KAM4D

A UNIVERSAL MULTILINGUAL DATA MATRIX FOR HUMAN REFERENCE AND NLP

Martin Benjamin and Jérôme Bâton, Kamusi Project International
UniDive COST Action: Universality, diversity and idiosyncrasy in language technology, Working Groups 3 & 4
1st General Meeting, Paris-Saclay University, 16 March 2023

Kam4D is a data matrix that has been implemented in an early form at http://kamusi.org for 44 languages. It is designed with the capacity to accommodate any spoken or signed language for which data can be acquired. Built upon a Neo4J graph database, the matrix charts morphological matters within a language, while uniting equivalents across languages on a semantic basis.

Surmountable Challenges: The Technology and the Linguistics

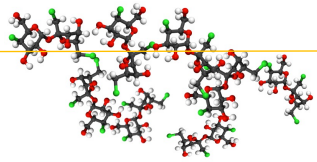
- Problem: Most linguistic data does not exist in digital form

Kamusi Labs solutions:
- Tools for language professionals and citizen linguists to gather and organize data based on human knowledge
- Games for crowdsourcing with built-in validation methods. People love word games, and people love giving to their community.
  - Bigger languages have waiting crowds
- Smaller languages have waiting researchers (graduate students everywhere) who can deploy Kam4D systems as a ready-to-roll toolkit

- Problem: Most existing data is not interoperable

impenetrable *adj.* ~ **(to)** 1 - siopenyeka; -siopitika ~ *to water* - siopenya maji. 2 *(unintelligible)* sioeleweka, -siotambulikana, -a fumbo. **impenetrability** *n.*

Kamusi Labs solutions:
- A tool aligns existing data based on human confirmation of a shared sense between a term in a new dataset and a concept in Kamusi
- Data that is stored in bricks of text needs additional pre-processing. Harvesting such data is a question of money, not technology



Obstacles and Limitations: The Ideology and the Money

- The colonial "mission civilisatrice" continues to guide much policy and education – that is, major European languages are seen as vessels for progress
- Many policymakers and grantmakers hold the opinion that other languages will gently and mercifully fade as the world flattens, or can be preserved with little sprinklings of four-figure fairy dust
- NLP methodological bias toward corpus and other techniques that are inapplicable for languages outside the top data tier
- Fixation on AI, neural networks, machine learning, that are impossible for languages without substantial data sets but attract all the funding
  - "Zero shot" has zero shot for most languages, and should be shot at zero range
- Free data can be an output, but creating it is not free
- Popular but absolutely false impression that Google has already solved language
- Speakers of embattled languages do not recognize that they have a common cause with common solutions, and thus do not come together as a united constituency for action
- Language Equity is not on the agenda of any funder, with the exception of "European Language Equality" that does not extend beyond EU borders
  - Language falls outside the scope of most organizations or grants, which target must-have causes like health or the environment while seeing language infrastructure to communicate their message as something that would be "nice to have" if someone else paid for it
- Money is in the hands of people who already have quality data and tools for their languages, and have not personally experienced linguistic exclusion. Thus, linguistic barriers are invisible to the people with the power to allocate the funds to surmount them

The Kam4D Solution Set: A Graph Database in 4 Dimensions

The Kam4D matrix is designed to catalogue human language systematically across time and space, as a consistent repository to collate and transmit complex linguistic data.

- The 3rd dimension is space. Data can be charted based on location, thus resolving "dialects" and other issues of variance within a language
- The 4th dimension is time. Historical data is enabled as interoperable, e.g. a contemporary English term could link to an Old English term that has a wardrobe of known costumes, and links to its Old High German progenitor, which also links to its contemporary German descendant

The architecture supports many additional elements that cannot fit on this poster, such as pronunciations (IPA and audio) including accent variations, a consistent method for marking tones, sense-annotated usage examples, ontologies, named entities, terminologies, a variety of relationships such as antonyms, video for sign languages, and audio for talking dictionaries of non-written languages and speech recognition for all

Fundamental Organizing Elements

Ducks

Data Unified Concept Knowledge Sets
- A concept with a shared meaning, however it is expressed within one or more languages
- Neat handling of exact, similar, or explanatory equivalents

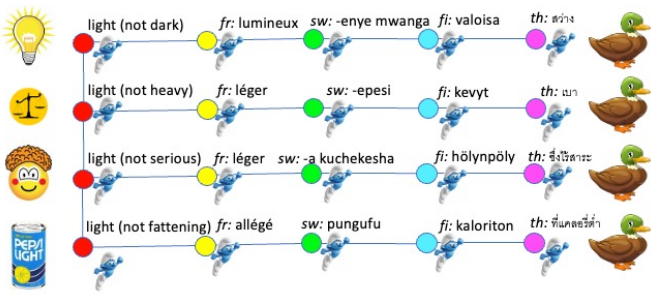
Lemurs

The lemmatic form, or primary spelling, for a listing in a dictionary
- One lemur can be associated with many meanings (polysemy)

Smurfs

Spelling/Meaning Unit Reference
- The intersection of a duck and a lemur
- One spelling with one meaning in one language

English Lemur



Key Supplemental Elements

Costumes & Wardrobes

The different inflections a lemur can wear
- A wardrobe for "see" holds costumes "sees", "saw", "seeing", and "seen"
- Some languages, e.g. Bantu, form costumes from rules, not data points

English Lemur

Costumes in the Wardrobe
Drives, Drove, Driving, Driven

Party Terms

Multiword Expressions
- lexicalizable lemurs that belong to one or more smurfs, and thus join ducks (e.g. "see the light" joins a duck with "comprendre")
- can be marked for separability
- all of these have meanings outside smurfs associated with the lemur "light", so they are their own lemurs: