

THE KAM4D LINGUISTIC KNOWLEDGE GRAPH:
PUTTING SMURFS, DUCKS, LEMURS, AND PARTY TERMS
TO THE SERVICE OF AFRICAN LANGUAGES



kamusi.org

Martin Benjamin

SADiLaR: DH Colloquium 17 November, 2021
South African Centre for Digital Language Resources



kamusi is Swahili for *dictionary*





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

- As a knowledge resource
- As a data resource
- As a basis for any-to-any translation



In service since 1994 - originally at **Yale Council on African Studies**

International NGO since 2009

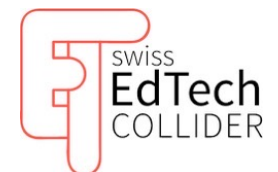
- Registered non-profit in  and 

Academic Home since 2013:

EPFL - Swiss Federal Institute of Technology in Lausanne

First at **LSIR** - Distributed Systems Information Laboratory

Now at the **Swiss EdTech Collider**





ACALAN (Intergovernmental language agency for 55 member states of the African Union):
Platform for African Language Empowerment development partner



Kamusi Jargon – essential terms for Kam4D

Lemur =

- Lemma
- Lemmatic form
- Dictionary form
- Canonical form
- Citation form



- Lemurs and Party Terms
- Smurfs and Ducks
- Costumes and Wardrobes



Party term =

- Multiword Expression
- MWE

Kamusi Jargon – essential terms for Kam4D

SMURF =
Spelling/ Meaning
Unit Reference



- Lemurs and Party Terms
- Smurfs and Ducks
- Costumes and Wardrobes

DUCKS = Data
Unified Concept
Knowledge Set



THE KAM4D LINGUISTIC KNOWLEDGE GRAPH: PUTTING SMURFS, DUCKS, LEMURS, AND PARTY TERMS TO THE SERVICE OF AFRICAN LANGUAGES

1. The problem with linguistic data
2. The Kam4D solution
3. Kamusi Labs projects



THE KAM4D LINGUISTIC KNOWLEDGE GRAPH: PUTTING SMURFS, DUCKS, LEMURS, AND PARTY TERMS TO THE SERVICE OF AFRICAN LANGUAGES

1. The problem with linguistic data
2. The Kam4D solution
3. Kamusi Labs projects



Data: WORDS that have been
digitized in a way that can be
used within computer processes

