ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Martin Benjamin & Paula Radetzky

Multilingual Lexicography with a Focus on Less-Resourced Languages: Data Mining, Expert Input, Crowdsourcing, and Gamification

Workshop on Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era
Language Resources and Evaluation Conference (LREC 2014)
26 May, 2014 - Reykjavik, Iceland

- **~7000 languages spoken worldwide** (Ethnologue 2014)
- Top 5: Excellent ICT resources
  - English + FIGS (French, Italian, German, Spanish)
- ~30: Significant ICT resources (speech recognition, machine translation, etc)
  - Many official EU languages + Russian, Arabic, Chinese, Japanese…
  - ~~Hindi~~, ~~Bengali~~, ~~Punjabi~~, ~~Indonesian~~, ~~Swahili~~… (100 million+ speakers)
- **~80: More than 10,000,000 speakers**
- <100: Basic ICT resources (corpora, machine-readable dictionaries, POS taggers, morphological analyzers, parsers, etc)
- **~300: More than 1,000,000 first language speakers**
- **~1500: More than 100,000 first language speakers**
- **~2000: Threatened – 10,000 to 100,000 speakers**
- ~2000: Some existing data (wordlists, Bibles, recordings)
- **~3000: Endangered**
  - Fewer than 3000 speakers
  - Not being passed to next generation – extinct within 100 years
  - Embedded human cultural patrimony

Multilingual Lexicography with a Focus on Less-Resourced Languages:

Multilingual Lexicography with a Focus on Less-Resourced Languages: Data Mining, Expert Input, Crowdsourcing, and Gamification

Workshop on Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era
Language Resources and Evaluation Conference (LREC 2014)
26 May, 2014 - Reykjavik, Iceland

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

kamusi is Swahili for *dictionary*

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Comprehensive linguistic data for every language

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Comprehensive linguistic data for every language

- For people
- For machines

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

In service since 1994 (originally at Yale Council on African Studies)
International NGO since 2009

- Registered non-profit in USA and Switzerland (and soon Tanzania)

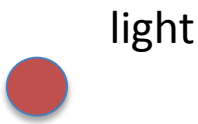
Academic Home since 2013:
EPFL - Swiss Federal Institute of Technology in Lausanne
LSIR - Distributed Systems Information Laboratory

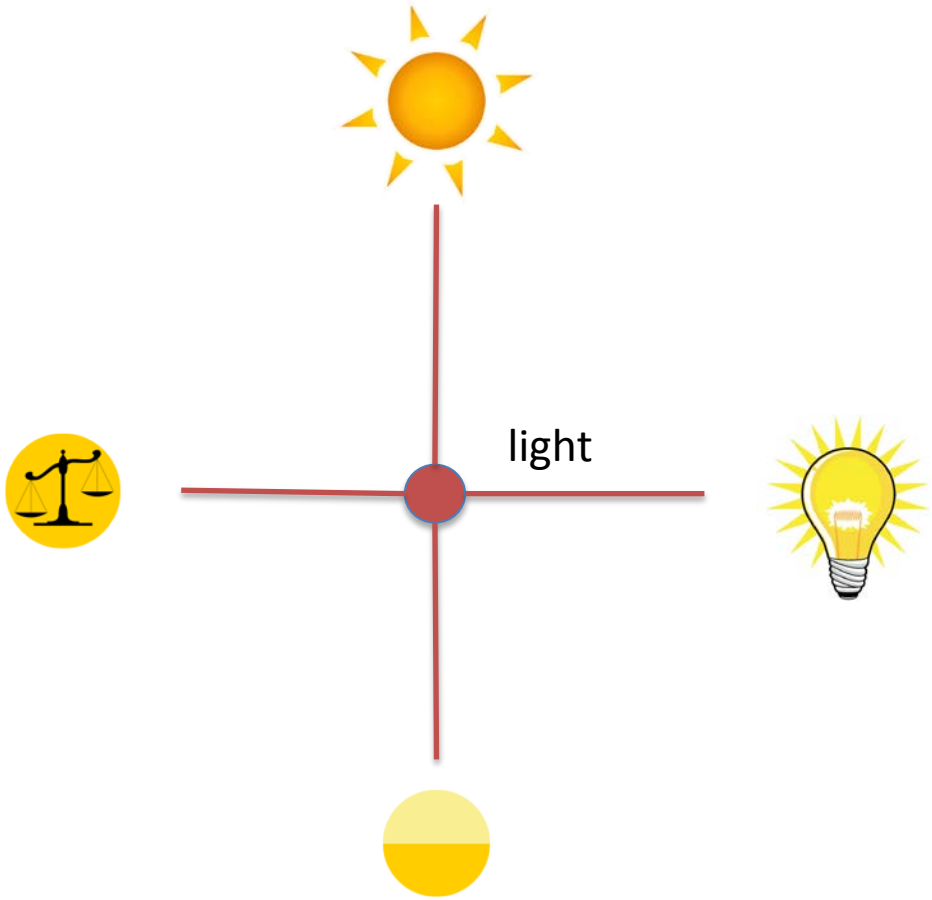
ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

1. Project objectives
2. Acquiring data
3. Crowdsourcing innovations

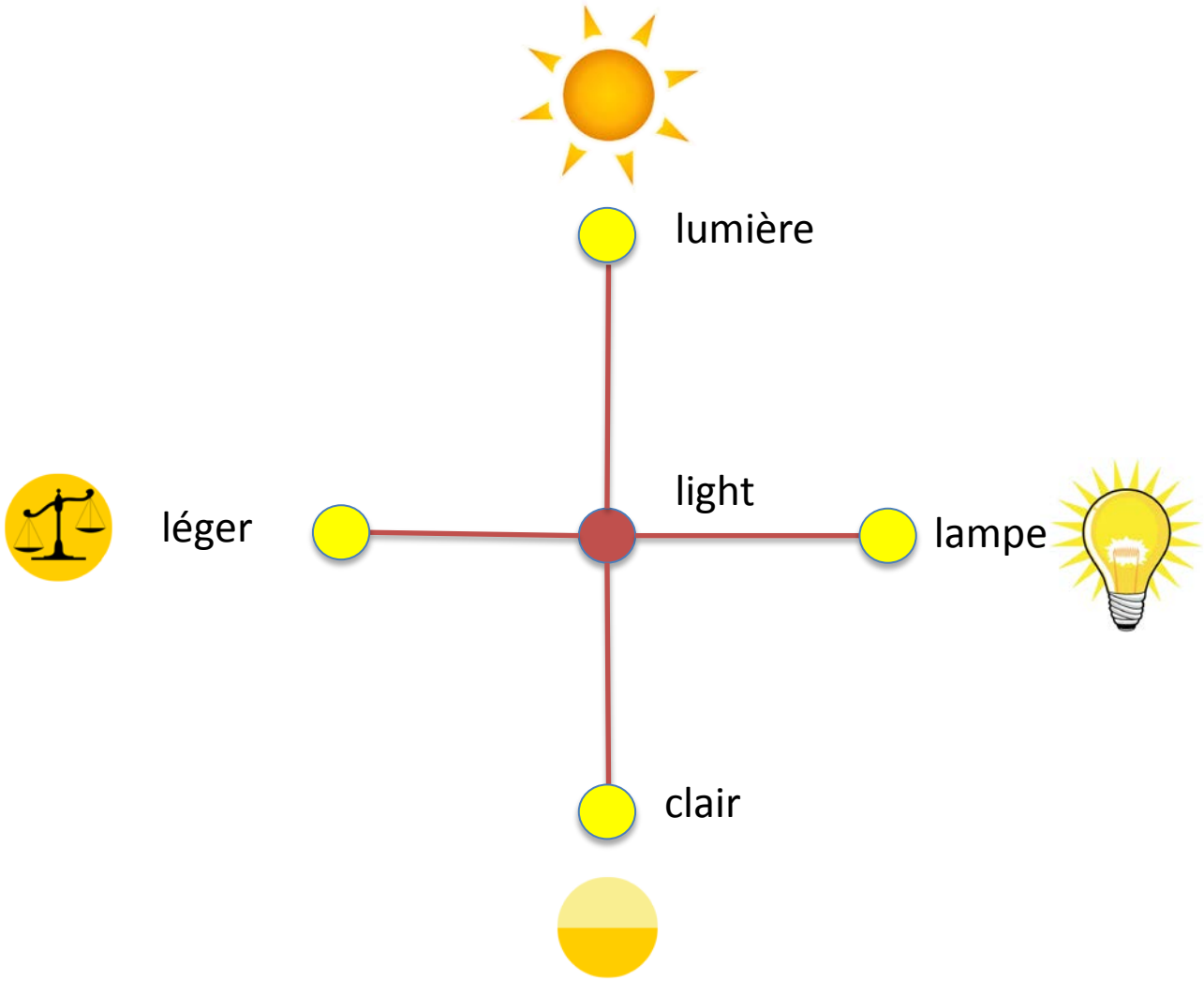
ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

