Elephant Beer and Shinto Gates: Managing Similar Concepts in a Multilingual Database

Martin Benjamin
EPFL Swiss Federal Institute of Technology
Lausanne, Switzerland
martin.benjamin@epfl.ch

Abstract

This paper addresses problems in equivalence among concepts, within and between languages. The Kamusi Project has begun building a massively multilingual dictionary that relates as many languages as possible for which data can be gathered. In the process, we have encountered numerous complexities that we attempt to address through the design of our data structure. This paper presents the issues we have encountered, and discusses the solutions that we have developed.

1 Introduction

True synonyms are rare within a language, if synonyms are taken to be words that can stand in each other's place in all contexts. Even if you cannot propose a whisper's difference between the ideas of "snuggling" and "cuddling", you can "snuggle against" someone, but you cannot "cuddle against" them. In Swahili, "ndovu" and "tembo" are completely interchangeable when talking about elephants, but bring you different brands of beer when you ask for them in a bar. Each word must thus be treated differently in a dictionary, so that its particular nuances can be elaborated.

Between languages, it is quite common that terms exist for exactly the same concept. When speaking of colors, English "red" evokes essentially the same bloody hue as "rouge" in French or "nyekundu" in Swahili. An "elephant" is an elephant, whether it is "éléphant" or "ndovu". "Beer" and "bière" and "bia" are all beer. However, we do not expect other senses of a word to map identically in translation; we anticipate that a "red" grape in English might be "noir" in French.

These issues are not new to lexicographers, and this paper will not claim to advance our understanding of synonymy; a trio of recent articles in the International Journal of Lexicography (2013) by Adamska-Sałaciak, Gouws, and Murphy provide the context from which this paper is launched. What is new is the system that Kamusi is developing to produce a global dictionary that can catalogue synonyms within and across languages, and account for their subtle differences.

2 Monolingual Pillars and Multilingual Beams

The basic architecture of Kamusi was developed to handle cases like the examples above, which we now think of as the easy ones. The initial structure is two dimensional, with vertical pillars and horizontal beams.

The vertical axis is the monolingual entry for a term. Within a language, each term is entitled to as many entries as that particular sequence of letters has senses; "light" (not dark) is a different entry from "light" (not heavy) or "light" (not serious) or "light" (low calorie). Those entries can then be segregated into groups, so that a "light" (flame for a cigarette) is grouped with the verb "light" (ignite a fire) while "light" (a lamp) is grouped with "light" (a traffic signal). Within groups, entries can be ranked, so that "light" (a lamp) is listed above "light" (a traffic signal). The groups themselves can be ranked on a scientific or whimsical basis – a corpus count would place the groups for light (energy) and light (lamps) high on the list, but the decision about where to rank light comedy versus light soda can only be arbitrary. This vertical structure provides all the space needed to engage in the lexicographical challenge of giving a definition to each of a term's different senses. In order to support the horizontal beams, no entry is deemed complete in Kamusi unless it includes a definition written in its own language.

The horizontal axis is the same concept as expressed in different languages. "Light" (not dark) can be expressed with some term in German, another term in Japanese, and another one in Songhay. Once a concept from one language has been determined to be equivalent to a particular entry for a different language that is already in the system, we take the relationship to be transitive across all the other equivalents in all the other languages in the system. Because "red" for colors and "red" for grapes are two different entries on the vertical pillar in English, they connect to different horizontal beams, and we can weld on terms in different languages that match those varying concepts:

Red (color of blood) ↔ rouge ↔ nyekundu
↓
Red (color of wine) ↔ rouge ↔ nyekundu
↓
Red (color of grapes) ↔ noir ↔ zambarau

In this schema, "rouge" in French has its own monolingual pillar (color of blood, color of wine, type of cosmetic), as does "noir". It is clear what terms gloss each other between languages – one would not look up "red" and mistakenly use the color of blood to talk about grapes in either French or Swahili. Horizontal beams work perfectly when concepts are essentially the same across languages.

3 Mapping Inexact Concepts

Unfortunately for our architecture, however, languages do not map on a simple one-to-one basis. We have had to address five major problems with the internal wiring of our edifice.

1) Partial equivalence
2) No equivalent term
3) Different forms
4) Different parts of speech
5) Synonyms within a language

1. Partial equivalence. In English, we have ten fingers and ten toes, and the Dutch have ten "vingers" and ten "tenen". Romanians, however, have twenty "degete", and Swahili speakers have twenty "vidole". Nowhere is this a problem for glove makers, but it wreaks havoc for a multilingual dictionary. English and Dutch are full equivalents, as are Romanian and Swahili, but those two sets only partially match each other. Thus, the flow of transitivity is broken, and the nature of the partial relationships is ambiguous.