This is a **WORD**



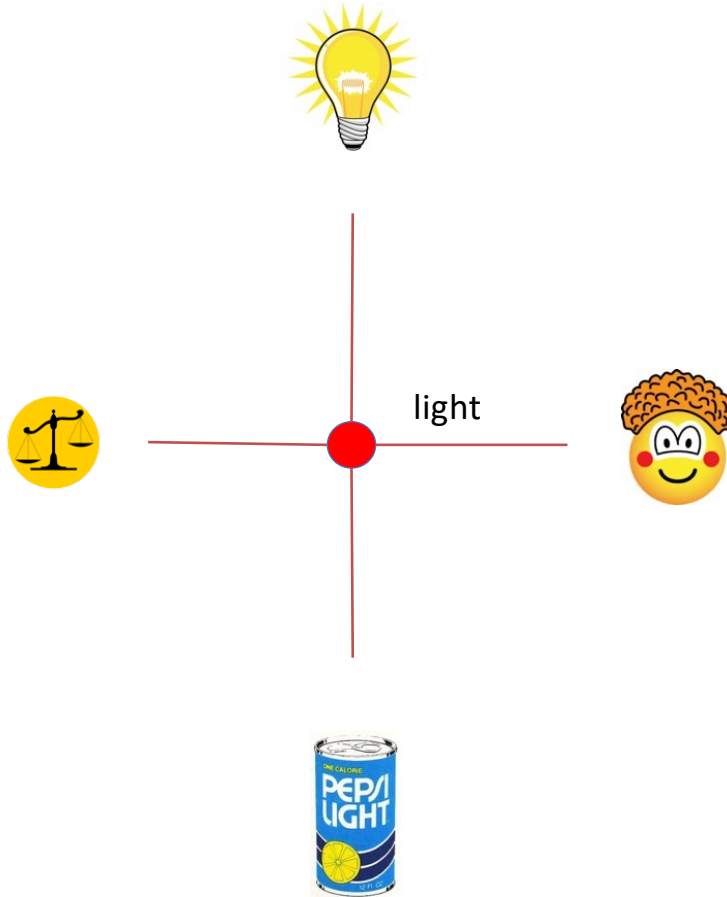
light

Lemur =

- Lemma
- Lemmatic form
- Dictionary form
- Canonical form
- Citation form

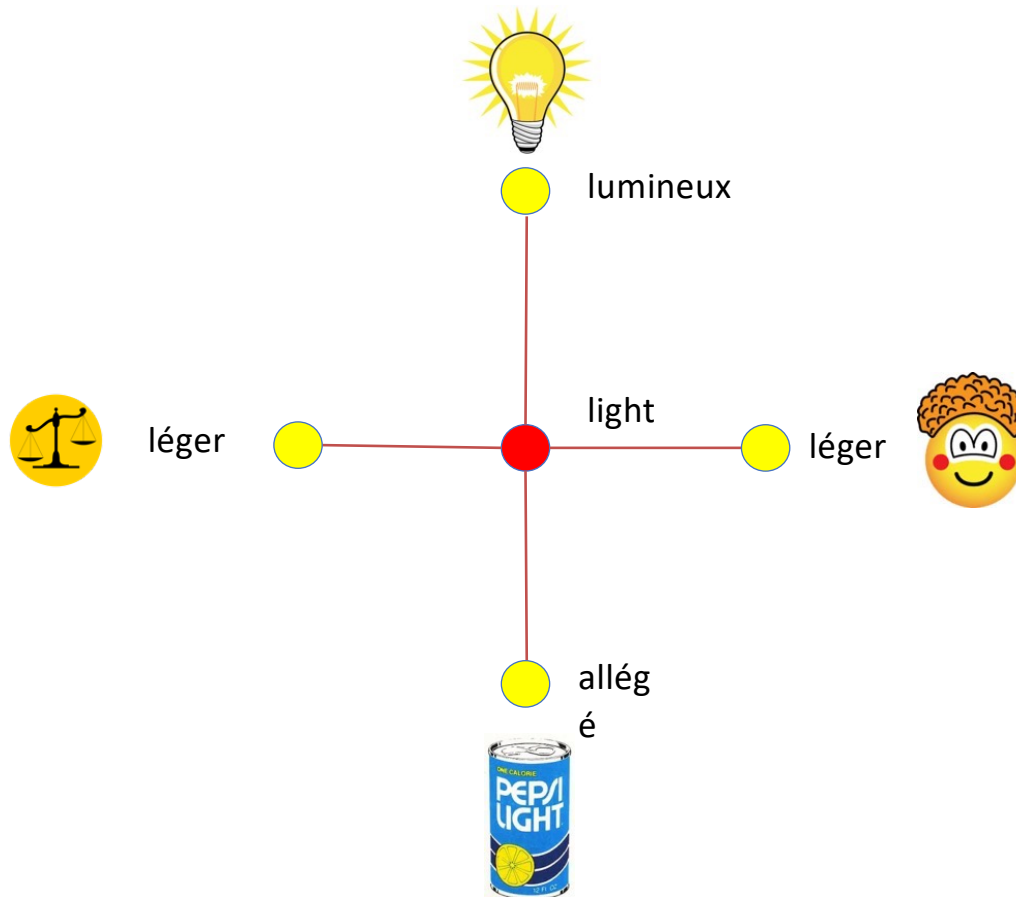


light



light

why multilingual dictionaries were impossible



why multilingual dictionaries were impossible



light up the town



red light district



give the green light



come to light



in light of



see the light



running light



light year



trip the light fantastic

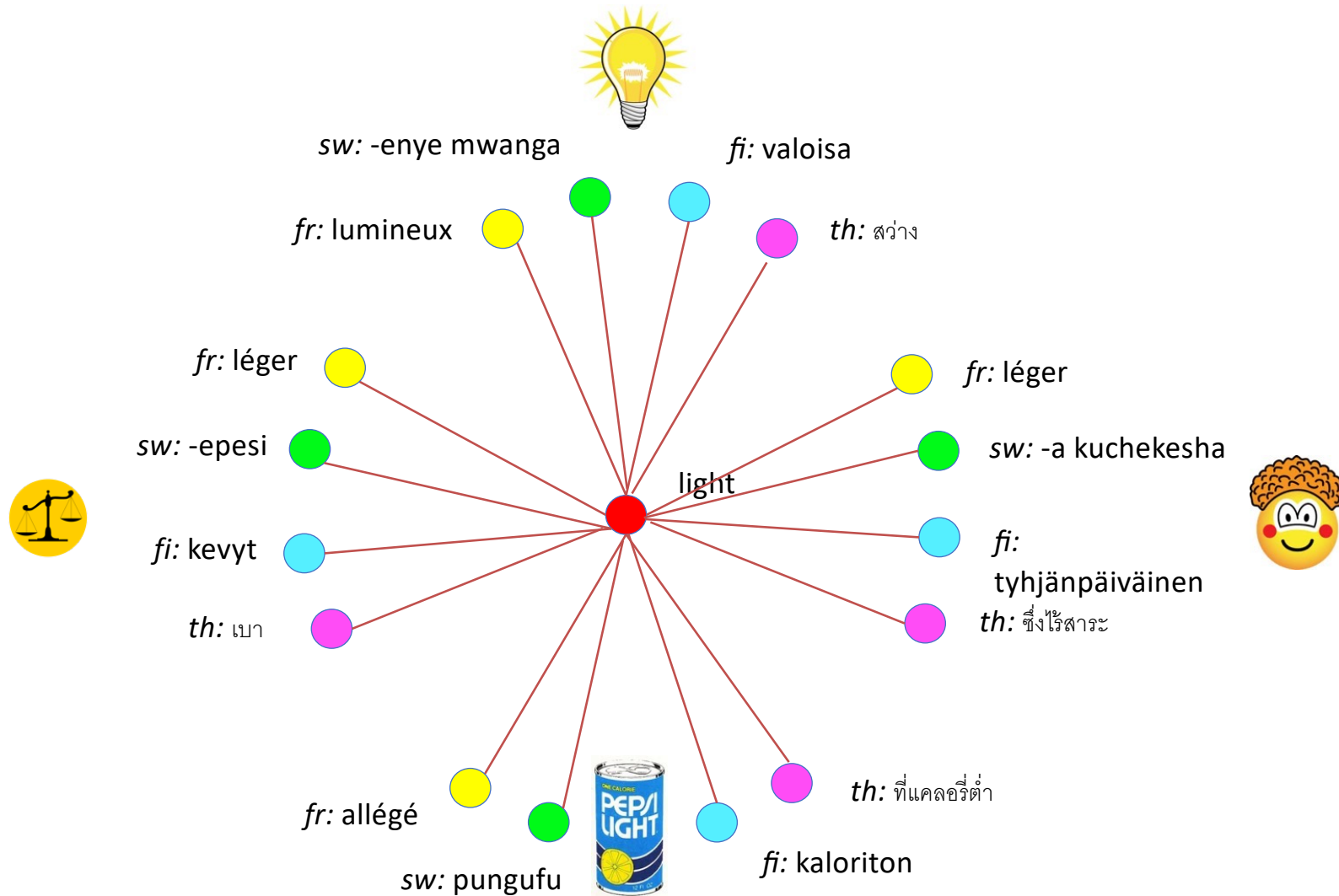


light

