View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by SZTAKI Publication Repository

Building basic vocabulary across 40 languages

Judit Ács **Katalin Pajkossy** **András Kornai**
HAS Computer and Automation Research Institute
H-1111 Kende u 13-17, Budapest
{judit.acs,pajkossy,kornai}@sztaki.mta.hu

Abstract

The paper explores the options for building bilingual dictionaries by automated methods. We define the notion 'basic vocabulary' and investigate how well the conceptual units that make up this language-independent vocabulary are covered by language-specific bindings in 40 languages.

Introduction

Globalization increasingly brings languages in contact. At the time of the pioneering IBM work on the Hansard corpus (Brown et al., 1990), only two decades ago, there was no need for a Basque-Chinese dictionary, but today there is (Saralegi et al., 2012). While the methods for building dictionaries from parallel corpora are now mature (Melamed, 2000), there is a dearth of bilingual or even monolingual material (Zséder et al., 2012), hence the increased interest in comparable corpora.

Once we find bilingual speakers capable of carrying out a manual evaluation of representative samples, it is relatively easy to measure the precision of a dictionary built by automatic methods. But measuring recall remains a challenge, for if there existed a high quality machine-readable dictionary (MRD) to measure against, building a new one would largely be pointless, except perhaps as a means of engineering around copyright restrictions. We could measure recall against Wiktionary, but of course this is a moving target, and more importantly, the coverage across language pairs is extremely uneven.

What we need is a standardized vocabulary resource that is equally applicable to all language pairs. In this paper we describe our work toward creating such a resource by extending the *4lang* conceptual dictionary (Kornai and Makrai, 2013) to the top 40 languages (by Wikipedia size) using a variety of methods. Since some of the resources studied here are not available for the initial list of 40 languages, we extended the original list to 50 languages so as to guarantee at least 40 languages for every method. Throughout the paper, results are provided for all 50 languages, indicating missing data as needed.

Section 1 outlines the approach taken toward defining the basic vocabulary and translational equivalence. Section 2 describes how Wiktionary itself measures up against the 4lang resource directly and after triangulation across language pairs. Section 2.3 and Section 2.4 deals with extraction from multiply parallel and near-parallel corpora, and Section 3 offers some conclusions.

1 Basic vocabulary

The idea that there is a *basic* vocabulary composed of a few hundred or at most a few thousand elements has a long history going back to the Renaissance – for a summary, see Eco (1995). The first modern efforts in this direction are Thorndike's (1921) *Word Book,* based entirely on frequency counts (combining TF and DF measures), and Ogden's (1944) *Basic English,* based primarily on considerations of definability. Both had lasting impact, with Thorndike's approach forming the basis of much subsequent work on readability (Klare 1974, Kanungo and Orr 2009) and Ogden's forming the basis of the Simple English Wikipedia[1]. An important landmark is the Swadesh (1950) list, which puts special emphasis on cross-linguistic definability, as its primary goal is to support glottochronological studies.

Until the advent of large MRDs, the frequency-based method was much easier to follow, and Thorndike himself has extended his original list of ten thousand words to twenty thousand (Thorndike

[1] http://simple.wikipedia.org

1931) and thirty thousand (Thorndike and Lorge 1944). For a recent example see Davies and Gardner (2010), for a historical survey see McArthur (1998). The main problem with this approach is the lack of clear boundaries both at the top of the list, where function words dominate, and at the bottom, where it seems quite arbitrary to cut the list off after the top three hundred words (Diedrich 1938), the top thousand, as is common in foreign language learning, or the top five thousand, especially as the frequency curves are generally in good agreement with Zipf's law and thus show no obvious inflection point. The problem at the top is perhaps more significant, since any frequency-based listing will start with the function words of the language, characterizing more its grammar than its vocabulary. For this reason, the list is highly varied across languages, and what is a word (free form) in one language, like English *the*, often ends up as an affix (bound form) in another, like the Romanian suffix *-ul*. By choosing a frequency-based approach, we inevitably put the emphasis on comparing grammars and morphologies, instead of comparing vocabularies.

