



Martin Benjamin

The Particles of Language:
"The Dictionary" as elemental data for 7000 languages across time and space



kamusi is Swahili for *dictionary*



Goal: A complete matrix of human expression across time and space

- As a knowledge resource
- As a data resource



In service since 1994 (originally at Yale Council on African Studies)

International NGO since 2009

- Registered non-profit in USA and Switzerland

Academic Home since 2013:



EPFL - Swiss Federal Institute of Technology in Lausanne

LSIR - Distributed Systems Information Laboratory



White House Big Data Initiative:

Launch Partner for Building the Data Innovation Ecosystem
Networking and Information Technology R&D Program
Office of Science and Technology Policy

What is the overlap between  and  ?

- Big goals, small particles
- Big collaboration
 - 7000 languages
 - “Human Languages Project”
- Pure science – data for knowledge
- Practical science – data for use
- High energy particle detectors

Problems for Lexicography

What are Concepts?

- How to explain an idea in its own language
- How to express an idea across languages
- How to account for variation



What are Words?

- A set of letters?
- A set of sounds?
- A “canonical” form?
- A single entity?

Problems for Lexicography

What are Concepts?

- How to explain an idea in its own language
- How to express an idea across languages
- How to account for variation



What are Words?

- A set of letters?
- A set of sounds?
- A “canonical” form?
- A single entity?

Problems for Lexicography

What are Concepts?

- How to explain an idea in its own language
- How to express an idea across languages
- How to account for variation



What are Words?

- A set of letters?
- A set of sounds?
- A “canonical” form?
- A single entity?

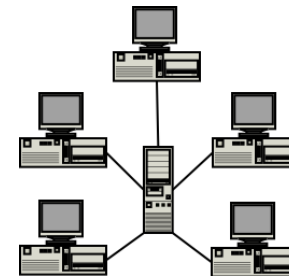
Problems for Lexicography

What are Concepts?

- How to explain an idea in its own language
- How to express an idea across languages
- How to account for variation



C-L-I-E-N-T



What are Words?

- A set of letters?
- A set of sounds?
- A “canonical” form?
- A single entity?



Problems for Lexicography

What are Concepts?

- How to explain an idea in its own language
- How to express an idea across languages
- How to account for variation

whined



wind



wined



What are Words?

- A set of letters?
- A set of sounds?
- A “canonical” form?
- A single entity?

Problems for Lexicography

What are Concepts?

- How to explain an idea in its own language
- How to express an idea across languages
- How to account for variation

SEE

sees

saw

seen

seeing

Kinyarwanda

900 million forms for every verb

What are Words?

- A set of letters?
- A set of sounds?
- A “canonical” form?
- A single entity?

Problems for Lexicography

What are Concepts?

- How to explain an idea in its own language
- How to express an idea across languages
- How to account for variation

African fish eagle



What are Words?

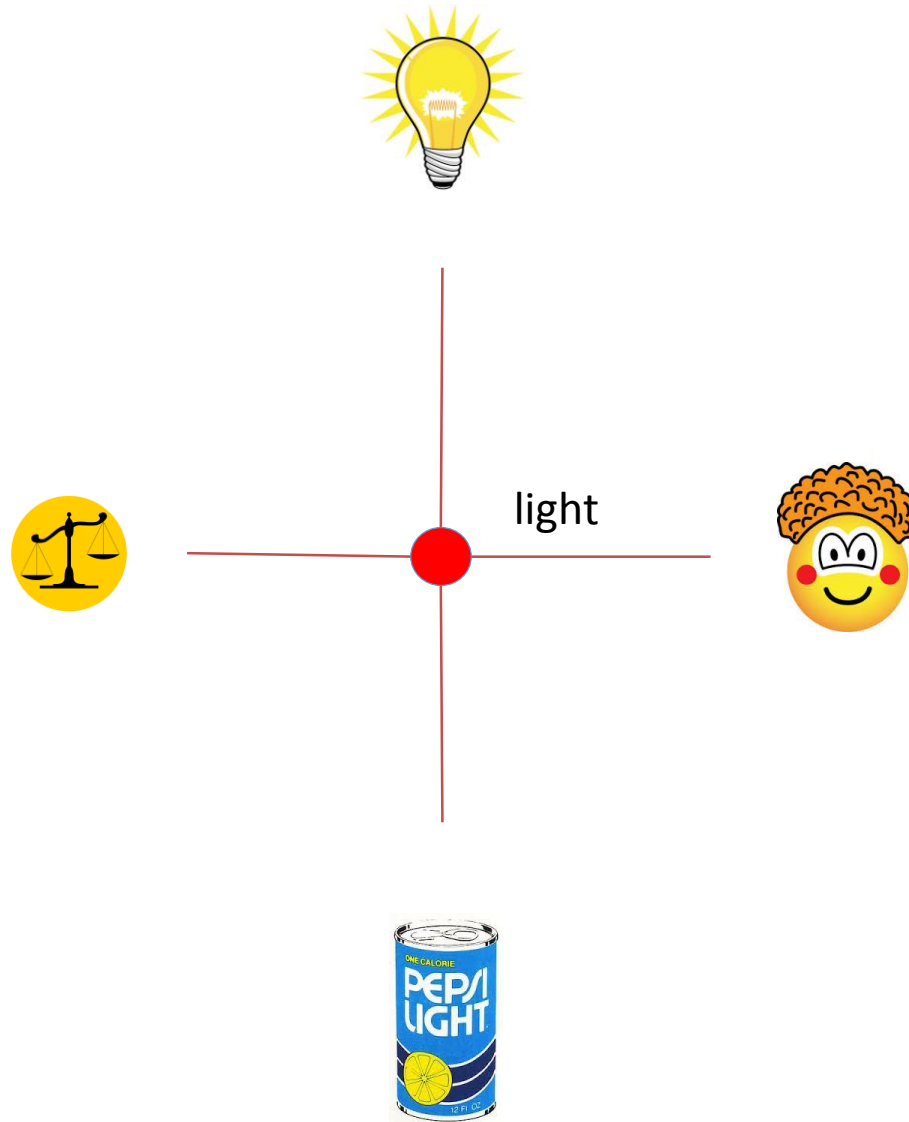
- A set of letters?
- A set of sounds?
- A “canonical” form?
- A single entity?