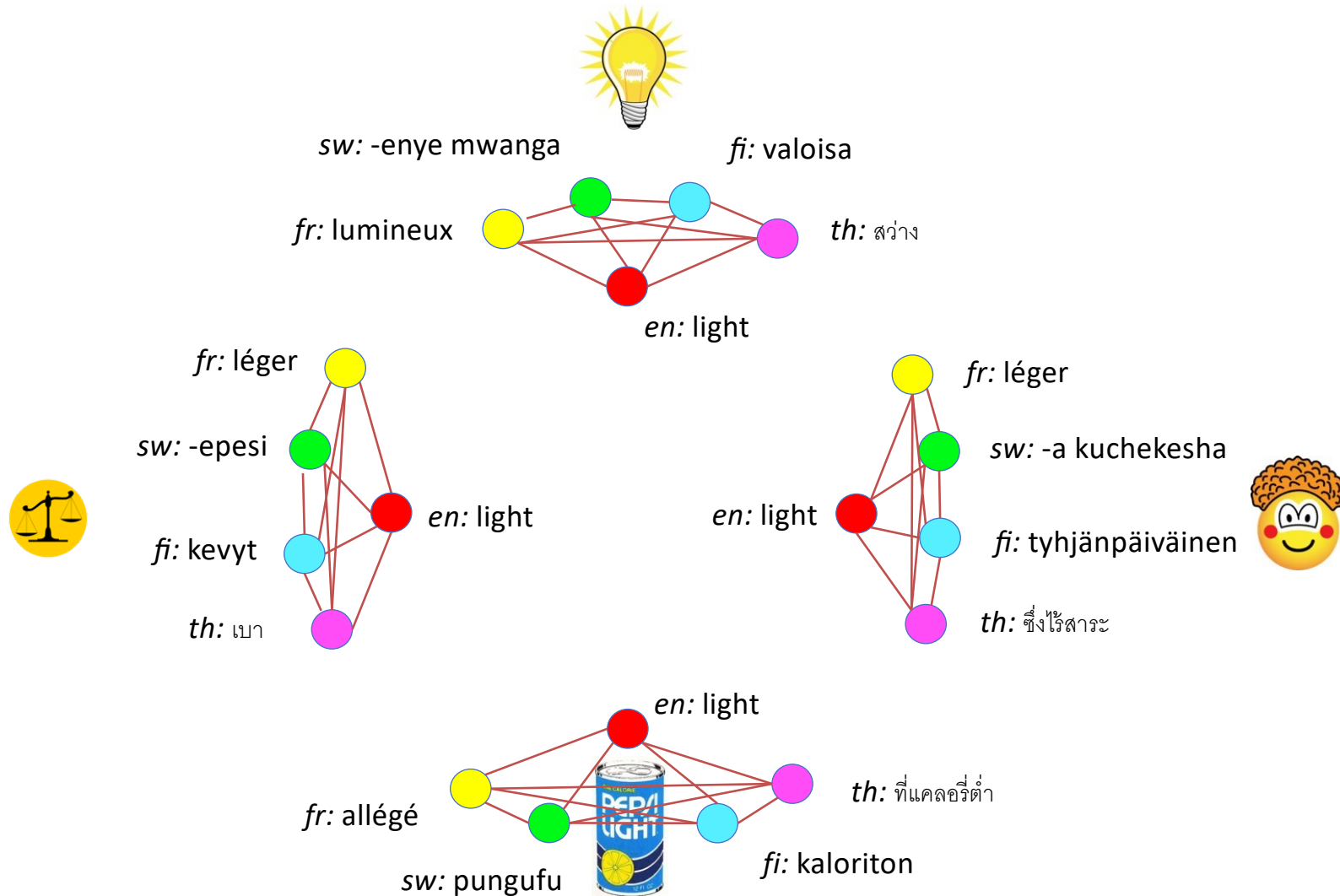


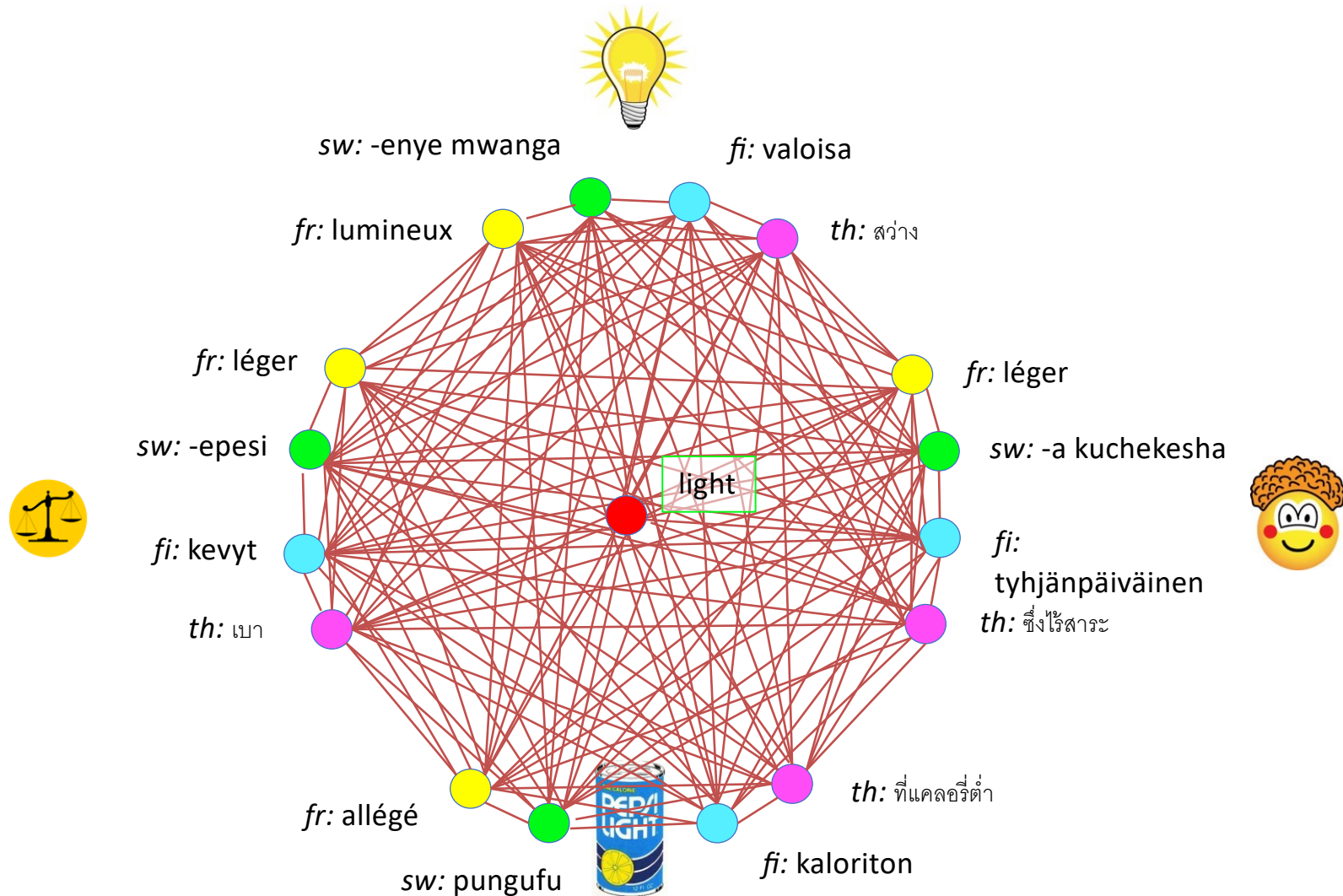
Party term =

- Multiword Expression
- MWE

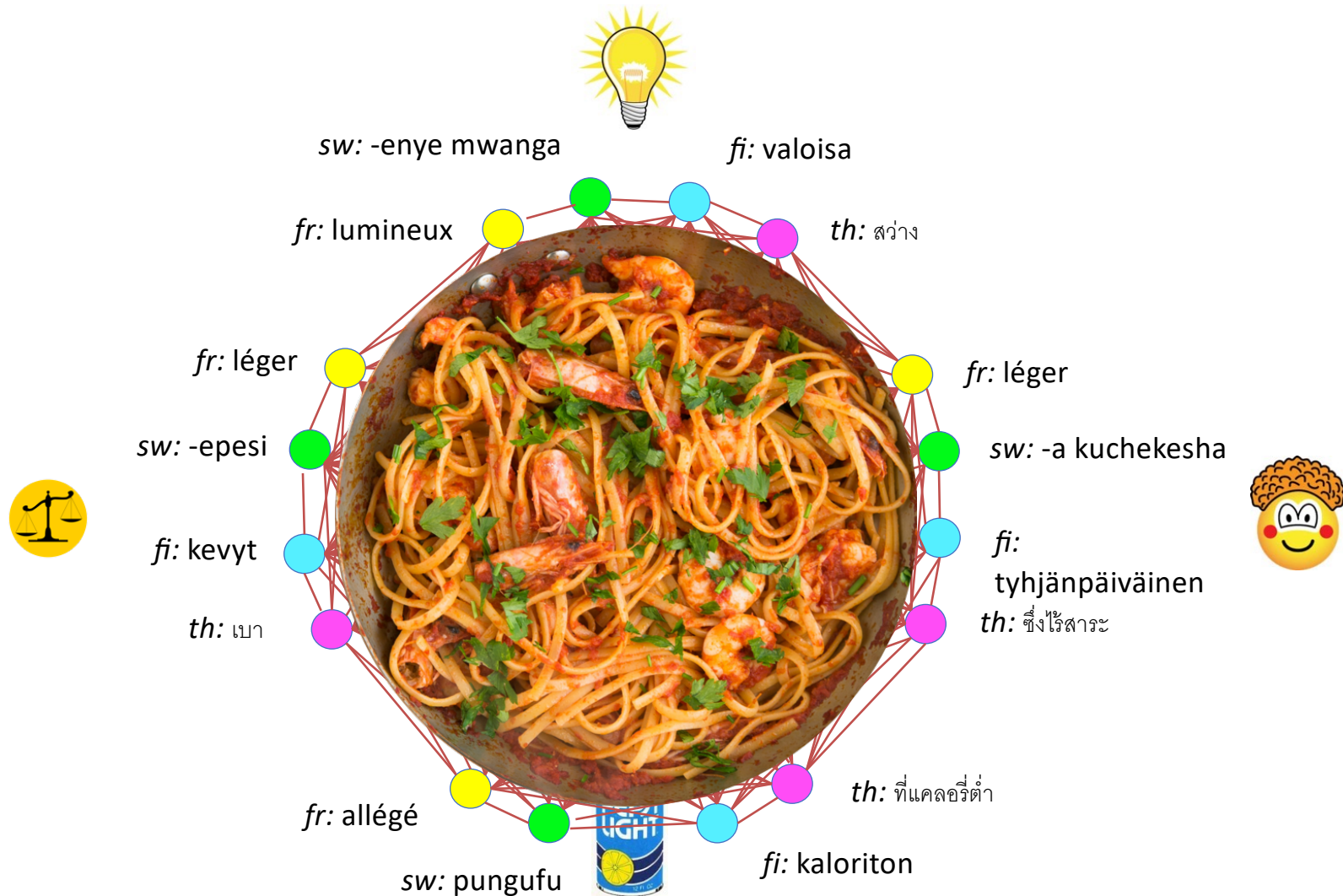


why multilingual dictionaries were impossible





why multilingual dictionaries were impossible



why multilingual dictionaries were impossible

THE KAM4D LINGUISTIC KNOWLEDGE GRAPH: PUTTING SMURFS, DUCKS, LEMURS, AND PARTY TERMS TO THE SERVICE OF AFRICAN LANGUAGES

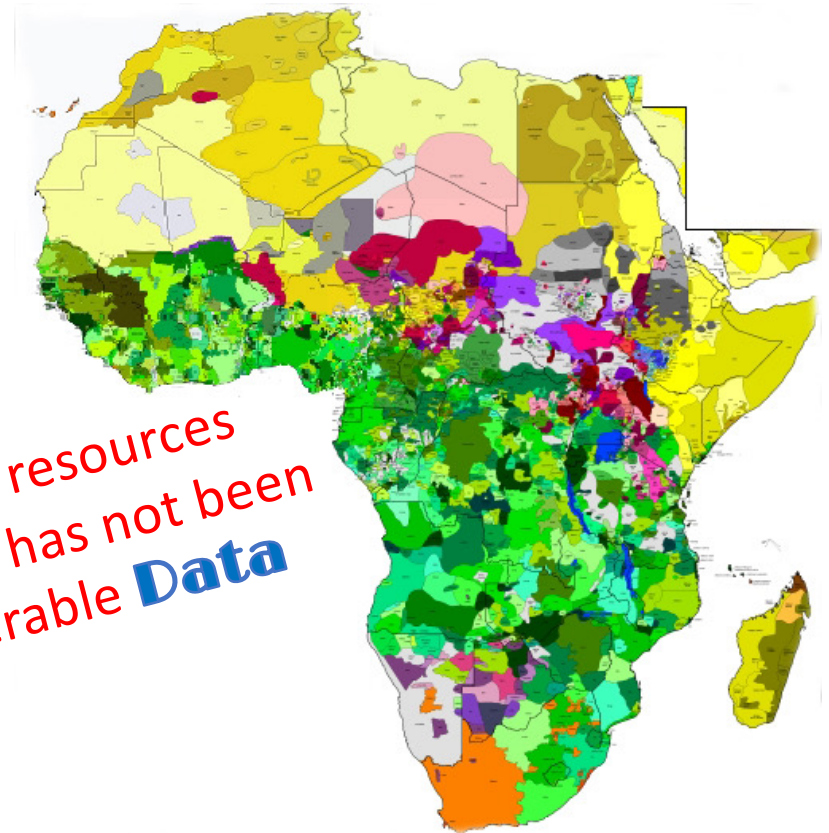
1. The problem with linguistic data
2. The Kam4D solution
3. Kamusi Labs projects



