Towards Language Acquisition Through Cross-Language Etymological Links in Linguistic Linked Open Data

Maxim Dužij
Prague University
of Economics and Business
Nám. W. Churchilla 4
130 67 Praha 3, Czech Rep.
maxim.duzij@hotmail.com

Vojtěch Svátek
Prague University
of Economics and Business
Nám. W. Churchilla 4
130 67 Praha 3, Czech Rep.
svatek@vse.cz

Petr Strossa
Prague University
of Economics and Business
Nám. W. Churchilla 4
130 67 Praha 3, Czech Rep.
petr.strossa@vse.cz

Abstract

We explore the possibility of using linguistic linked open data for supporting a foreign language acquisition application through cross-language links. The links in the used LLOD resource, the Etytree knowledge graph, are primarily of etymological nature. Through a questionnaire survey we explore what interval of an edit distance measure may be suitable as guidance for offering word pairs (in an unknown and known language), connected with an etymological chain, that are too dissimilar to immediately remind of the learned word when encountering the known word but allowing to establishing a mental association between them when seeing both. A proof-of-concept application was also designed and tested for usability. While the principles of the approach look viable after this initial study, our conclusion is that large-scale enhancement of the underlying LLOD resources will be needed before tools could be delivered for real use. An edit distance measure, particularly one sensitive to cross-language character mapping, may be useful for selecting training cases with respect to the language-acquisition proficiency of the learner.

1 Introduction

One of the important aspects of linguistic linked open data (LLOD) is the consideration of cross-language links. While many efforts have been centred on semantic equivalence links, useful for tasks such as search or translation, less attention has been paid to etymological links (whether cross- or intra-language ones). A prominent recent project is Etytree (Pantaleo et al., 2017), which produced a tool for interactively exploring etymologically related words. Its target user group are the researchers and public interested in the study of etymology, who can benefit from intuitive graph-based visualization of etymological links.

We hypothesize that another beneficiary of LLOD with etymology coverage could be *foreign language learners*. Experts generally agree that etymology is one of language aspects (together with phonology, morphology, semantics and syntax) relevant for language acquisition (Rothstein and Rothstein, 2008). However, the studies have so far been focused on classroom educational setting, and largely agnostic of support that could be provided by online databases.

Presumably, the benefits of etymology would vary across several dimensions of language learning, such as: the prior knowledge of the target (to-be-learned) and background (native or better commanded) language/s by the learner; the closeness of those languages as such; active vs. passive vocabulary acquisition setting; written vs. spoken form of the language; personal characteristics of the learner. As a promising case we want to primarily focus on is that of *passive* acquisition of (primarily) *written* form of words in the target language that has *observable* but *not strikingly obvious* etymologically justified surface similarity to words in a background language the learner knows better. Since the probability of finding such background language words increases with the number (and, perhaps, taxonomic variety) of mastered background languages, the gain might be highest for learners moderately or highly equipped with prior knowledge of languages, who at the same time experience limitations in pure memorization of words and their meanings by heart. Let us consider the following scenario:

1. The learner is exposed to a word in the target language.
2. S/he acquires the meaning of the word using a dictionary or thesaurus.
3. In the course of time, s/he encounters the word repeatedly, and has to look the meaning up again and again – until the bond between the written word and its meaning becomes firm enough.

The key question is whether showing the word together with a *personalized* etymological context, in step 2, would reduce the number of repeated look-ups in the next phase. Obviously, while showing a given word with its generic etymological context (as performed by the Etytree application) is not much different from what even paper-based etymological resources can provide, the power of LLOD knowledge graphs might nicely manifest through such dynamically generated, personalized views.

Imagine two foreign visitors to Sweden, A and B, whose mother tongue has no manifested similarity to Swedish, and none of them has any knowledge of Swedish yet. A only knows her/his mother tongue, while B knowledge a bit of English and German. They both come across the words[1] “Akta huvudet!” on a sign, and acquire its meaning via translation to their mother tongue, which is “Mind your head!”. As regards A, for the future comprehension of these or related lexemes s/he only depends on memorization. In contrast, B could benefit from her/his prior knowledge as follows:

- ‘huvud/et’ has a surface similarity to its English equivalent, ‘head’
- ‘akta/r’, in turn, does not such an obvious link for English – where instead, *false friends* such as ‘acting’ pop up. However, it does have them for German, where the ‘*achten’ family of verbs and the ‘Achtung’ noun are a part of the basic vocabulary for foreign learners.