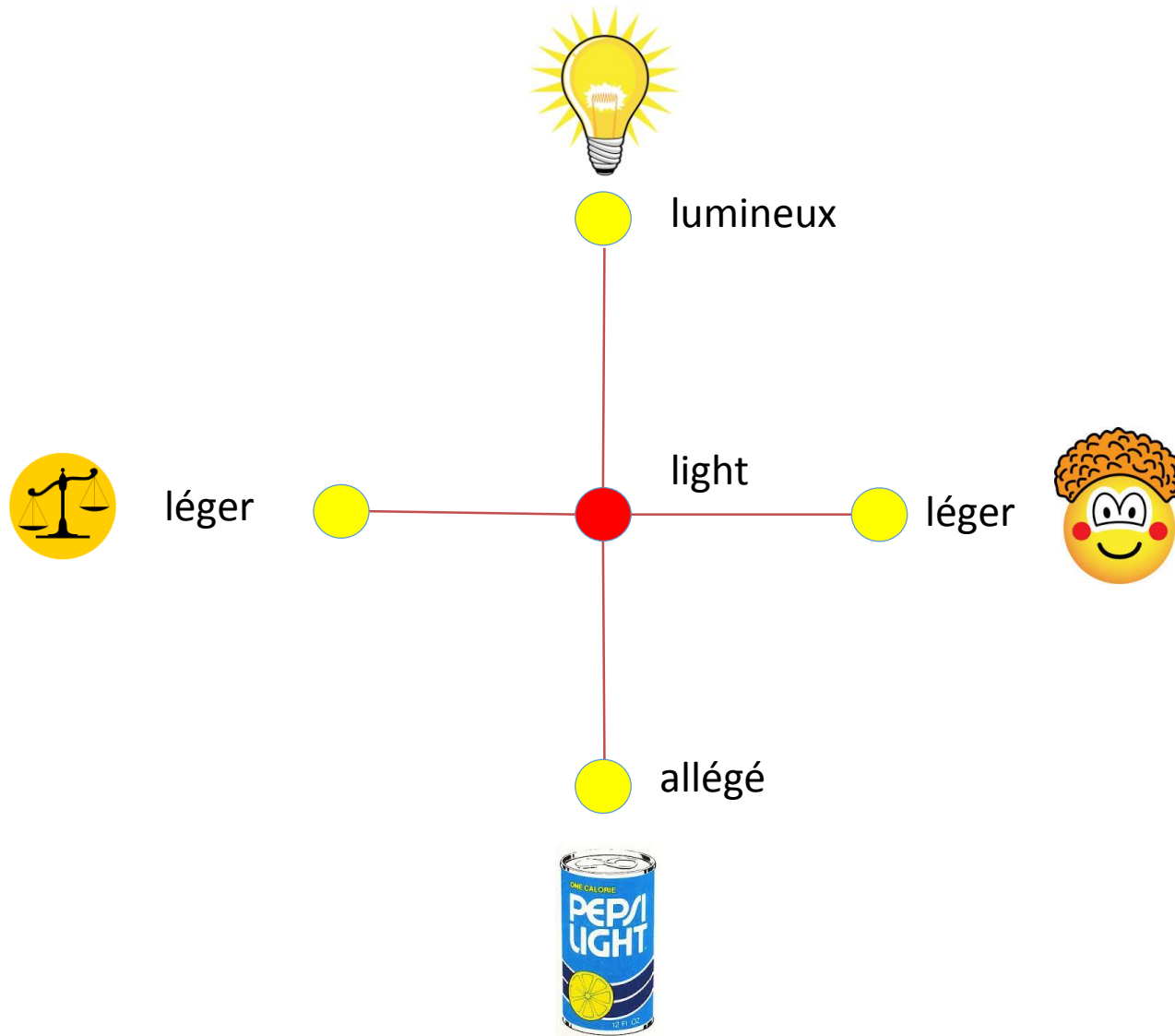
drive up the wall



light



why multilingual dictionaries were impossible

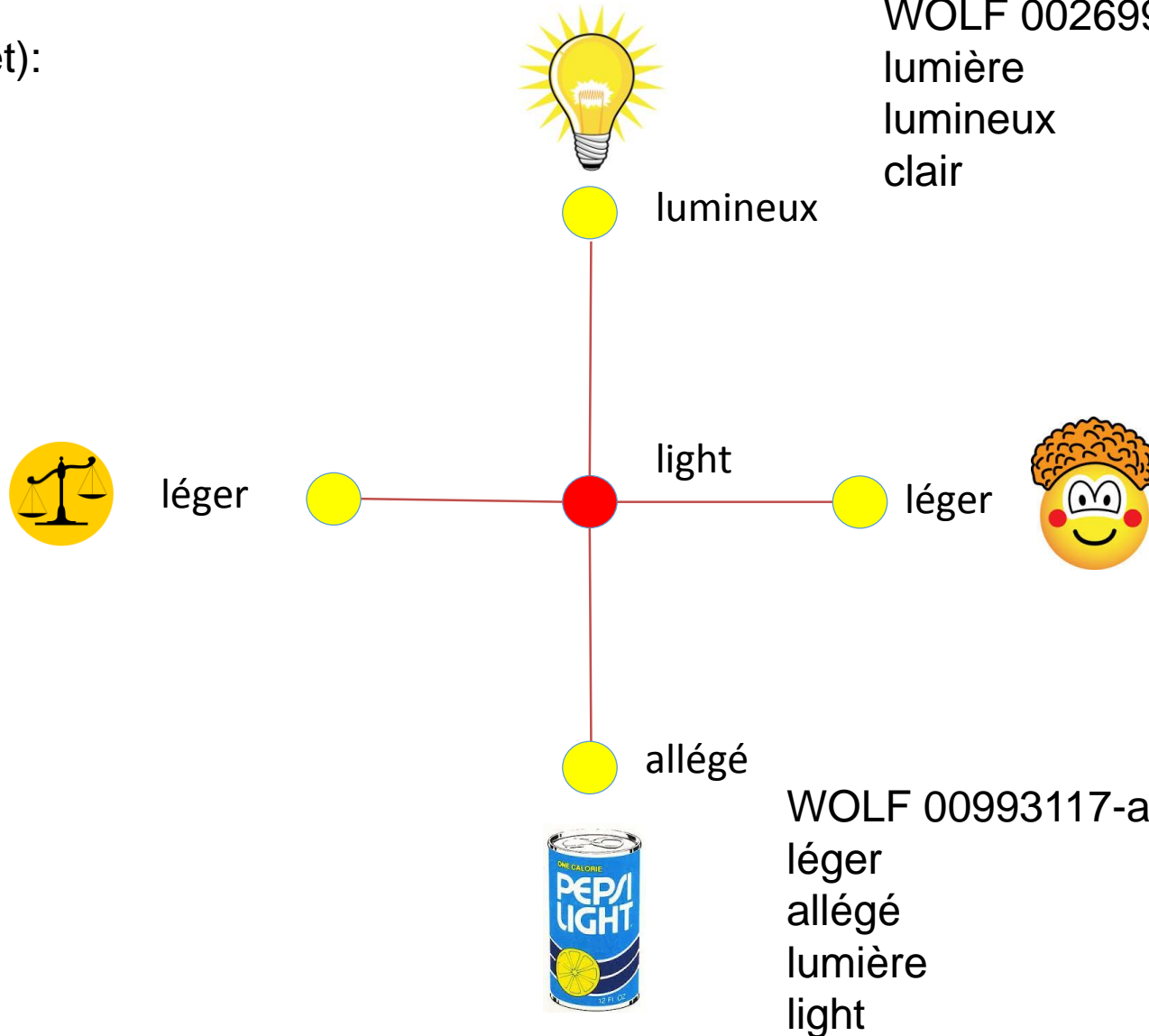


why multilingual dictionaries were impossible

PWN (English Wordnet):
light x 47

WOLF (French Wordnet):
light = lumière x 44
light = léger x 37

WOLF 00269989-a:
lumière
lumineux
clair

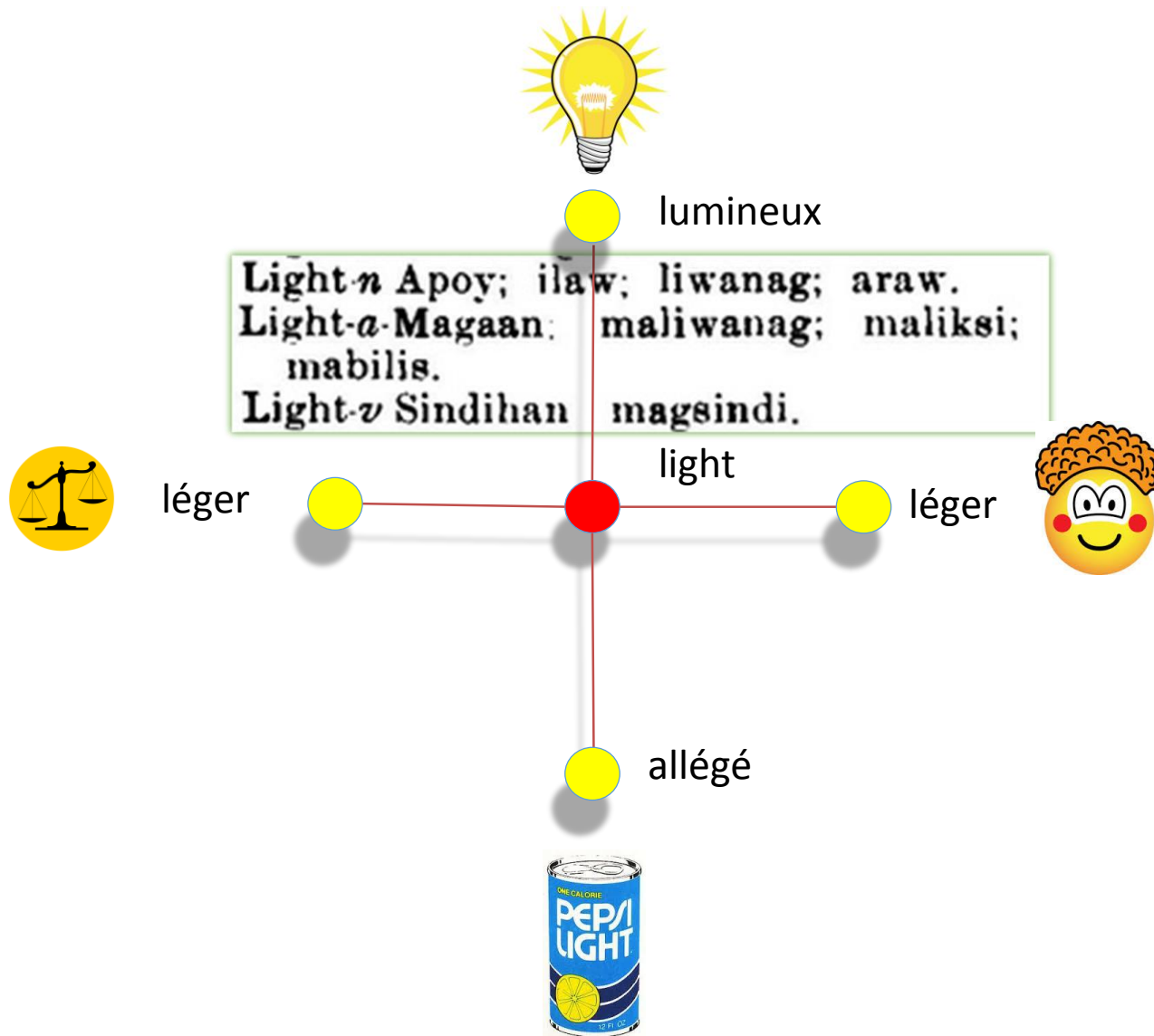


WOLF 01186408-a:
léger

WOLF 02121424-a:
léger
lumière

WOLF 00993117-a:
léger
allégé
lumière
light

why multilingual dictionaries were impossible



why multilingual dictionaries were impossible



light¹ *adj* 1 (of colour) -siokoza, -sioiva angazi, -a ~ brown kahawia isioiva, hafifu. 2 (of a place) -enye mwanga. ~ **coloured** *adj* -enye rangi isioiva. *n* 1 nuru, mwanga *the* ~ *begins to fail* mwanga unaanza kufifia *day* ~ mchana. **in a good/bad** ~ (of picture etc) -a kuonekana vizuri/vibaya; (*fig*) eleweka vizuri/vibaya. **see the** ~ (*liter or rhet*) zaliwa; baini; tangazwa; tambua; -okoka. **be/stand in one's** ~ kinga nuru; (*fig*) zuia mtu anikio/maendeleo kwake. **stand in one's own** ~ zuia kazi yako isionekane; fanya kinyume na matakwa yako. ~ **year** *n* (*astron*) kipimo cha umbali kati ya nyota. 2 taa. ~ **s out** muda wa kuzima taa. **the northern/southern** ~ *n* miali ya mwanga katika ncha za kaskazini na kusini. 3 mwako wa moto; kiberiti *strike a* ~ washa moto; washa kiberiti. 4 uchangamfu (usoni mwa mtu). **the** ~ **of somebody's countenance** (*biblical*) kupendezwa kwake. 5





lumineux

ગુજરાતીલેક્સિકોન.કોમ
Ratilal Chandara's Gujarati Language Resources

Gujaratilexicon » Dictionary » English To Gujarati

Dictionary | Opposites | Thesaurus | Idioms | Proverbs | Phrases

Dictionary

English to Gujarati | Gujarati to Gujarati | Gujarati to English | Hindi to Gujarati

light SUBMIT SUGGEST A WORD

Exact Phrase Like

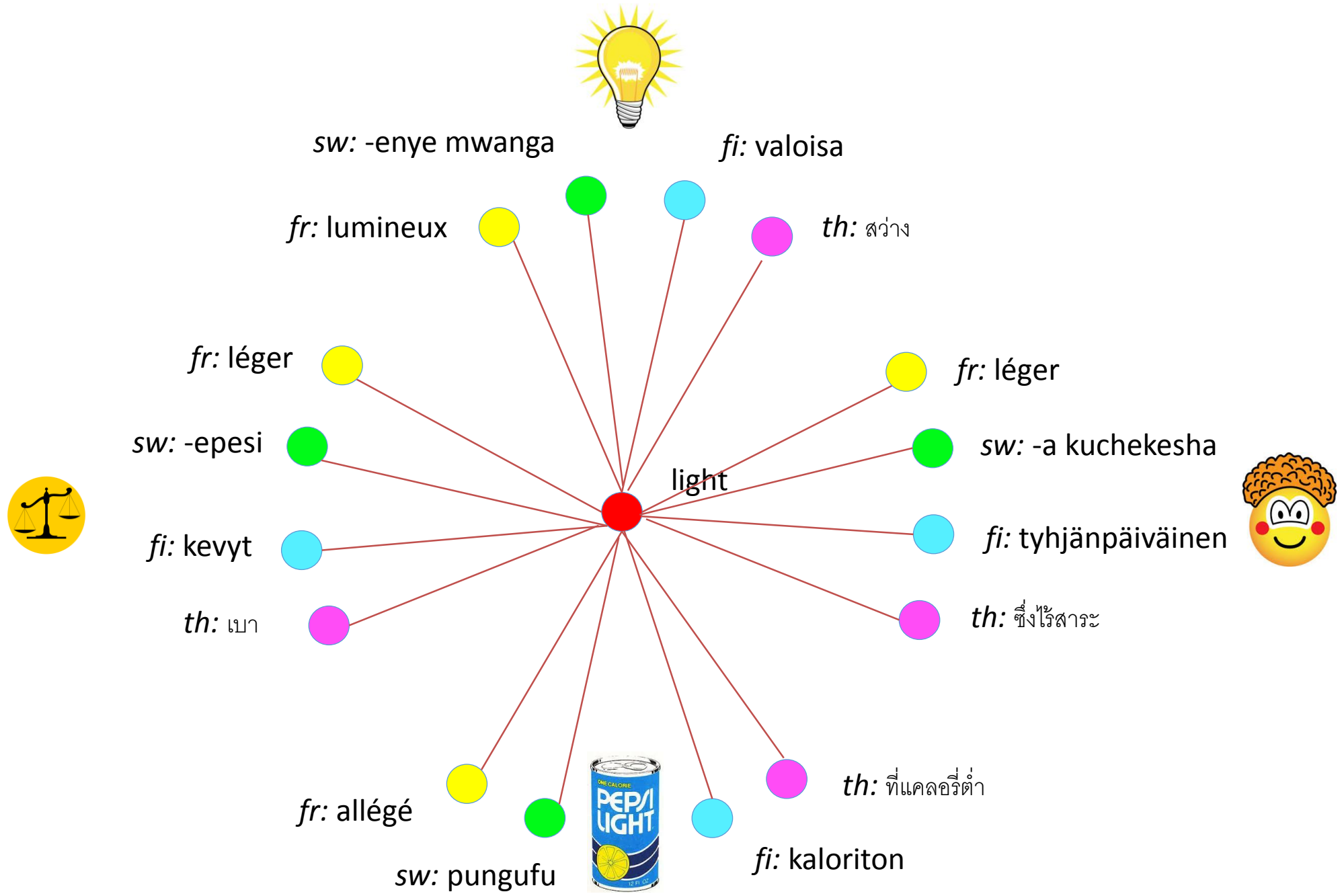
A
Word List

light

No.	Type	Pronunciation	Meaning
1	-લો	લાઈટ	તેજ, પ્રકાશ, અજવાળું, જ્યોતિ, દીપ, દીવો, અવસ્થાવર નિર્મલક દીવો, જેના વડે વસ્તુ દુષ્કમાન થાય છે તે સાધન-પ્રકાર, દેવતા સજાવાવવાની દીવાસળી કે કાકરો, તેજસ્વિતા, આંખનું તેજ, પ્રકાશનું ઘસોઈ ઉદ્ભવસ્થાન સૂર્ય, મીઠાખતી, ઈ, જેનું ચમક, કચાકની ઊજાળી ખાજુ, હાકિગોણ, ફાકિ, ઓપ, સાન, કોઈ ખાખતને સ્પષ્ટ કરનાર નવું સાન, માસિની ઈ, અંખડું-નારિ, (ફેલ અંગે) કીકું, ચાંચું, જાંબું, (ભાવ ઈ.) સજાવાવવું, પ્રકાશવું, પેટવું, ખાખતું, દીવાથી સ્ત્રનો ખતાવવો, -ને પ્રકાશ આપવો, પ્રસાર થવું કે કરવું, ઉદાસિત થવું કે કરવું