THE KAM4D LINGUISTIC KNOWLEDGE GRAPH: PUTTING SMURFS, DUCKS, LEMURS, AND PARTY TERMS TO THE SERVICE OF AFRICAN LANGUAGES

1. More problems with linguistic data!
2. The Kam4D solution
3. Kamusi Labs projects

- 2000 languages in Africa
- Very few have any digital resources
- What has been digitized has not been harmonized as interoperable **Data**



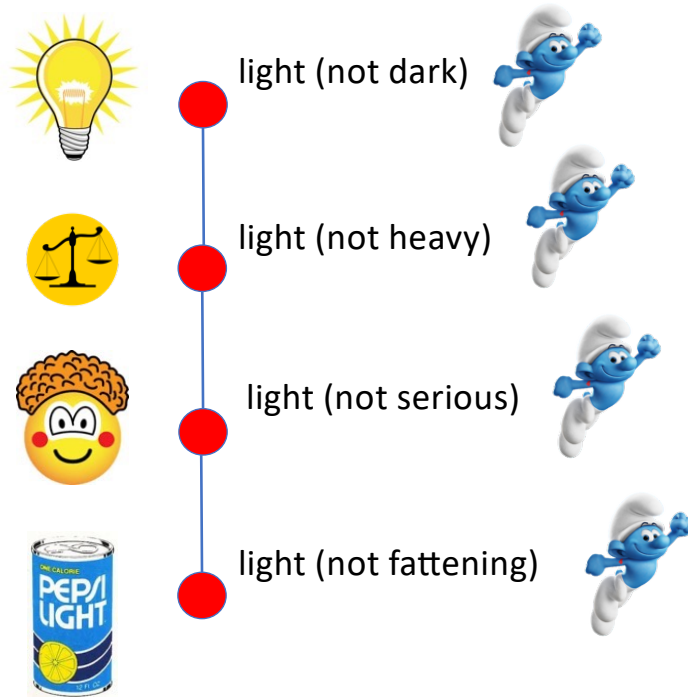
THE KAM4D LINGUISTIC KNOWLEDGE GRAPH: PUTTING SMURFS, DUCKS, LEMURS, AND PARTY TERMS TO THE SERVICE OF AFRICAN LANGUAGES

1. The problem with linguistic data
2. The Kam4D solution
3. Kamusi Labs projects



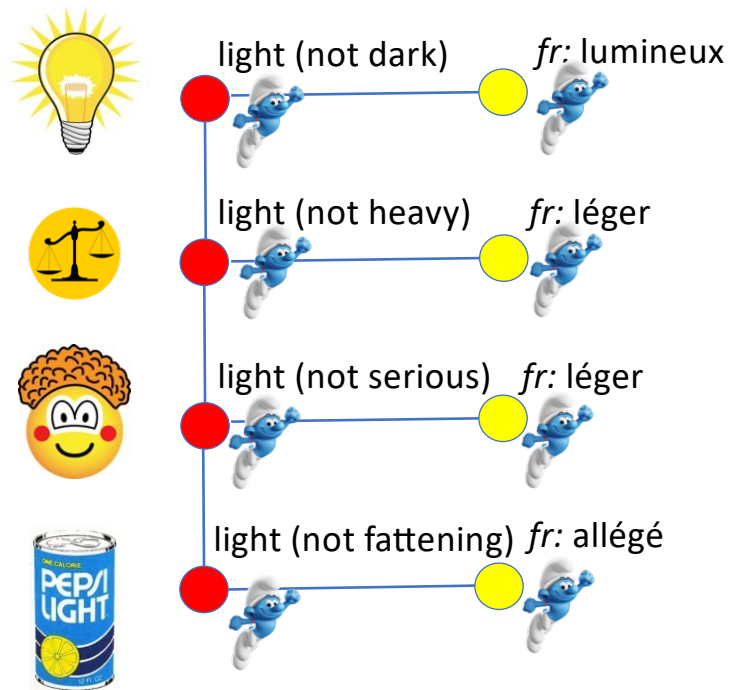


light



SMURF =
Spelling/ Meaning
Unit Reference





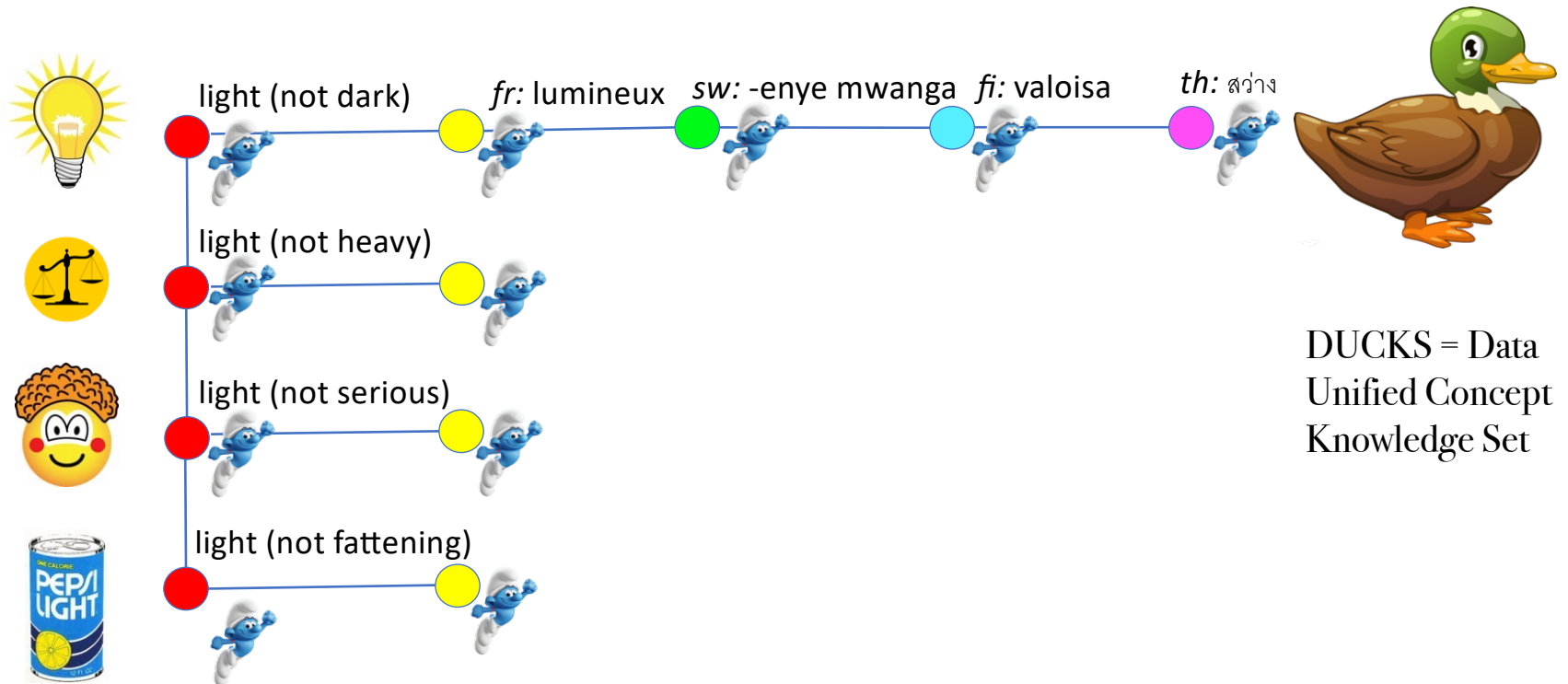
SMURF =
Spelling/ Meaning
Unit Reference

