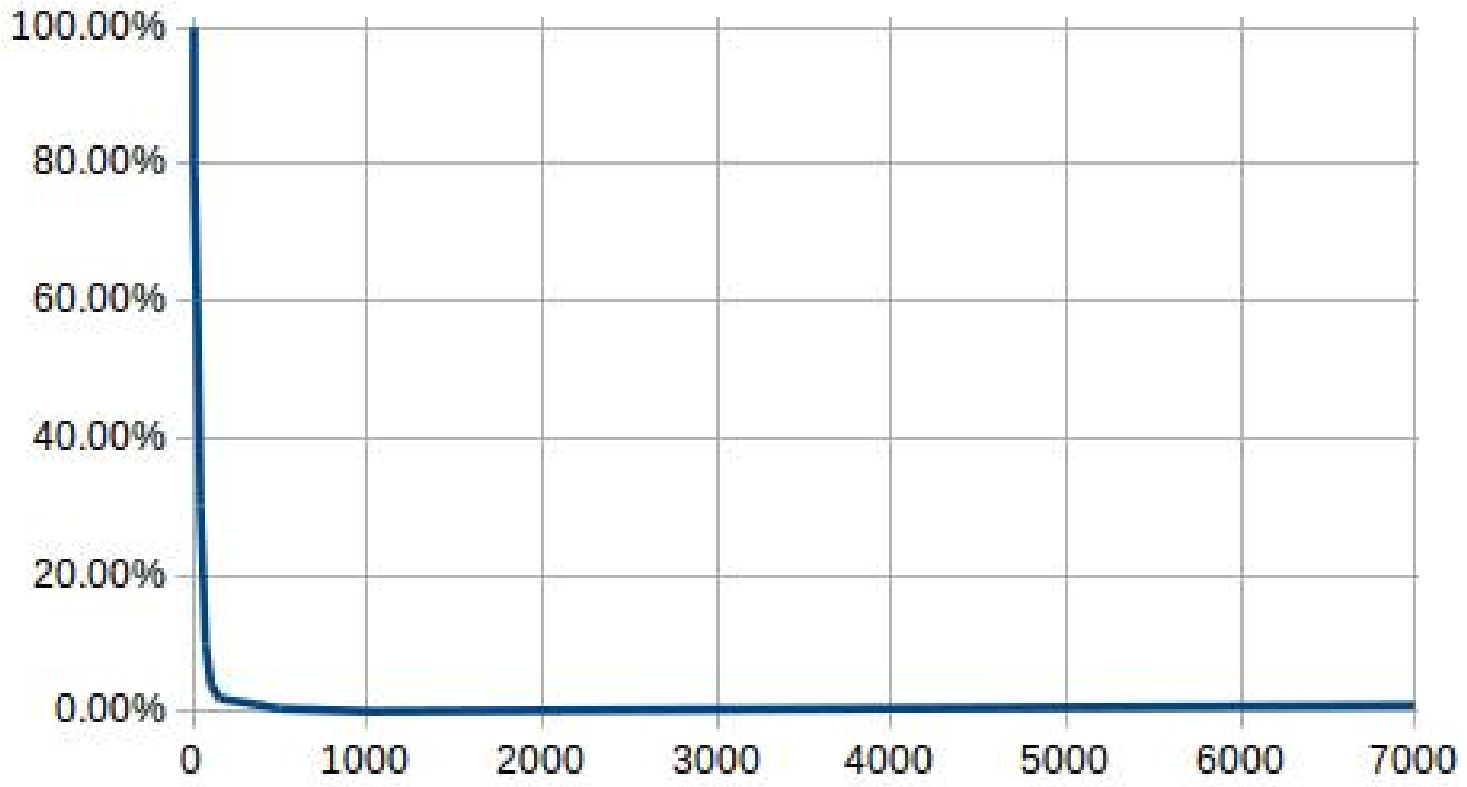
Catalan	*James Cook*
English	*Cook , James Cook , Captain Cook , Captain James Cook*
Finnish	*Cook , kapteeni James Cook , kapteeni Cook , James Cook*
French	*cuisinier , chef , Philippe Lefebvre , cook , cuisinière , James Cook*
Galician	*James Cook*
Indonesian	*Cook , Captain James Cook , Captain Cook , James Cook*
Japanese	*ジェームズ・クック*
Portuguese	*James cook , James Cook*
Slovene	*kuhar , cook , kuharica , James Cook*
Spanish	*James Cook*
Thai	*เจมส์ คุก , กัปตันเจมส์ คุก , กัปตันคุ๊ก*
Malaysian	*Captain James Cook , Captain Cook , James Cook , Cook*

