A Universal Multilingual Data Matrix for Human Reference and NLP – Description and Early Implementation

Martin Benjamin
Executive Director, Kamusi Project International
Lausanne, Switzerland
martin@kamusi.org

Jérôme Bâton
Chief Data Officer, Kamusi Project International
Paris, France
jerome@kamusi.org

We discuss Kam4D, a data matrix that has already been implemented in an early form at http://kamusi.org for 44 languages, 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.

The chief limitation to the matrix is the quantity of data itself, and the chief limitation to data is the negligible funding available to languages that do not have large market potential or the interest of governments with generous coffers. Most linguistic data does not exist in digital form, and most digitized data does not exist in forms that are readily interoperable. The project has mechanisms for gathering data that is not yet digitized, particularly through crowdsourcing and gamification, and for harmonizing data from existing sources, discussed in (Benjamin 2014, 2015a, 2015b, 2016; Benjamin and Radetzky 2014a, 2014b). Thus, the matrix has the theoretical capacity to richly catalogue every word in every language, but the journey there is long, and not interesting to funders who are content to focus on lucrative languages.

The matrix is designed to chart language across four dimensions, including the capacity to document time (when a term came into or out of usage, and its etymology) and space (where a term occurs geographically, in the case of dialects and variants). The starting point, though, is the two-dimensional intersection of what, for purposes of being understandable to non-specialists, we call "Ducks" and "Lemurs" (which have "costumes" and "wardrobes"), at a locus we call a "Smurf". Those terms need unpacking:

"DUCKS" are "Data Unified Concept Knowledge Sets" – that is, all items that share a common semantic sense. For example, the thing that is "a motor vehicle with four wheels" is "car" in English, and also "automobile". That same concept is shared with "coche" and "automóvil" in Spanish, "imoto" in Zulu, and "车" in Mandarin Chinese. By aligning these terms across the shared concept, we get our "ducks" in a row, so to speak.

"Lemurs" are the lemmatic, or "dictionary", form of a word. In and of themselves, lemurs are devoid of meaning. For example, the English lemur "see" could mean many things, from noticing something with one's eyes, to dating someone romantically, to understanding something. Further, that lemur could have many inflected forms, in this case "sees", "saw", "seeing", and "seen", that are the same for some or all of the different senses. We call those inflected forms "costumes", because they are different clothing that a lemur can wear without changing its ultimate meaning. Furthermore, costumes belong to "wardrobes"; for example, the lemur "hang" has two verb wardrobes, {hang, hangs, hanging, hung} and {hang, hangs, hanging, hanged}, with the former applying to hanging laundry, art, or computer programs, and the latter applying to hanging people. In the graph, these inflections are data elements that can be linked to any sense of the lemur, thus detailing the forms a term can take without regard to its meaning. Within NLP applications, this

reduces the computational burden, since the machine only needs to know that "seeing" maps to "see", and can then deal with the underlying meaning in a separate step:

A "SMURF" is a "Spelling/ Meaning Unit Reference" – that is, the intersection of one lemur (the spelling) with one duck (the meaning). For example, "car" with the meaning of "a motor vehicle with four wheels" is one smurf in English, while "car" with the same meaning as a railway wagon is a different smurf. The smurfs "car/ motor vehicle" and "automobile/ motor vehicle" join the same duck as "coche/ motor vehicle" and "imoto/ motor vehicle", while the smurfs "car/ rail carriage" and "wagon/ rail carriage" join a different duck that has different equivalents in other languages. This semantic alignment across languages greatly improves the ability to translate between languages that have not been individually mapped by humans, versus computer inferences in MT that are largely based on the happenstance of a shared spelling of a known polysemic English term in the middle.

Ideally, each concept should have a definition in its own language, in addition to the English sense that we use as the starting point for linking smurfs within ducks. Producing own-language definitions where they do not yet occur is future work that will largely depend on crowdsourcing and the funding to manage it. Again, this is a feature that is essential for linguists and ordinary users, and pie in the sky for most funding agencies for most languages.