Now, the key questions are:

1. Is it likely that B would *fail* to *directly* see the cross-language link/s?
2. Is it likely that B would *understand an etymological explanation* of the link/s if it were served to him/her?
3. Would the awareness of the etymological link positively influence the *remembering* of the meaning of the words by B, in long term? (Would B on the next occasion bow her/his head instead of invoking the translation service again prior to entering the building...?)

If the answers to all these questions are positive then the example witnesses the relevance of the research line started in this paper.

In the presented preliminary research we thus aim at exploring various issues related to the prospects of using *personalized etymological context* of words, provided via *LLOD knowledge graphs*, in *foreign passive written vocabulary* acquisition. The main axes of this research are:

- Analysis of *LLOD resources* with respect to coverage of etymological links
- Study of *cross-language word pairs* returned via such links, with respect to their ‘adequate’ adoption through etymology, in terms of the first two questions above – i.e., not too trivial (which would make the etymological explanation redundant), but not too hard either (as the words may then elude adoption even with such an explanation).
- Study of actual (longer-term) learnability of word pairs, through a *prototype application*.

Those three axes roughly correspond to the next three sections of the paper.

2 Etymological Linked Data Sources and their Limitations

By a brief analysis of the available resources, it appears that LLOD sources covering etymology have been partially or fully created using an extractor from Wiktionary, since other etymological resources are typically copyright-protected.[2] Note however that Wiktionary itself, being one of the biggest online sources of word etymology, is essentially an unstructured source and cannot be used directly for our purposes. We identified two relevant: *Dbnary* (Sérasset, 2015) and *Etytree* (Pantaleo et al., 2017). The former is a generic approach to Wiktionary extraction, while the latter specifically focusee on etymology and employs relatively advanced NLP-based extractors. Because of our focus on etymological relations between the languages, Etytree was selected as our primary source of data for the language acquisition (micro-)study.

It is not possible to straightforwardly interlink the two sources, as they employ each its specific set of unique identifiers and are not directly interlinked. The only connection are the *seeAlso* links that lead from Etytree entities to Wiktionary pages.

[1] We use an example in the form of a phrase in order to make the example more comprehensive. Admittedly, the research described later in the paper does not attempt to go from isolated words to the meaning of phrases.

[2] This is probably the reason why data from `https://starlingdb.org/` have not been published, although their RDF converter (Abromeit et al., 2016) exists.

	English	Latin	German	French
English	**2157076**	46624	3220	13910
Latin	46624	**230754**	4166	24700
German	3220	4166	**328340**	3442
French	13910	24700	3442	**214958**

Table 1: Number of *:etymologicallyRelatedTo* and *:etymologicallyDerivesFrom* predicate occurrences in selected languages

Prior to starting the study, we computed the number etymology links in Etytree and its proportion wrt. the number of entities, for a subset of language, in order to be able to estimate the exploitability of this resource. The result, for four major languages, is in Tab. 1. It is apparent that there the majority of etymological links hold just within a language, and only few hold between different languages.

3 Cross-language Word Pair Analysis

Our goal was to correlate the surface similarity of etymologically related words with their perceived learnability. For this purpose, we needed to express this *surface similarity* using a suitable metric. Since our target was the written vocabulary, we had preference for *edit distance* measures over pronunciation-oriented measures such as Soundex[3] (which are also more language-dependent). Edit distances count the number (or sum up the costs) of operations that must be performed to transform one string into another, see e.g. an overview (Navarro, 2001). Probably the most widely used one is the Levenshtein distance, which counts the least number of single-character insertions, deletions, and replacements. Other known measures or algorithms are e.g. Hamming distance, Jaro-Winkler distance or Damerau–Levenshtein.

We eventually opted for the *Cross-Language Levenshtein Distance* (CLLD) (Medhat et al., 2015), which supports matching names across different writing scripts and uses many-to-many mapping characters. If the mapping is successful, the partial Levenshtein distance for a specific character is ignored. The intended target for this technique had indeed been the mapping between different scripts. We have however transferred the mapping-character heuristic to a somewhat different target. Namely, our intuition was that *etymologically grounded character mappings* (an example of which is, e.g., the orthographic reflection of the well-known High-German consonant shift) between the target and background language/s can be to some degree appropriated by the learners (even without full understanding of the etymological circumstances). Thus words differing along such mappings should have a smaller distance than those differing in other ways. Since we were unable to easily find a structured resource of cross-language character mappings, we provisionally created *ad hoc mappings* analytically, based on our speaker experience, namely, between English and two other major languages, German and French. Examples of such mappings are “th → d” or “p → f” for English vs. German. There were 22 pairs overall, of which 15 for German and 7 for French.