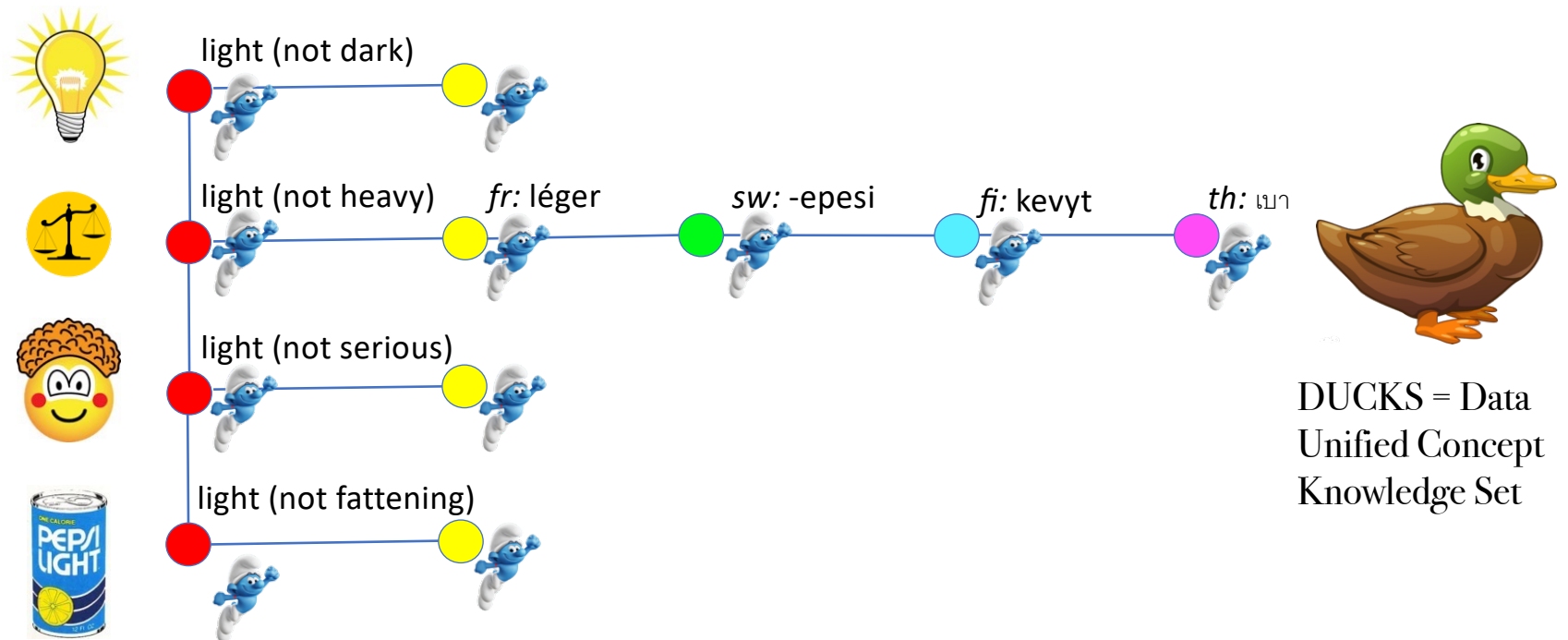


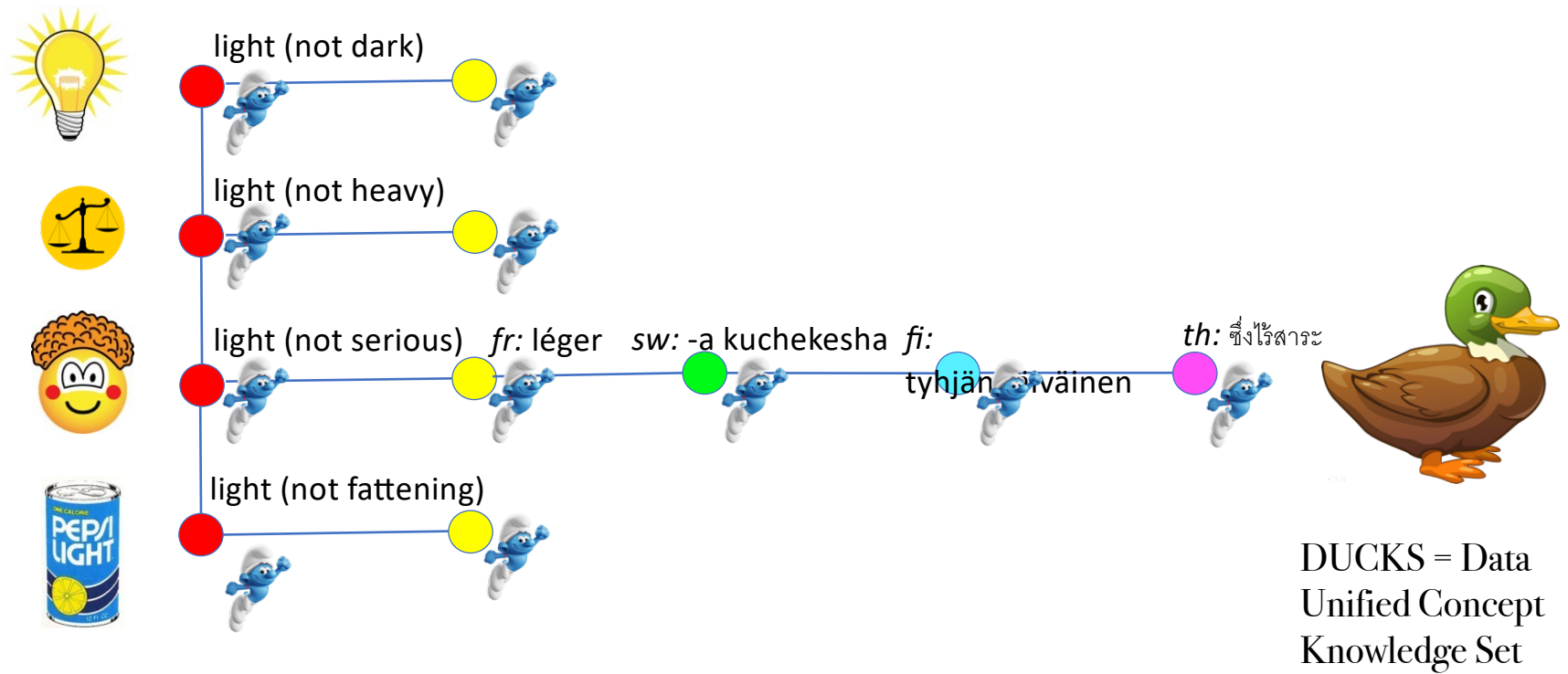
DUCKS = Data
Unified Concept
Knowledge Set

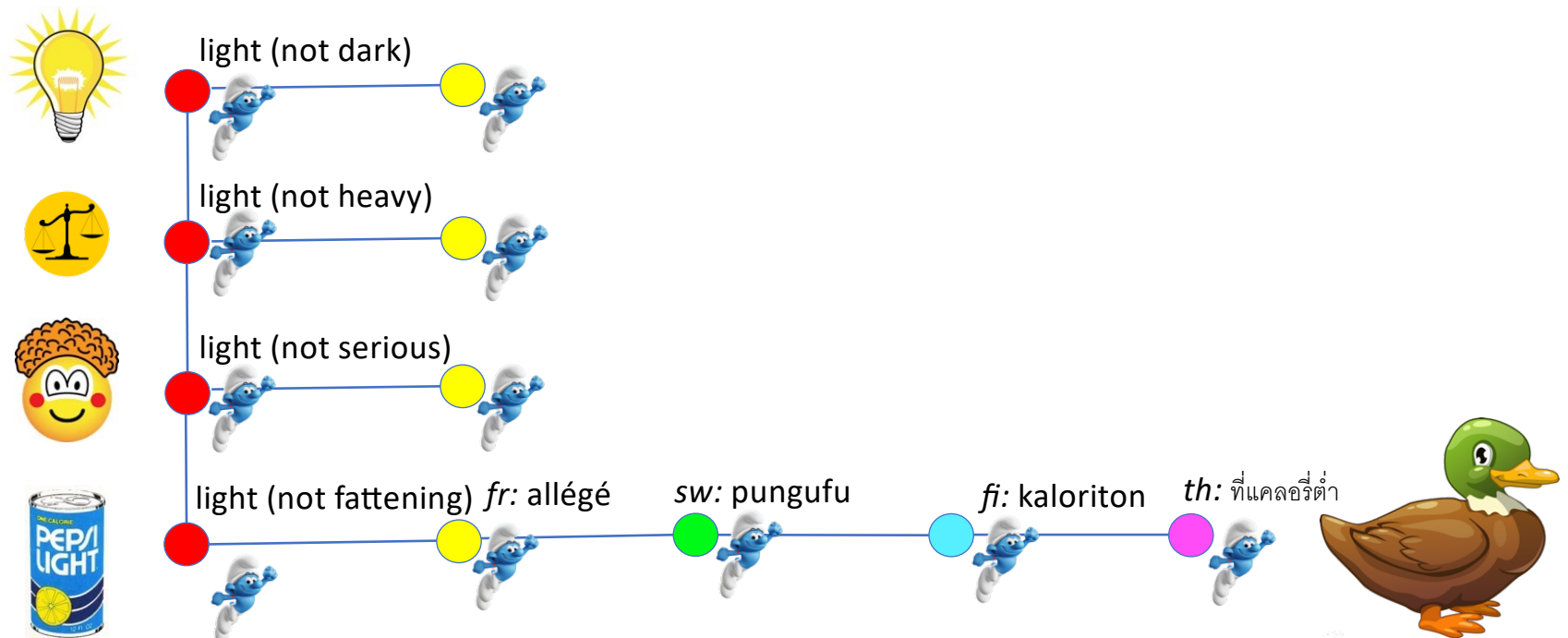


how Kamusi makes a multilingual dictionary possible

















DUCKS = Data
Unified Concept
Knowledge Set

	light (not dark)	<i>fr:</i> lumineux	<i>sw:</i> -enye mwanga	<i>fi:</i> valoisa	<i>th:</i> สว่าง	
	light (not heavy)	<i>fr:</i> léger	<i>sw:</i> -epesi	<i>fi:</i> kevyt	<i>th:</i> เบา	
	light (not serious)	<i>fr:</i> léger	<i>sw:</i> -a kuchekesha	<i>fi:</i> hölynpöly	<i>th:</i> ขี้เล่น	
	light (not fattening)	<i>fr:</i> allégé	<i>sw:</i> pungufu	<i>fi:</i> kaloriton	<i>th:</i> ที่แคลอรีต่ำ	





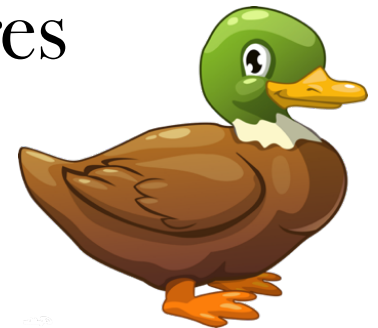
Lab Server Server

- 2,099,419 Smurfs
- 122 Languages



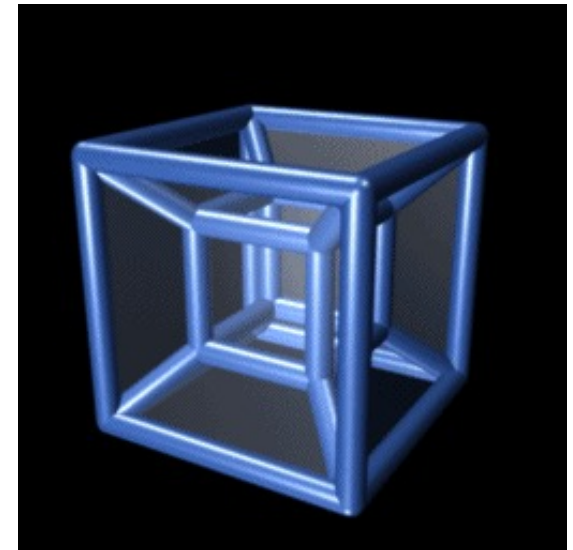
Production

- ~138,000 Ducks
- 44 Languages



Kam4D - kamu.si/kam4d

- 4D = Four Dimensional
 - Time is the fourth dimension - capacity to treat language change and historical languages
- Graph database structure for a complete matrix of human expression across time and space
 - the structure is realistic; the final goal is an impossible aspiration
- Molecular lexicography design