The definitional method is based on the assumption that dictionaries will attempt to define the more complex words by simpler ones. Therefore, starting with any word list L, the list $D(L)$ obtained by collecting the words appearing on the right-hand side of the dictionary definitions will be simpler, the list $D(D(L))$ obtained by repeating the method will be yet simpler, and so on, until we arrive at an irreducible list of basic words that can no longer be further simplified. Modern MRDs, starting with the Longman Dictionary of Contemporary English (LDOCE), generally enforce a strict list of words and word senses that can appear in definitions, which guarantees that the basic list will be a subset of this defining vocabulary. This method, while still open to charges of arbitrariness at the high end, in regards to the separation of function words from basic words, creates a bright line at the low end: no word, no matter how frequent, needs to be included as long as it is not necessary for defining other words.

In creating the 4lang conceptual dictionary (Kornai and Makrai, 2013), we took advantage of the fact that the definitional method is robust in terms of choosing the seed list L, and built a seed of approximately 3,500 entries composed of the Longman Defining Vocabulary (2,200 entries), the most frequent 2,000 words according to the Google unigram count (Brants and Franz 2006) and the BNC, as well as the most frequent 2,000 words from Polish (Halácsy et al 2004) and Hungarian (Kornai et al 2006). Since Latin is one of the four languages supported by 4lang (the other three being English, Polish, and Hungarian), we added the classic Diederich (1938) list and Whitney's (1885) *Roots*.

The basic list emerging from the iteration has 1104 elements (including two bound morphemes but excluding technical terms of the formal semantic model that have no obvious surface reflex). We will refer to this as the *basic* or *uroboros* set as it has the property that each of its members can be defined in terms of the others, and we reserve the name *4lang* for the larger set of 3,345 elements from which it was obtained. Since 4lang words can be defined using only the uroboros vocabulary, and every word in the Longman Dictionary of Contemporary English can be defined using the 4lang vocabulary (since this is a superset of LDV), we have full confidence that every sense of every non-technical word can be defined by the uroboros vocabulary. In fact, the Simple English Wikipedia is an attempt to do this (Yasseri et al., 2012) based on Ogden's Basic English, which overlaps with the uroboros set very significantly (Dice 0.527).

The lexicographic principles underlying 4lang have been discussed elsewhere (Kornai, 2012; Kornai and Makrai, 2013), here we just summarize the most salient points. First, the system is intended to capture everyday vocabulary. Once the boundaries of natural language are crossed, and goats are defined by their set of genes (rather than an old-fashioned taxonomic description involving cloven hooves and the like), or derivative is defined as $\lim_{\Delta \to 0}(f(x+\Delta) - f(x))/\Delta$, the uroboros vocabulary loses its grip. But for the non-technical vocabulary, and even the part of the technical vocabulary that rests on natural language (e.g. legal definitions or the definitions in philosophy and discursive prose in general), coverage of the uroboros set promises a strategy of gradually extending the vocabulary from the simple to the more complex. Thus, to define *Jupiter* as 'the largest planet of the Sun', we need to define *planet*, but not *large* as this item is already listed in the uroboros set. Since *planet* is defined 'as a large body in space that moves around a star', by substitution we will obtain for Jupiter the definition 'the

largest body in space that moves around the Sun' where all the key items *large, body, space, move, around* are part of the uroboros set. Proper nouns like *Sun* are discussed further in (Kornai, 2010), but we note here that they constitute a very small proportion (less than 6%) of the basic vocabulary.

Second, the ultimate definitions of the uroboros elements are given in the formal language of machines (Eilenberg, 1974), and at that level the English words serve only a mnemonic purpose, and could in principle be replaced by any arbitrary names or even numbers. Because this would make debugging next to impossible, as in purposely obfuscated code, we resort to using English printnames for each concept, but it is important to keep in mind that these are only weakly reflective of the English word. For example, the system relies heavily on an element `has` that indicates the possessive relation both in the direct sense, as in *the Sun's planet, the planet of the Sun* and in its more indirect uses, as in *John's favorite actress* where there is no question of John being in possession of the actress. In other languages, `has` will generally be translated by morphemes (often bound morphemes) indicating possession, but there is no attempt to cross-link all relevant uses. The element *has* will appear in the definition of Latin *meus* and *noster* alike, but of course there is no claim that English *has* underlies the Latin senses. If we know how to express the basic vocabulary elements in a given language, which is the task we concentrate on here, and how to combine the expressions in that language, we are capable of defining all remaining words of the language.