why multilingual dictionaries were impossible



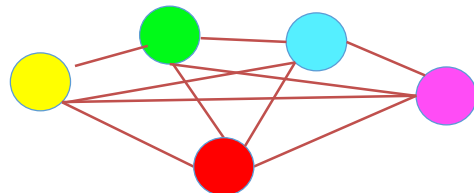
why multilingual dictionaries were impossible



sw: -enye mwanga

fi: valoisa

fr: lumineux



th: สว่าง

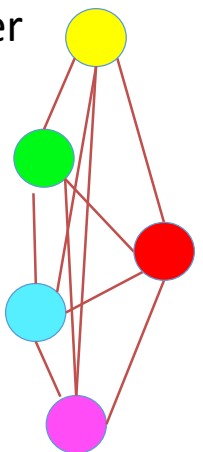
en: light

fr: léger

sw: -epesi

fi: kevyt

th: เบา



en: light

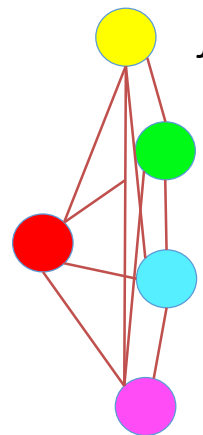
fr: léger

sw: -a kuchekesha

fi: tyhjämpäiväinen

th: ว่างไร้สาระ

en: light



fr: allégé

sw: pungufu

en: light

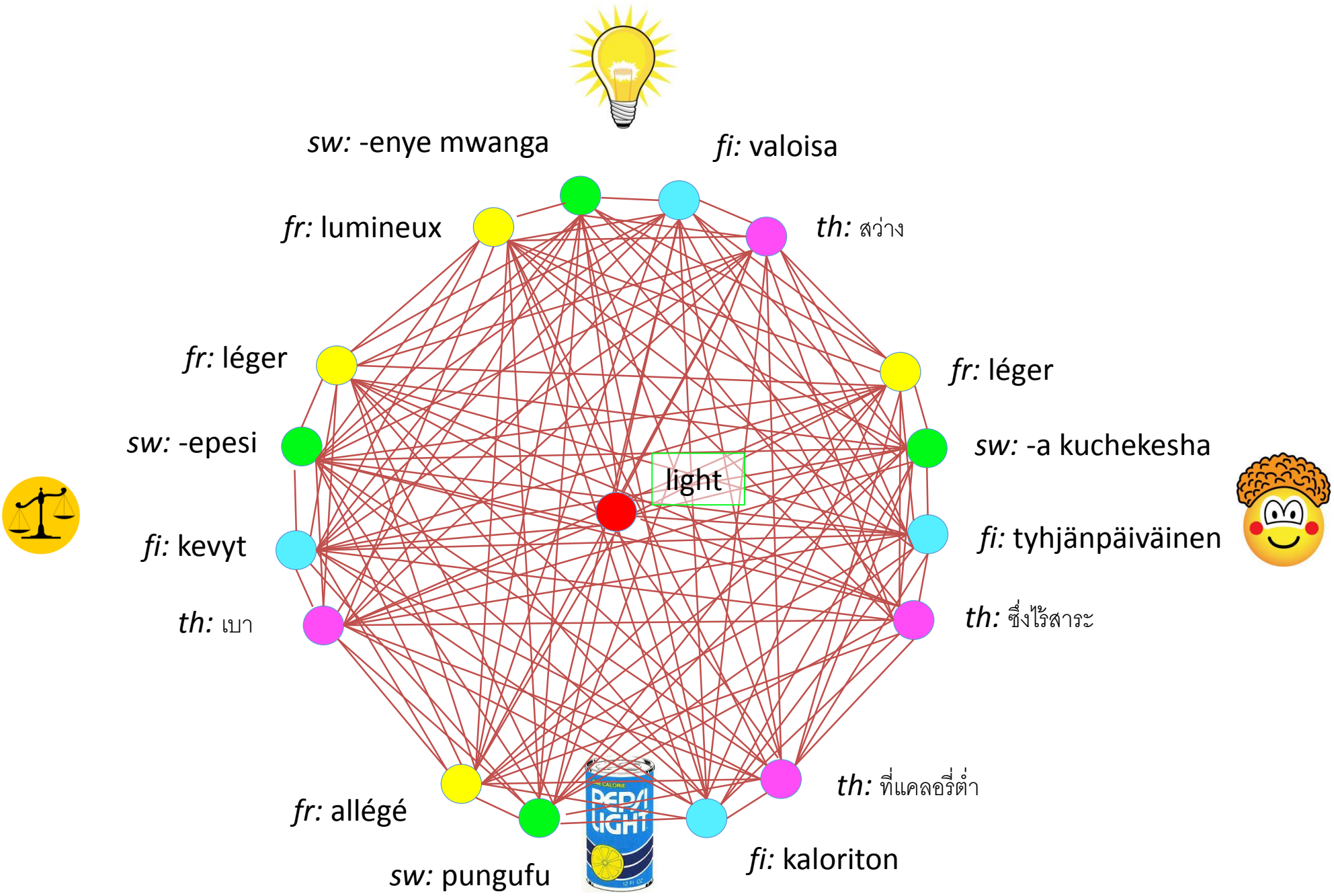
th: ที่แคลอรีต่ำ



fi: kaloriton



why multilingual dictionaries were impossible



why multilingual dictionaries were impossible



why multilingual dictionaries were impossible

● light



light (not dark)



light (not heavy)



light (not serious)



light (not fattening)



light (not dark) *fr: lumineux*



light (not heavy) *fr: léger*



light (not serious) *fr: léger*



light (not fattening) *fr: allégé*





light (not dark)

fr: lumineux

sw: -enye mwanga

fi: valoisa

th: สว่าง



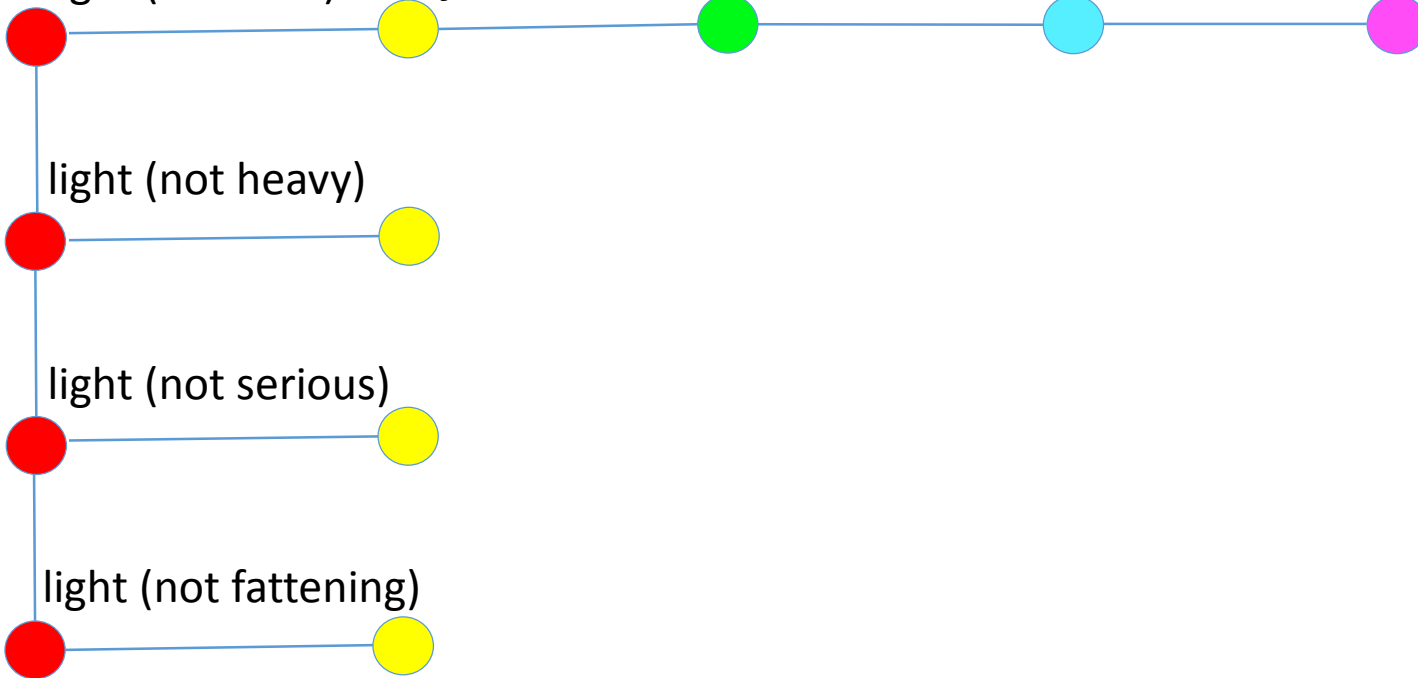
light (not heavy)



light (not serious)



light (not fattening)





light (not dark)



light (not heavy)

fr: léger



sw: -epesi



fi: kevyt



th: เบา



light (not serious)



light (not fattening)



how Kamusi makes a multilingual dictionary possible



light (not dark)



light (not heavy)



light (not serious) *fr:* léger *sw:* -a kuchekecha *fi:* tyhjänpäiväinen *th:* شيءไรสาระ



light (not fattening)





light (not dark)



light (not heavy)



light (not serious)