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

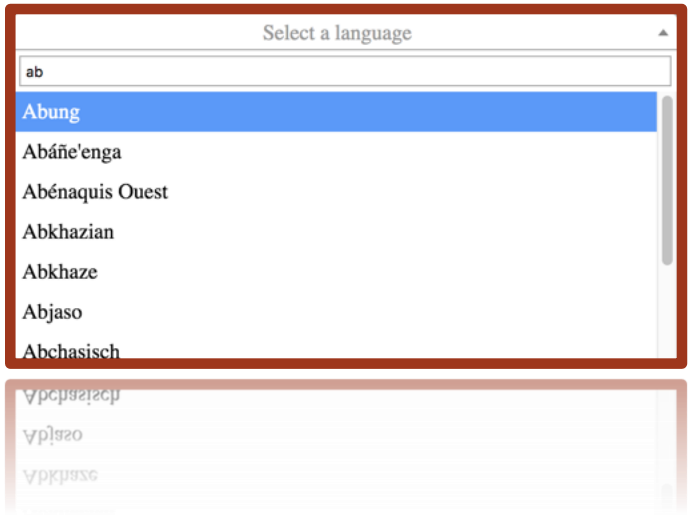
Wordnet and the Crowd

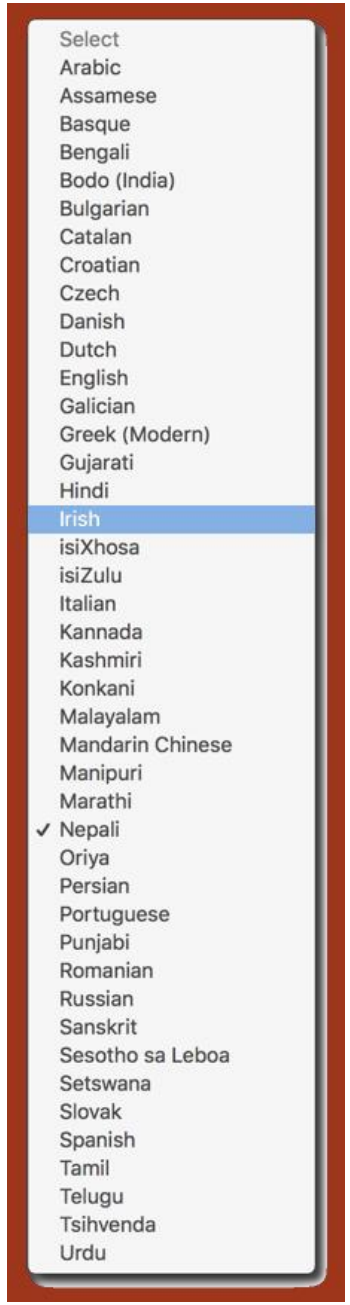
- Untreated languages
- Poorly-treated languages
- Well-treated languages



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
aai	Arifama-Miniafia
aak	Ankave
aal	Afade
aam	Aramanik
aan	Anambé
aao	Arabic (Algerian Sahara)
aap	Arára, Pará
aaq	Abnaki (Eastern)
aar	Afar
aas	Aasáx
aat	Albanian (Arvanitika)
aau	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