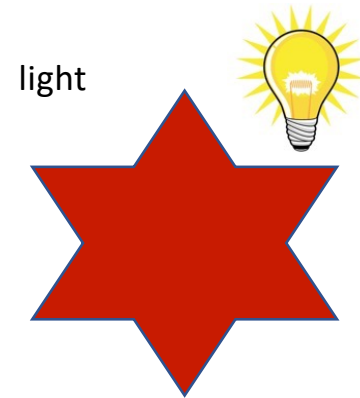
recommended reading:
➤ kamu.si/kam4d



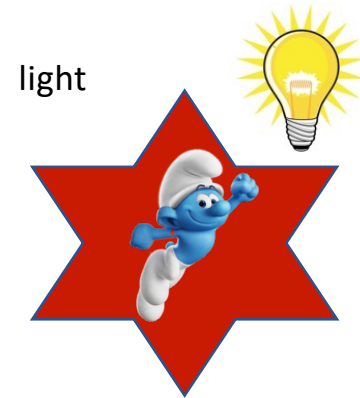
light



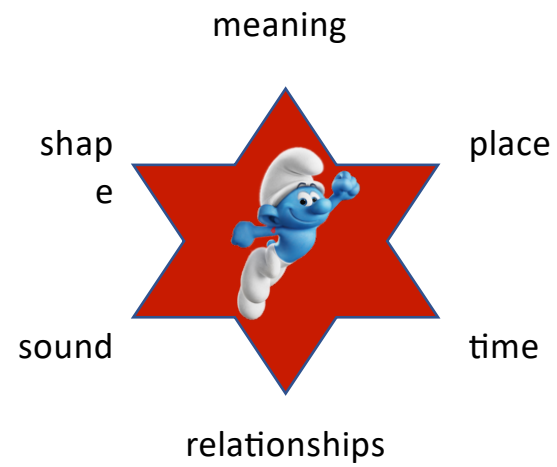
recommended reading:
➤ kamu.si/kam4d



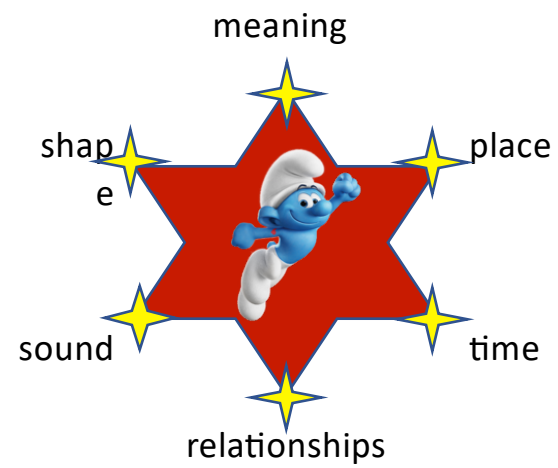
recommended reading:
➤ kamu.si/kam4d



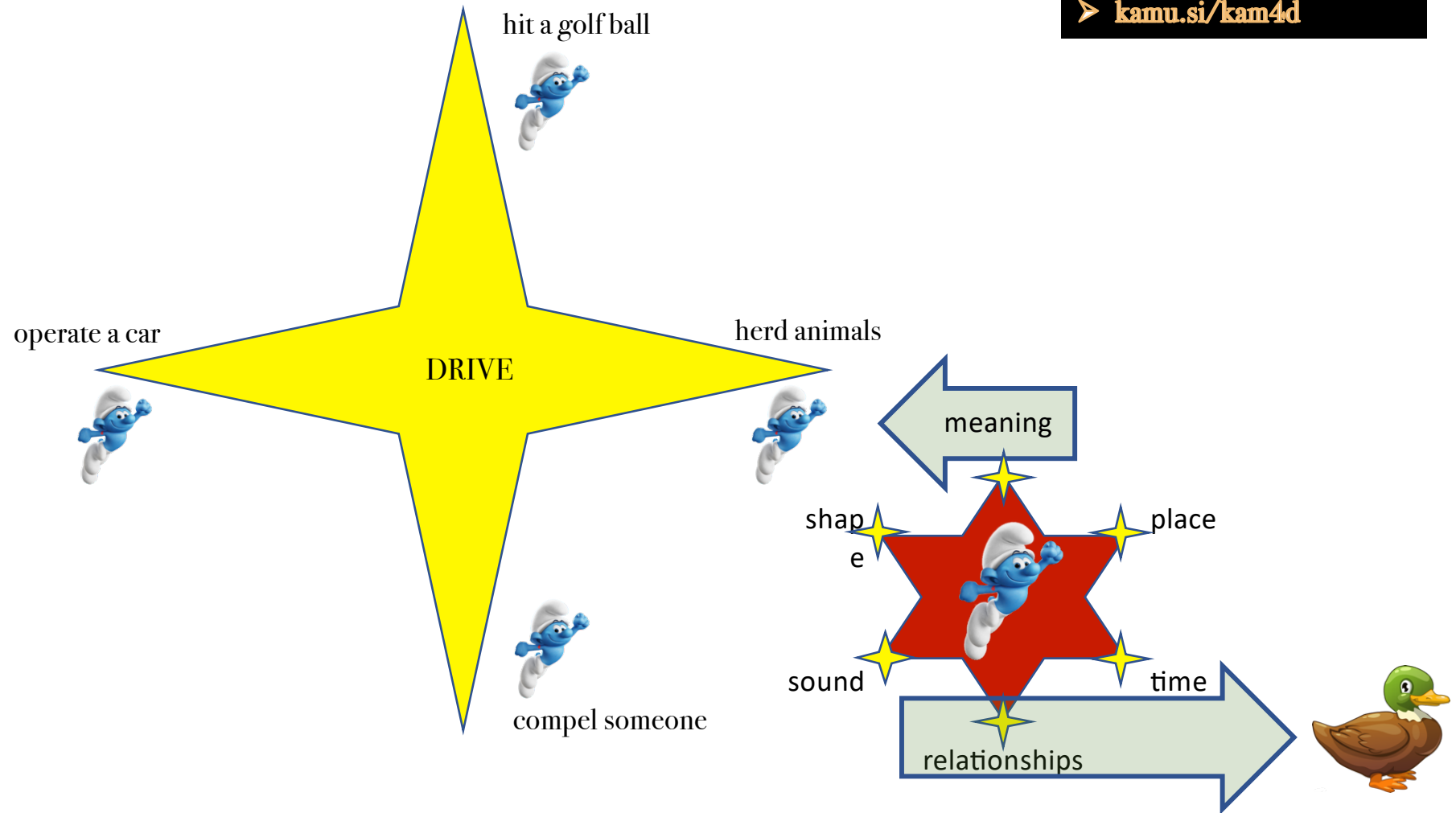
recommended reading:
➤ kamu.si/kam4d



recommended reading:
➤ kamu.si/kam4d



recommended reading:
 ➤ kamu.si/kam4d



Search

From



To

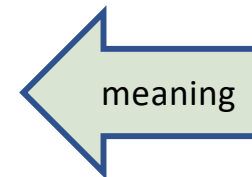
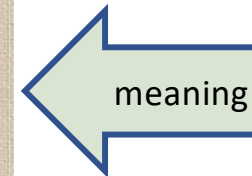
 Go

kamusi GOLD
Global Online Living Dictionary

hubby (n)		
Portuguese	homem (n)	Definition: uma pessoa adulta que é masculina (ao contrário de uma mulher)
isiZulu	indoda (n)	Definition: umuntu wesilisa osequinile nje
English	adult male (n) man (n)	Definition: an adult person who is male (as opposed to a woman) Example: there were two women and six men on the bus.
Portuguese	homem (n)	Definition: qualquer membro vivo ou extinto da família Hominidae caracteriza-se por uma inteligência superior, articular o discurso e erigir a carruagem
isiZulu	umuntu (n)	Definition: ilunga eliphilayo nelingasaphili lomndeni wabantu
English	homo (n) human (n) human being (n) man (n)	Definition: any living or extinct member of the family Hominidae characterized by superior intelligence, articulate speech, and erect carriage



hubby (n)		
Portuguese	 homem (n)	Definition: uma pessoa adulta que é masculina (ao contrário de uma mulher)
isiZulu	 indoda (n)	Definition: umuntu wesilisa osequinile nje
English	adult male (n) man (n)	Definition: an adult person who is male (as opposed to a woman) Example: there were two women and six men on the bus.
Portuguese	 homem (n)	Definition: qualquer membro vivo ou extinto da família Hominidae caracteriza-se por uma inteligência superior, articular o discurso e erigir a carruagem
isiZulu	 umuntu (n)	Definition: ilunga eliphilayo nelingasaphili lomndeni wabantu
English	homo (n) human (n) human being (n) man (n)	Definition: any living or extinct member of the family Hominidae characterized by superior intelligence, articulate speech, and erect carriage



Kamusi Jargon – essential terms for Kam4D

- Lemurs and Party Terms
- Smurfs and Ducks
- Costumes and Wardrobes

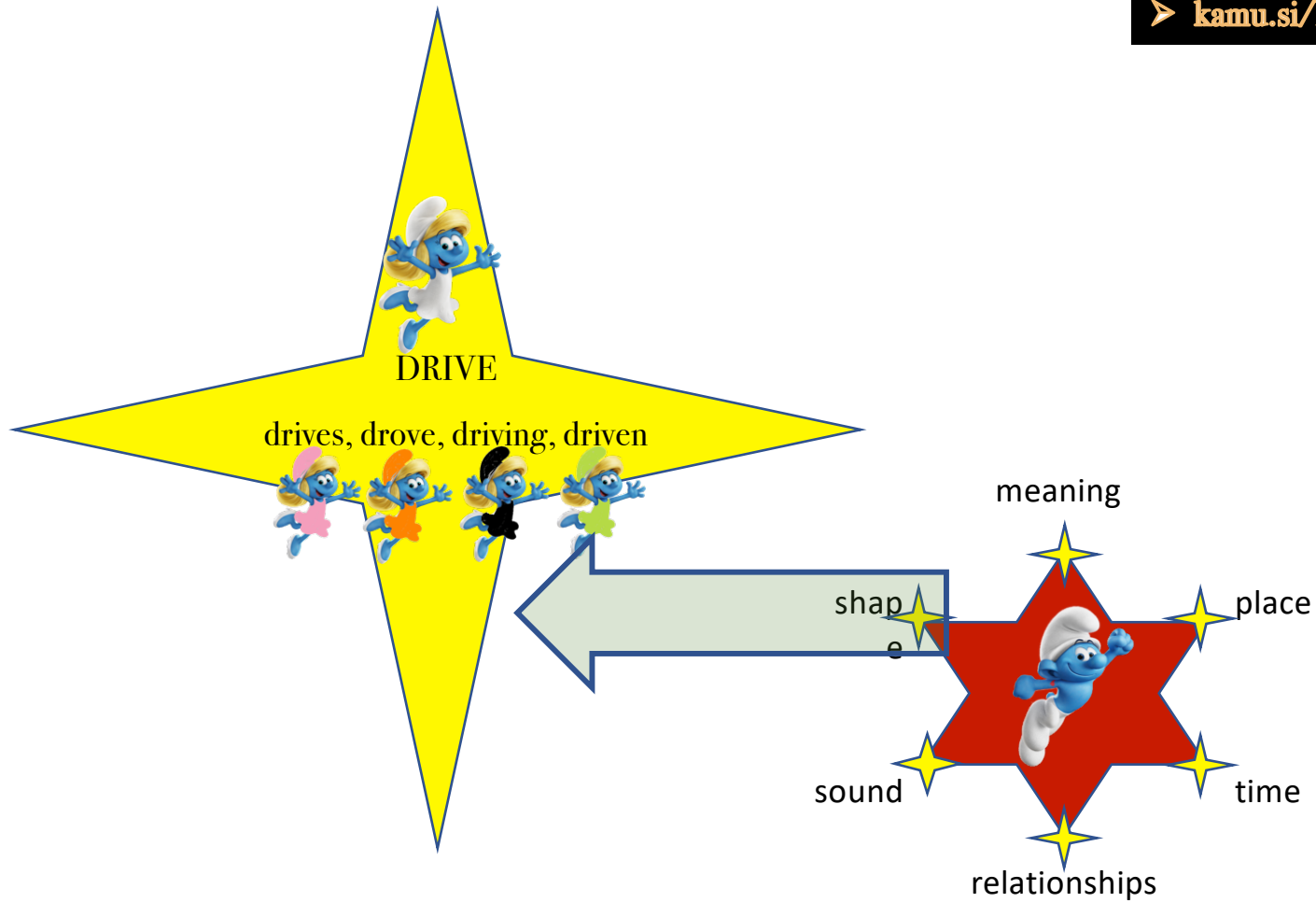
Costume =
A single form (a.k.a.
inflection) that might
be used by a smurf



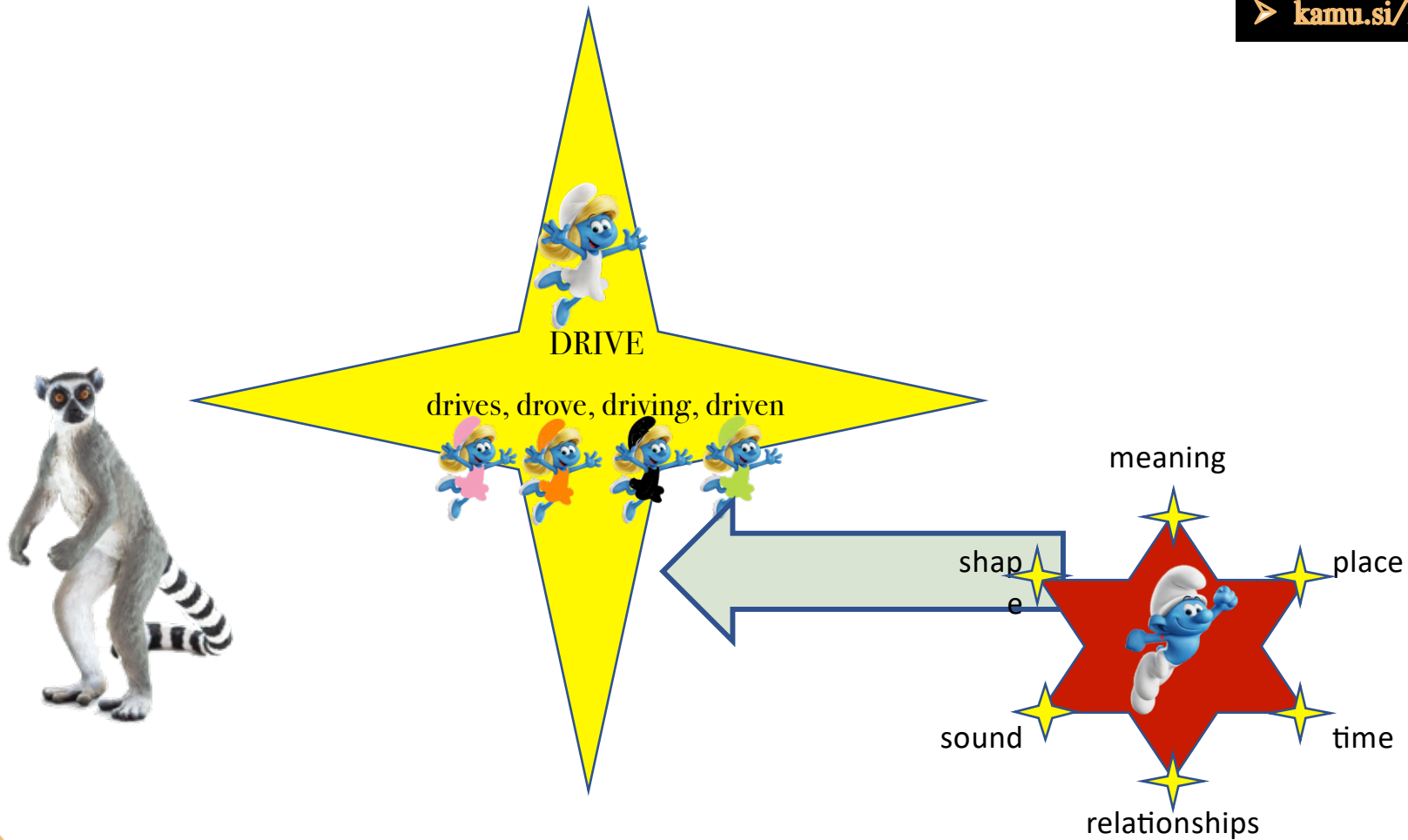
Wardrobe =
A set of forms (a.k.a.
inflections) that might
be used by a smurf