1. **Project objectives**
2. Acquiring data
3. Crowdsourcing innovations

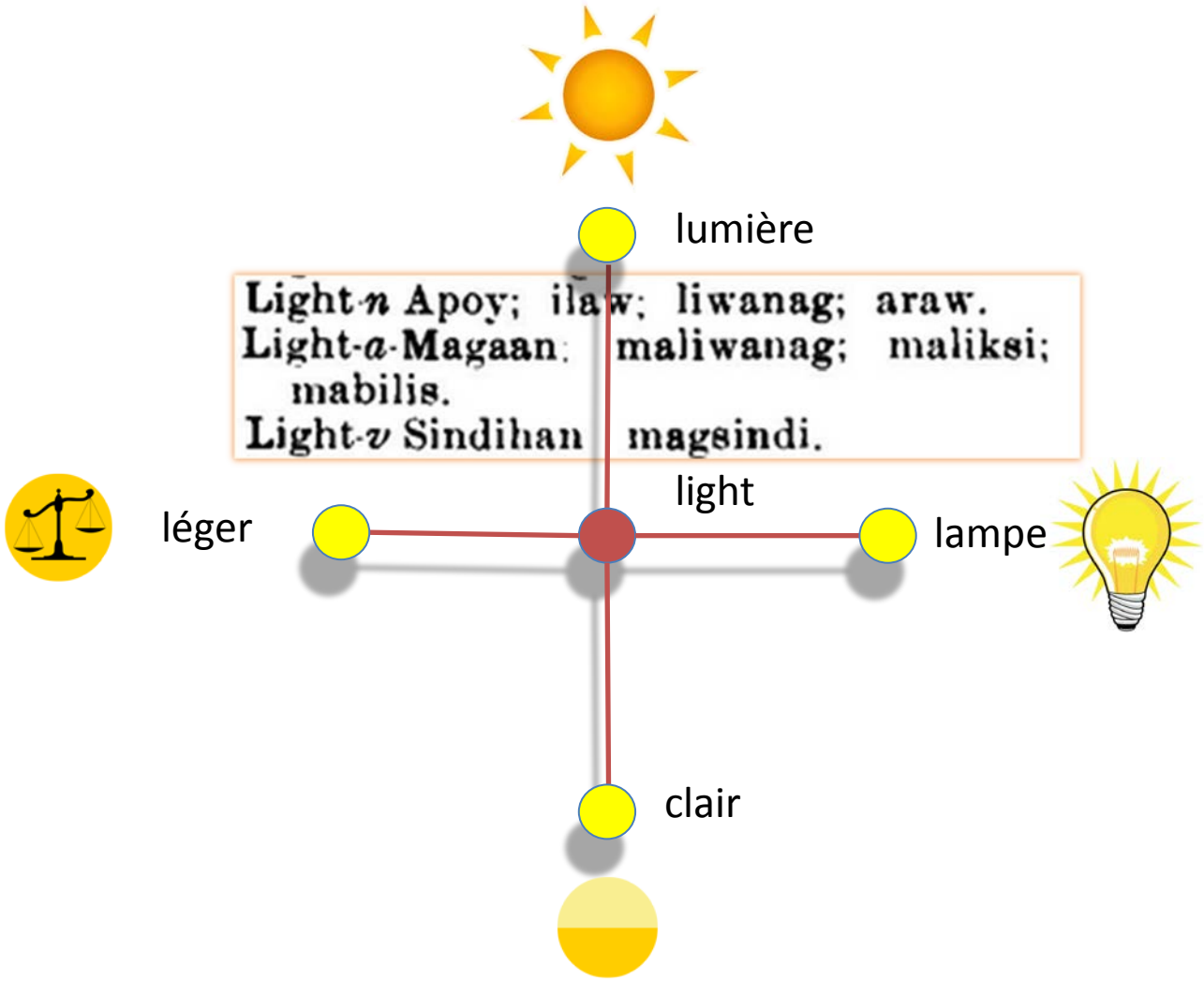

light



why multilingual dictionaries were impossible

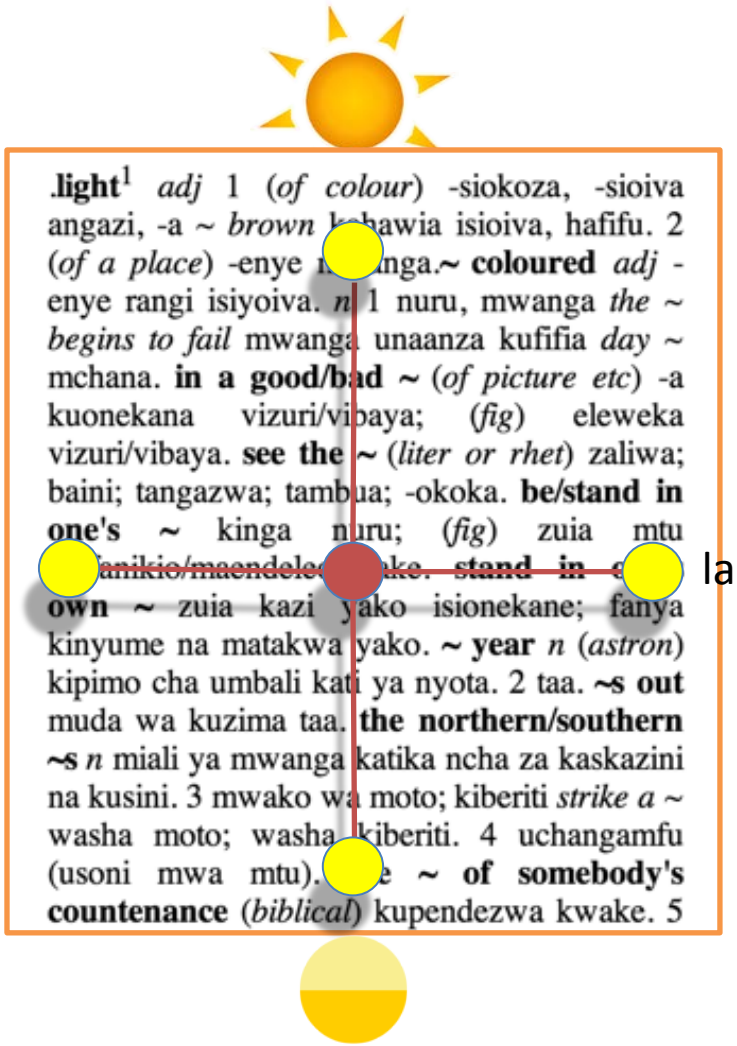


why multilingual dictionaries were impossible



why multilingual dictionaries were impossible

léger

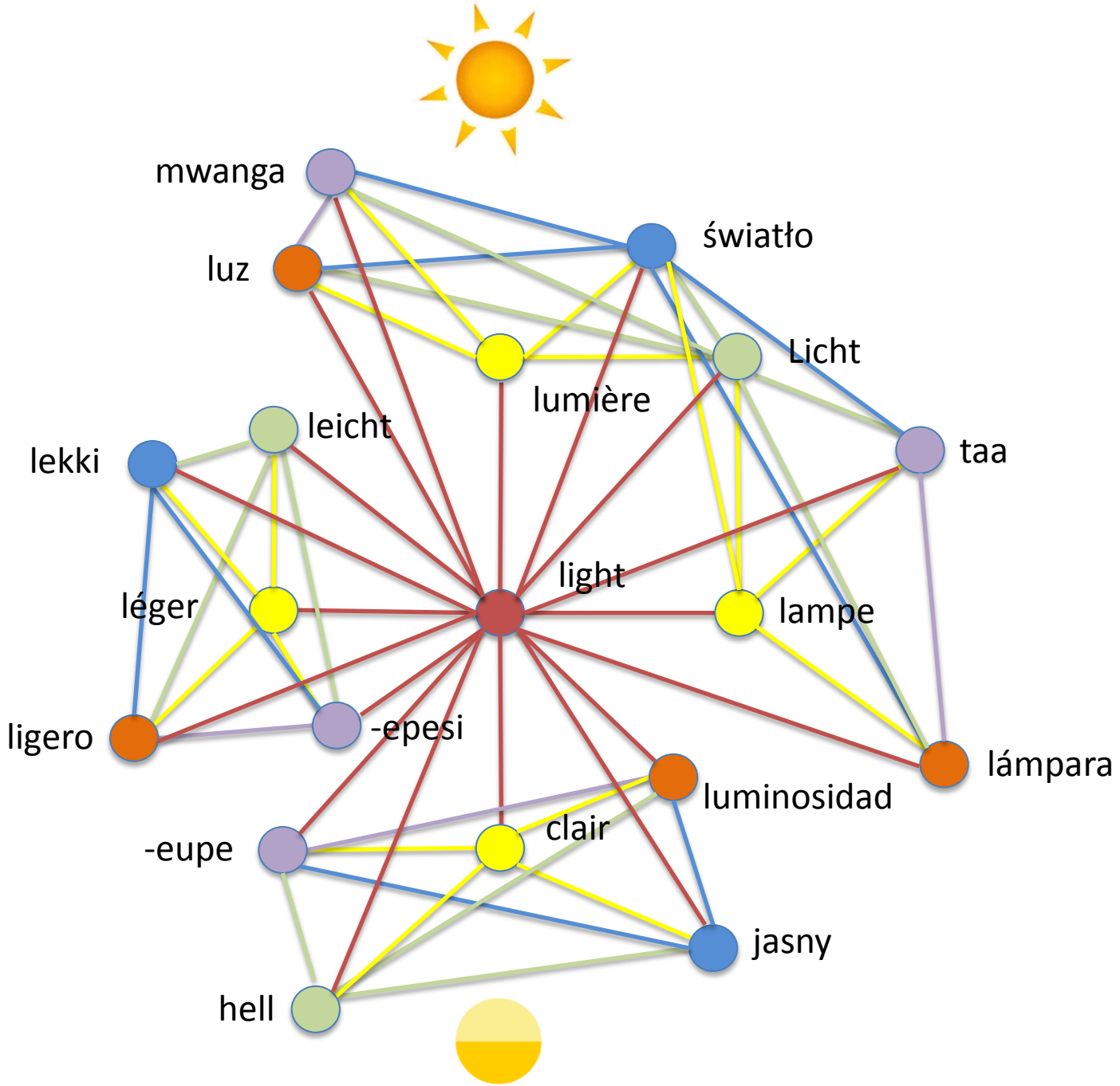


.light[1] *adj* 1 (*of colour*) -siokoza, -sioiva angazi, -a ~ *brown* kahawia isioiva, hafifu. 2 (*of a place*) -enye nuru, -enye mwanga. **~ coloured** *adj* -enye rangi isiyoiva. *n* 1 nuru, mwanga *the ~ begins to fail* mwanga unaanza kufifia *day ~* mchana. **in a good/bad ~** (*of picture etc*) -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 ~~anikie/maendele~~ ~~ke~~. ~~**stand in**~~ **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). **~ of somebody's countenance** (*biblical*) kupendezwa kwake. 5

lampe

why multilingual dictionaries were impossible

why multilingual dictionaries were impossible



why multilingual dictionaries were impossible

light[1] *adj* 1 (*of colour*) -siokoza, -sioiva angazi, -a ~ *brown* kahawia isioiva, hafifu. 2 (*of a place*) -enye mwanga.**~ coloured** *adj* -enye rangi isiyoiva. *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

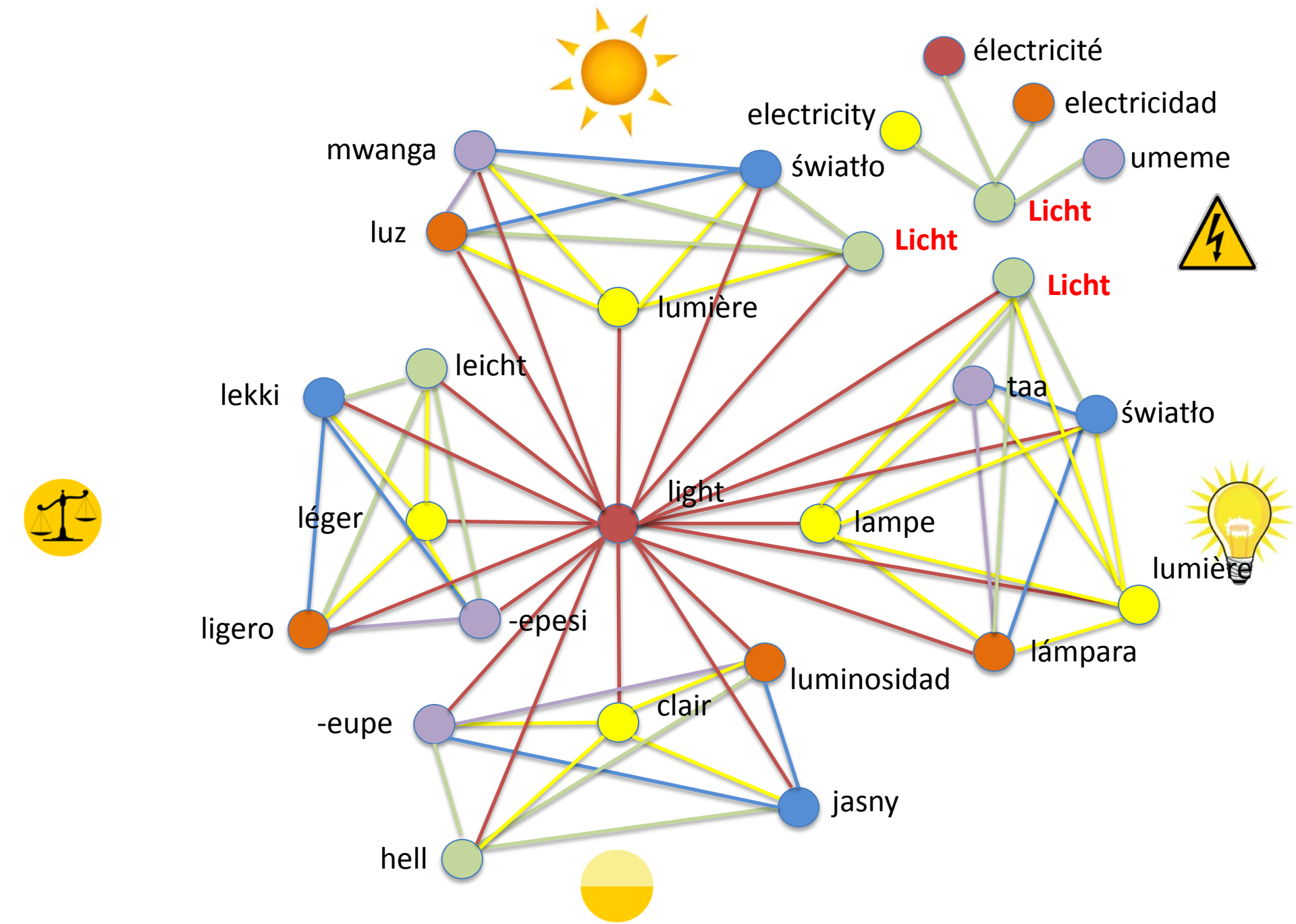
Meaning

તેજ, પ્રકાશ, અજવાળું, જ્યોતિ, દીપ, દીવો, અવરજવર નિયંત્રક દીવો, જેના વડે વસ્તુ દૃશ્યમાન થાય છે તે સાધન-પ્રકાશ, દેવતા સળગાવવાની દીવાસળી કે કાકડો, તેજસ્વિતા, આંખનું તેજ, પ્રકાશનું હલકોઈ ઉદ્ગમસ્થાન સૂર્ય, મીણબત્તી, ઇ. જેવું, ચમક, કશાકની ઊજળી બાજુ, દૃષ્ટિકોણ, દૃષ્ટિ, બોધ, જ્ઞાન, કોઈ બાબતને સ્પષ્ટ કરનારું નવું જ્ઞાન, માહિતી ઇ., અંધારું નહિ, (રંગ અંગે) ફીકું, પાંખું, ઝાંખું, (આગ ઇ.) સળગાવવું, પ્રકાશવું, પેટવવું, ચેતાવવું, બળવું, દીવાથી રસ્તો બતાવવો, –ને પ્રકાશ આપવો, પ્રસન્ન થવું કે કરવું, ઉલ્લસિત થવું કે કરવું