Next we created a *questionnaire*, aimed at general public, to which we manually selected *word pairs* such that:

- The target language word was always a *German* one and the background language word was always an *English* one.
- The words in the pair were connected by an *etymological link* in Etytree, i.e., they were chosen from the set of 3 220 linked words as indicated in Table 1.
- The *CLLD distance* of the pair varied between 1-6.

The choice of German and English was motivated by the following. English is a known language for a high number of learners. It is also the hub language of Etytree, with the highest number of cross-language links. German, in turn, features many word-level etymological links with English due to their partially shared roots. It is also an official language of several EU countries, thus many people learn it as a foreign language.

In total, seven-word pairs were manually selected, see Table 2. The questionnaire displayed for each pair[4] the following question: “*After reviewing this etymologically related word pair, do you think a learner can later remember the meaning of the **foreign word** when seeing it in written form?*”. The answer was a choice among three options (plus the possibility to provide one’s own answer):

- *Yes, the learner will surely remember it. The words are almost the same. Upon seeing the*

[3] https://www.archives.gov/research/census/soundex

[4] German nouns, except for proper nouns, were displayed as decapitalized.

German word, its English equivalent will immediately occur to the learner.

- *Unsure if the learner will remember it. The words are somewhat different. Seeing the German word might or might not "ring the bell" with reference to the English word.*

- *It is unlikely that the learner will remember it. The words are too different.*

The foreword to the questionnaire also suggested the users to always abstract from their familiarity with either word and provide feedback relative to their expectation of a learner who would know the English word but wouldn't know the German word.

The design of the study already revealed some limitations of the current setting. First and foremost, the number of etymological links was not only small with respect to the total vocabulary of both languages (less than 1% wrt. German and less than 0.15% wrt. English, see Table 1), but it was also biased towards words with *very high visual similarity*, such as #1 and #2. Finding 'interesting' pairs with manifestation of mapping rules, such as #5-#7, was not easy. There are also many *proper names* among the linked words (such as #3 and #4). Those might be less useful in language acquisition, first, because their translation between languages is not essential for communication, and second, because their frequency of occurrence is on average lower than that of common nouns. This also leads us to the suggestion that etymological resources should be used for suggesting word pairs in combination with a source of *word occurrence frequency* information. Finally, #3 also possibly manifests three natural deficiencies of the CLLD metric: (1) setting the contribution of the mapped characters to the CLLD to *zero* is an overshot; (2) *very short words* exhibit low distance despite being apparently rather dissimilar; (3) CLLD also (contrary to the commonsense of word similarity perception) does not distinguish the first letter in the calculation.

In this respect it should be noted that the scope of our word pair analysis was intentionally bound to pairs that *truly originate from our LLOD resource*. This on the one hand limits the variety of cases considered, but on the other hand contributes to the assessment whether benefits to language acquisition can be obtained even for the present-day, modest, availability of etymological links in LLOD.

The questionnaire was sent to members of general public; most audience were young university students or graduates. It returned filled by 29 respondents. Only the first three answer options (we will nick them 'Yes', 'Unsure' and 'Unlikely') were used overall. By the distribution of these answers, the cases (word pairs) can be relatively clearly ranged into three apparent clusters:

#	English	German	CLLD
1	transphenomenal	transphänomenal	1
2	heuristic	heuristisch	2
3	Vaud	Waadt	3
4	Nuremberg	Nürnberg	3
5	ravenstone	rabenstein	3
6	oversightly	übersichtlich	5
7	sharpshooter	scharfschütze	6

Table 2: Questionnaire word pairs and their CLLD

- #1 and #2 (CLLD ≤ 2) got 'Yes' from over 90% of respondents. We hypothesize that for such pairs the etymological links might help less-proficient language learners, but would be of limited value for experienced learners, since they could see the correspondence even without having been pointed to it.

- #4 got 'Yes' from over 60% of respondents, and 'Unsure' from the remaining ones. We hypothesize that for such pairs the etymological links might help the majority of language learners. Note that, however, #4 is inseparable from #3 and #5 through CLLD. Its shifted score might be influenced by the proper name nature of the word/s, which reduces the space of notions to be matched, as well as by the match at the beginning and end of the strings.

- #3, #5, #6 and #7 got 'Yes' from 7-20% of respondents, 'Unsure' from 34-52%, and 'Unlikely' from 34-48%. We can hypothesize that for such word pairs the etymological links might help advanced learners who would possibly either be explicitly aware of or intuitively adopt some of the mapping rules.