light (not fattening) *fr:* allégé



sw: pungufu



fi: kaloriton



th: ทีแคลอรีต่ำ





light (not dark) *fr:* lumineux *sw:* -enye mwanga *fi:* valoisa *th:* สว่าง



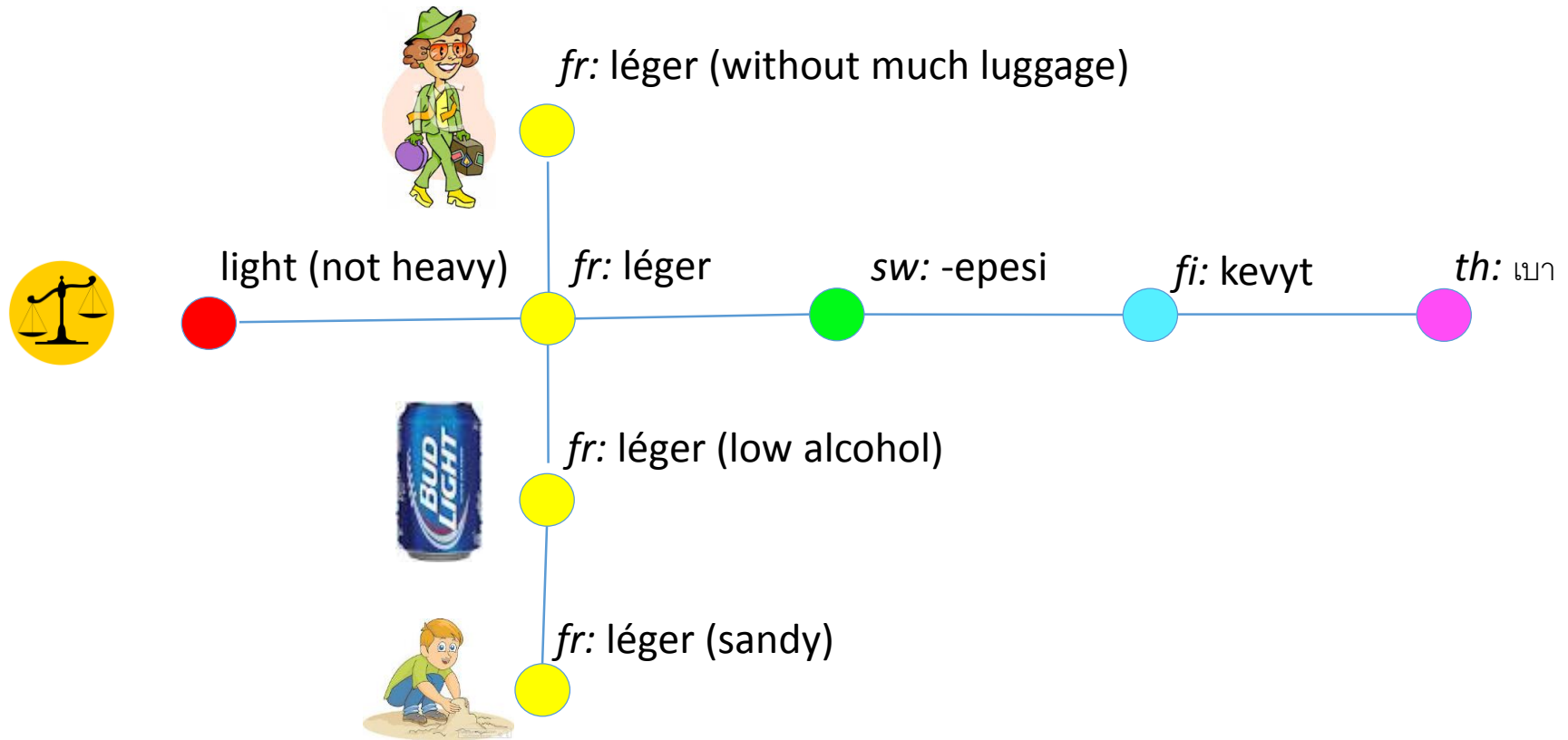
light (not heavy) *fr:* léger *sw:* -epesi *fi:* kevyt *th:* เบา



light (not serious) *fr:* léger *sw:* -a kuchekesha *fi:* tyhjänpäiväinen *th:* شيءไร้สาระ



light (not fattening) *fr:* allégé *sw:* pungufu *fi:* kaloriton *th:* ที่แคลอรีต่ำ





light (not dark)



light (not heavy)



light (not serious)



light (not fattening)





light (not dark)



light (not heavy)



light (not serious)



light (not fattening)





light (not dark)



light (not heavy)



light (not serious)



light (not fattening)





light (not dark)



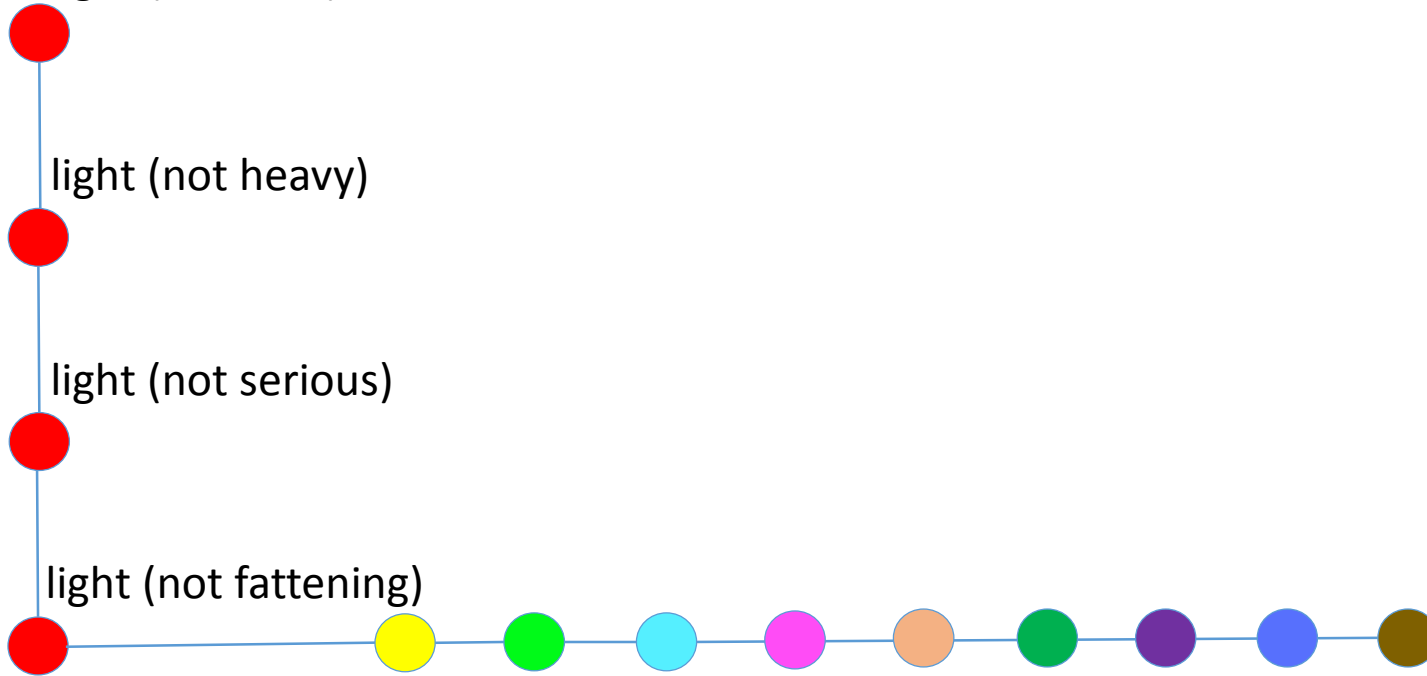
light (not heavy)



light (not serious)



light (not fattening)





light (not dark)



light (not heavy)



light (not serious)



light (not fattening)



how Kamusi makes a multilingual dictionary possible



light (not dark)

fr. lumineux

sw: -enye mwanga

fi: valoisa

th: สว่าง



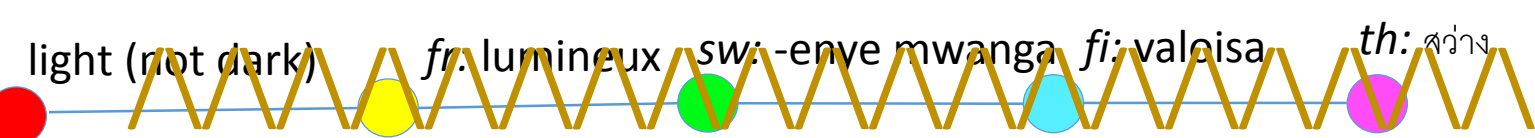
light (not heavy)

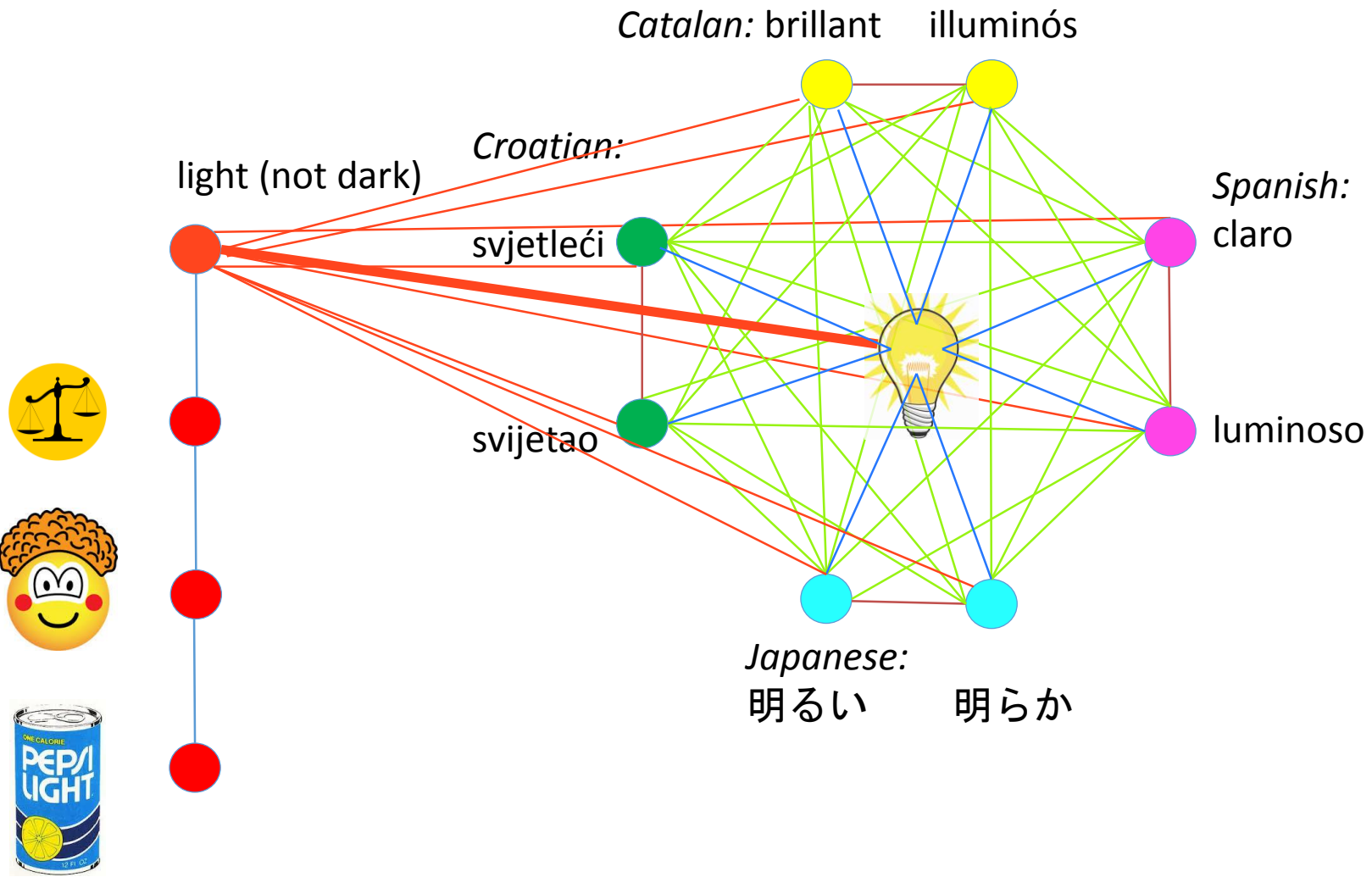


light (not serious)