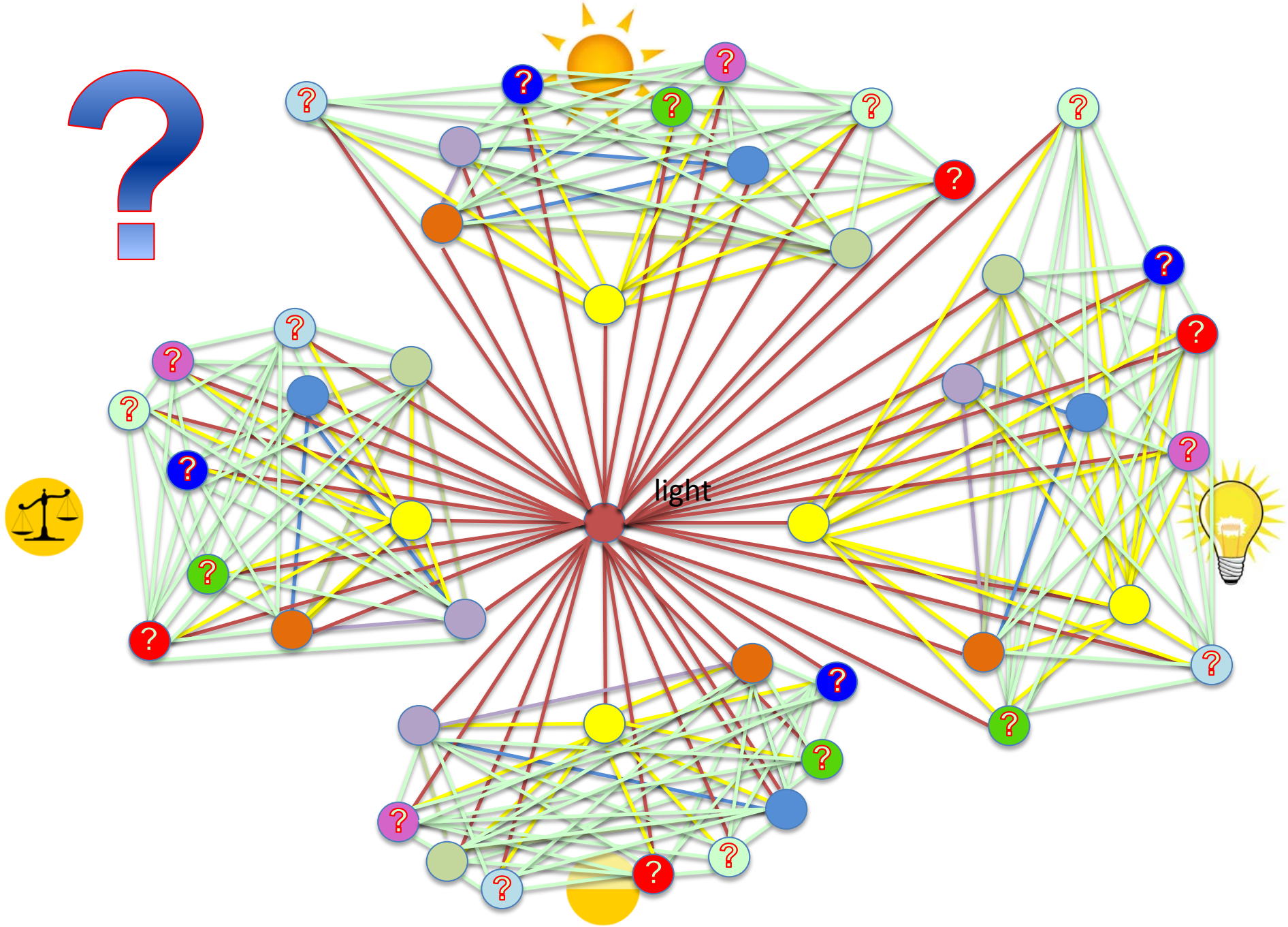
mwanga
światło
luz
Licht
lumière
leicht
lekki
taa
light
léger
lampe
ligero
-epesi
lámpara
luminosidad
clair
-eupe
jasny
hell

Light-*n* Apoy; ilaw; liwanag; araw.
Light-*a*-**Magaan**. maliwanag; maliksi; mabilis.
Light-*v* Sindihan magsindi.

why multilingual dictionaries were impossible



why multilingual dictionaries were impossible

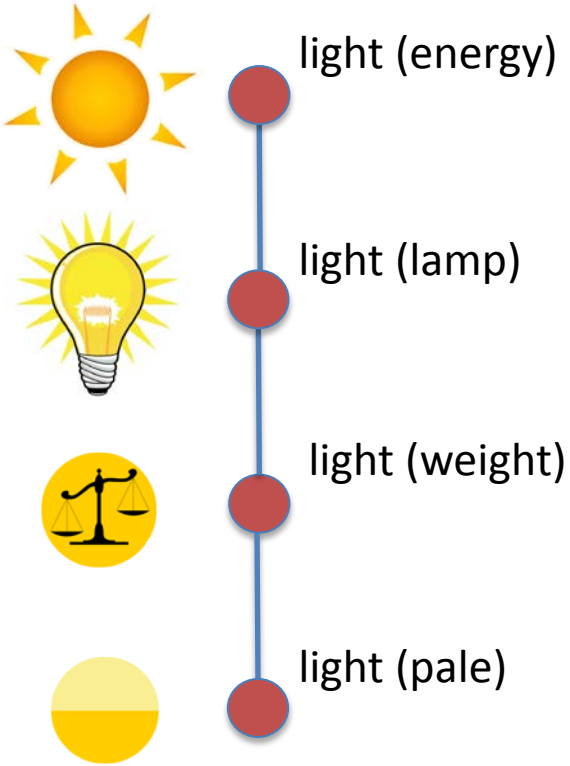


why multilingual dictionaries based on spelling are impossible

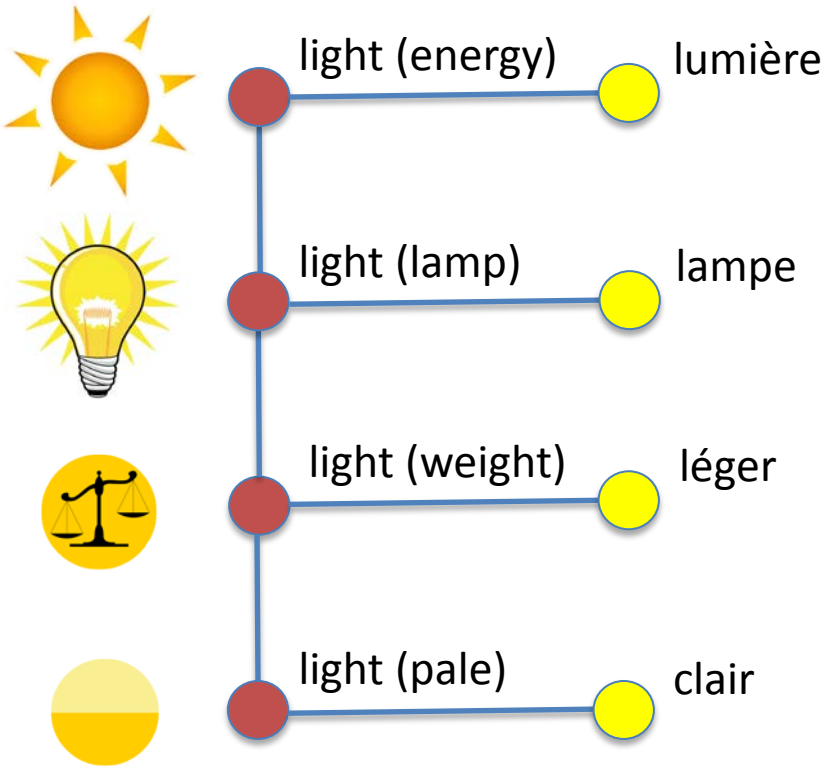
why multilingual dictionaries based on spelling are impossible

impossible

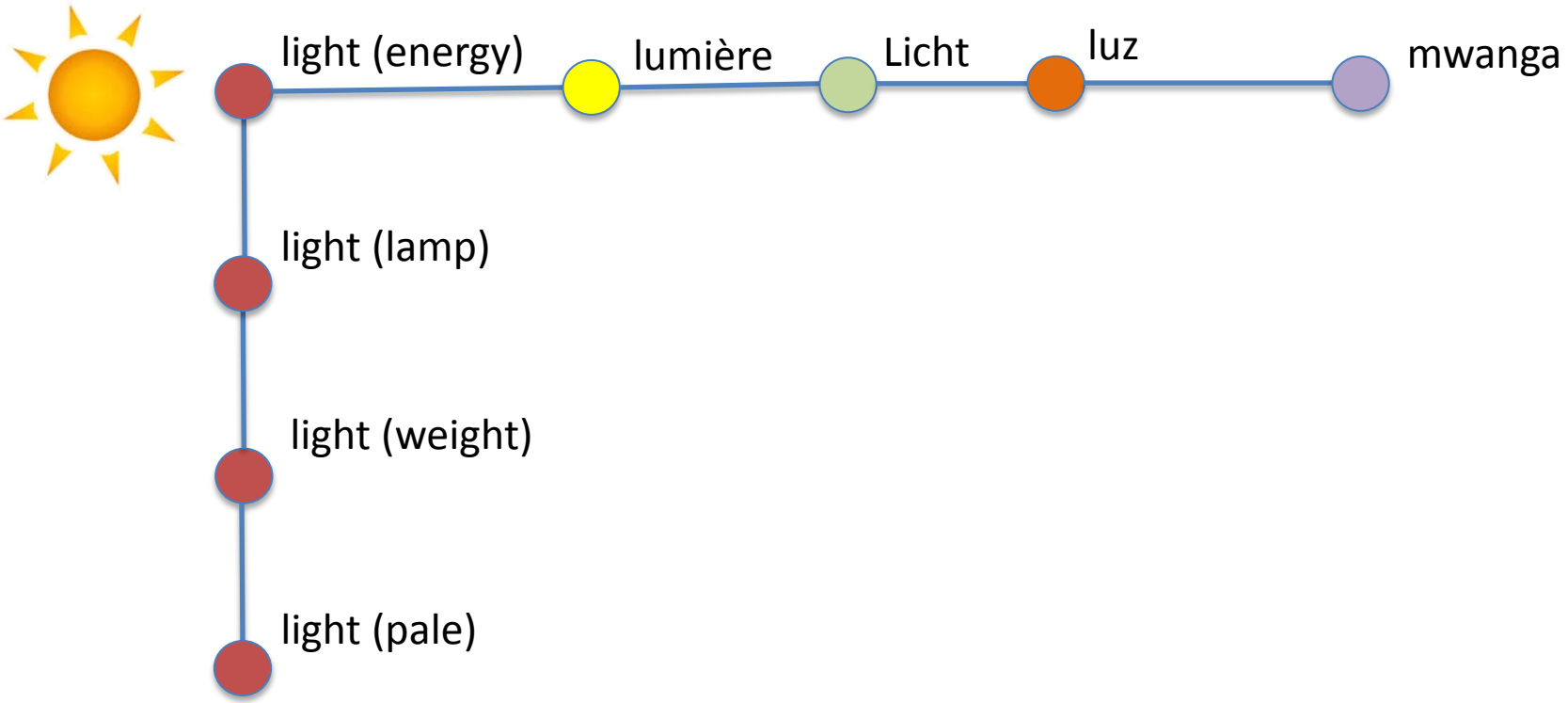
how Kamusi makes a multilingual dictionary **possible**



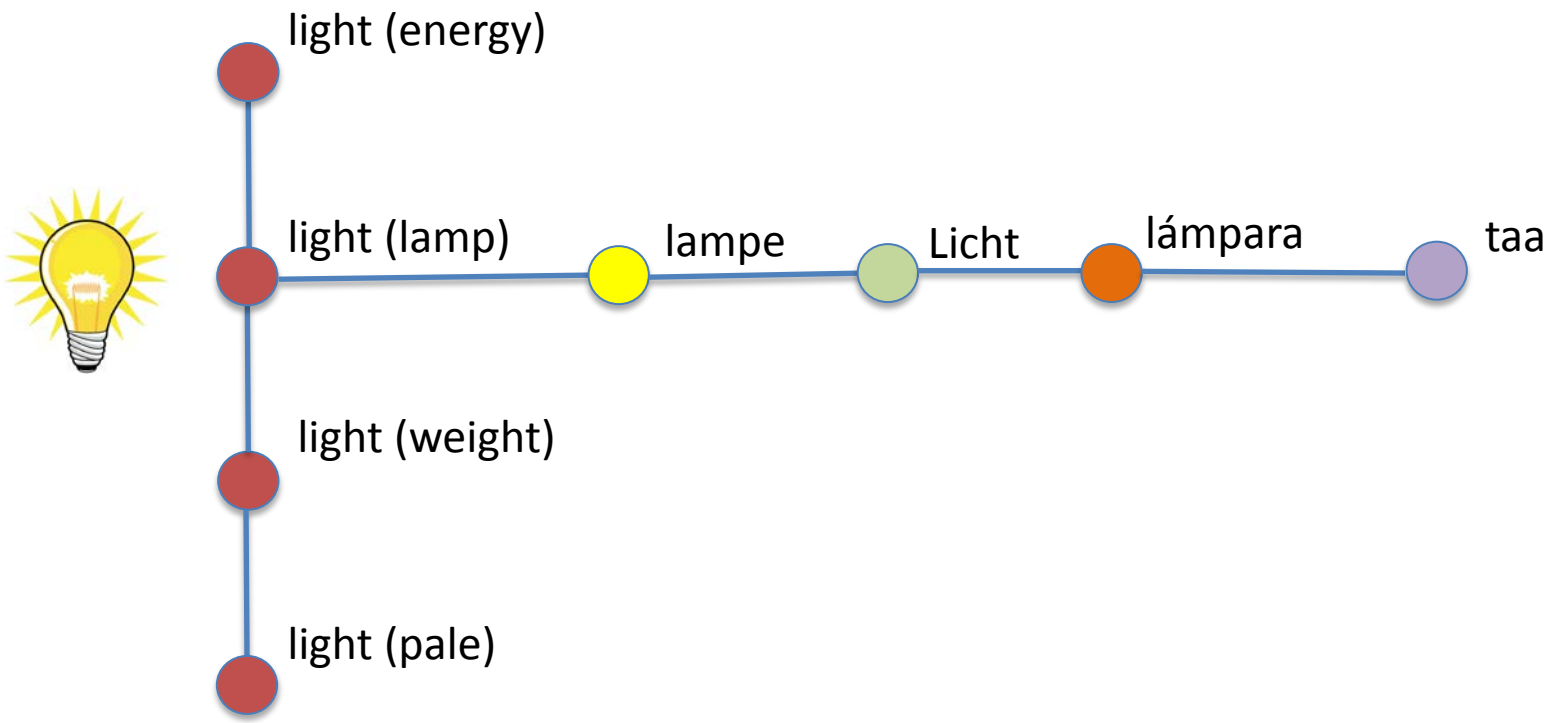
how Kamusi makes a multilingual dictionary possible



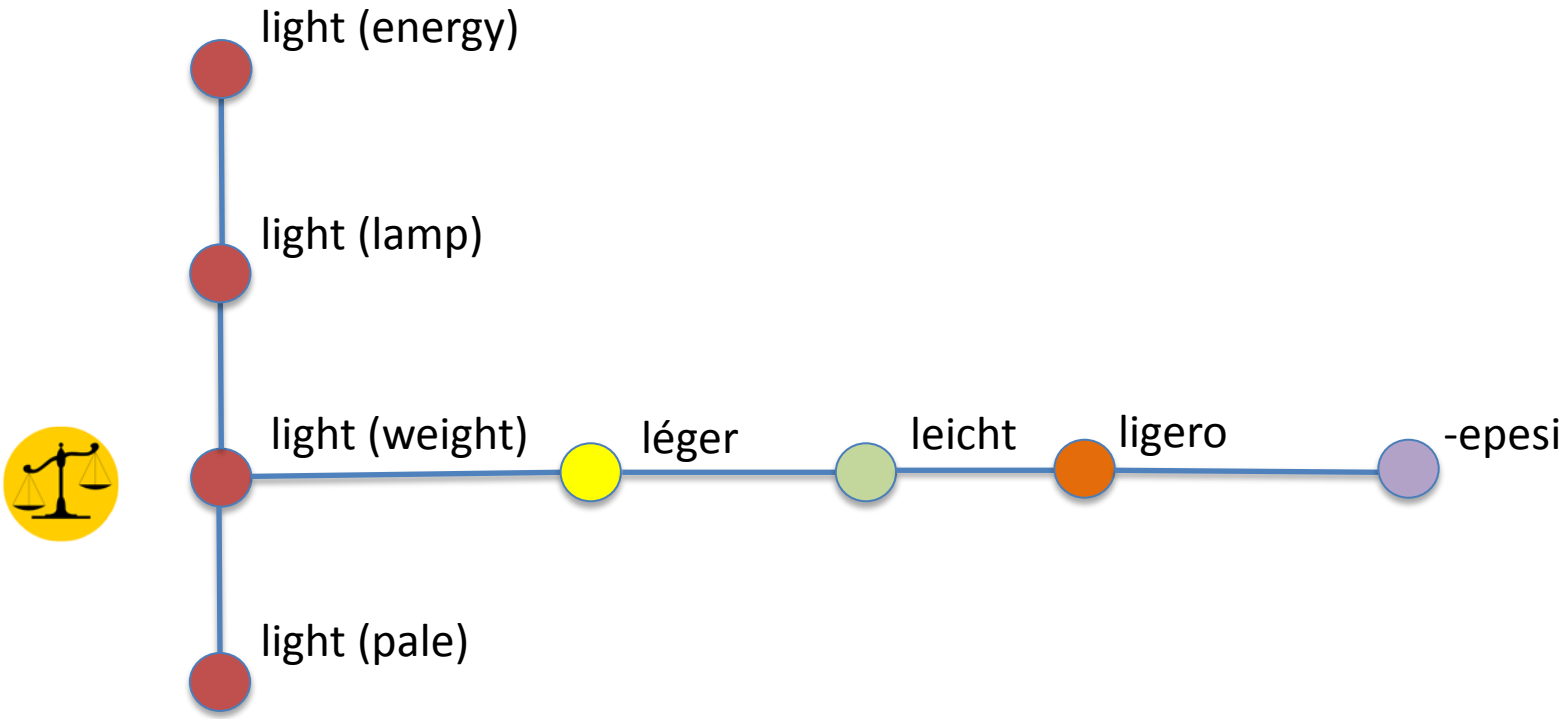
how Kamusi makes a multilingual dictionary possible



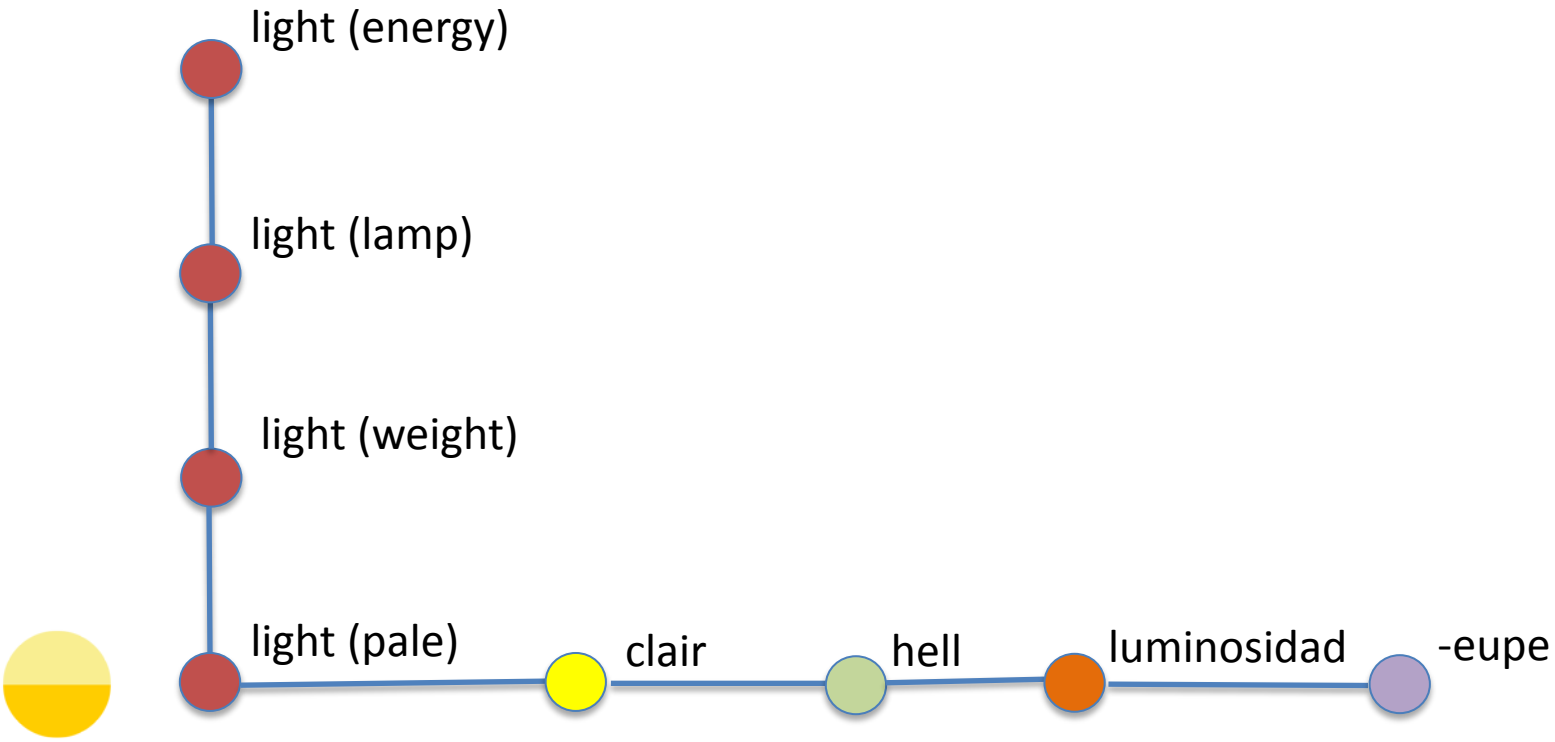
how Kamusi makes a multilingual dictionary possible



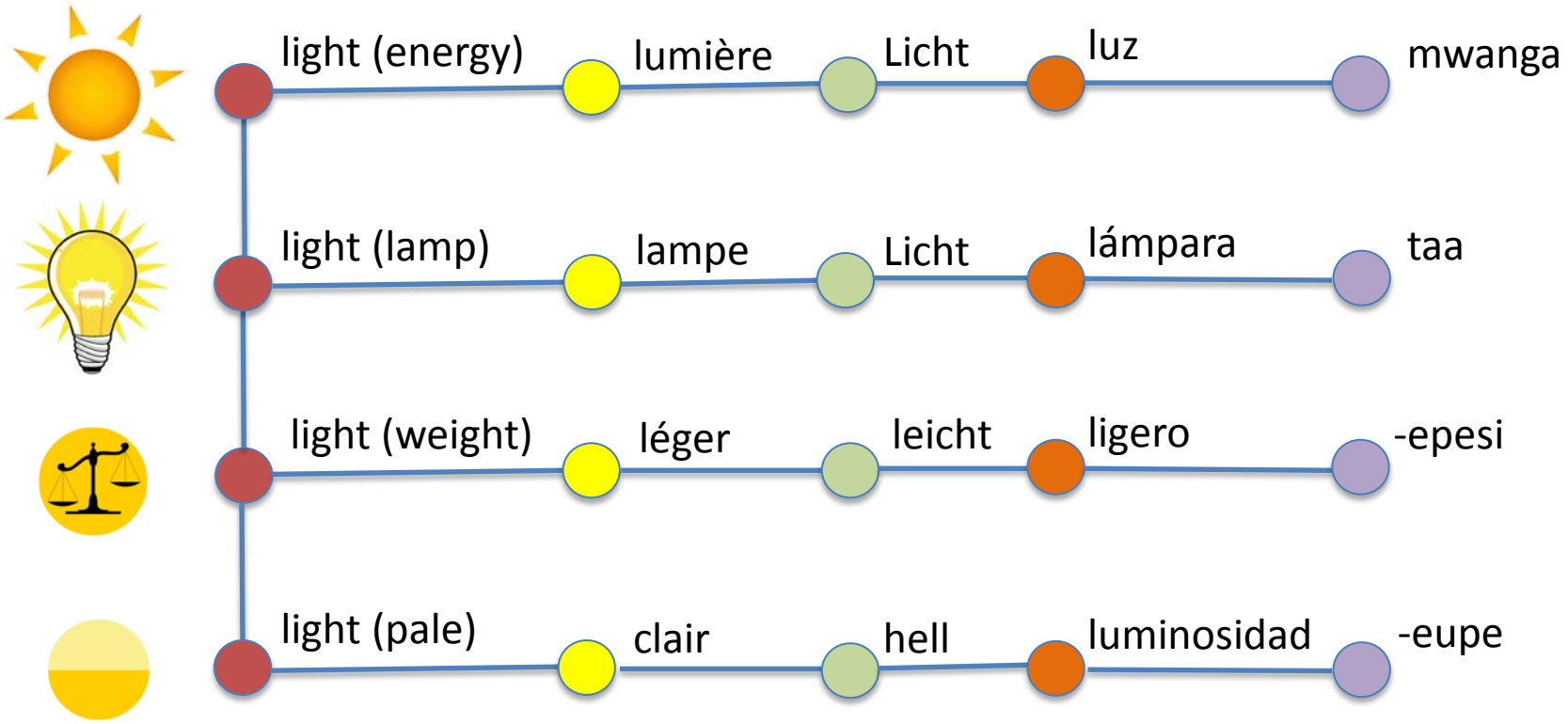
how Kamusi makes a multilingual dictionary possible



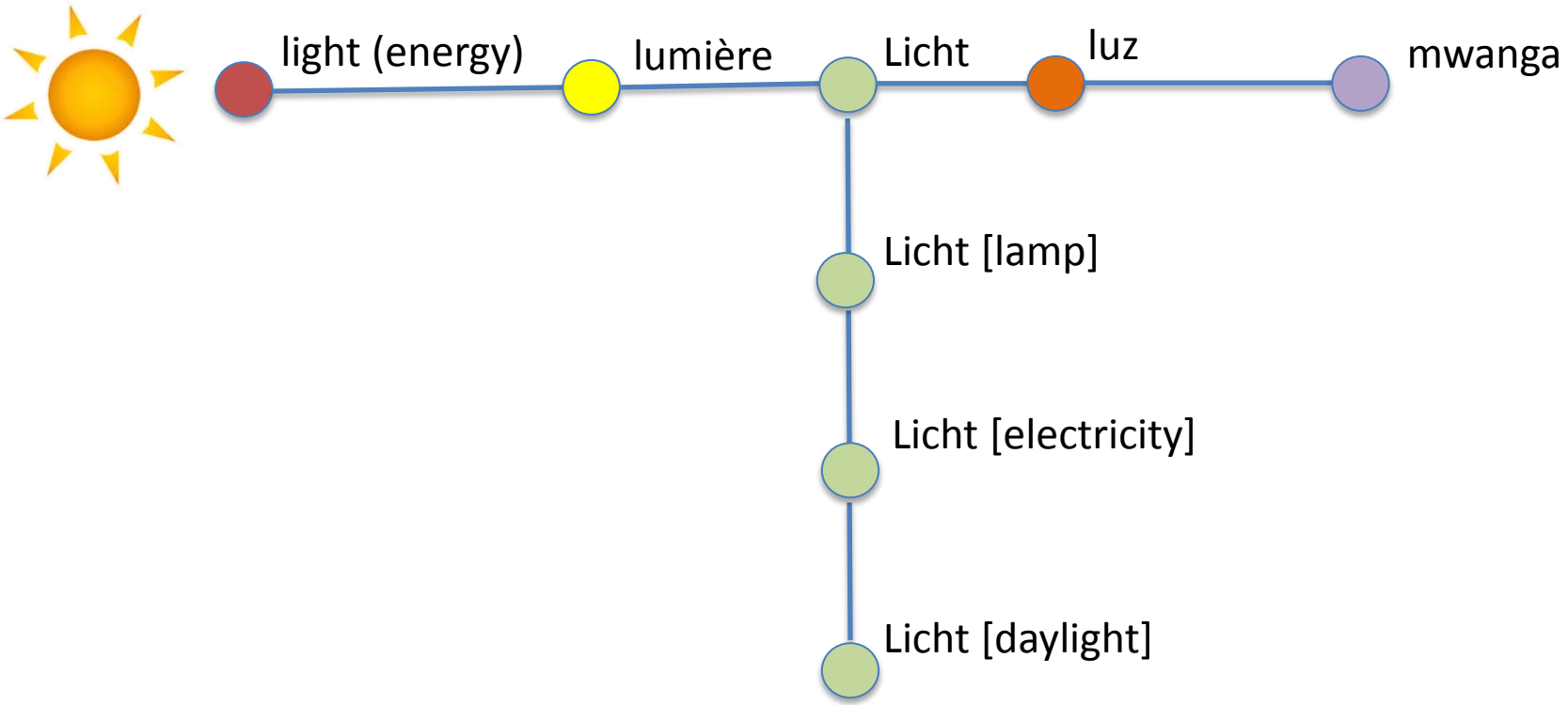
how Kamusi makes a multilingual dictionary possible



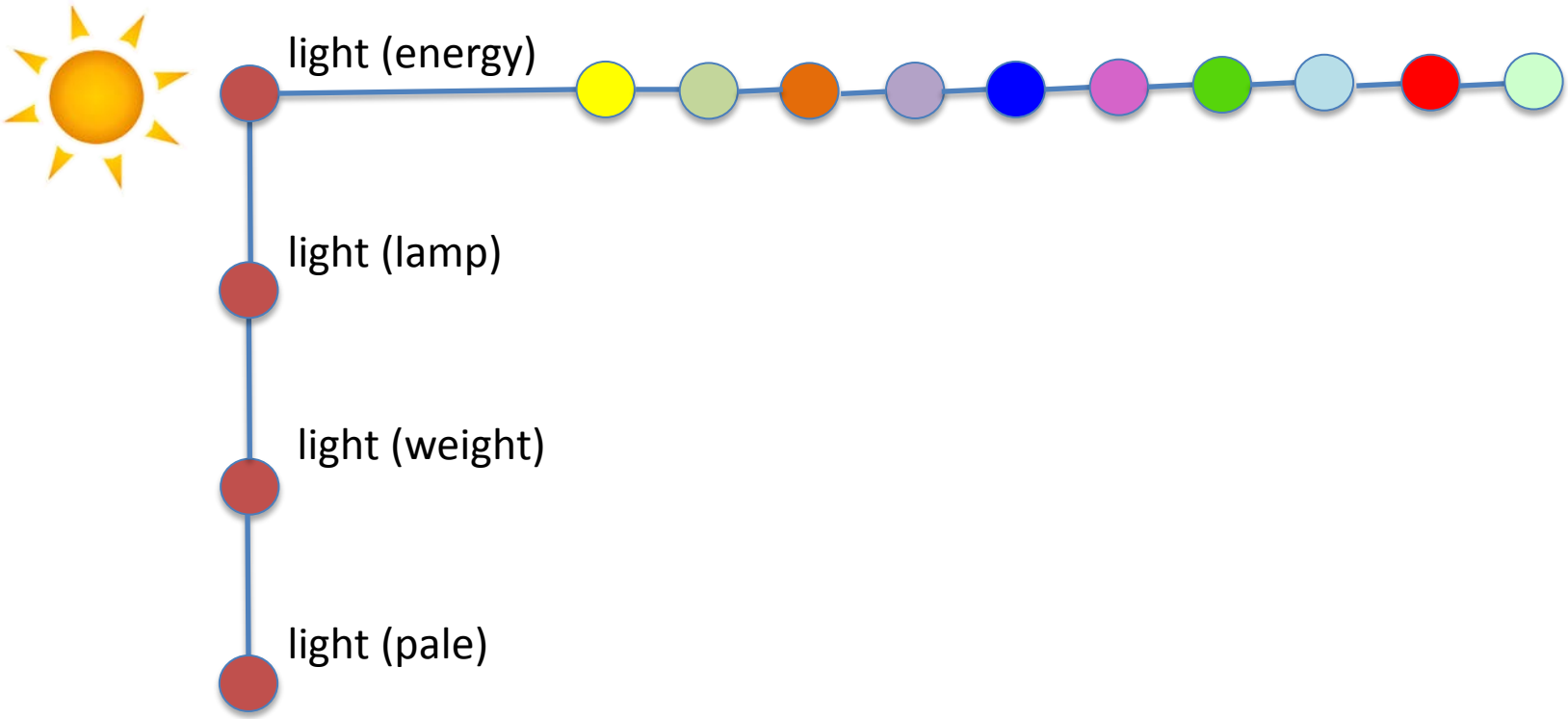
how Kamusi makes a multilingual dictionary possible



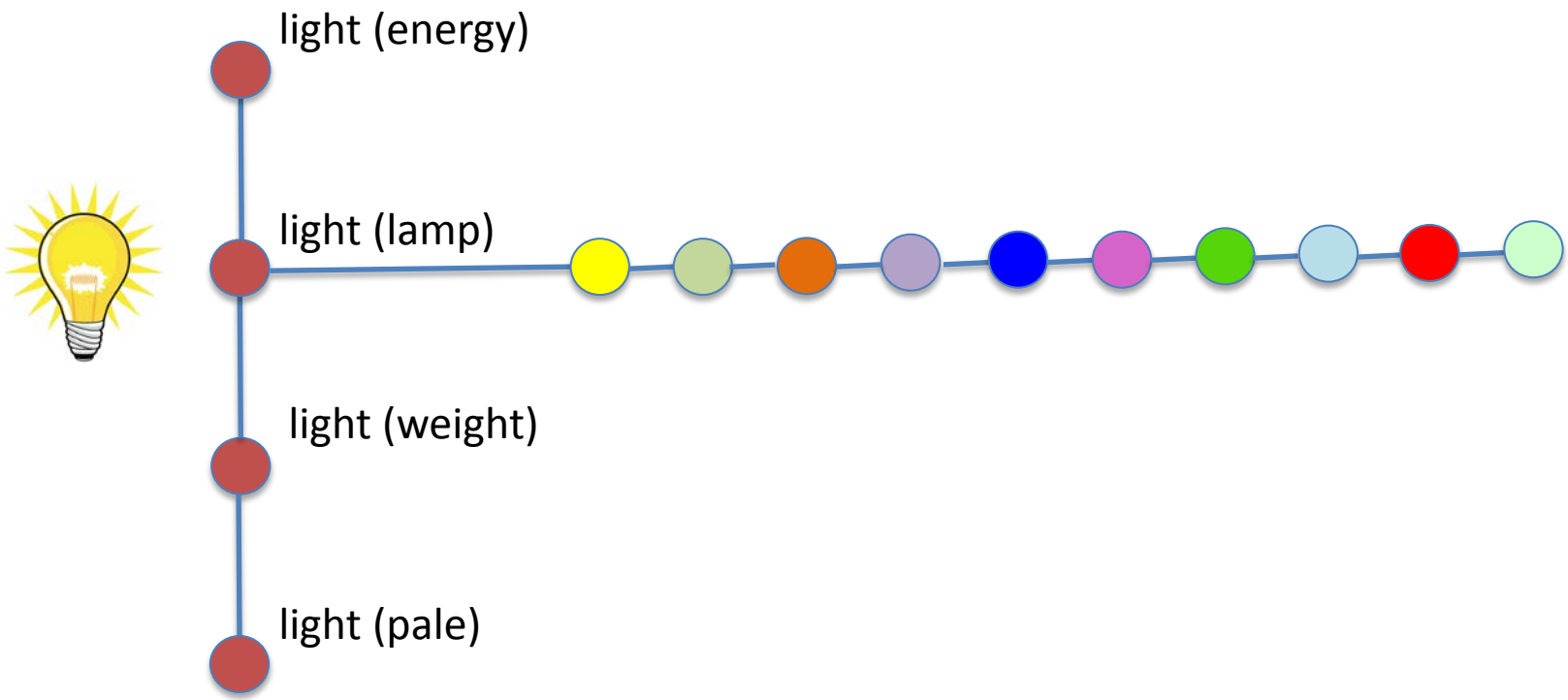
how Kamusi makes a multilingual dictionary possible



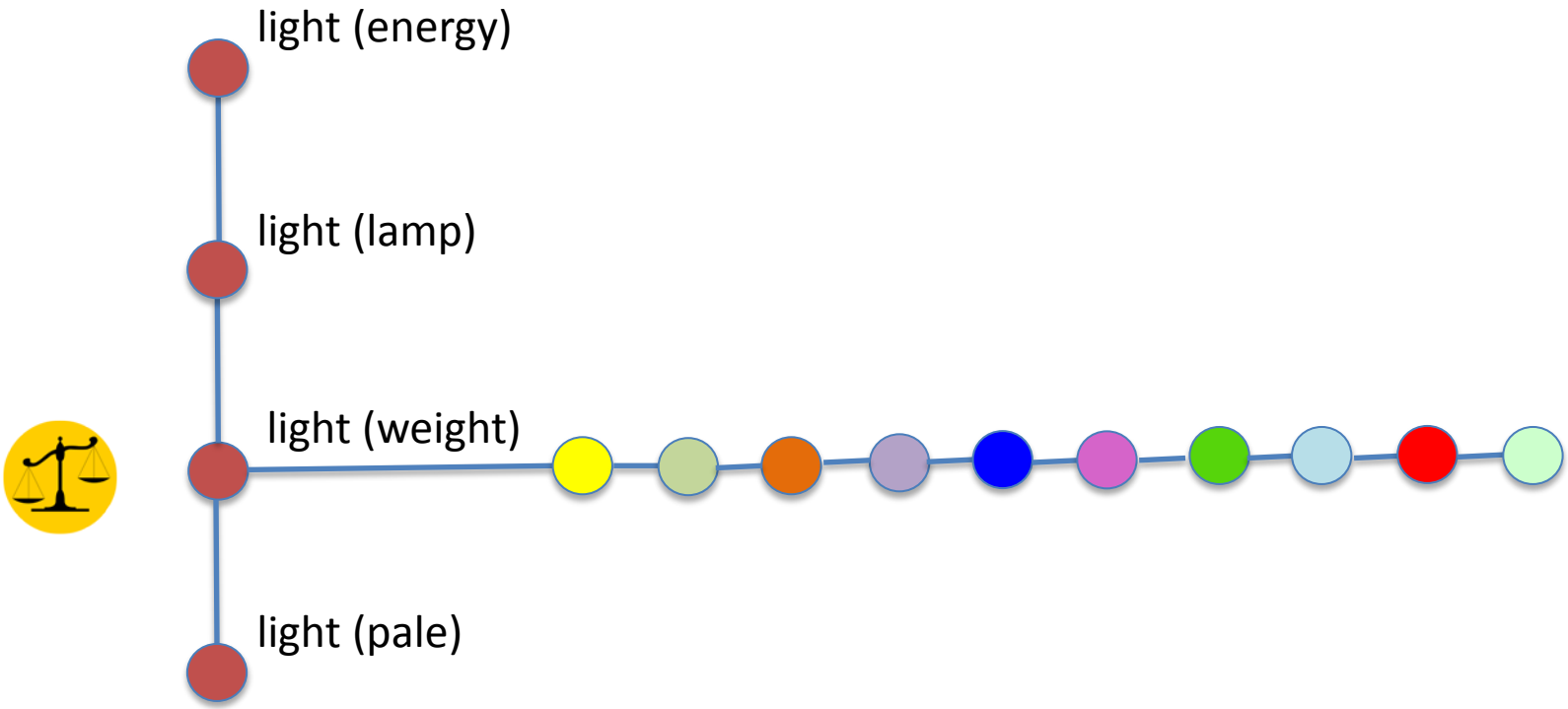
how Kamusi makes a multilingual dictionary possible