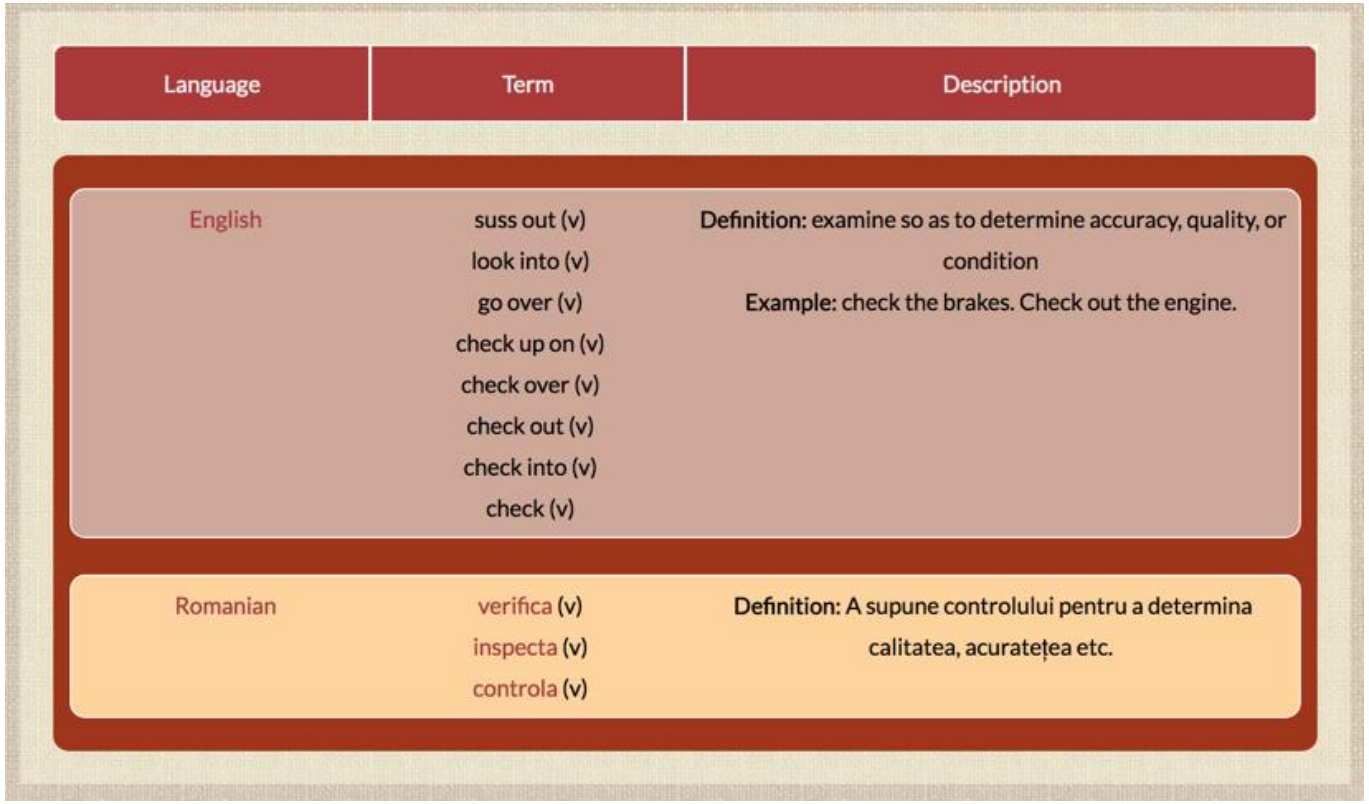
- Untreated languages
- Poorly-treated languages
  - 4,476 (dan) → 117,659 (PWN) = 3.8%
  - Merge with other datasets (DUCKS)
- Well-treated languages

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

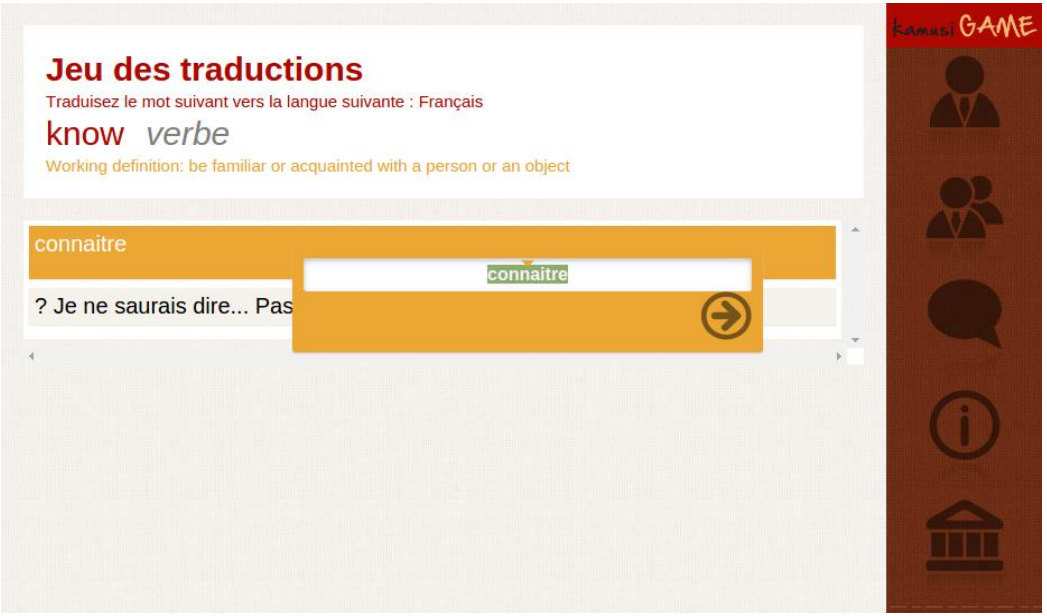
- Warnings
- Terms
- Definitions
- Usage examples
- Inflections
- Alignment
- Grouping and ranking
- Unknown indigenous terms
- Prioritization