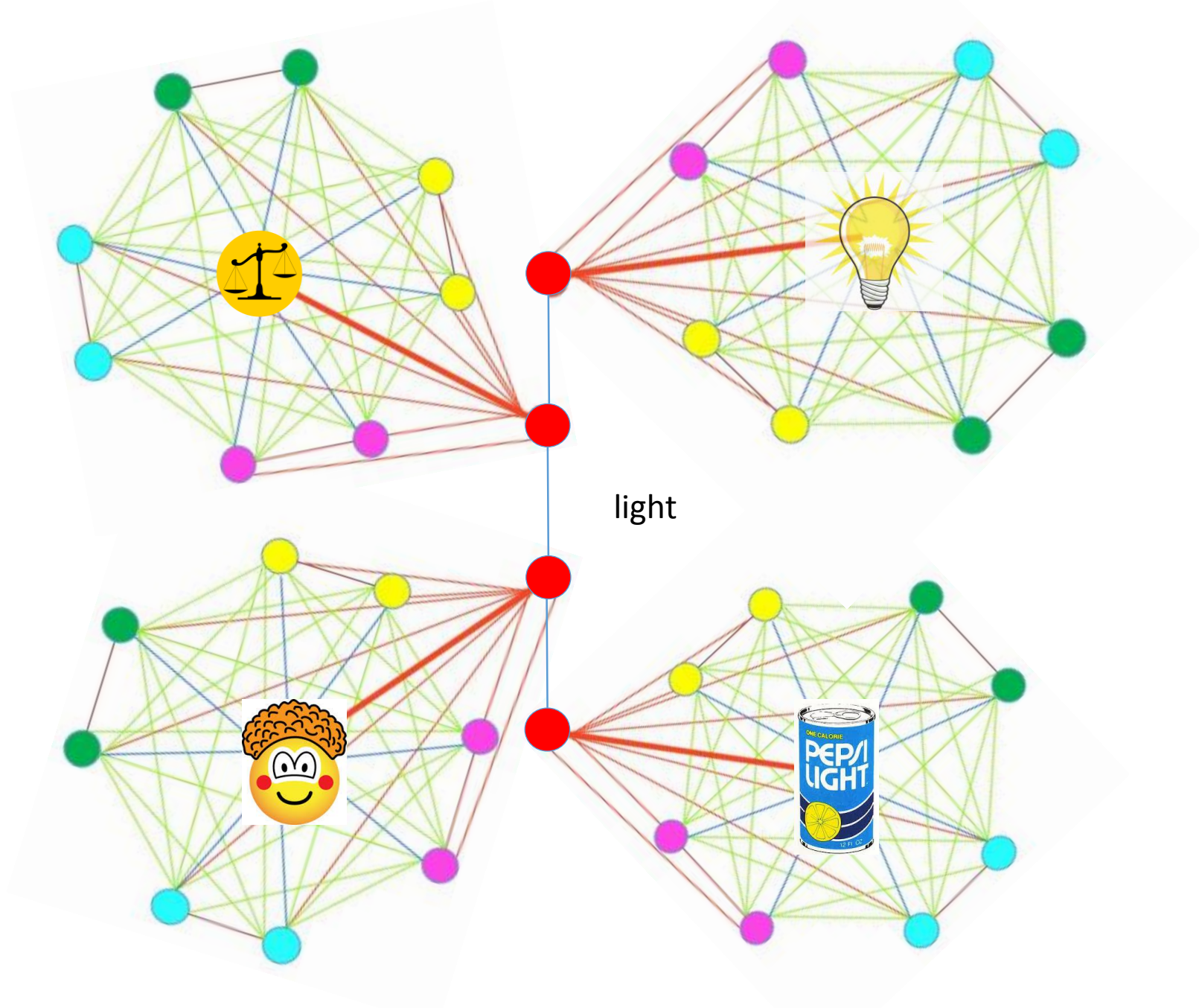


light (not fattening)





how Kamusi makes a multilingual dictionary possible



how Kamusi makes a multilingual dictionary possible

بارش

Urdu noun (Urdu) اسم

Definition: وہ پانی جو آسمان میں بادلوں سے گرتا ہے اور زمین تک قطروں کی شکل میں پہنچتا ہے۔

plural (Urdu) جمع

بارشیں

Translations

1° (verified)

rain (English)

2° (predicted: high)

invura (Kirundi)

invura (Yeyi)

embura (Gusii)

ploaie (Romanian)

あめ (Japanese 日本語)

右

[View in context](#) [Add an example](#) [Add a translation](#)

Mandarin (Chinese) 中文 adjective (Chinese) 形容词

Definition: 面朝北时东边的方向，与“左”相对

Examples: 右边有个人。

右

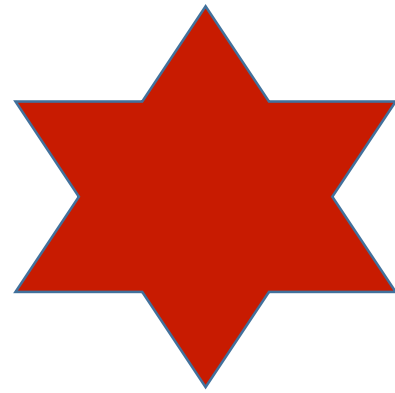
- 1° right (English) → 2° iburyo (Kirundi)
- 2° -tunganya (Kirundi)
- 2° kulila (Hehe)
- 2° dekstra (Esperanto)
- 2° derecho (Spanish)
- 2° droit (French)
- 2° gumo (Songhay) → 3° dreapta (Romanian)
- 2° kulia (Swahili)
- 2° ñaamo (Pulaar)
- 2° prawy (Polish)
- 2° -ddyo (Luganda)
- 2° -ink'amu (Yeyi)
- 2° みぎ (Japanese 日本語)



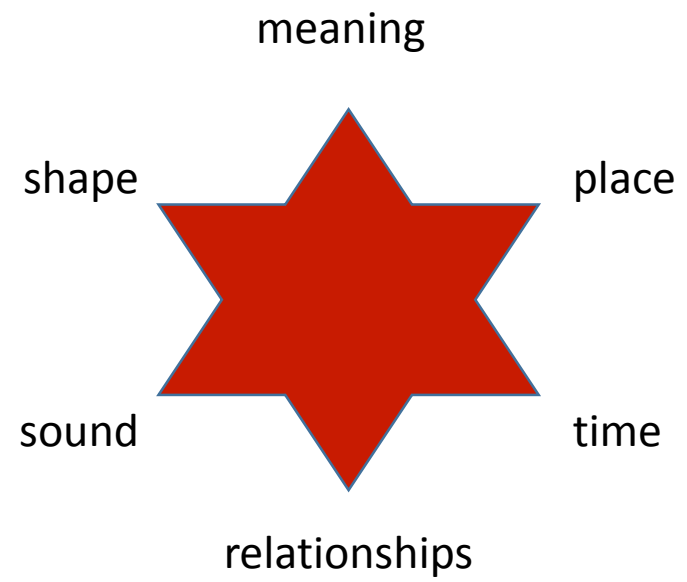
light



light



light



light



meaning

shape

place

sound

time

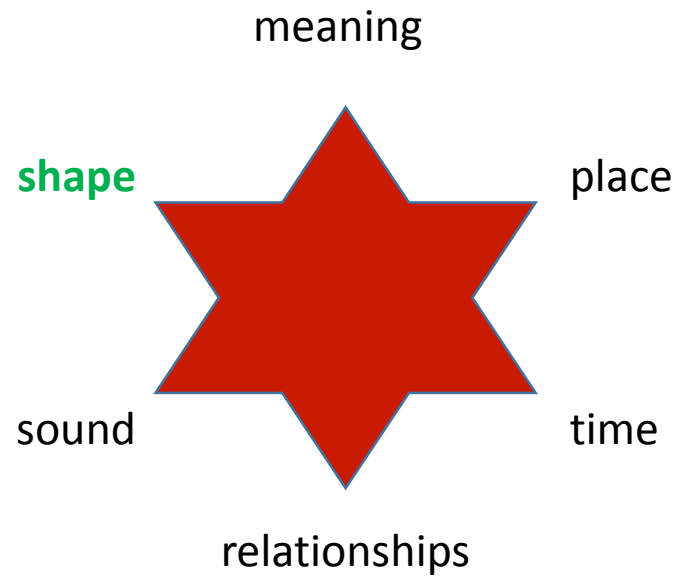
relationships

light

lights

lighted
lit

lighting



light

lighter

lightest

light



robot



meaning

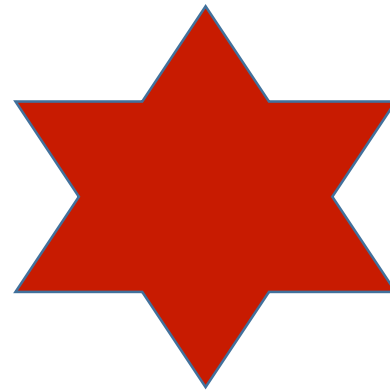
shape

place

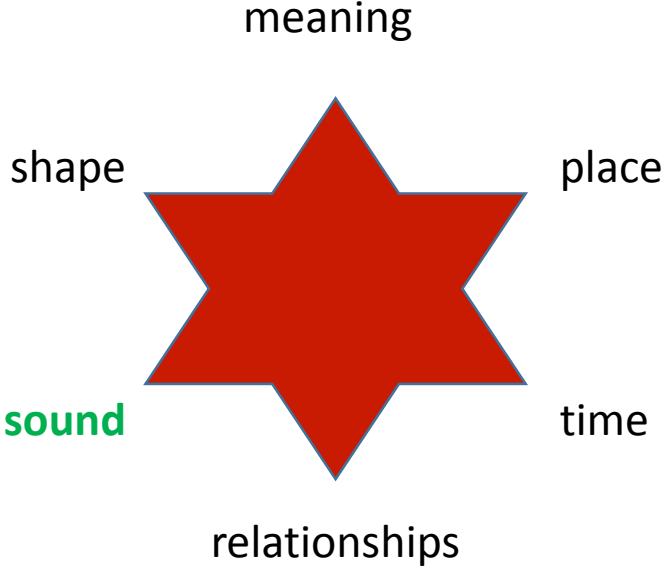
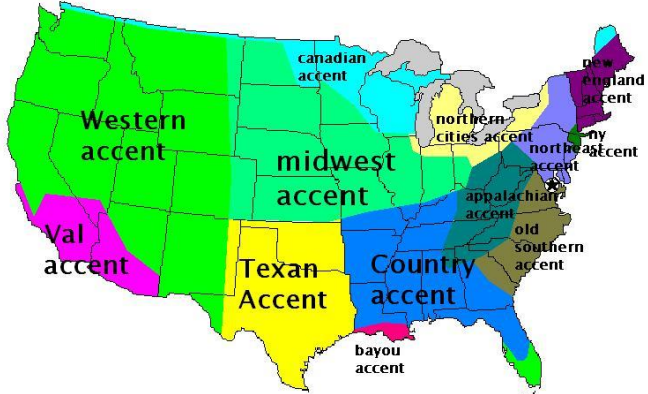
sound

time

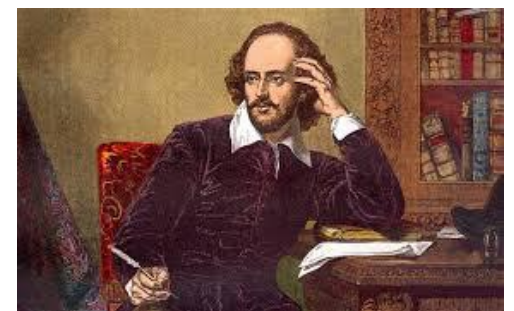
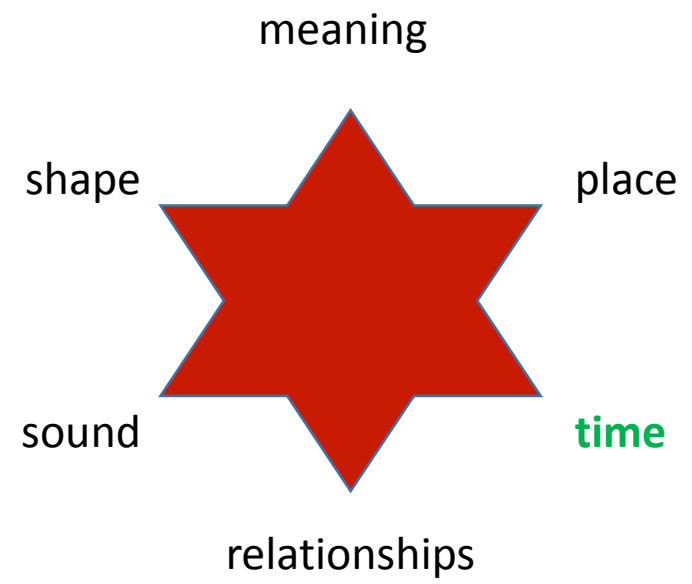
relationships



light



light



linhtaz

light



lighthouse
(spawn)