In general, matching up function words cross-linguistically is an extremely hard task, especially as they are often expressed by inflectional morphology and our workflow, which includes stemming, just strips off the relevant elements. Even across languages where morphological analysis is a solved task, it will take a great deal of manual work to establish some form of translational equivalence, and we consider the issue out of scope here. But for content words, the use of language-independent concepts simplifies matters a great deal: instead of finding $\binom{40}{2}$ translation pairs for the 3,384 concepts that already have manual bindings in four languages (currently, Latin and Polish are only 90% complete), our goal is only to find reasonable printnames for the 1,104 basic concepts in all 40 languages. Translation pairs are only obtained indirectly, through the conceptual pivot, and thus do not amount to fully valid bilingual translation pairs. For example, *he-goat* in one language may just get mapped to the concept *goat*, and if *billy-goat* is present in another language, the strict translational equivalence between the gendered forms will be lost because of the poverty of the pivot. Nevertheless, rough equivalence at the conceptual level is already a useful notion, especially for filtering out candidate pairs produced by more standard bilingual dictionary building methods, to which we now turn.

2 Wiktionary

Wiktionary is a crowdsourced dictionary with many language editions that aim at eventually defining 'all words'. Although Wiktionary is primarily for human audience, since editors are expected to follow fairly strict formatting standards, we can automate the data extraction to a certain degree. While not a computational linguistic task par excellence, undoing the MediaWiki format, identifying the templates and simply detecting the translation pairs requires a great deal of scripting. Some Wiktionaries, among others the Bulgarian, Chinese, Danish, German, Hungarian, Korean, and Russian, are formatted so heterogeneously that automated extraction of translation pairs is very hard, and our results could be further improved.

Table 1 summarizes the coverage of Wiktionary on the basic vocabulary from the perspective of translation pairs with one manual member, English, Hungarian, Latin, and Polish respectively. The last column represents the overall coverage combining all four languages. As can be seen, the better resourced languages fare better in Wiktionary as well, with the most translations found using English as the source language (64.9% on the smaller basic set, and 64% on the larger 4lang vocabulary), Polish and Hungarian faring about

Table 1: 4lang coverage of Wiktionary data.

	Based on				
	en	hu	la	pl	all
4lang	59.43	22.09	7.9	19.6	64.01
uroboros	60.29	22.88	9.11	21.09	64.91

equally well, although the Polish list of 4lang has more missing bindings, and the least resourced Latin faring the worst.

Another measure of coverage is obtained by seeing how many language bindings are found on the average for each concept: 65% on 4lang and 64% for the basic set (32 out of the 50 languages considered here).

2.1 Triangulating

Next we used a simple triangulation method to expand the collection of translation pairs, which added new translation pairs if they had been linked with the same word in a third language. An example, the English:Romanian pair *guild:breaslă*, obtained through a Hungarian pivot, is shown in Figure 1.

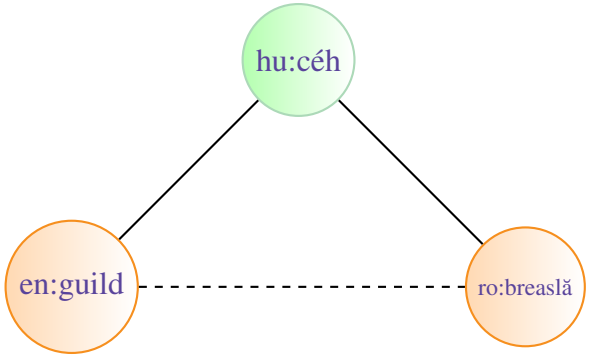

Figure 1: The non-dashed edge represents translation pairs extracted directly from the Wiktionaries. The pair *guild–breaslă* were found via triangulating.