Finally, speaking of "pie in the sky", the matrix handles multiword expressions (MWEs, which we call "party terms" as a way to describe to ordinary users and crowd-source contributors, rather than linguists, words that combine to form meanings when they dance together that cannot be discerned from the sum of their parts) as data elements in their own right. Once a party term such as "pie in the sky" has been identified (through processes outside the scope of this paper), it becomes lexicalized with the same definition as shared by the smurfs "unrealistic" and "ludicrous", and joins their duck along with equivalents in any other language for which we have a term for that concept. For MT, this provides the opportunity to translate party terms as coherent units, rather than individual words. Even when inflected and separated, such as "she drove everyone at the office up the wall", the matrix allows the machine to see that "drove" is a costume for the lemur "drive", and that "drive" is the beginning of certain party terms, and that "drive up the wall" is one such smurf, with a meaning that has equivalents in other languages that are members of the same duck.

The matrix includes many other elements, including features regarding the sounds of spoken words and relationships such as antonyms, that are described in a detail at http://kamu.si/kam4d. Sounds can be mapped to costumes, and deployed in speech recognition and speech synthesis. Of note, we have access to video data for about 25 sign languages that can be easily aligned within ducks when and if funding becomes available – but we have no manpower to dig under rocks for support for languages most people don't care about like Nicaraguan Sign Language, so the potential of the matrix in this regard will remain indefinitely unfulfilled. Similarly, the matrix holds a large and growing set of concepts that can be used to elicit terms for endangered and embattled languages, without much technical complexity to create an app for field lexicography that feeds back into the universal repository, were the funding environment to support data collection and preservation in diverse languages.

The initial implementation is online for 44 languages with varying amounts of data (more than 130 languages if you count a small set of parallel terms for Covid-19, and about 1500 concepts across 72 languages that pivot through UNICODE emojis), with limited features that include about 1.5 million smurfs across about 120,000 ducks. Most initial data came by disaggregating individual elements (lemmatic single words or MWEs) within the synsets of existing Wordnets and aligning those elements across languages. An API enables access to smurf-to-smurf translations; future work will allow remote queries that explore throughout the graph, providing consistent references to enable interoperability across diverse NLP projects. More than 60 partners have signed letters of intent, from an indigenous organization wishing to preserve their endangered language in western Congo (DRC), to the intergovernmental language organ of the African Union that sees this as the technological base for a platform for the continent's major languages. The AU, however, does not

have cash on hand to support its goal of language equality (https://kamu.si/au-language-action-plan), and, despite platitudes about international cooperation and the central importance of digitalization, the EU commitment to language equality does not extend beyond European borders. Expanding to more languages and building out the feature set depends on recognition by funders that language inequity excludes most people from equal participation in the global economy. Also of importance, the computer science community that reviews grant proposals is currently fixated on trendy topics like AI, neural networks, and zero-shot translation, which are fantabulous distractions with little or no chance of success for the super-majority of languages that have little or no systematically digitized data. By contrast with the unsupported optimism that computers will somehow pluck data for diverse languages from the ether, and make sense of it all with the wave of a magical AI wand, the matrix described in this paper already collates, houses, and disseminates a substantial amount of human-validated linguistic data in ways that can be read by people for knowledge purposes and by machines for NLP. Filling the matrix with data thus becomes a challenge of logistics and financing to build on a platform that, as developmental technology, is fully operational for core early features.

References:

Benjamin, M, 2016, Problems and Procedures to Make Wordnet Data (Retro)Fit for a Multilingual Dictionary , Proceedings of the Eighth Global WordNet Conference, Global WordNet Conference 2016, Bucharest, Romania

Benjamin, M, 2015 (a), Crowdsourcing and Gamification for Multilingual Linguistic Data, EnetCollect WG3/WG5 Meeting, Leiden, The Netherlands, European Network for Combining Language Learning with Crowdsourcing Techniques

Benjamin, M, 2015 (b), Excluded Linguistic Communities and the Production of an Inclusive Multilingual Digital Language Infrastructure, 11th Language and Development Conference, New Delhi, India

Benjamin, M, 2014, Collaboration in the Production of a Massively Multilingual Lexicon [PDF] By Martin Benjamin, LREC 2014 Proceedings, Language Resources and Evaluation Conference, Reykjavik, Iceland

Benjamin, M and Radetzky, P, 2014 (a), Multilingual Lexicography with a Focus on Less-Resourced Languages: Data Mining, Expert Input, Crowdsourcing, and Gamification, Language Resources and Evaluation Conference, Reykjavik, Iceland

Benjamin, M and Radetzky, P, 2014 (b), Small Languages, Big Data: Multilingual Computational Tools and Techniques for the Lexicography of Endangered Languages, Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages, 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, Maryland