When establishing a relationship between terms in Kamusi, a contributor specifies whether they are "parallel", "similar", or "explanatory" (see the next section). Terms that are designated as "similar" disrupt the welding of the horizontal beam. We know that items that are added as parallel to "finger" will be transitive to the first set, and items that are parallel to "vidole" are parallel to the second set, and we can also infer the same similarity between new terms on either side of the divide. However, we cannot infer any inherent relationship between similarities that have not been documented; a language that had terms for each individual finger but no overriding category term, for example, would be similar to finger and vinger in a different way than it is similar to kidole and deget, and differently than the similarity between finger/vinger ↔ deget/kidole. The programming to chart similarities between transitive groups is not complete as of this writing.

Forthcoming programming will include two new features for similarities. First, each relationship pair will have a descriptive field in which differences can be explained in writing, in multiple languages. Second, users will be able to vote on the level of similarity (close, distant, barely comparable), and the votes can be aggregated into a graphic such as a Venn diagram to alert dictionary users about potential dangers in equivalence.

Partial equivalence is also addressed within Kamusi's vertical pillars. As discussed, each sense should have a definition of a term written in its own language. Each of these definitions can be further translated into any other language. Thus, an English definition of "finger" would refer to the ten digits of the hand, and the Romanian and Swahili translations of that definition, stored within the English concept of "finger", would also discuss the ten digits of the hand. Conversely, the Romanian definition of "deget" would refer to both hands and feet, and the English translation of that definition within the Romanian entry would contain that clarifying information for English readers.

2. No equivalent term. Numerous concepts that exist in one language do not exist in another. For example, Japanese has a term “torii” (鳥居) for the ceremonial gate to a Shinto shrine seen in the images above. “Torii” is not an English word, but we need a way to describe it in a Japanese-English dictionary. Our solution is to create an entry on the English side that is labeled “explanatory” of the Japanese term: “Shinto gate”. This term does not become part of the larger English lexicon, but will be visible when a user looks up “torii” in Japanese or conducts a direct English-Japanese search.

Explanatory phrases come with their own complications. “Shinto gate” is an endpoint on the horizontal beam; one can add a French explanatory phrase for “torii”, but that will not link to the English explanation. However, Okinawan does have the concept, and uses the term トリイ (torii). In this case, the relationship between Japanese and Okinawan is transitive, so we assume that English “Shinto gate” is explanatory of the Okinawan and any other languages that enter the parallel set.

Parallel relationships cannot be automatically inferred from explanations in the current Kamusi system. For example, “-simulia” in Swahili and “a povesti” in Romanian are both explained in English with the phrase “tell a story”, but that relationship is not easily discoverable. Future programming will address this gap.

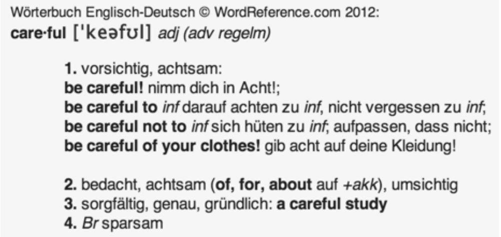
3. Different forms. Two languages might have the same concept, represented by the same part of speech, but approached from different directions. For example, placing a passive suffix on the Swahili verb “-abiri” (travel as a passenger) produces the verb “-abiriwa”, which can translate to English as “be crossed” in the sense that a river is crossed by a ferry. Such misalignments occur ad infinitum between Bantu languages and English, and similar form differences occur throughout the data.

Kamusi has a tidy system for handling different forms of a word (although we do not have a tidy term, since neither “morphemes” nor “inflections” cover the concept; our current candidate is the coinage “morphlections”). When a language has a manageable number N of morphlections, such as the four possible forms of a Portuguese adjective, we create N minus one additional input boxes for that part of speech, which we label during the setup process (e.g., feminine singular, masculine plural, and feminine plural). A more automated system for large conjugation sets such as Romance verbs is on the agenda, and a fully automated system for machine-predictable agglutination parsing has already been developed for Swahili and should be transferable (not without tears) to languages from German to Xhosa.

This morphlection system makes it possible to list forms that do not normally appear in dictionaries, such as the passive verb form in English. “-Abiriwa” can then be linked to “be crossed” within the correct sense entry of “cross” (not betray, nor intersect, etc.). Everything that one needs to know in order to make sense of “be crossed” is contained within the English entry (it is the passive form of a verb meaning “to pass from one side to the other”), without having to create a full separate English entry to accommodate the Swahili formation. It also becomes possible to link morphlections from one language to morphlections in another, such as mapping the English past participle “crossed” to the French past participle “traversé”. A search for a morphlection will pull up the full result for the canonical form, but show any relevant interlanguage links for the morphlection as well.