While direct translation pairs come from the manually built Wiktionaries and can be considered gold (not entirely without reservations, but clearly over 90% correct in most language pairs we could manually spot-check), indirect pairs must be viewed with considerable suspicion, as multiple word senses bring in false positives quite often. Using 3,317,861 pairs extracted from 40 Wiktionaries, we obtained a total of 126,895,236 indirect pairs, but in the following table we consider only those that were obtained through at least two different third-language pivots with the pairs originating from different Wiktionaries, and discarded the vast majority, leaving 5,720,355 pairs that have double confirmation. Manual checking proved that the quality of these pairs is comparable to that of the original data (see Table 7). A similar method, within one dictionary rather than

Table 2: 4lang coverage of triangulating.

	Based on				
	en	hu	la	pl	all
4lang	76.09	64.91	43.25	53.74	85.81
basic	77.81	64.74	48.07	58.55	86.97

Table 3: 4lang coverage of Wiktionary data and triangulating.

	Based on				
	en	hu	la	pl	all
4lang	80.77	65.69	43.63	54.30	86.80
basic	82.07	65.47	48.41	59.13	87.81

across several, was used in (Saralegi et al., 2012) to remove triangulation noise. Since recall would be considerably improved by some less aggressive filtering method, in the future we will also consider improving the similarity scores of our corpus-based methods using the single triangles we now discard.

Triangulating by itself improves coverage from 65% to 85.8% (4lang) and from 64% to 87% (basic), see Table 2. Table 3 shows the combined coverage which is not much different from Table 2 but considering that the triangulating used the Wiktionary data as input, we expected a very large intersection (it turned out to be more than 40% of the pairs acquired through triangulating). The average number of language bindings also improves significantly, to 43.5/50 (4lang) and 44/50 (basic).

2.2 Wikipedia titles

Another crowdsourced method that promises great precision is comparing Wikipedia article titles across languages: we extracted over 187m potential translation pairs this way. Yet the raw data is quite noisy, for example French *chambre* points to English *Câmara*, an article devoted to the fact that 'Câmara (meaning 'chamber') is a common surname in the Portuguese language' rather than to some article on *bedroom, room,* or *chamber*. We filtered this data in several ways. First, we discarded all pairs that contain words that appear five or fewer times in the frequency count generated from the language in question. This reduced the

Table 4: 4lang coverage of Wikipedia interwiki links (langlinks).

	Based on				
	en	hu	la	pl	all
4lang	21.51	14.4	9.54	12.26	31.74
basic	20.7	13.0	10.22	13.43	31.32

number of pairs to 15m. Most of these, unfortunately, are string-identical across languages, leaving us with a total of 6.15m nontrivial translation pairs. A large portion of these are named entities that do not always add meaningfully to a bilingual dictionary.

The average number of language bindings is 16.5 and 12.6 respectively. The combined results improve slightly as shown in Table 8.

2.3 Parallel texts

Using the Bible as a parallel text in dictionary building has a long tradition (Resnik et al., 1999). Somewhat surprisingly in the age of parallel corpora, the only secular text available in all our languages is the Universal Declaration of Human Rights, which is simply too short to add meaningfully to the coverage obtained on the Bible. In addition to downloading the collection at *http://homepages.inf.ed.ac.uk/s0787820/bible*, we used *http://www.jw.org* (for Dutch, Armenian and Korean), *www.gospelgo.com* (for Catalan, Kazakh, Macedonian, Malay and Persian), *http://www.biblegateway.com* (for Czech), *http://biblehub.com* (for English) and *http://www.mek.oszk.hu* (for Hungarian). To the extent feasible we tried to use modern Bible translations, resorting to more traditional translations only where we could not identify a more contemporary version.

The average number of languages with translations found is 19 (basic) and 17.8 (4lang). These numbers are considerably weaker than the crowdsourced results, suggesting that the dearth of multiply parallel texts, even in the best resourced group of 40 languages, needs to be addressed.

Table 5: 4lang coverage of the Bible data.

	Based on				
	en	hu	la	pl	all
4lang	19.64	15.17	13.78	14.13	35.49
basic	21.47	17.12	15.67	15.78	38.13

2.4 Comparable texts

Comparable corpora were built from Wikipedia articles in the following manner. For each language pair, we considered those articles that mutually linked each other, and took the first 50 words, excluding the title itself. Article pairs whose length differed drastically (more than a factor of five) were discarded.