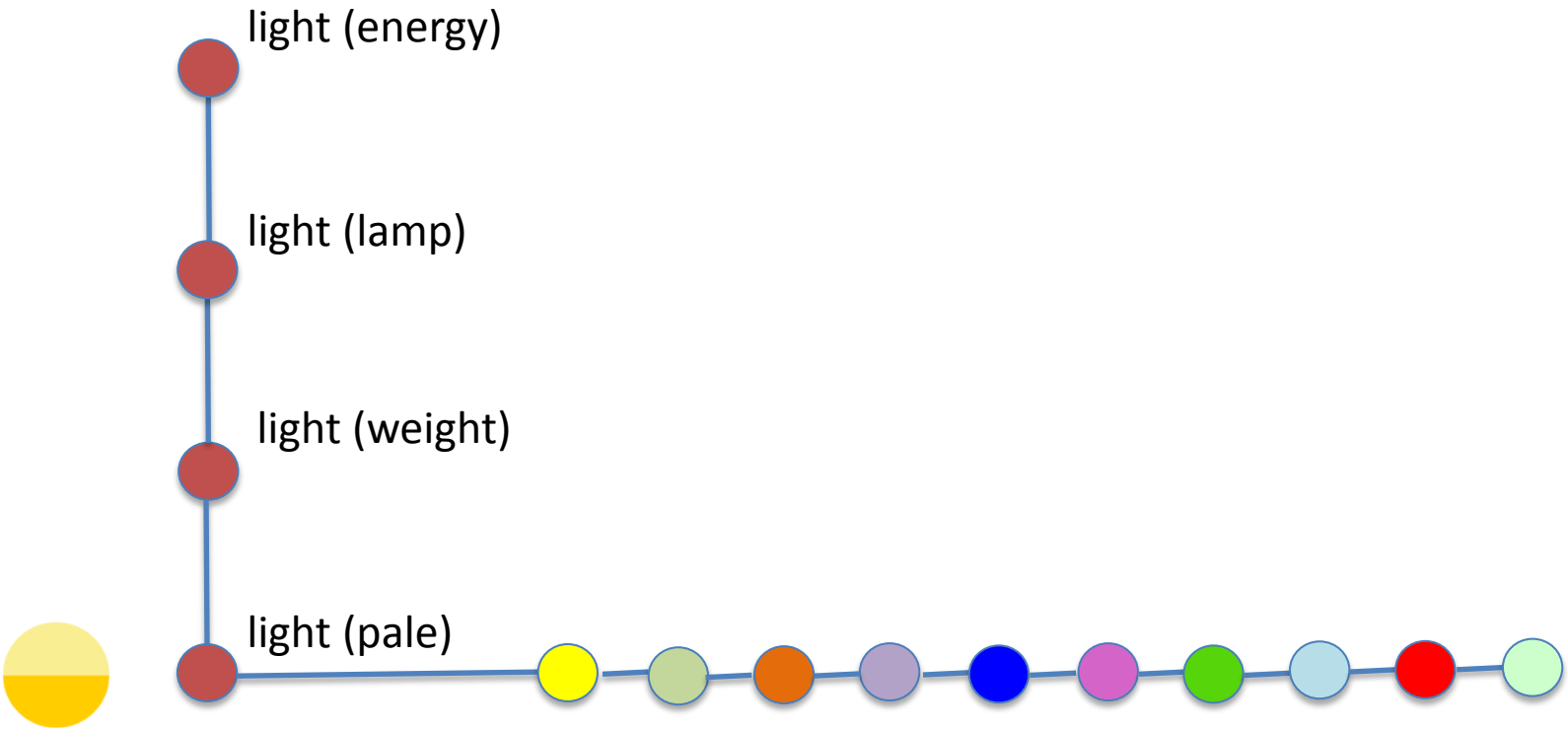
how Kamusi makes a multilingual dictionary possible



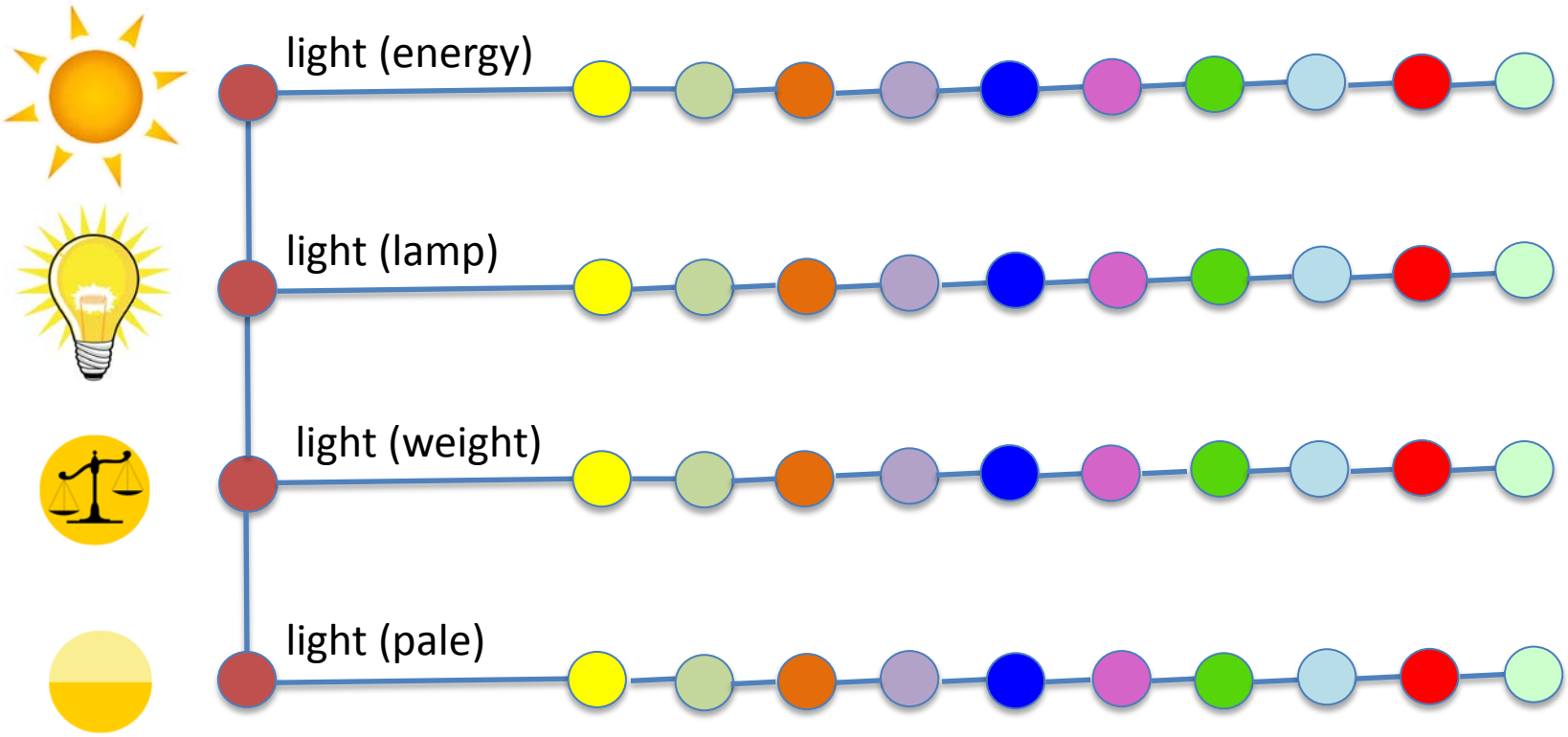
how Kamusi makes a multilingual dictionary possible



how Kamusi makes a multilingual dictionary possible

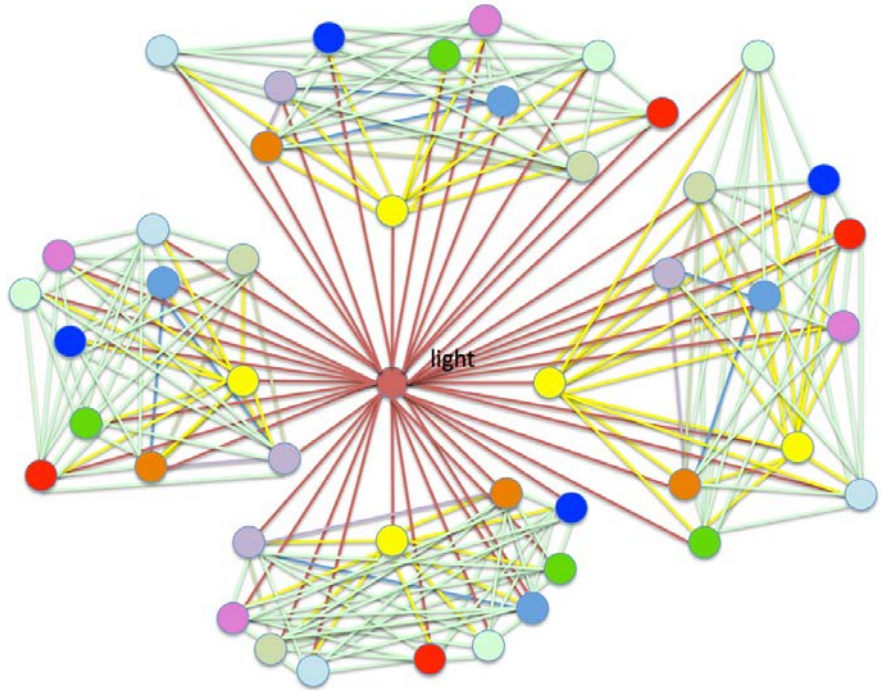


how Kamusi makes a multilingual dictionary possible



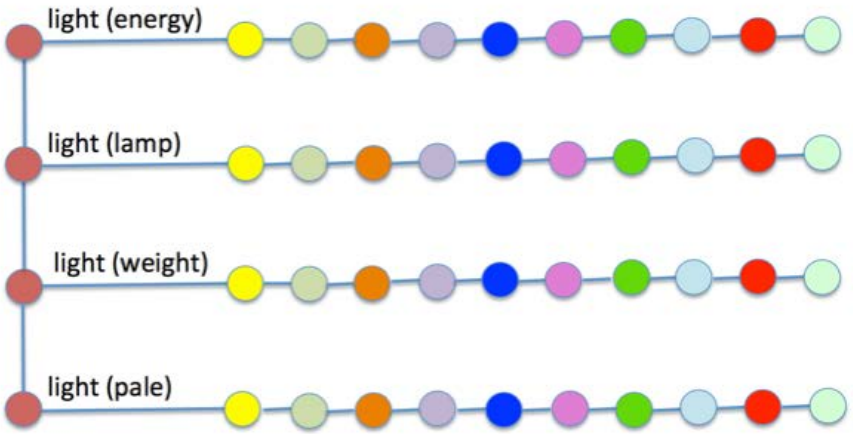
how Kamusi concept linking makes a multilingual dictionary possible

impossible



light

possible



light

View current | Edit latest | Revisions | Devel

light

View in context | Add an example | Add a translation

English noun

Definition: A device that provides illumination by burning electricity, fuel, or wax.

Terminology: general

Plural	lights

My Languages

- **itara** in Kirundi
- **taa** in Swahili
- **lumière** in French
- **lumină** in Romanian

Other Languages

- **taala** in Luganda

licht

View current | Edit current | Revisions | Devel

licht

View in context | Add an example | Add a translation

Dutch zelfstandig naamwoord - noun (Dutch) het-woord, onzijdig (het)

Definition: Voorwerp dat licht produceert, gevoed door elektriciteit of door verbranding van was of een andere brandstof.

Examples:

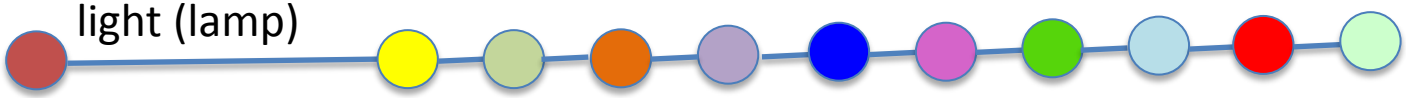
We gaan slapen. Doe jij het licht (de lichten) uit?

meervoud onzijdig (de)	lichten

Other Languages

- **lámpara** in Spanish
- **bugum** in Mampruli
- **-tara** in Kinyarwanda
- **بتی** in Urdu
- **oborabu** in Gusii
- **Licht** in German

light (lamp)