We also consequently prepared another questionnaire, this time addressing *linguistics/lexicography experts* (members of the Language Acquisition workgroup of the Nexus Linguarum COST Action[5]). It contained the same word pairs, but provided additional background information (e.g., about the nature and values of the CLLD measure),

[5] https://nexuslinguarum.eu/

prompted at entering qualitative responses on the word pairs, and also featured a set of general questions such as: "*Do you think it is more beneficial to learn etymologically connected short words rather than long words?*" or "*Do you think it is more beneficial to learn a pair of words that have the same meaning or, rather, a pair of words that have different meanings? The meaning will be shown during the learning process. Different meanings: gift (present) (en) - Gift (poison) (de). Same meaning: house (en) - Haus (de)*".

We collected answers from four respondents. The feedback provided through the expert questionnaire largely confirmed the quantitative findings from the first ('lay person') questionnaire. Intersting insights were, e.g., the following:

- If the mapping rules are applied on multiple *neighboring* characters (as 'w→ v', 'aa→ au' and 'dt→ d' in 'Waadt vse. Vaud'), they might be more difficult to identify.
- For compound terms affected by mapping rules (#4–#7), it might be even difficult to correctly *tell* the different compounds *apart*.

Answers to the general questions also indicated that: both *long* and *short* words are worth learning via etymology; while pairs with the *same meaning* are a most suitable learning input for beginners, advanced learners will also benefit from pairs with *different meaning*; the coupling of written-form and *pronunciation* learning was also raised as a possible future agenda.

4 Experiment with a Proof-of-concept Vocabulary Acquisition Application

A proof-of-concept *web application*[6] was developed (in .NET with a React front end), which leverages on SPARQL[7] queries to the Etytree database for selecting word pairs from ten available languages (the mappings rules are however only used for English, German and French, as described above). Only word pairs with CLLD distance 3 or smaller are considered by the application; pairs whose strings were either identical or only differing in diacritics are also ignored. Word *meanings* are also retrieved and presented to the user; this among other helps identify words that are 'false friends' despite being etymologically related.

The users are required to create their account and to select their known and unknown languages. The *learning phase* then consists in accepting/rejecting word pair candidates for later testing, see Fig. 1. The system relies on an SQL Server Database to cache the results of the SPARQL endpoint, and this, in turn, enables a more tailored user experience. New word pairs are retrieved from the SPARQL endpoint only in case all word pairs from cache have been used. Such an architectural decision enables *collaborative filtering*: word pairs rejected by too many users are filtered out for new users. Then the user proceeds to the *testing phase*, when the previously approved word pairs are presented, but the word in the known language is left blank; the user is to complete the pair. If s/he fails to do so, the correct answer is revealed. The number of words revealed is a metric for overall test success.

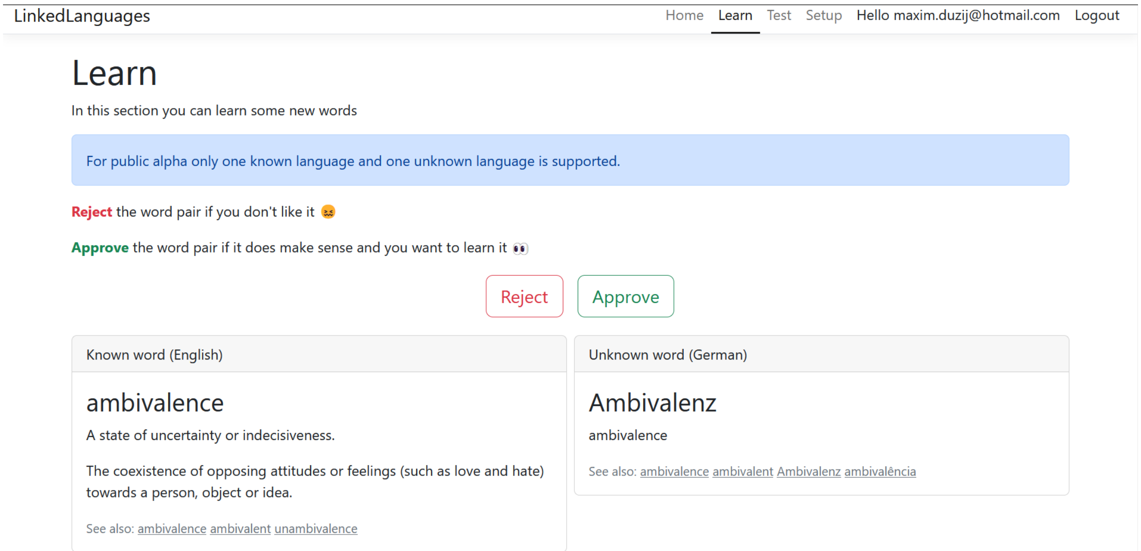
During a weeklong user testing phase, 20 users used the application, and 1 725 times word pairs were either rejected or approved by users; 391 of these were either learned or revealed. Eventually, the application was formally evaluated via a *questionnaire*, which was filled by 11 users. Their responses were collected both for the common *System Usability Scale* (SUS) (Brooke et al., 1996) and for a few application-specific questions. The feedback was generally positive; the main issue reported was the fact that the application proposed 'niche word pairs' that were not beneficial for an average learner. This is however related to the issues with the word pair source. The average SUS score was 69.5, which corresponds to grade B – "Good".