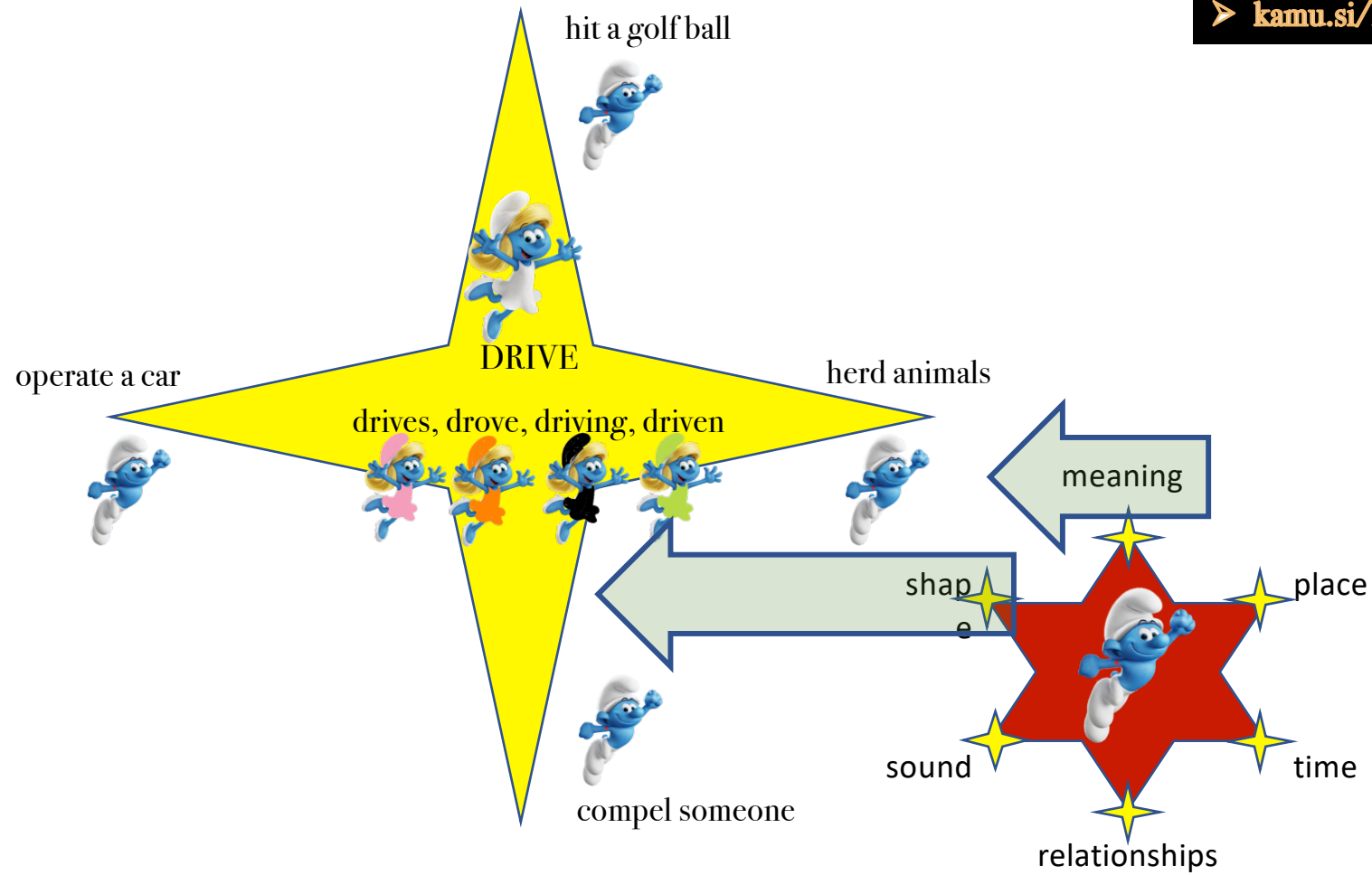
recommended reading:
 ➤ kamu.si/kam4d



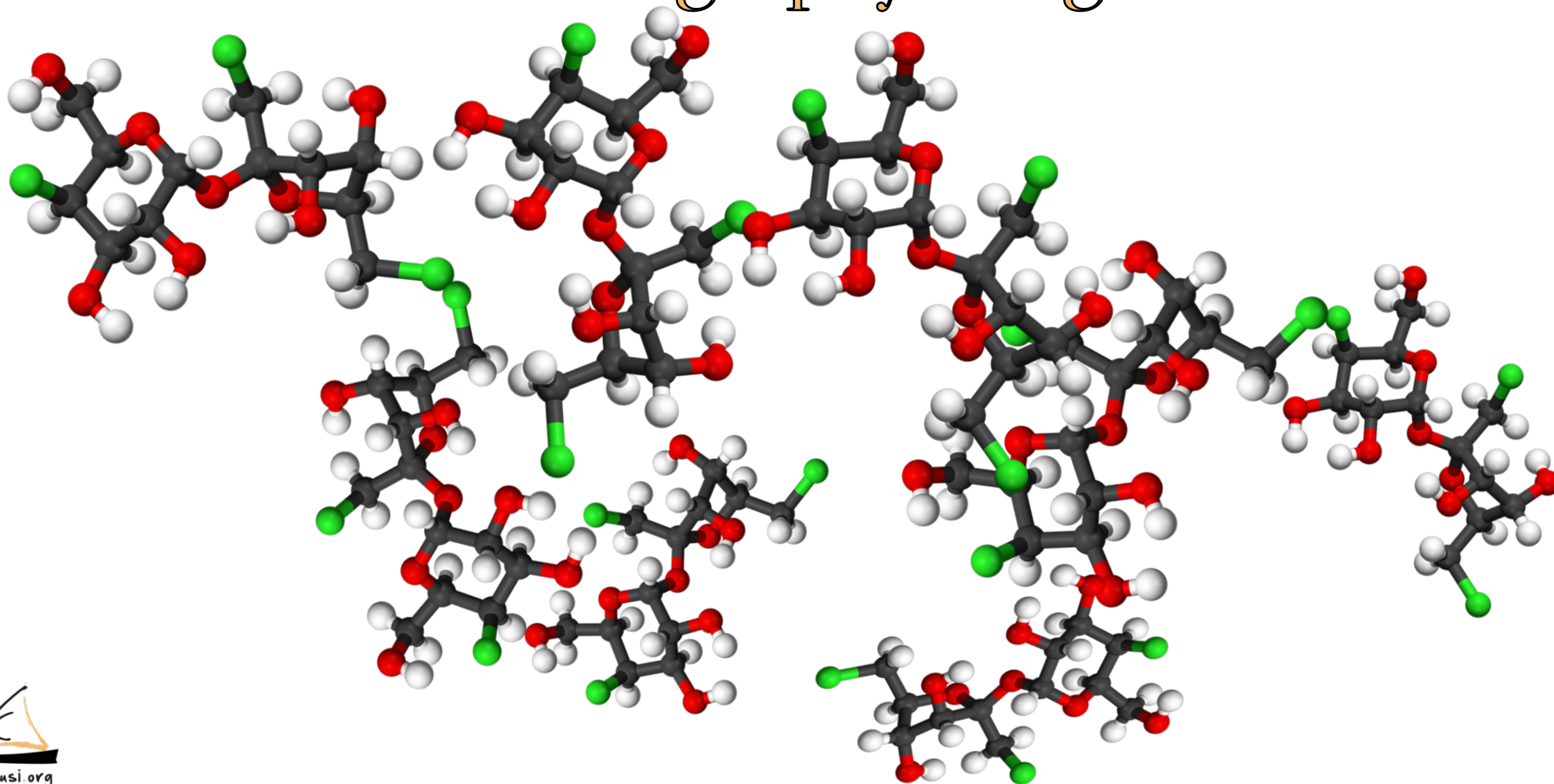
recommended reading:
 ➤ kamu.si/kam4d



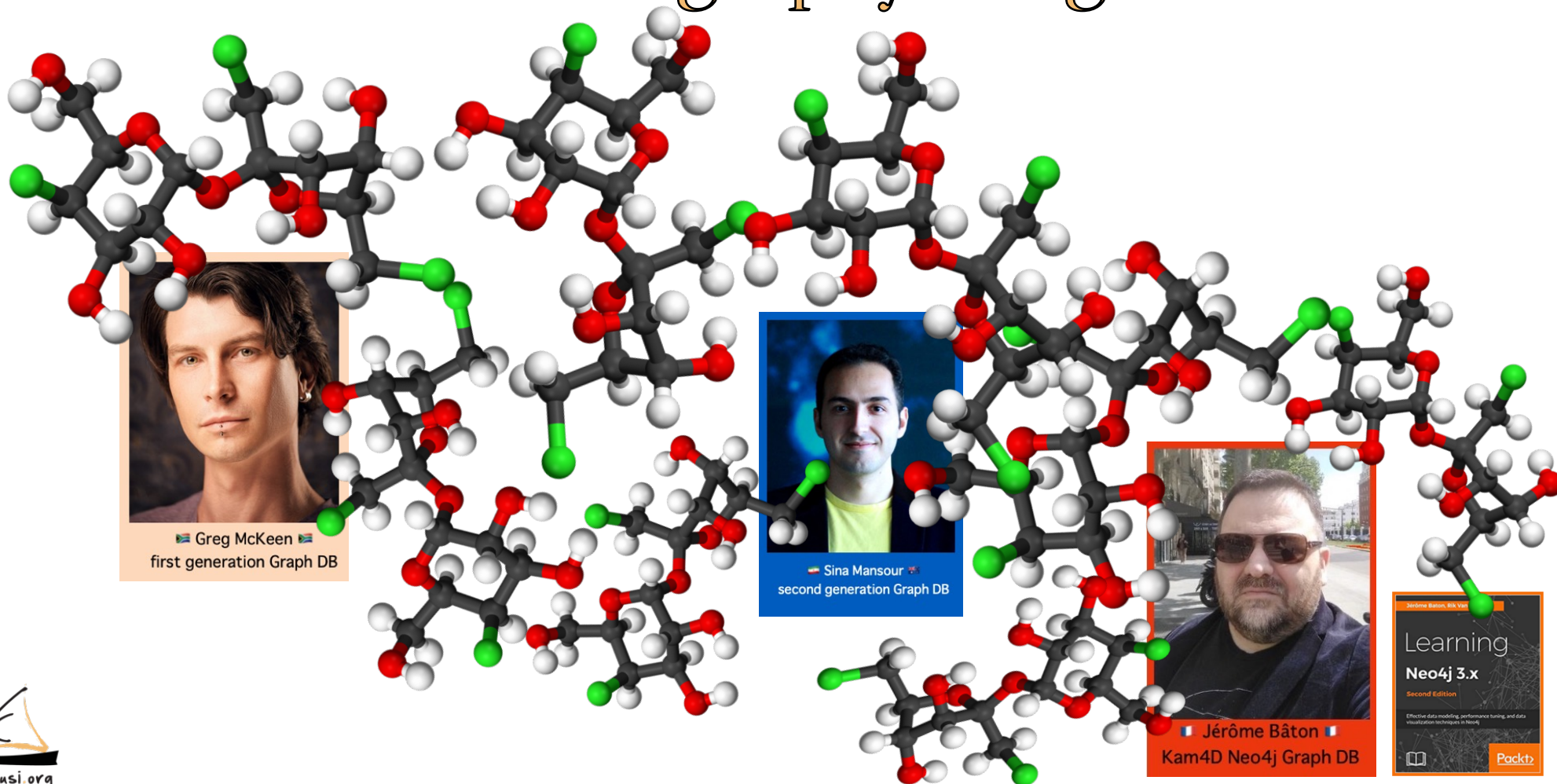
recommended reading:
 ➤ kamu.si/kam4d



Molecular lexicography design



Molecular lexicography designers



THE KAM4D LINGUISTIC KNOWLEDGE GRAPH: PUTTING SMURFS, DUCKS, LEMURS, AND PARTY TERMS TO THE SERVICE OF AFRICAN LANGUAGES

1. The problem with linguistic data
2. The Kam4D solution
3. Kamusi Labs projects



KAMUSI LABS PROJECTS

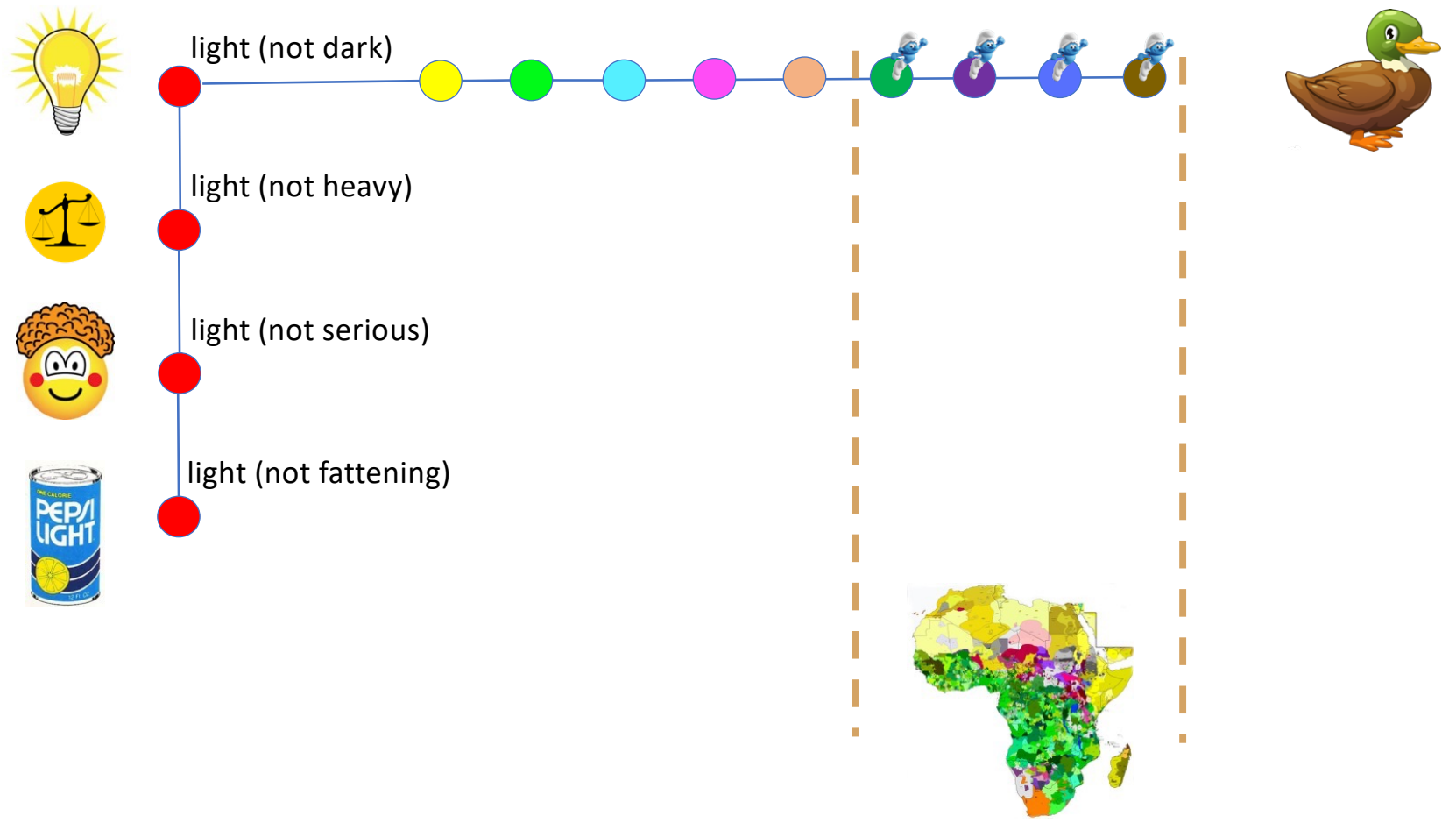
1. Gathering data for African languages
2. SlowBrew assisted translation
3. PALE: Platform for African Language Empowerment
4. Many more projects...