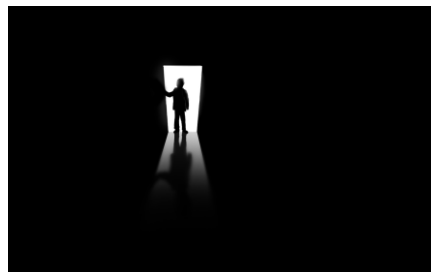
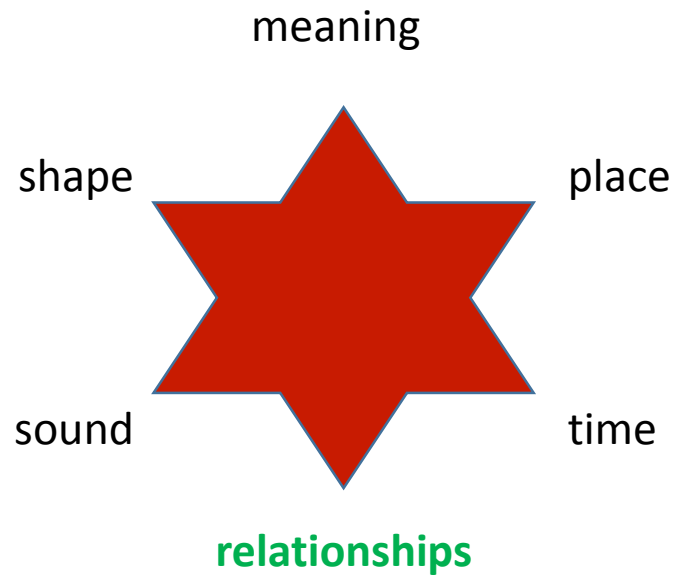
lamp
(synonym)



torch
(hyponym)



car
(holonym)



dark
(antonym)

light



(difference)

meaning

shape

place

sound

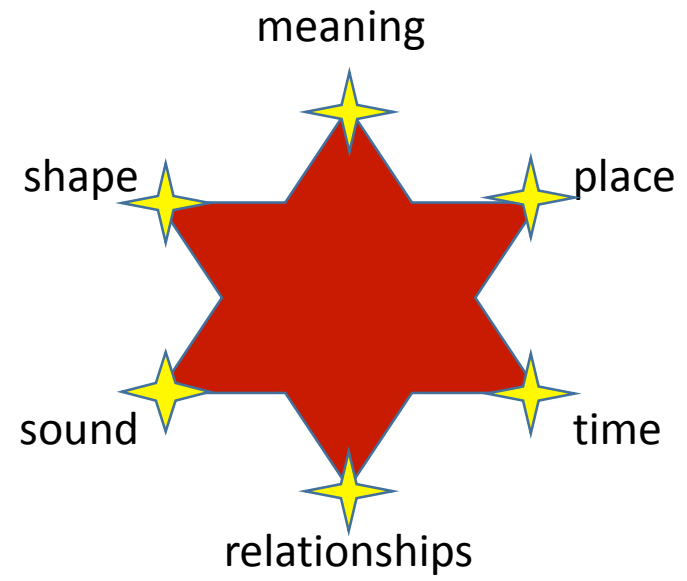
time

relationships

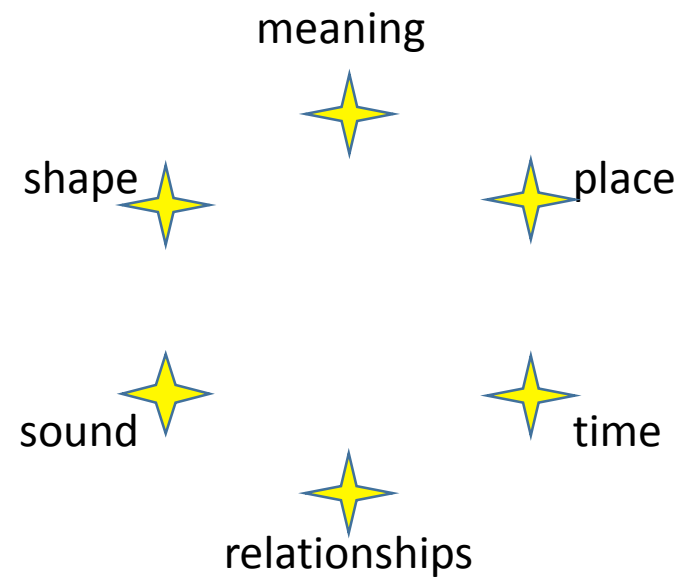
lamp
(synonym)



light



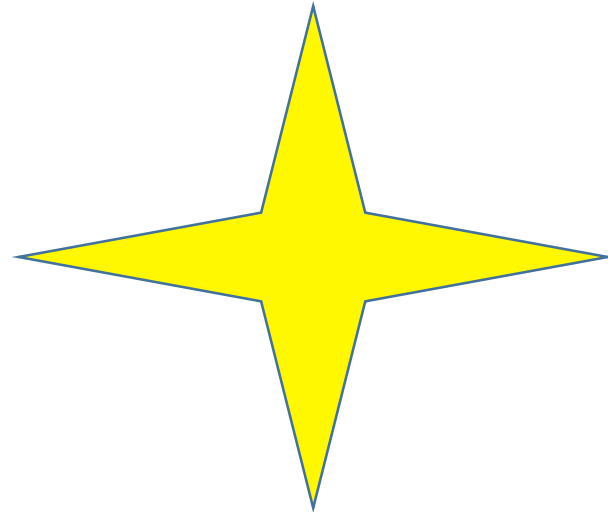
light



light



meaning



definition

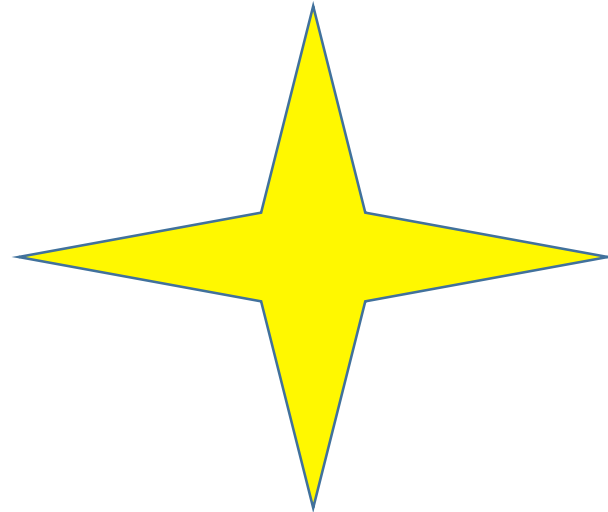
examples

translations

light



meaning

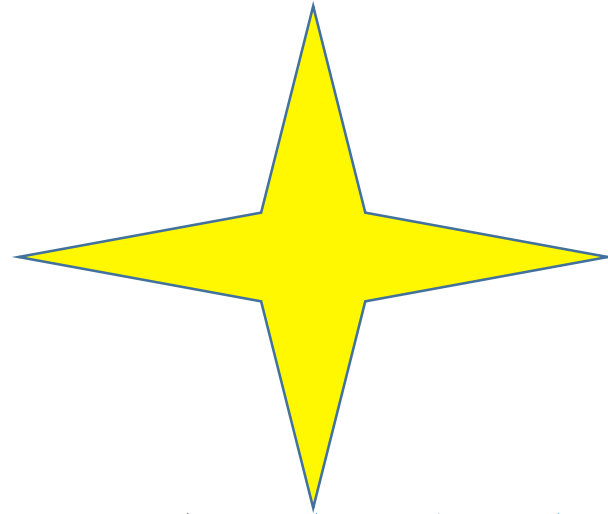


translations

light



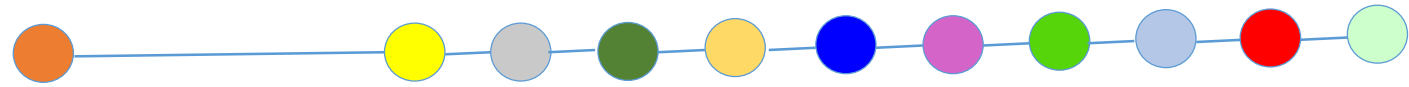
meaning



translations

equivalence

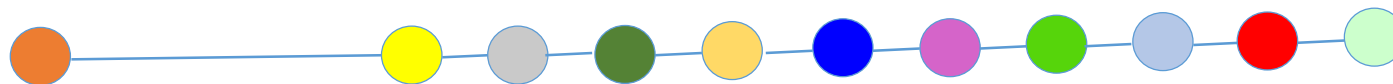
- Parallel
- Similar
- Explanatory



translations

equivalence

- **Parallel**
- Similar
- Explanatory



hand (English) = *main* (French)

✓: transitive across languages



translations

equivalence

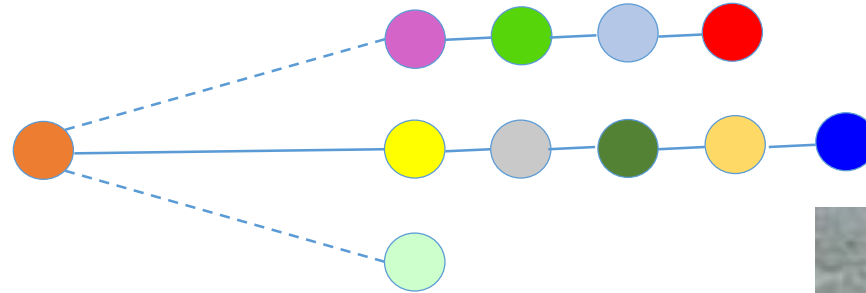
- Parallel
- **Similar**
- Explanatory



difference



difference translation



mkono (Swahili) = *hand* + *arm* (English)

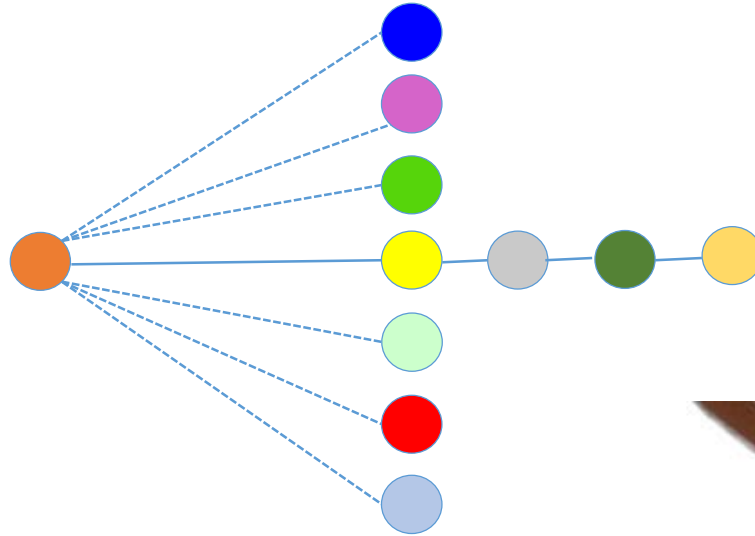
?? : might be transitive across languages



translations

equivalence

- Parallel
- Partial
- **Similar**



hand (English) = 10.2 cm (most languages)