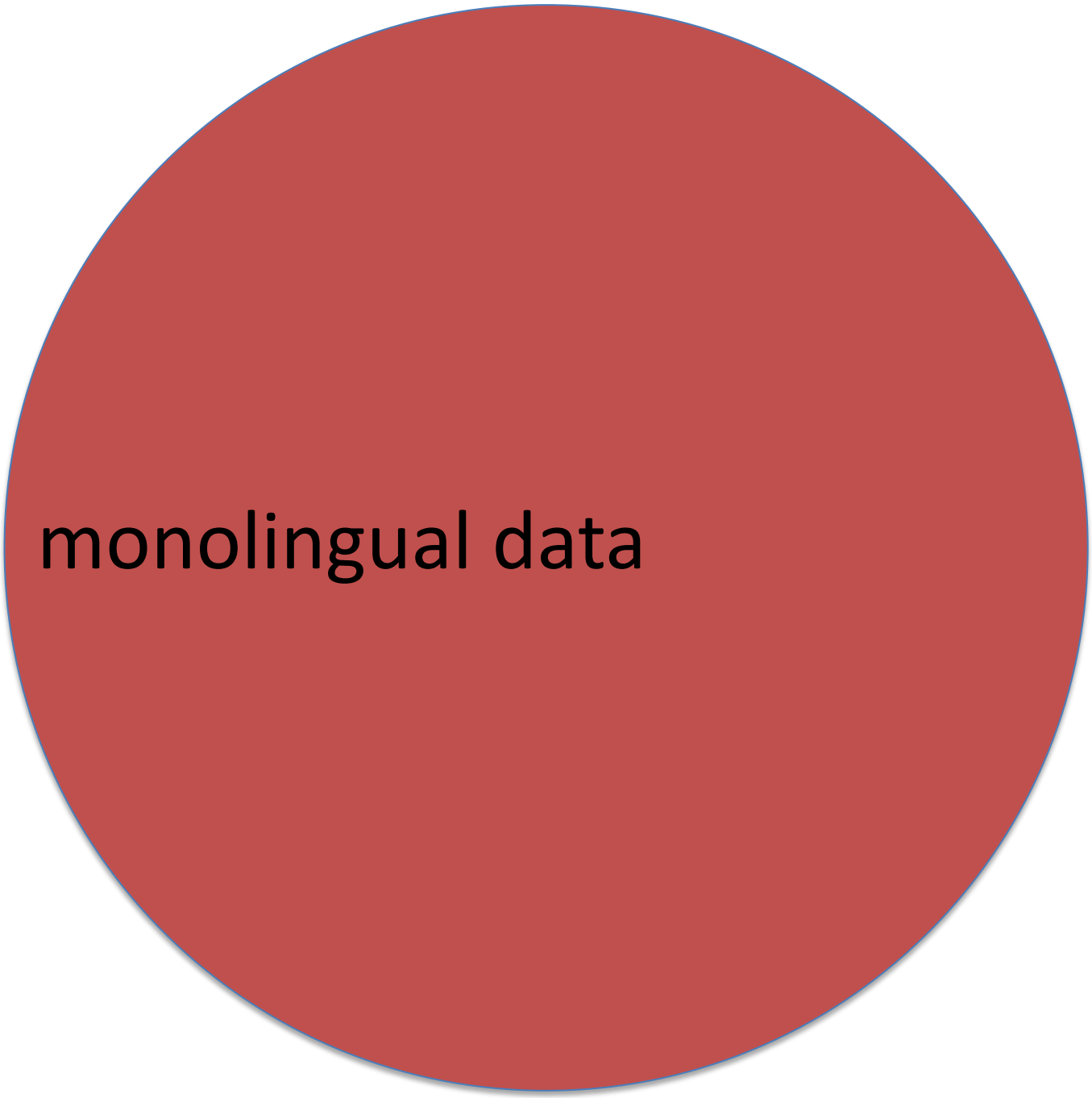


lumină

→ 1° **lumière** (French)

→ 2° **light** (English) → 3° **taa** (Swahili)

→ 3° **Licht** (German) → 4° **światło** (Polish)

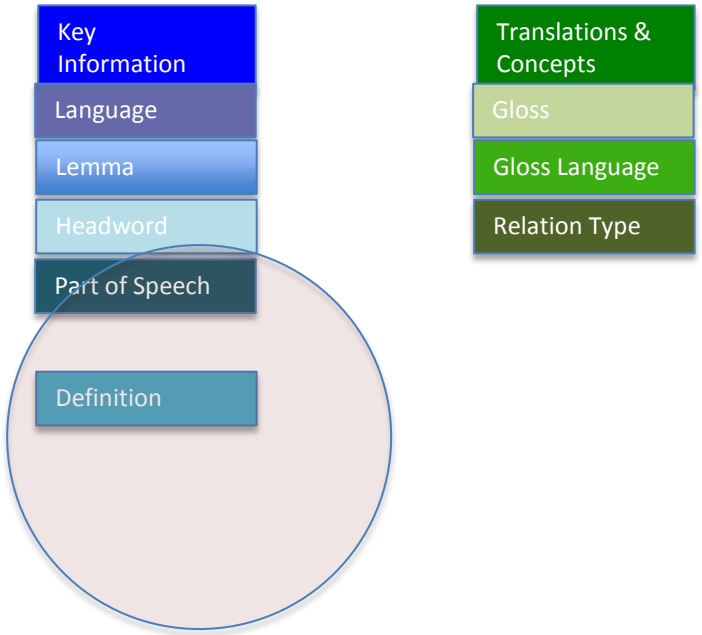

monolingual data

depth: kamusi (unlimited)

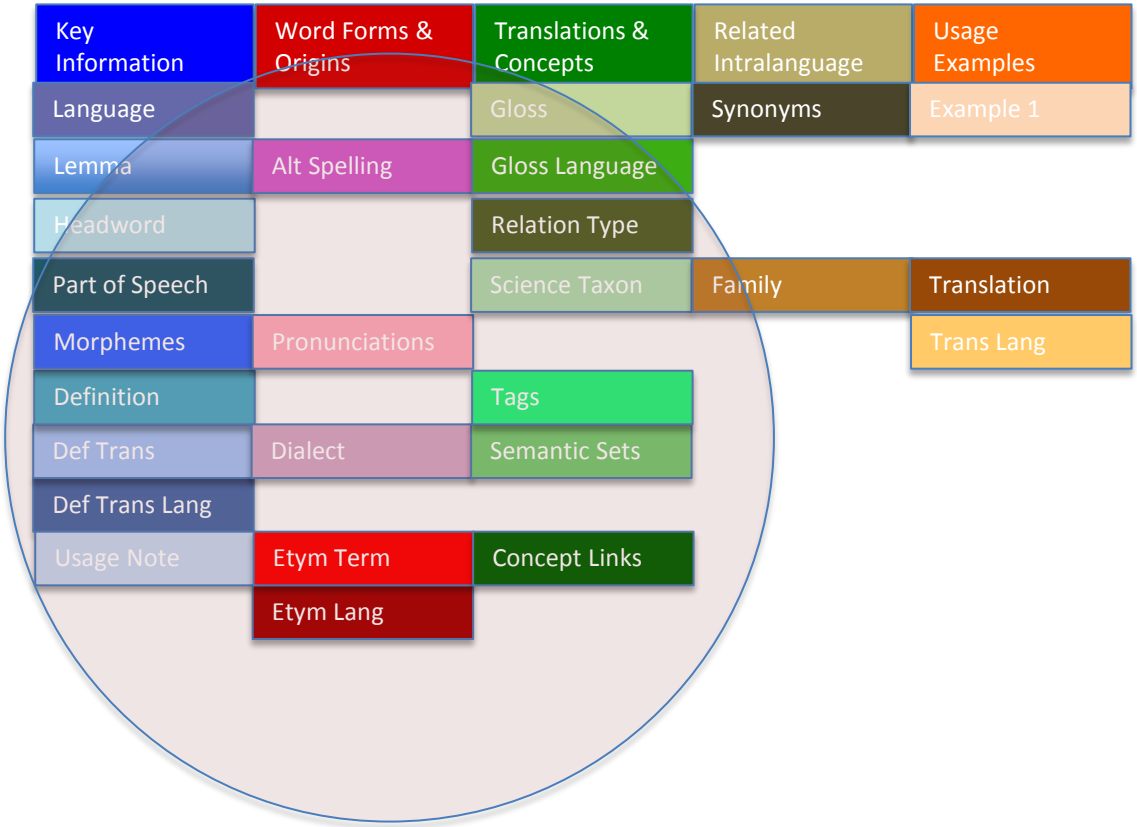
Key Information	Word Forms & Origins	Translations & Concepts	Related Intralanguage	Usage Examples
Language	Phonetic IPA	Gloss	Synonyms	Example 1
Lemma	Alt Spelling	Gloss Language	Antonyms	Source
Headword	Tone Spelling	Relation Type	Spawn	Audio
Part of Speech	Alt Script	Science Taxon	Family	Translation
Morphemes	Pronunciations	Term Set	Misspellings	Trans Lang
Definition	Pron Geo-Tag	Tags		Example X
Def Trans	Dialect	Semantic Sets		Source
Def Trans Lang	Word Geo-Tag	Image		Audio
Usage Note	Etym Term	Concept Links		Translation
Cultural Note	Etym Lang			Trans Lang
Special Note	Etym Note			

growth: a “living” dictionary

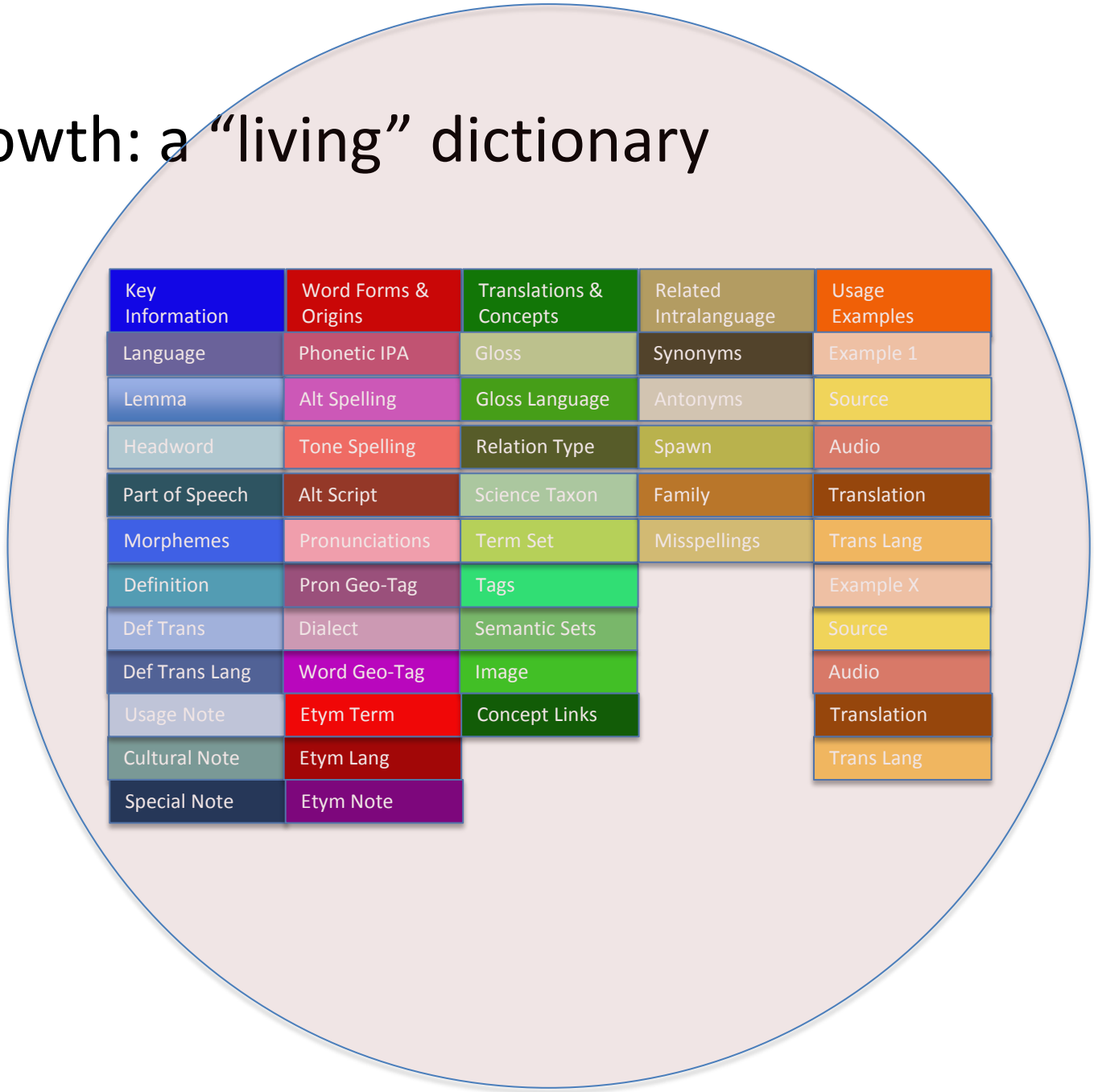
growth: a “living” dictionary



growth: a “living” dictionary



growth: a “living” dictionary



Edit Dictionary Term licht

View current | Edit current | Revisions | Devel

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

Concept

Term (lemma) *
licht

Term Language *
Dutch

Headword *
(?) licht

Part of Speech *
zelfstandig naamwoord - noun

Part Attribute
het-woord, onzijdig (het)

Morpheme Information

Collected Morphemes

meervoud mannelijk (de)

enkelvoud mannelijk de-woord (de-m)

enkelvoud onzijdig het-woord (het)

enkelvoud vrouwelijk de-woord (de-v)

meervoud vrouwelijk (de)

meervoud onzijdig (de)
lichten [nid:263256]

Definition

Definition of the term in the language to which it belongs
(?)
Voorwerp dat licht produceert, gevoed door elektriciteit of door verbranding van was of een andere brandstof.