4. Different parts of speech. Although you may think your watch is on your left wrist, with “left” as an adjective, in Kirundi it is on your wrist leftly, with “bubamfu” as an adverb. Similarly, the verb “achtgeben” in German is expressed in English as an auxiliary verb plus an adjective, “be careful”, and in French as an auxiliary verb plus a noun, “faire attention”. A green cigar may be just a cigar with an adjective, but it greens as a verb (“guun”) in the Aukan language of Suriname.

A monolingual dictionary should contain only the terms that exist in that language; “careful” is an English term, whereas “be careful” is a non-problematic construction of two terms that has no home in any English dictionary consulted for this paper, nor in the Princeton WordNet. Bilingual dictionaries, however, need ways to show how terms in one language are expressed in the other. As shown in the example below from WordReference.com, showing equivalence between languages in such cases is a struggle; “achtgeben” is glossed as “be careful”, but “be careful” is shown on the English side as a usage example that does not track back to “achtgeben”.

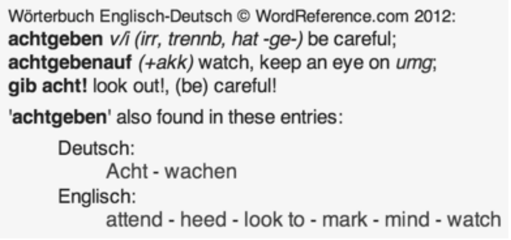


Wörterbuch Englisch-Deutsch © WordReference.com 2012:
care·ful [ˈkeəfʊl] *adj (adv regelm)*

1. vorsichtig, achtsam:
be careful! nimm dich in Acht!;
be careful to *inf* darauf achten zu *inf*, nicht vergessen zu *inf*;
be careful not to *inf* sich hüten zu *inf*; aufpassen, dass nicht;
be careful of your clothes! gib acht auf deine Kleidung!

2. bedacht, achtsam (**of, for, about** auf *+akk*), umsichtig
3. sorgfältig, genau, gründlich: **a careful study**
4. *Br* sparsam

Figure 1: "careful" in English-German translation, http://www.wordreference.com/ende/careful



Wörterbuch Englisch-Deutsch © WordReference.com 2012:
achtgeben *v/i (irr, trennb, hat -ge-)* be careful;
achtgebenauf *(+akk)* watch, keep an eye on *umg*;
gib acht! look out!, (be) careful!

'**achtgeben**' also found in these entries:

Deutsch:
Acht - wachen
Englisch:
attend - heed - look to - mark - mind - watch

Figure 2: "achtgeben" in German-English translation, http://www.wordreference.com/deen/achtgeben

The Kamusi solution is to provide fields for "bridges". Though not implemented as of this writing, the monolingual entry for a term will also include the option for a contributor to "add a bridge" for a part of speech. The English adjective "careful" can be augmented with the verb bridge "be careful", and the French noun "attention" can have the verb bridge "faire attention". The English and French items can then be linked to German and become connected transitively along the horizontal beam, or they can be linked directly without the German intermediary. In either case, we do not crowd the monolingual side of a dictionary with unnecessary entries for differently-structured concepts from other languages, but we include the necessary information and make it discoverable.

5. Synonyms within a language. The Kamusi structure makes it easy to attach a synonym to a single sense of a word, such as matching "traverse" only to the sense of "cross" as passing from one side to the other. However, we face three additional challenges: a) whether the terms are exact equivalents, b) whether one term is preferred to another, and c) how they act in transitive translation sets.

a. When presenting glosses between languages, one has some latitude to stretch the notion of exact equivalence between terms; English "stool" can be linked as equivalent to Swahili "kigoda" even though a typical stool is much higher above the ground than a typical kigoda. Within a language, though, the subtle differences between terms arguably take on more significance. "Think" and "ponder" are synonyms in the WordNet sense of "reflect deeply on a subject", but there is a nuanced difference of degree.

As with bilingual glosses, forthcoming programming will provide the opportunity to categorize a synonym relationship as parallel or similar. Users will have the opportunity to rate the closeness of similar terms, and a comment field will provide the opportunity to stipulate the ways that synonyms differ. In the above example, "think" and "ponder" would likely be shown as parallel for the specific sense, but a comment that adheres to the relationship might explain that pondering is a somewhat more intense activity.

b. Within a group of terms that are listed as synonyms, a system is needed to rank those that are more prevalent. This is especially important when showing the set within a translation result, because language learners will have little independent basis to judge which term to use. A student of English would be hard pressed to select a best choice from among the options in the WordNet synset: "chew over, think over, meditate, ponder, excogitate, contemplate, muse, reflect, mull, mull over, ruminate, speculate". A chief complaint that Swahili teachers have about the Kamusi Project is that students tend to use the first entry of a search result, even if the display is alphabetical because the result has not yet been ranked, so essays are often submitted with some rather strange choices of vocabulary. Without a ranking system, English students worldwide will chew over problems more often than they ponder them, and they will excogitate more than they contemplate.