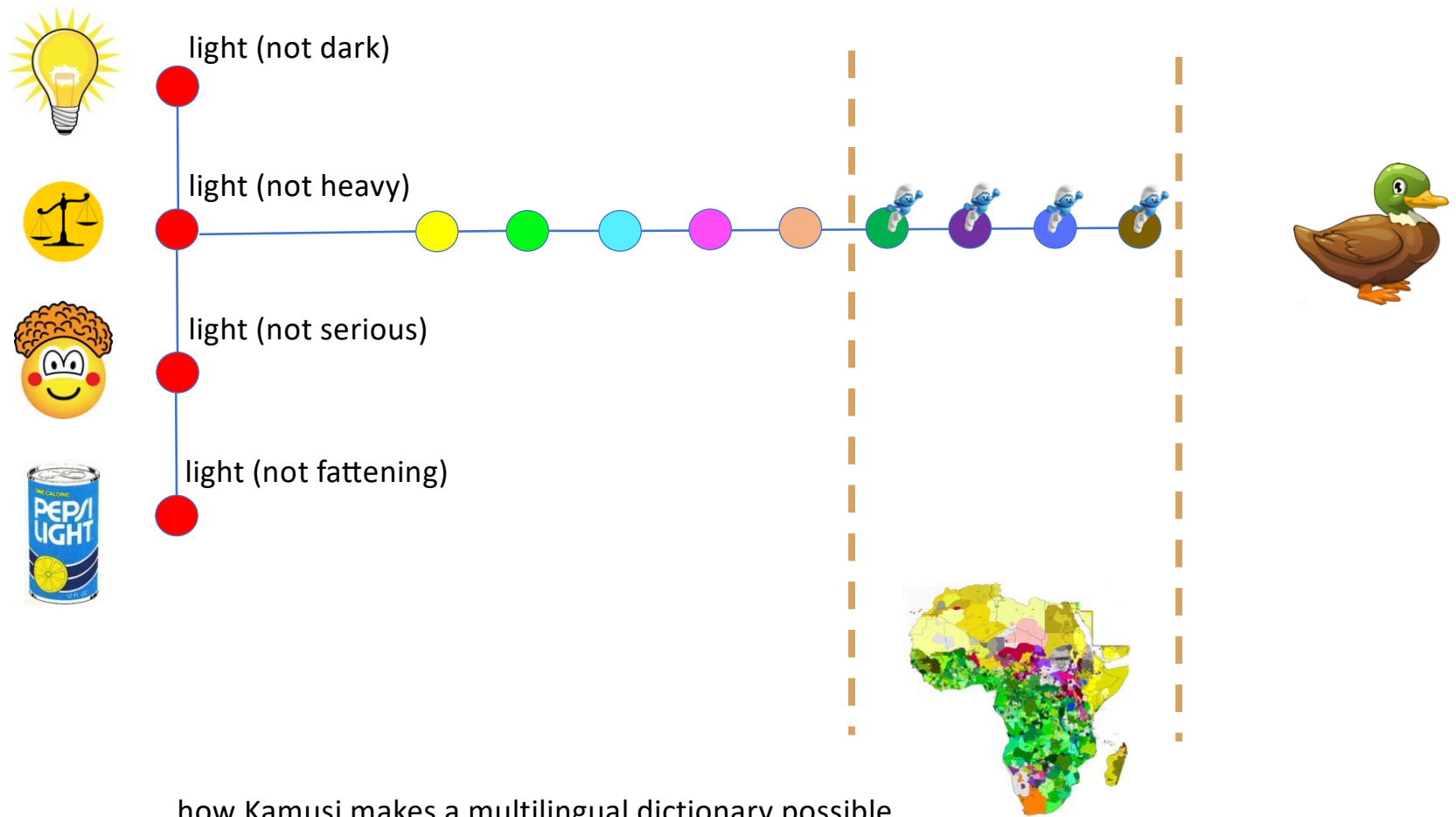


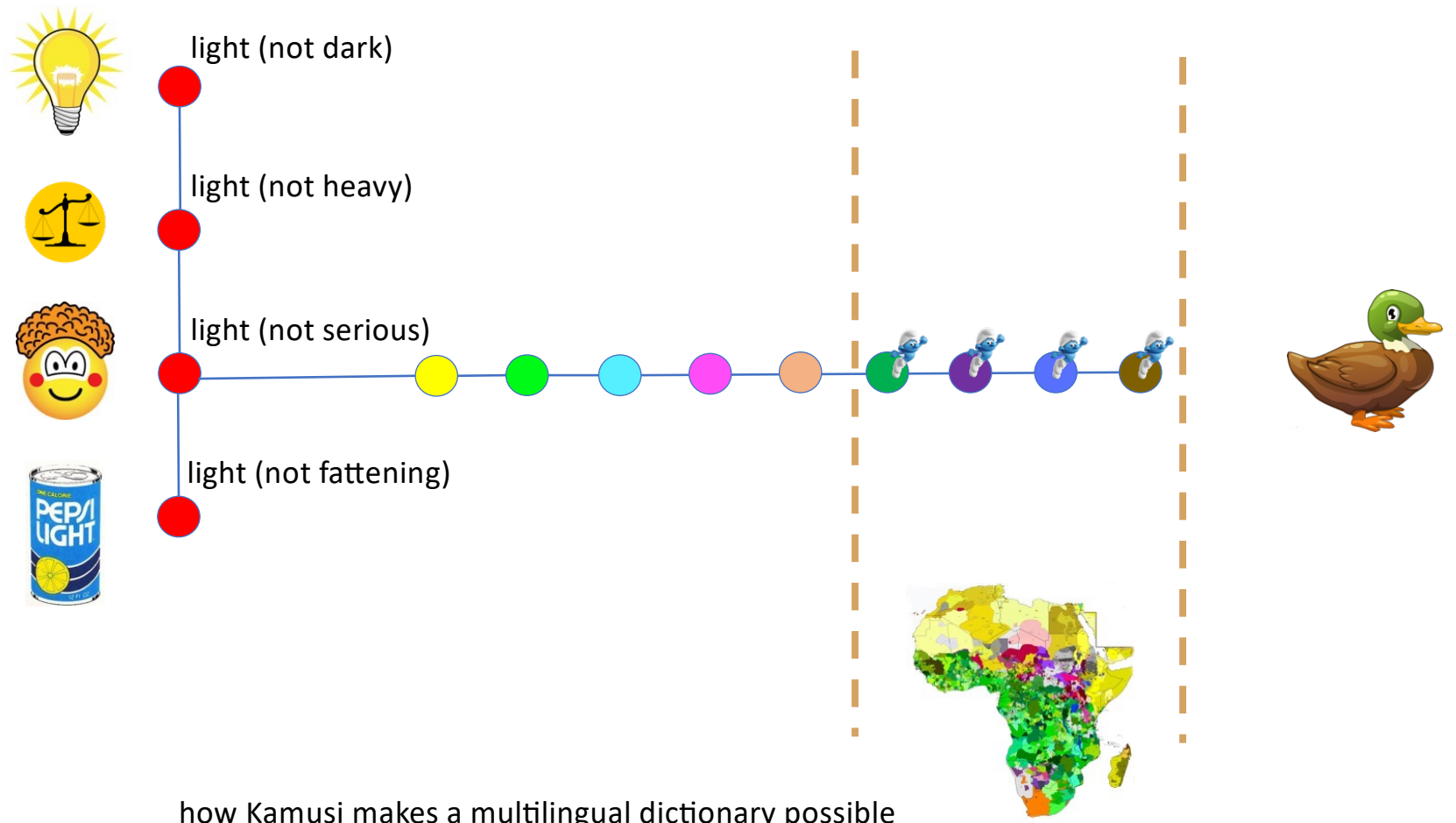
KAMUSI LABS PROJECTS

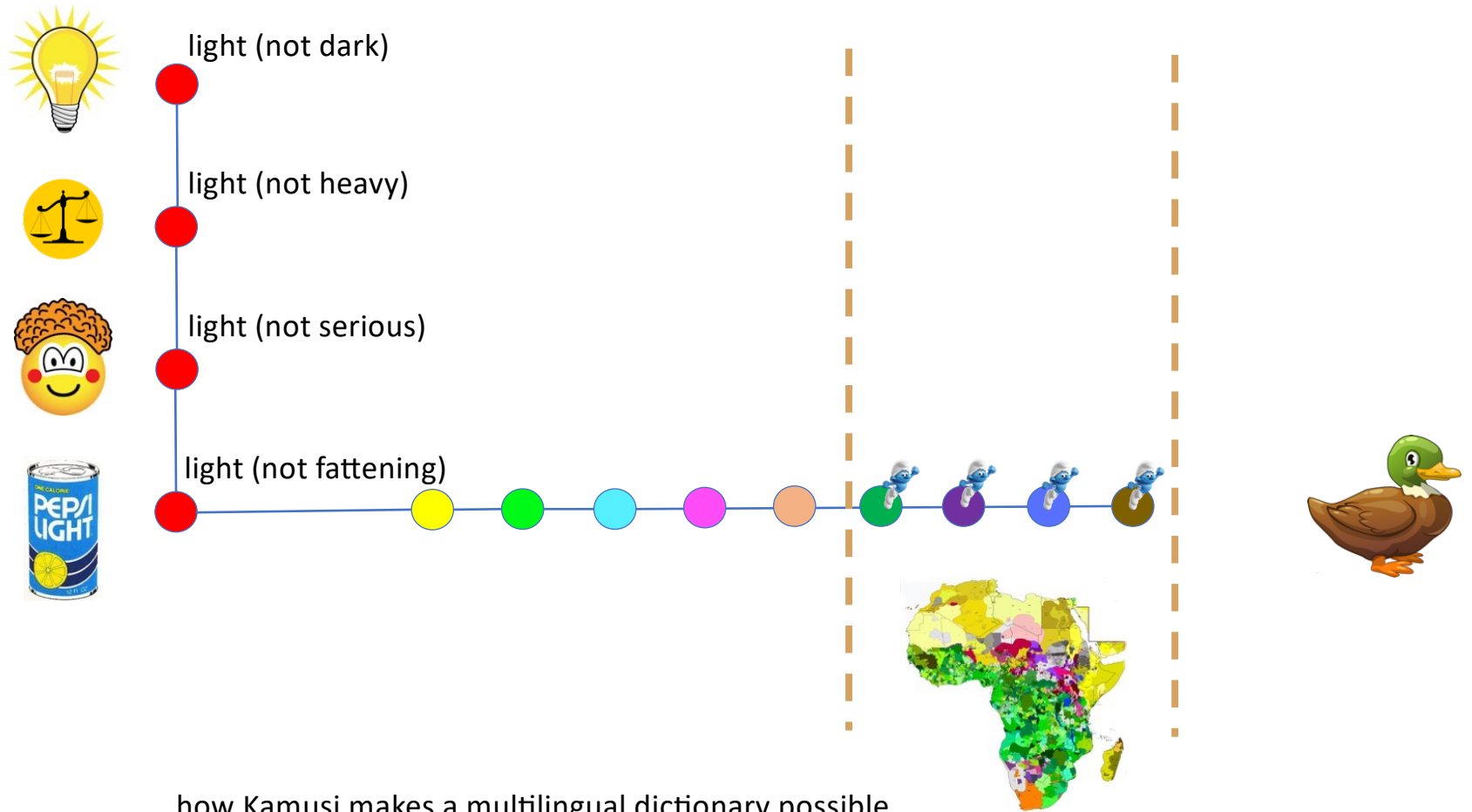
1. Gathering data for African languages
2. SlowBrew assisted translation
3. PALE: Platform for African Language Empowerment
4. Many more projects...

















light (not dark)



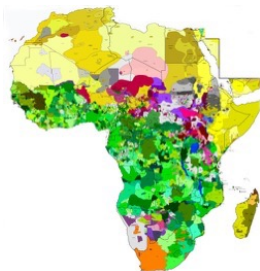
light (not heavy)







light (not serious)

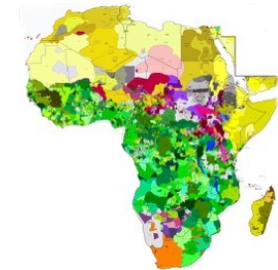


light (not fattening)





GATHERING **Data** FOR AFRICAN LANGUAGES



- Duck Duck Kamus – aligning existing datasets
- Crowdsourcing games for new microdata
- GOLDdigger – engine for expert editors

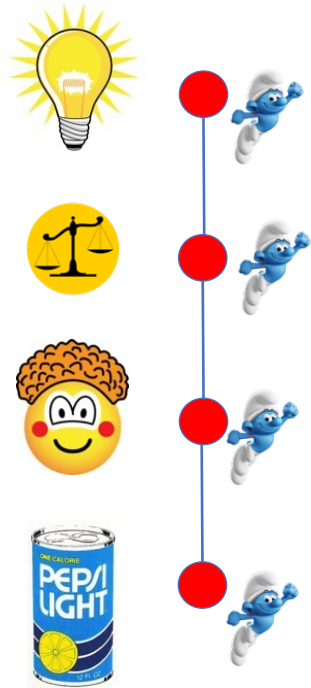
KAMUSI LABS PROJECTS

1. Gathering data for African languages
2. SlowBrew assisted translation
3. PALE: Platform for African Language Empowerment
4. Many more projects...

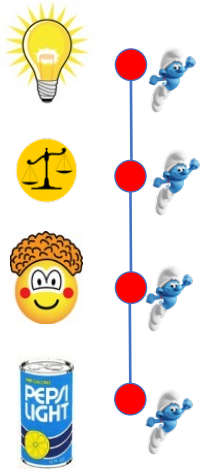


SLOWBREW ASSISTED TRANSLATION

- User selects their meaning on the source side (predisambiguation)
 - Users can suggest missing senses



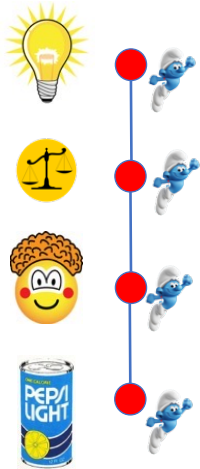
SLOWBREW ASSISTED TRANSLATION



- User selects their meaning on the source side (predisambiguation)
 - Users can suggest missing senses
- SlowBrew suggests Party Terms (MWEs), or users can mark their own
 - Party Terms are treated as Smurfs in Kam4D
 - Separated expressions easily conjoined (unlike NMT)



SLOWBREW ASSISTED TRANSLATION



- User selects their meaning on the source side (predisambiguation)
 - Users can suggest missing senses
- SlowBrew suggests Party Terms (MWEs), or users can mark their own
 - Party Terms are treated as Smurfs in Kam4D
 - Separated expressions easily rejoined (unlike NMT)



- Smurfs and Ducks
- Kam4D – kamu.si/kam4d
- SlowBrew

She **drove** everyone in her class at school **up the wall** last night

• annoy

drove → drive

- through Kam4D “costumes”
- NLP toolkits not available for most languages
- * “drive” triggers search for downstream party terms

• Social stratus

• School room

• Style

• Group of fish

• Group of thinkers

• Educational institution

• previous

• final

• night before

Microsoft Bing

English (detect) French

She drove everyone in her class at school up the wall last night

Elle a conduit tout le monde dans sa classe à l'école jusqu'à la nuit dernière

• Really?
• How - human reviewers or as is?
• Why do you trust me?
• Aren't I asking you?

Your submission will be used by Microsoft translator to improve translation quality

Submit Cancel

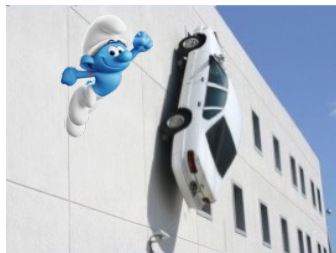
SYSTRAN translate

English - Detected French

English

She drove everyone in her class at school up the wall last night

Elle a fait grimper le mur tous les élèves de sa classe hier soir



Nice! NMT Victory!

Try DeepL Pro DeepL

Translate from English (detected) Into French Formal/informal ON Glossary

She drove everyone in her class at school up the wall last night

Elle a conduit tous les élèves de sa classe à l'école jusqu'au mur hier soir

Wrong! NMT Defeat.

She drove everyone in her class at school up the wall last night

• annoy

drove → drive

- through Kam4D "costumes"
- NLP toolkits not available for most languages
- * "drive" triggers search for downstream party terms

• Social stratus

• School room

• Style

• Group of fish

• Group of thinkers

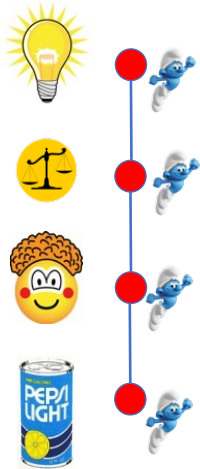
• Educational institution

• previous

• final

• night before

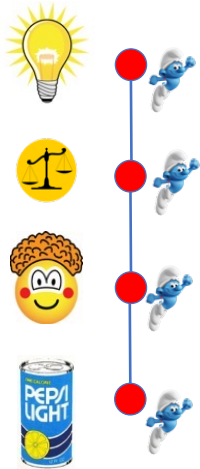
SLOWBREW ASSISTED TRANSLATION



- User selects their meaning on the source side (predisambiguation)
 - Users can suggest missing senses
- SlowBrew suggests Party Terms (MWEs), or users can mark their own
 - Party Terms are treated as Smurfs in Kam4D
 - Separated expressions easily rejoined (unlike NMT)
- DUCKS finds equivalent term in Language B



SLOWBREW ASSISTED TRANSLATION



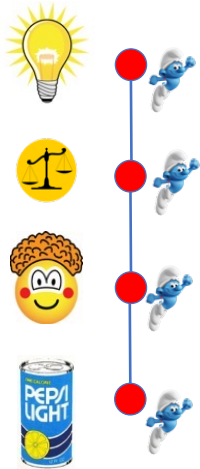
- User selects their meaning on the source side (predisambiguation)
 - Users can suggest missing senses
- SlowBrew suggests Party Terms (MWEs), or users can mark their own
 - Party Terms are treated as Smurfs in Kam4D
 - Separated expressions easily rejoined (unlike NMT)
- DUCKS finds equivalent term in Language B



- Machine learns from context-specific user selections
 - Crowdsourced dataset of spelling/meaning annotations
 - AI builds from human intelligence on the source-side



SLOWBREW ASSISTED TRANSLATION



Unanswered Questions:

- Will users take the time to predisambiguate?
 - People take time to choose images
 - People take time to spellchick
- Syntax on the target side?
 - Outside Kamusi wheelhouse – partners needed
- How to pay for it?



KAMUSI LABS PROJECTS

1. Gathering data for African languages
2. SlowBrew assisted translation
3. PALE: Platform for African Language Empowerment
4. Many more projects...



PALE PLATFORM FOR AFRICAN LANGUAGE EMPOWERMENT



- Kam4D will serve as the linguistic data core for the ACALAN-AU platform
- Integrated with Kamusi systems for gathering and disseminating data
- Initial focus on 20 VCBLs (Vehicular Cross-Border Languages)



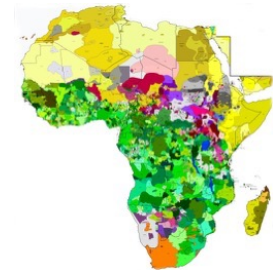
KAMUSI LABS PROJECTS

1. Gathering data for African languages
2. SlowBrew assisted translation
3. PALE: Platform for African Language Empowerment
4. Many more projects...



MANY MORE KAMUSI LABS PROJECTS...

- KamuSee 🕶️ visual dictionary
- Sign languages 🧑🏾🗣️ gesture video dictionary
- Logikamusi 🏹 ontological dictionary
- Kamedicine 💊 medical terminology translator
- Kamigrate 🦶👤 refugee and immigrant services translator
- Kamergency 🚑🧑🏾🚒 phrasebook for accident and disaster first responders
- Kamuseum 🏛️ guides for public spaces
- Box-o-Lex 🧰 field lexicography toolkit
- Talkamusi 🗣️💬📖 talking dictionary
- KamHoosi 🦉 named entities
- EdTech Trio 🎵🎓 for learning IN African languages, learning FROM African languages, and learning OF African languages
- **AND EVEN MORE: <http://kamu.si/big-picture-playbook>**



THE KAM4D LINGUISTIC KNOWLEDGE GRAPH: PUTTING SMURFS, DUCKS, LEMURS, AND PARTY TERMS TO THE SERVICE OF AFRICAN LANGUAGES



kamus*i*.org

Martin Benjamin

SADiLaR: DH Colloquium 17 November, 2021
South African Centre for Digital Language Resources

martin@kamusi.org

recommended reading:

- kamu.si/kam4d
- kamu.si/big-picture-playbook
- teachyoubackwards.com