Table 6: 4lang coverage of the dictionary extracted from Wikipedia as comparable corpora.

	Based on				
	en	hu	la	pl	all
4lang	5.58	5.66	4.30	4.96	16.00
basic	5.70	5.86	4.93	5.39	16.77

The 4lang coverage based solely on the translations acquired from comparable corpora is presented in Table 6. The average number of languages with translations found is 8 (basic) and 8.4 (4lang).

2.5 Evaluation

We used manual evaluation for a small subset of language pairs. Human annotators received a sample of 100 translation candidate-per-method. The samples were selected from translations that were found by only one method, as we suspect that translations found by several methods are more likely to be correct. Using this strict data selection

Table 7: Manual evaluation of extracted pairs that do not appear in more than one dictionary.

	Wikt	Tri	Title	Par	Comp
cs-hu	82	81	95	41	40
de-hu	92	87	96	46	68
fr-hu	76	80	89	43	54
fr-it	79	79	92	43	36
hu-en	87	75	92	28	63
hu-it	94	93	93	35	61
hu-ko	87	85	99	N/A	N/A
avg	85.3	82.9	93.7	39.3	53.7

criterion we evaluated the *added quality* of each method. Results are presented in Table 7. It is clear that set next to the crowdsourced methods, dictionary extraction from either parallel or comparable corpora cannot add new translations with high precision. When high quality input data is available, triangulating appears to be a powerful yet simple method.

3 Conclusions and future work

The major lesson emerging from this work is that currently, crowdsourced methods are considerably more powerful than the parallel and comparable corpora-based methods that we started with. The reason is simply the lack of sufficiently large parallel and near-parallel data sets, *even among the most commonly taught languages*. If one is actually interested in creating a resource, even a small resource such as our basic vocabulary set, with bindings for all 40 languages, one needs to engage the crowd.

Table 8: Summary of the increase in 4lang coverage achieved by each method. Wikt: Wiktionary, Tri: triangulating, WPT: Wikipedia titles, Par: the Bible as parallel corpora, WPC: Wikipedia articles as comparable corpora

Src	Set	Based on				
		en	hu	la	pl	all
Wikt	4lang	59.43	22.09	7.90	19.6	64.01
	basic	60.29	22.88	9.11	21.09	64.91
Tri	4lang	80.77	65.69	43.63	54.3	86.8
	basic	82.07	65.47	48.41	59.13	87.81
WPT	4lang	81.39	66.27	44.2	54.66	87.39
	basic	82.51	65.86	48.89	59.53	88.17
Par	4lang	82.22	67.35	45.99	55.4	88.22
	basic	83.27	67.04	50.62	60.25	88.91
WPC	4lang	81.56	66.49	44.42	54.77	87.58
	basic	82.66	66.06	49.14	59.62	88.33

The resulting *40lang* resource, currently about 88% complete, is available for download at *http://hlt.sztaki.hu*. The Wiktionary extraction tool is available at *https://github.com/juditacs/wikt2dict*. 40lang, while not 100% complete and verified, can already serve as an important addition to existing MRDs in several applications. In comparing corpora the extent vocabulary is shared across them is a critical measure, yet the task is not trivial even when these corpora are taken from the same language. We need to compare vocabularies at the conceptual level, and checking the shared 40lang content between two texts is a good first cut. Automated dictionary building itself can benefit from the resource, since both aligners and dictionary extractors benefit from known translation pairs.

Acknowledgments

The results presented here have improved since Ács (2013). Ács did the work on Wiktionary and Wikipedia titles, Pajkossy on parallel corpora, Kornai supplied the theory and advised. The statistics were created by Ács. We thank Attila Zséder, whose HunDict system (see *https://github.com/zseder/hundict*) was used on the (near)parallel data, for his constant support at every stage of the work. We also thank our annotators: Klára Szalay, Éva Novai, Angelika Sándor, and Gábor Recski.

The research reported in the paper was conducted with the support of the EFNILEX project *http://efnilex.efnil.org* of the European Federation of National Institutions for Language *http://www.efnil.org*, and OTKA grant #82333.

References

Judit Ács. 2013. Intelligent multilingual dictionary building. *MSc Thesis, Budapest University of Technology and Economics*.