We have developed a simple tool (currently not online due to a change in our programming platform) that allows contributors to slide entries in a set up or down in relation to each other. A set can ultimately be locked down by a moderator, but we see the ranking tool as lightweight work that is a good use of crowd-source energy. Synonyms cannot easily be ranked based on corpus frequency results, because the work of determining the specific senses of homonyms is prohibitive. Future programming will simplify crowdsourcing even more, posing questions to users such as, "'Ponder' and 'think' are both defined as 'to reflect deeply on a subject.' Which do you use more often?" Without digressing into our plans for building Kamusi data through tightly-controlled input from the crowd, we can still propose that aggregated voting results will provide a somewhat scientific method to rank terms within a set of synonyms.

c. Monolingual synonyms within multilingual translation sets. “Ndovu” and “tembo” are both translations of “elephant” and “éléphant”, but they are not translations of each other. In future, were we to link “ndovu” to “elephant” as a parallel translation, and then link “tembo” and “elephant”, Kamusi will be savvy enough to recognize that “ndovu” and “tembo” are the same language, and therefore synonyms rather than translations. Conversely, if we have a set of synonyms in one language, and we link one of those terms to a term in another language, then we can create a transitive translation relationship for each of those synonyms. The coding for this feature will follow significant refinements to the behavior of translation sets that have just been completed as of this writing, with ramifications described in the conclusion.

4. Concluding thoughts: Integration with the Global WordNet

The questions of synonymy raised above are, of course, not new to WordNet. What is new is the potential that the Kamusi system offers for fine-tuning relationships identified as synonymous within a language, and for charting those identified semantic links across language WordNets.

As an example, the Princeton WordNet contains the synset: car/ auto/ automobile/ machine/ motorcar. UWN/MENTA maps the sense of that synset to the following French equivalents: automobile/ auto/ bagnole/ voiture/ wagon, and similar clusters or single terms in many other languages.

Tying five terms identified as synonyms in English with five terms identified as synonyms in French creates 25 pairs, each of which needs to be differentiated from homonyms on both sides. When the programming resources are available, Kamusi proposes to address this challenge through a process that engages the crowd to validate synsets within a language, and their glosses across languages. In the above example, crowd consensus might push “machine” out of the English synset, or bring “wheels” into the group. A similar process would be in effect on the French side. When a link is established between any item within a set of synonyms in one language, and another item within a set of synonyms in another language, then the computer establishes the existence of a relationship among all the entities.

What is significant about these links from one synonym to the next, and from one translation to the next, however, is that they are not absolute. With programming completed just in time for this paper to go to press, Kamusi charts degrees of separation between links that have been validated by humans and those that have been inferred by transitivity algorithms. Those degrees of separation will track through intra-language synonyms. Thus, if “wheels” is human-linked to “car”, “car” is linked by hand to “voiture”, and “voiture” is manually linked to “bagnole”, then “wheels” and “bagnole” will be shown to be separated by three degrees. This will enable readers to make an educated judgment about the tightness of the association between any two terms. In addition, knowledgeable users will be able to help confirm, reject, or add nuance to computer-predicted linkages.

The programming to implement a smooth integration of WordNet data within the Kamusi framework has not yet advanced out of the conceptual stage, for two reasons. First, a variety of other tasks must be completed in order for working with WordNet data to be practical, particularly reestablishment of the grouping tool in a multilingual context, certain behaviors of morphlections, and the big upcoming task of developing an effective system of working with the crowd. Second, finances. Once those elements are in order, and we have had further conversations with members of the WordNet community to refine our approach, we look forward to seeing what can happen when we connect the extensive multilingual WordNet data sets with the lexicographical potential that the Kamusi framework makes possible.

5. References

Arleta Adamska-Sałaciak. 2013. Equivalence, Synonymy, and Sameness of Meaning in a Bilingual Dictionary, *Int J Lexicography* 26 (3): 329-345

Rufus H. Gouws. 2013. Contextual and Co-Textual Guidance Regarding Synonyms in General Bilingual Dictionaries, *Int J Lexicography* 26 (3): 346-361

M. Lynne Murphy. 2013. What We Talk About When We Talk About Synonyms: (and What it Can Tell Us About Thesauruses), *Int J Lexicography* 26 (3): 279-304