X: not transitive across languages



translations

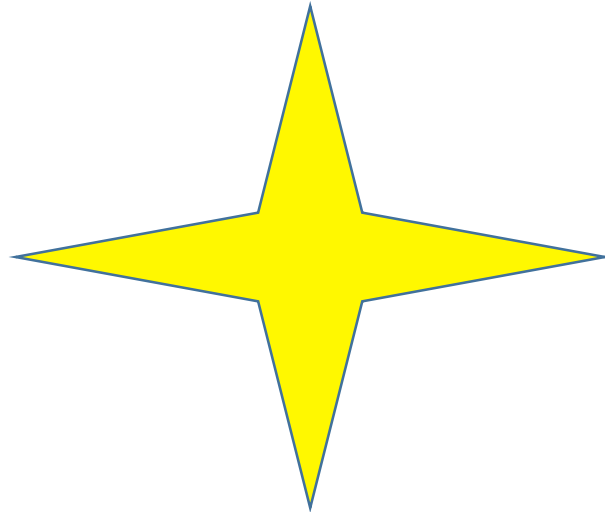
light



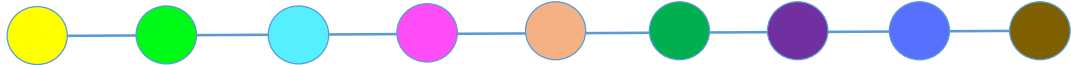
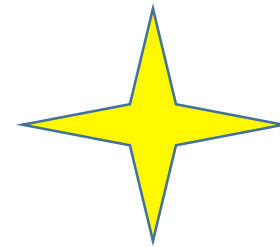
definition



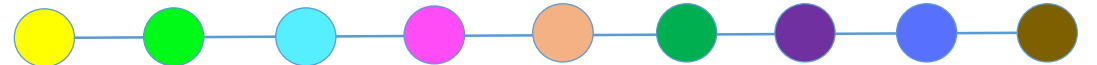
meaning



examples



definition
translations



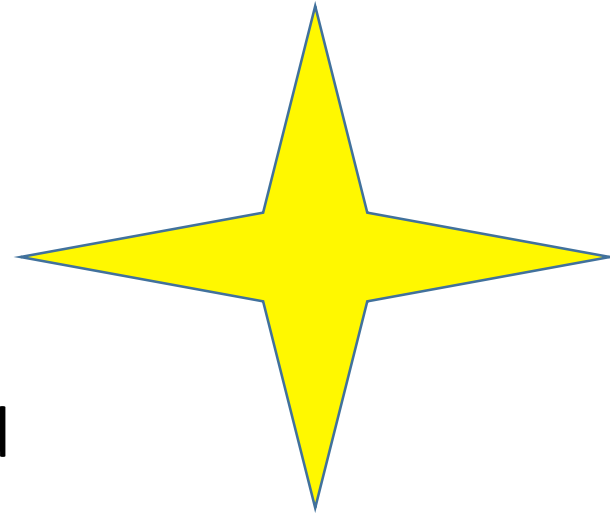
translations

example
translations

light



meaning



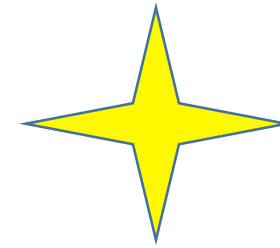
definition



easy

hard

examples



time

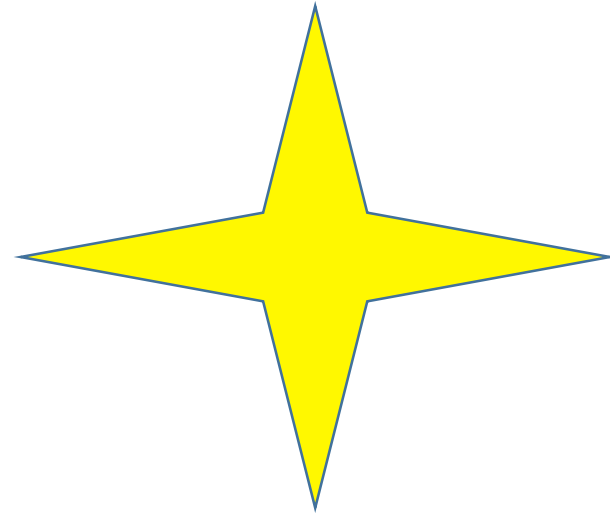
place

notes

light



shape



inflections

multiple words

alternates

light



sound

lighter



translation
shape

separability (MWEs)

alternate spellings

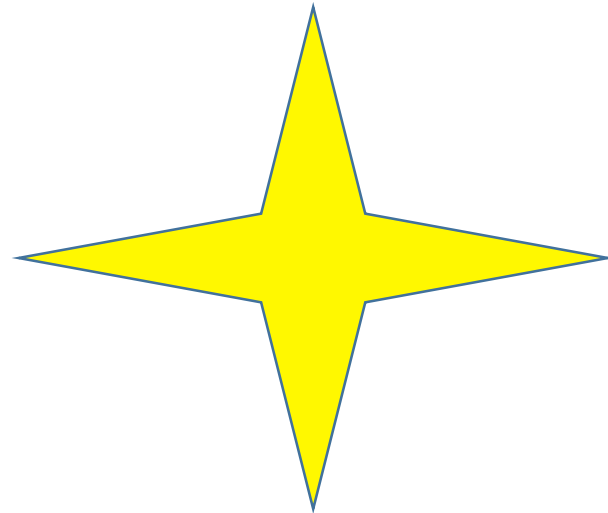
lightest



place

inflections

shape



spelling sets:
polysemous terms
often have the same
inflections.

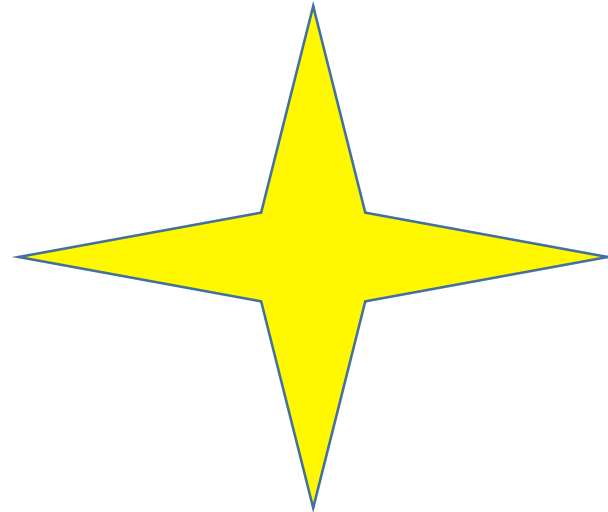
- Simple
Configurable form
e.g., English verbs
- Complex
Fixed table
e.g., French verbs
- Agglutinative
Rule-based coding
e.g., Swahili verbs

light



lite

shape



Kanji	Hiragana	Katakana	Rōmaji	English
金魚	きんぎょ	キンギョ	<i>kingyo</i>	goldfish

alternates

https://en.wikipedia.org/wiki/Japanese_writing_system

separability

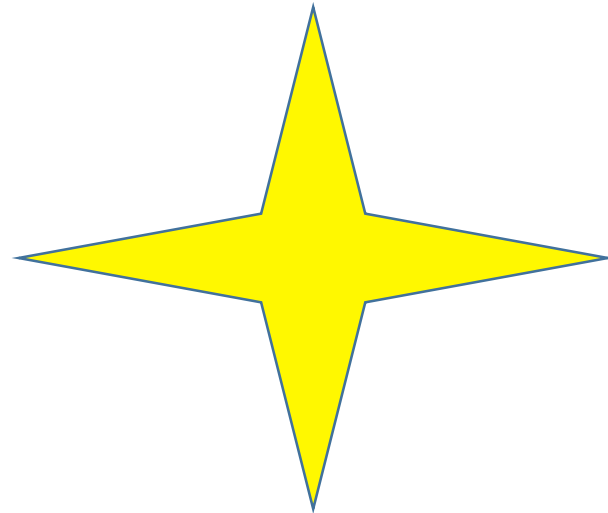
drive || up the wall

inflections
(+separability)

drives || up the wall
drove || up the wall
driven || up the wall
driving || up the wall

Research question:
Can we determine
Separability Sets?

shape



multiple words



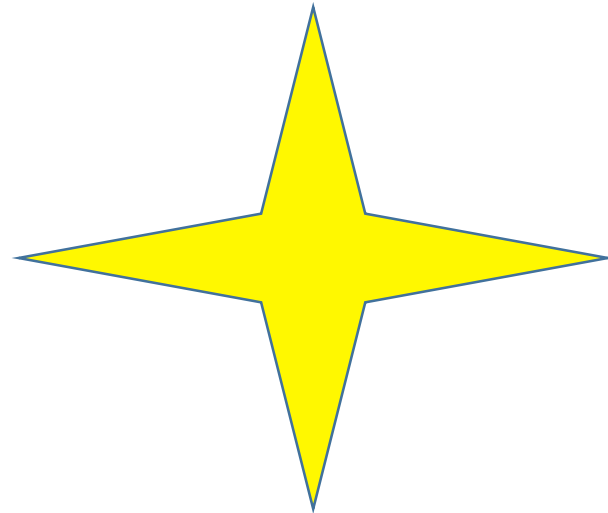
sign languages

e.g. Uganda Sign Language
Solomon Islander Sign Language

- no sound
- no spelling
- need for gesture recognition
(future research)



shape



ideograms

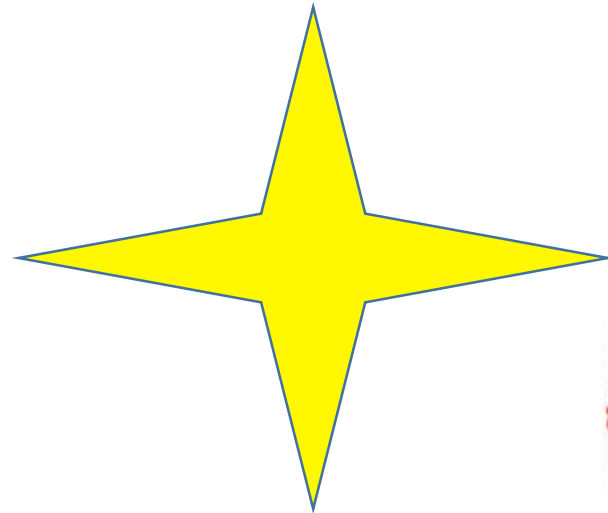


- no relation between shape and sound
- no sequencing
- ontological relationships