What the Crowd can Give

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

What the Crowd can Give

- Warnings
- Terms
- Definitions



What the Crowd can Give

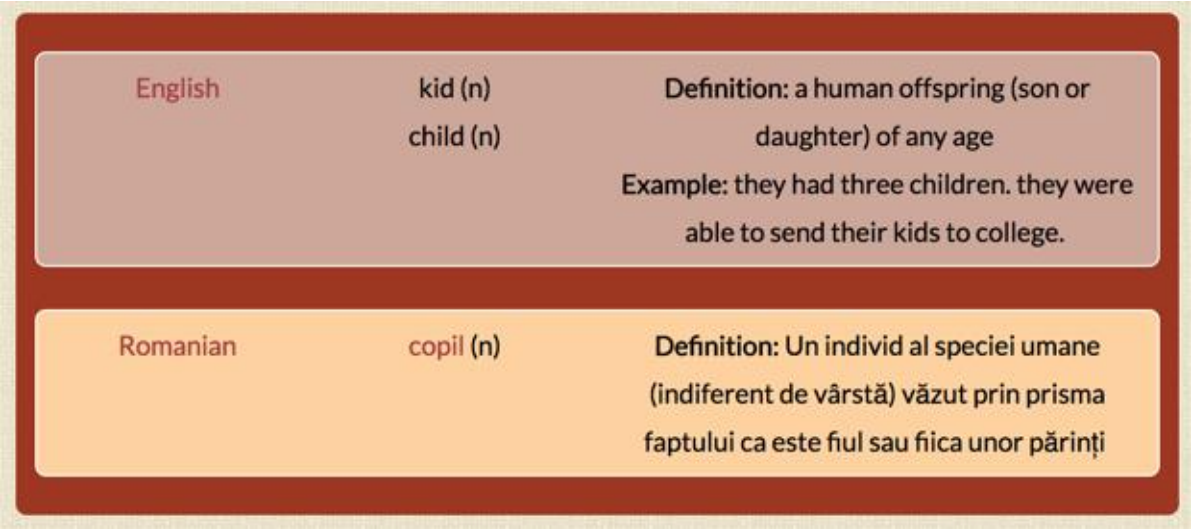
- Warnings
- Terms
- Definitions
- Usage examples
- Inflections
- Alignment
- Grouping and ranking
- Unknown indigenous terms
- Prioritization