Definition Source URL

Definition Translations

and more…

- Morphemes and inflections
- Attributes
- Behaviors
- Grouping and ranking
- Agglutination parsing
- Grammatical bridges
- Family relations
- Etymology and temporality
- Geography
- And more…

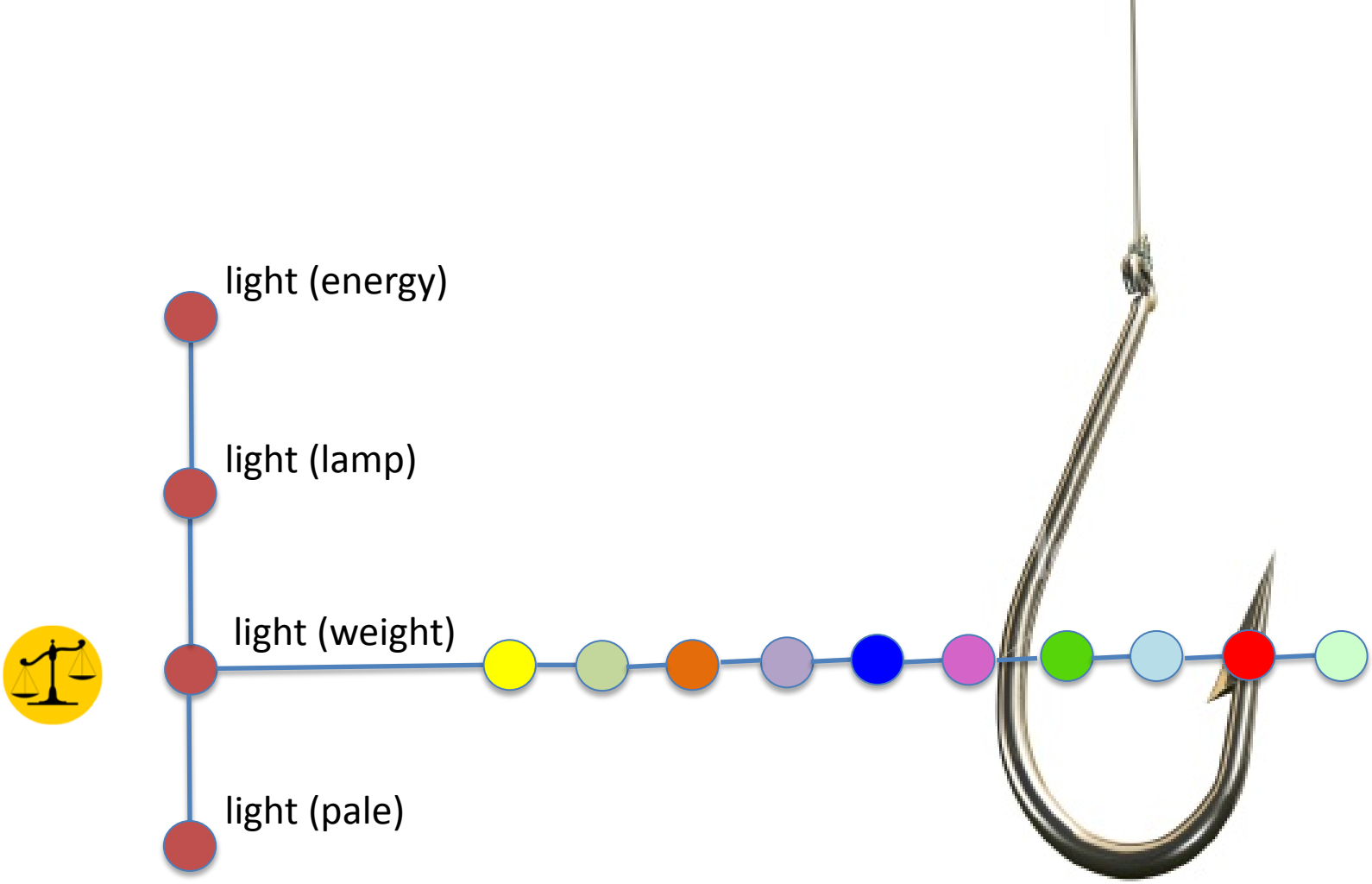
hooks

hooks

Resource
Description
Framework

hooks

Open Language Resources


light (energy)
light (lamp)
light (weight)
light (pale)

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

1. Project objectives
2. **Acquiring data**
3. Crowdsourcing innovations

- For people
- For machines

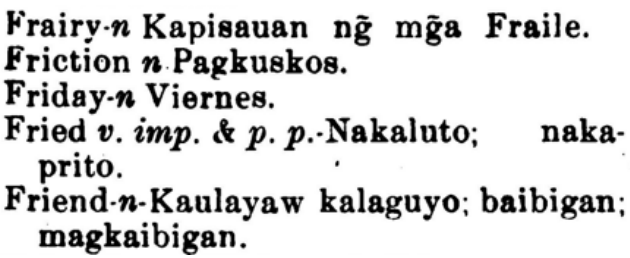
acquiring data

- With people
- With machines

acquiring data

acquiring data

- From existing data
- From experts
- From the crowd



Frairy-*n* Kapisauan nḡ mḡa Fraile.
Friction *n* Pagkuskos.
Friday-*n* Viernes.
Fried *v. imp.* & *p. p.*-Nakaluto; naka-
prito.
Friend-*n*-Kaulayaw kalaguyo; baibigan;
magkaibigan.

```
Frairy-n Kapisauan ng mga Fraile.
Friction n Pagkuskos.
Friday-n Viernes.
Fried v. imp. & p. p.-Nakaluto; naka-

prito.
Friend-n-Kaulayaw kalaguyo; baibigan;

magkaibigan.
```

acquiring data

- existing data

Challenges of existing data for LRLs:

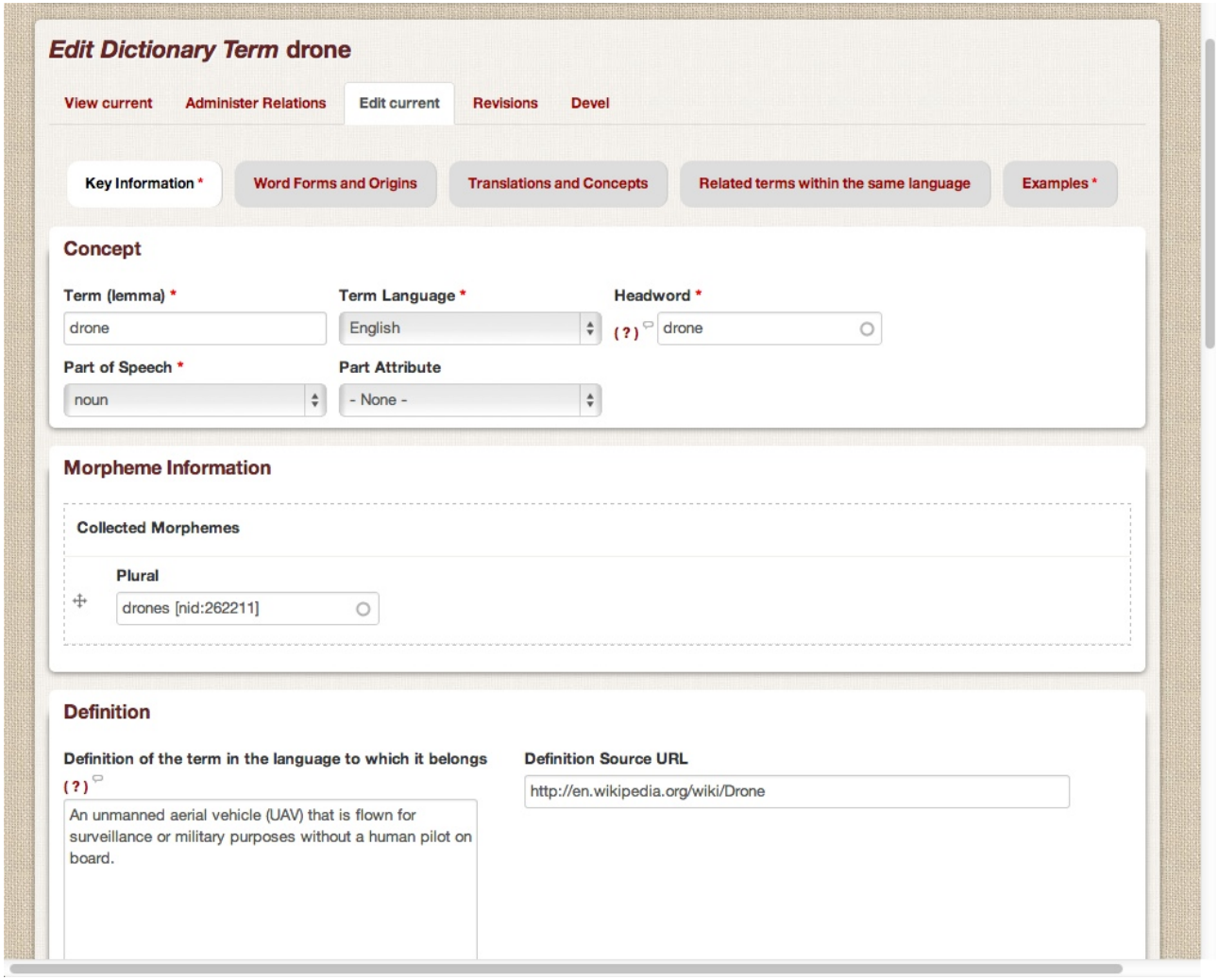
- Paltry
  - Few data items
  - Few data elements
- Poorly structured
- Inconsistently structured
- No OCR training
- IP restrictions

And then:

- Non-disambiguated

Solution – Merging Engine:

- Get data into consistent fields
- Human review for sense alignment



Challenges of language specialists for LRLs:

- Limited # of languages
- Experts have limited time
- Expertise costs money
  - 10,000 entries = 1 year of labor

Expert Tool – Edit Engine:

- Easy
- Consistent
- Comprehensive

acquiring data • experts

acquiring data • the crowd

Challenges of crowdsourcing for LRLs:

- Eliciting specific data
- Keeping things simple
- Validating data without "ground truth"
  - Minor errors
  - Wrong data
- Malicious users
  - Lots of ways to make mischief
- Sparking participation
- Crowd size
- Maintaining participation

Solutions:

- Fidget Widget
- Facebook Game

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

1. Project objectives
2. Acquiring data
3. **Crowdsourcing innovations**

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Purposes:

- Gather original data
- Manipulate existing data
- Validate existing data
- Validate new data

crowdsourcing

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Methods:

- Targeted questions
- Short questions
- Engaging games
- Motivational incentives

crowdsourcing

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Tools:

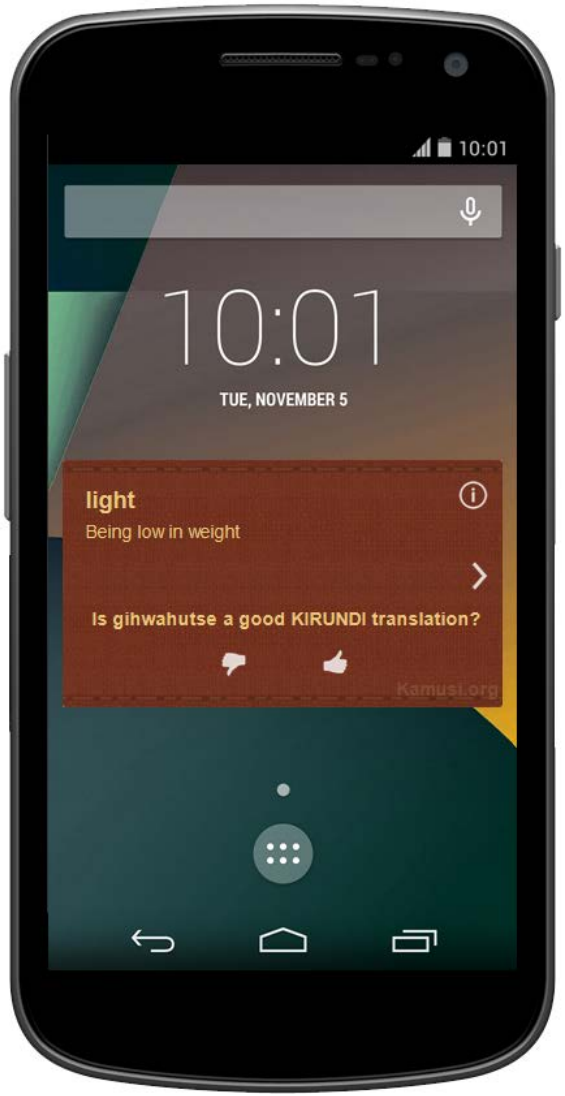
- Play to Pay
- Fidget Widget (mobile app)
- Facebook Game

crowdsourcing

Fidget Widget - Logic:

- **Single word answers**
- Thumbs up/ down
- Longer answers for advanced users

crowdsourcing



Fidget Widget - Logic:

- Single word answers
- **Thumbs up/ down**
- Longer answers for advanced users

crowdsourcing

Fidget Widget – Logic:

- Single word answers
- Thumbs up/ down
- **Longer answers for advanced users**

crowdsourcing

crowdsourcing

Fidget Widget - Issues:

- Synchronization
  - Offline in LRL settings
- Security
  - User accounts
- Competing answers
  - Noise vs. signal
- Publication
  - Deciding a quality threshold

Fidget Widget – v 2.0:

Talking dictionary

- Voice recording
- Spoken vignettes as definitions
- Spoken translations of definitions

crowdsourcing

Gamification (Fall 2014)

Facebook

- Where the people are
- Where their friends are

The Games

- Pyramid of points for winning answers
- Team points for fastest languages
- Difficulty points for stumpers

crowdsourcing

partners

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Contribution analysis

- Good answers – build trust
- Bad answers – change the game
- Malicious answers – remove the user's history

crowdsourcing

languages

5 year goal:
100 languages worldwide

words

5 year goal:
10,000 parallel terms
in each language

data

5 year goal:
100,000,000 searches/ month

peopl

5 year goal:
1,000,000 registered users

e

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Martin Benjamin, EPFL

Network: http://www.linkedin.com/in/martinbenjamin
Write: martin.benjamin@epfl.ch
Meet: http://meetme.so/kamusi