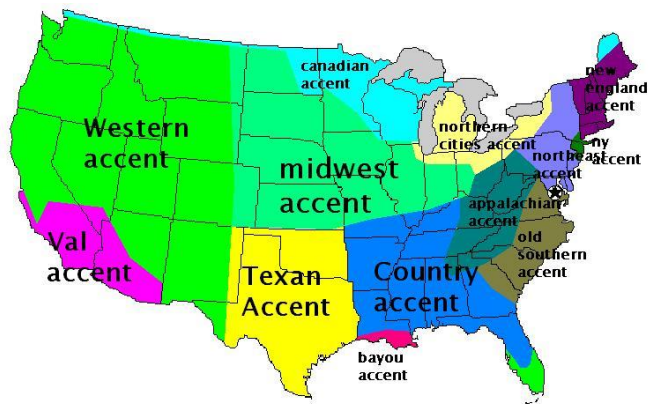
light



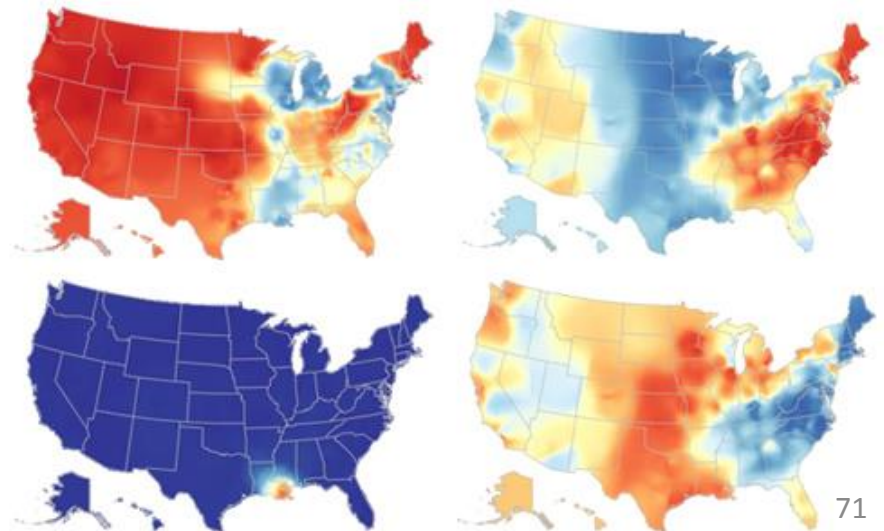
place



dialect



dialect word sightings

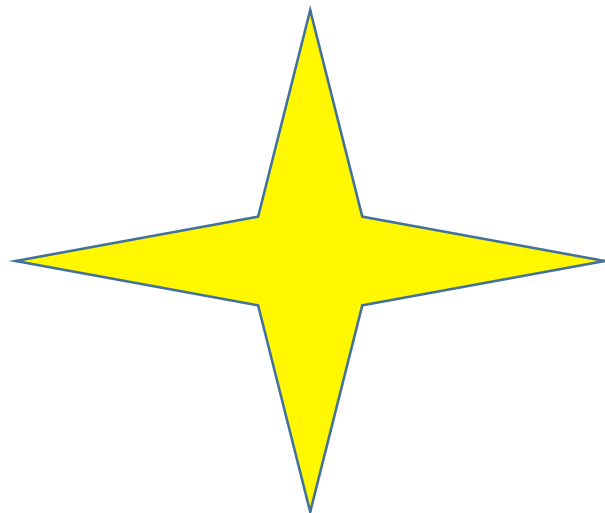


sound sightings

light



sound



audio



place

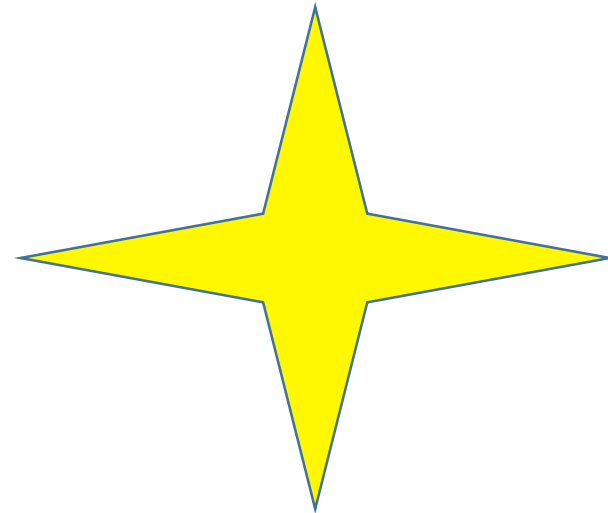
tone

IPA (phonetics)

light



time



ancestors
(other languages)

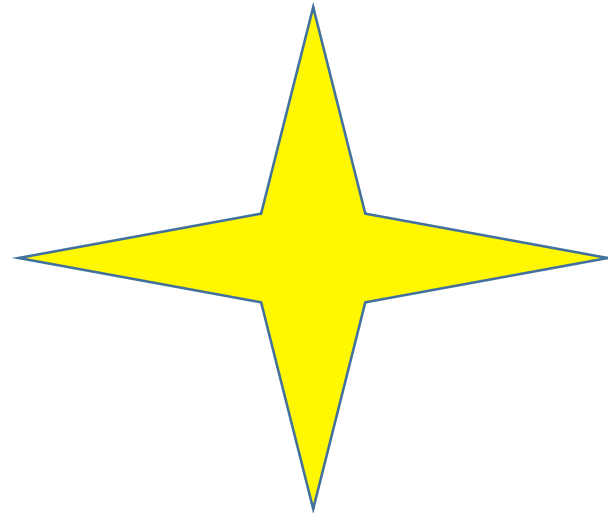
ancestors
(own language)

datings (examples)

light



relationships



synonyms



transitivity
with
translations

ontologies



hierarchies
or
reciprocity

terminologies

Lexicography vs. Terminology

Lexicography:

- General terms
- Variability of concepts among languages
- Describes indigenous words



Terminology

- Domain-specific terms
- Fixed meaning within context
- Prescribes words



Collecting Data

- Gathering new data
 - For languages with zero digitized data (most world languages)
 - For languages with incomplete data (all languages)
- Aligning existing data
 - To separate terms at concept level
 - To match concepts across languages

Collecting Data

Existing Data

- Copyright restrictions
- Data structure
- Data alignment

.light¹ *adj* 1 (*of colour*) -siokoza, -sioiva angazi, -a ~ *brown* kahawia isioiva, hafifu. 2 (*of a place*) -enye mwanga. ~ **coloured** *adj* -enye rangi isioiva. *n* 1 nuru, mwanga *the* ~ *begins to fail* mwanga unaanza kufifia *day* ~ mchana. **in a good/bad** ~ (*of picture etc*) -a kuonekana vizuri/vibaya; (*fig*) eleweka vizuri/vibaya. **see the** ~ (*liter or rhet*) zaliwa; baini; tangazwa; tambua; -okoka. **be/stand in one's** ~ kinga nuru; (*fig*) zuia mtu mafanikio/maendeleo yake. **stand in one's own** ~ zuia kazi yako isionekane; fanya kinyume na matakwa yako. ~ **year** *n* (*astron*) kipimo cha umbali kati ya nyota. 2 taa. ~ **s out** muda wa kuzima taa. **the northern/southern** ~ **s** *n* miali ya mwanga katika ncha za kaskazini na kusini. 3 mwako wa moto; kiberiti *strike a* ~ washa moto; washa kiberiti. 4 uchangamfu (usoni mwa mtu). **the** ~ **of somebody's countenance** (*biblical*) kupendezwa kwake. 5

Collecting Data

Expert Interface: Edit Engine

Key Information * Word Forms and Origins Translations and Concepts Related terms within the same language Examples *

Concept

Term (lemma) * **Term Language *** **Headword ***

Part of Speech * **Part Attribute**

Morpheme Information

Collected Morphemes

Plural

Definition

Definition of the term in the language to which it belongs

Definition Source URL

Crowdsourcing Lexicography

- Gathering new data
 - For languages with zero digitized data (most world languages)
 - For languages with incomplete data (all languages)
 - Aligning existing data
 - To separate terms at concept level
 - To match concepts across languages
-
- People are very good at these tasks
 - Machines are very bad
 - Scholars are very busy

Crowdsourcing with Games

- Engage the public in producing raw data
- Data can be built upon and refined over time
- Collecting “facts” that
 - can best come from native informants
 - can be verified by consensus as fulfilling a communicative role
- Wrong data and bad actors can be removed

Game Architecture

- Simple tasks the public can understand
- “Word” questions to stimulate the mind
- Competition elements to stimulate the heart
- Answers validated by consensus
- Starts with English concept set to have a shared realm of ideas
- Grows progressively – winning answers in one mode generate more advanced questions in the next



Games



Martin

Home

FBP



Find Games

Your Games

Activity

Search for games...



kamusi GAME

Global Online Living Dictionary

Kamusi GAME



Trivia & Word Game · 24 players



ABOUT KAMUSI GAME

Play Now

Send to Mobile

- Share
- App Website
- Remove
- Block
- Report a Problem

Available on Mobile Web, Facebook.com



Nina Kaplan #Hillary, #Gaga & #TBennett at fundraising event with #johnbarrettny



Rebecca Goff likes Emily Morningstar's photo.



Patrick Hall likes Bob Kennedy's link on his own Timeline.



Jennifer Bower Knowles likes Episcopal Diocese of Vermont's video.



Anders Halverson likes Taylor Keen's photo.



Sandra Foyt likes Carol Cain's photo.



Funya Gleason shared a link.



Prisca M Kikwabha likes TIME's link.



Anna Sawyer commented on her own post.



Mary Doyle Collins Habich likes Dream Design Flooring.



Arky Arky Arky likes Chia-liang Kao's photo.



Caroline Old Co... Mobile



Truong Anh Tuan Mobile

GROUP CONVERSATIONS

MORE FRIENDS (8)



Anders Halverson Web

62%

Search



Game Modes

1. Translation
2. Synonyms
3. Word Forms
4. Definitions
5. Examples
6. Alignment
7. Equivalence
8. Difference

Translation Game

The screenshot shows a game interface with a main content area and a vertical sidebar on the right. The main area contains the following text:

Игра в переводики ⓘ
Translate the following word to : Russian
out *прилагательное*
Working definition: **excluded from use or mention**

Below this is a yellow input field containing the text: я могу перевести это слово!

Underneath the input field is a button with the text: ? I can't say - skip this one...

The sidebar on the right is dark brown and contains the following elements from top to bottom:

- A red header with the text "Kanssi GAME" in white.
- Game statistics: "В игре 0" and "Banked: 0".
- Five icons: a single person, two people, a speech bubble, an information icon (i in a circle), and a building icon.

Translation Game

The screenshot shows a game interface for translating the word "out" to Russian. The main content area is titled "Игра в переводики" (Translation Game) and includes the instruction "Translate the following word to : Russian". The word "out" is shown with its part of speech "прилагательное" (adjective) and a working definition: "excluded from use or mention". A yellow button with a microphone icon says "я могу перевести это слово!" (I can translate this word!). Below it is a text input field with a cursor and a right-pointing arrow icon. A grey button below the input field says "? I can't say - skip this one...". On the right side, a vertical sidebar contains the game title "Kamus! GAME", the score "В игре0 Banked: 0", and five icons: a person, a group of people, a speech bubble, an information icon, and a building icon.

Игра в переводики ⓘ
Translate the following word to : Russian
out *прилагательное*
Working definition: **excluded from use or mention**

я могу перевести это слово!

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

Kamus! GAME
В игре0
Banked: 0

Icons: Person, Group of People, Speech Bubble, Information, Building

Definition Game

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!

? 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: Person, Group of People, Speech Bubble, Information, Building

Definition Game

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!

? 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

Definition Game

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

👤
👥
💬
ℹ️
🏛️

Example Game


Tweet Game ⓘ

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

every *adjective_satellite*






Working definition: **each and all of a series of entities or intervals as specified**

- tried to get tickets to meet Jonah on tour with Grant but every place is to far from Florida :(#HelpCaseyMeetJonah @JonahMarais 🍷💜
- @Kade0416 lol..you should have seen ghost where you could shoot through about every wall with an lmg
- That being said, some lives aren't as long as others. Agreed, but every life is short compared to what comes after death
- Get Free Bitcoin Automatically Every hour. No Need To Do Anything - \$ <http://t.co/fBITTqsdln> \$ EARN #Bitcoin HERE
- Lmao I didn't even know this picture existed until a few days ago and I die every time I look at my face 😂 <http://t.co/J4IK350Lf0>
- I'm fine w/ the fact my team (76ers) are horrendously rebuilding cause honestly watching them be a 8 or 9 seed every year was worse.
- There is a shooting in Chicago every 3 hours. Let that sink in.
- Some "investigation" found that @WholeFoods was overcharging customers although they every store has certain things priced lower or higher.



Kanvaal GAME

In Play: 189
Banked: 71

- 
- 
- 
- 
- 



Martin Benjamin

The Particles of Language:
"The Dictionary" as elemental data for 7000 languages across time and space