What the Crowd can Give

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

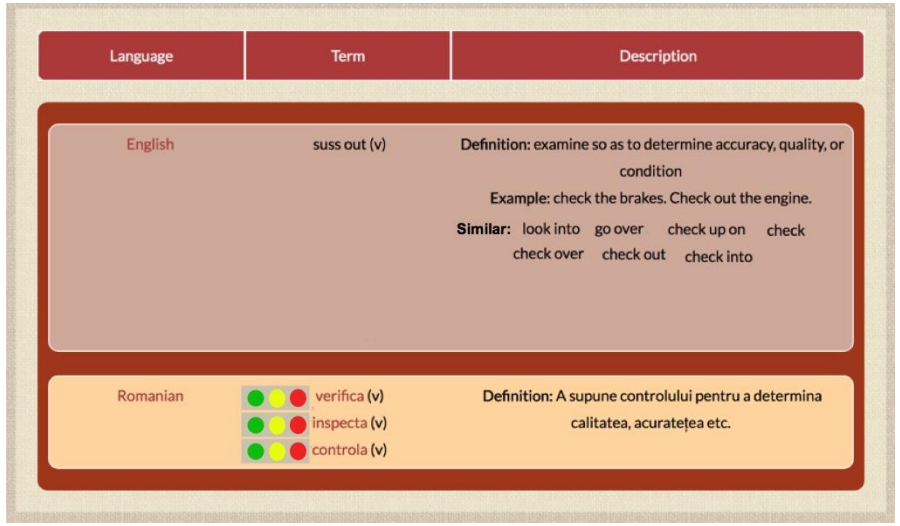
What the Crowd can Give

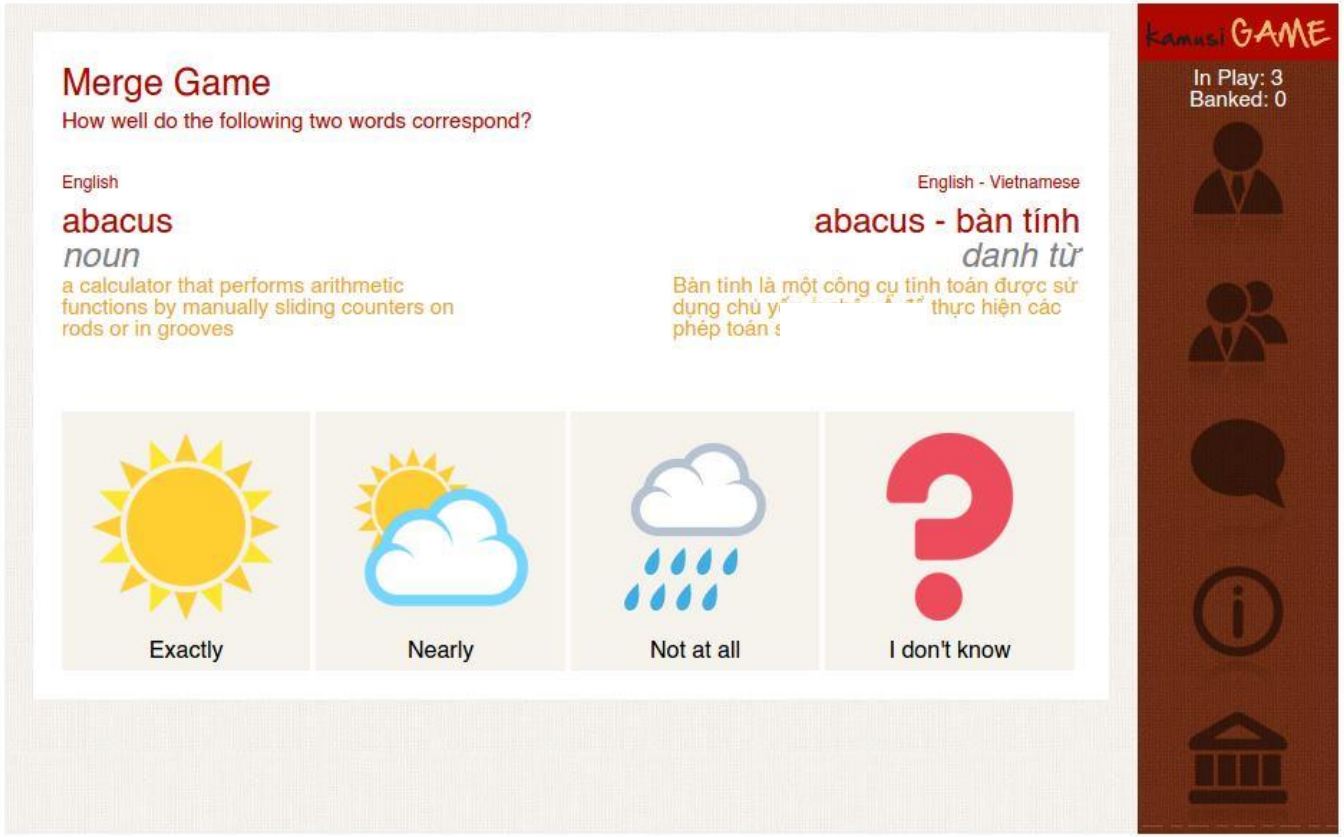
- Warnings
- Terms
- Definitions
- Usage examples
- Inflections
- Alignment
- Grouping and ranking
- Unknown indigenous terms
- Prioritization