Thorsten Brants and Alex Franz. 2006. Web 1T 5-gram version 1.

Peter Brown, John Cocke, Stephen Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16:79–85.

M. Davies and D. Gardner. 2010. *A Frequency Dictionary of Contemporary American English: Word Sketches, Collocates, and Thematic Lists*. Routledge Frequency Dictionaries Series. Routledge.

Paul Bernard Diederich. 1939. *The frequency of Latin words and their endings*. The University of Chicago press.

Umberto Eco. 1995. *The Search for the Perfect Language*. Blackwell, Oxford.

Samuel Eilenberg. 1974. *Automata, Languages, and Machines*, volume A. Academic Press.

Péter Halácsy, András Kornai, László Németh, András Rung, István Szakadát, and Viktor Trón. 2004. Creating open language resources for Hungarian. In

Proceedings of the 4th international conference on Language Resources and Evaluation (LREC2004), pages 203–210.

Tapas Kanungo and David Orr. 2009. Predicting the readability of short web summaries. In *2nd ACM Int. Conf. on Web Search and Data Mining*.

George R. Klare. 1974. Assessing readability. *Reading Research Quarterly*, 10(1):62–102.

András Kornai and Márton Makrai. 2013. A 4lang fogalmi szótár [the 4lang concept dictionary]. In A. Tanács and V. Vincze, editors, *IX. Magyar Számitógépes Nyelvészeti Konferencia [Ninth Conference on Hungarian Computational Linguistics]*, pages 62–70.

A. Kornai, P. Halácsy, V. Nagy, Cs. Oravecz, V. Trón, and D. Varga. 2006. Web-based frequency dictionaries for medium density languages. In A. Kilgarriff and M. Baroni, editors, *Proc. 2nd Web as Corpus Wkshp (EACL 2006 WS01)*, pages 1–8.

András Kornai. 2010. The algebra of lexical semantics. In Christian Ebert, Gerhard Jäger, and Jens Michaelis, editors, *Proceedings of the 11th Mathematics of Language Workshop*, LNAI 6149, pages 174–199. Springer.

András Kornai. 2012. Eliminating ditransitives. In Ph. de Groote and M-J Nederhof, editors, *Revised and Selected Papers from the 15th and 16th Formal Grammar Conferences*, LNCS 7395, pages 243–261. Springer.

Tom McArthur. 1998. *Living Words: Language, Lexicography, and the Knowledge Revolution*. Exeter Language and Lexicography Series. University of Exeter Press.

I Dan Melamed. 2000. Models of translational equivalence among words. *Computational Linguistics*, 26(2):221–249.

C.K. Ogden. 1944. *Basic English: A General Introduction with Rules and Grammar*. Psyche miniatures: General Series. Kegan Paul, Trench, Trubner.

Philip Resnik, Mari Broman Olsen, and Mona Diab. 1999. The bible as a parallel corpus: Annotating the 'Book of 2000 Tongues'. *Computers and the Humanities*, 33(1-2):129–153.

Xabier Saralegi, Iker Manterola, and Iñaki San Vicente. 2012. Building a basque-chinese dictionary by using english as pivot. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Morris Swadesh. 1950. Salish internal relationships. *International Journal of American Linguistics*, pages 157–167.

Edward L. Thorndike and Irving Lorge. 1944. *The teacher's word book of 30,000 words*. Teachers College Bureau of Publications.

Edward L. Thorndike. 1921. *The teacher's word book*. New York Teachers College, Columbia University.

E.L. Thorndike. 1931. *A teacher's word book*. New York Teachers College, Columbia University.

William Dwight Whitney. 1885. The roots of the Sanskrit language. *Transactions of the American Philological Association (1869-1896)*, 16:5–29.

Taha Yasseri, András Kornai, and János Kertész. 2012. A practical approach to language complexity: a wikipedia case study. *PLoS ONE*, 7(11):e48386. doi:10.1371/journal.pone.0048386.

Attila Zséder, Gábor Recski, Dániel Varga, and András Kornai. 2012. Rapid creation of large-scale corpora and frequency dictionaries. In *Proceedings to LREC 2012*.