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

What the Crowd can Give

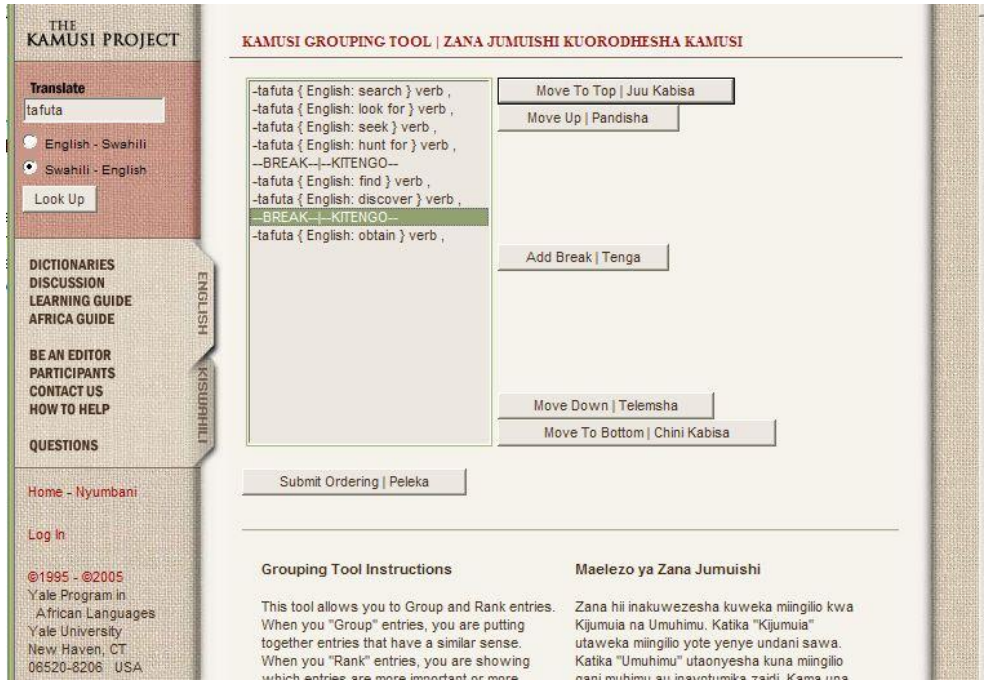


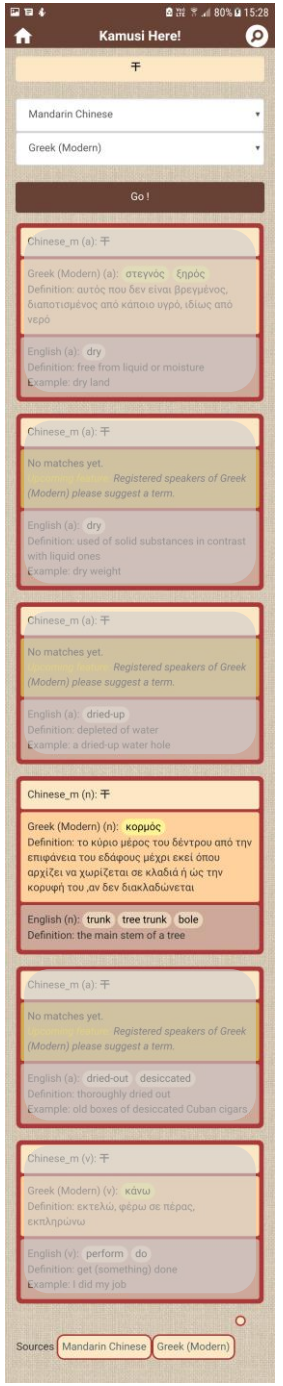


- Alignment
- Grouping and ranking
- Unknown indigenous terms
- Prioritization

What the Crowd can Give

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





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	*fly$_1$* (▻) , *fly ball*
Finnish	*korkea pallo*
French	*mouche , voler*
Japanese	*飛球 , フライ*
Spanish	*pelota elevada , elevado*
Thai	*ฟลายบอล*

02190166–n ✪ 'two–winged insects characterized by active flight'	
Arabic	*ذُبابة*
Bulgarian	*муха*
Catalan	*mosca*
Chinese (simplified)	*蝇 , 苍蝇*
Danish	*flue*
Greek	*μύγα*
English	*fly$_6$* (▻)
Basque	*euli*
Finnish	*kärpänen*
French	*mouche , voler*
Galician	*mosca*
Croatian	*muha*
Indonesian	*lalat*
Icelandic	*fluga*
Italian	*Mosca , mosca*
Japanese	*蝿 , フライ , ハエ*
Dutch	*vlieg*
Nynorsk	*fluge*
Bokmål	*flue*
Polish	*mucha*
Portuguese	*Moscas , mosca , Diptera*
Romanian	*muscă*
Slovene	*muha*
Spanish	*mosca*
Swedish	*fluga*
Thai	*แมงวัน , แมลงวัน*
Malaysian	*lalat*

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