5 Conclusion and Further Work

The presented research is, to our knowledge, the very first study relating language acquisition to an open etymology source on the web. It revealed that the coverage of etymological links in LLOD is so far (despite the commendable efforts in DBnary and Etytree) modest, which hinders their usage in real-world language acquisition. The major takeaway message is thus an encouragement to the community to push forward the (automated, as much as possible) *RDF-ization* of etymological paths that could become part of LLOD resources, whether bootstrapped from Wiktionary or also considering other, perhaps more even more rigorously collected database resources. Aside mere increase of *word*

[6] Source code available at https://github.com/Duzij/LinkedLanguages; online demo at https://linkedlanguages.azurewebsites.net.

[7] https://www.w3.org/TR/sparql11-query/



Learn

In this section you can learn some new words

For public alpha only one known language and one unknown language is supported.

Reject the word pair if you don't like it 😖

Approve the word pair if it does make sense and you want to learn it 👀

Known word (English)

ambivalence

A state of uncertainty or indecisiveness.

The coexistence of opposing attitudes or feelings (such as love and hate) towards a person, object or idea.

See also: ambivalence ambivalent unambivalence

Unknown word (German)

Ambivalenz

ambivalence

See also: ambivalence ambivalent Ambivalenz ambivalência

Figure 1: Proof-of-concept application interface: the learning phase

coverage, additional information on the given pairs would be beneficial, e.g., indicating whether the etymologically related word pairs are semantically *equivalent* or merely *related*. As another resource that could be of use if available within LLOD we identified *cross-language character mappings*, allowing to properly shrink the distance between etymologically related words that could be quite useful for learning that from the target language. Finally, another dimension to be considered in language acquisition is the frequency of word occurrence in the given languages – both the target and background ones. Therefore, *word frequency dictionaries* might also be exploited in future etymology-driven language acquisition applications.

In parallel, however, experiments can be undertaken even with manually constructed etymological explanations independent of LLOD, in order to study the *psychology of etymology adoption* (especially in the presence of mapping rules) in more depth – though, in contrast to earlier pure-domain-driven studies by language acquisition scholars, now also with the idea of the possible computational (LLOD-based) support in mind.

By the questionnaire (albeit limited in size), the *CLLD measure* seems to be reasonably correlated with the word pair learnability. It should be however, most likely, modified in the partial distance computation. The distance of mapped characters should be *non-zero* in general, and possibly higher at the *start* (maybe also end) of the word or for *neighboring* mapped characters, since these settings likely make the learning more difficult.

The research has been supported by the Nexus Linguarum COST Action (no. CA18209). We are indebted to G. Sérasset, E. Pantaleo and T. Di Noia for their assistance regarding Dbnary and Etytree, and to G. Hrzica, G. Valunaite Oleskevicienė, O. Dontcheva-Navrátilová and others from the LA team of Nexus Linguarum for their feedback.

References

Frank Abromeit, Christian Chiarcos, Christian Fäth, and Maxim Ionov. 2016. Linking the tower of babel: modelling a massive set of etymological dictionaries as rdf. In *LDL 2016, at LREC*, pages 11–19.

John Brooke et al. 1996. SUS-a quick and dirty usability scale. *Usability evaluation in industry*, 189(194):4–7.

Doaa Medhat, Ahmed Hassan, and Cherif Salama. 2015. A hybrid cross-language name matching technique using novel modified Levenshtein Distance. In *2015 Tenth International Conference on Computer Engineering & Systems (ICCES)*, pages 204–209. IEEE.

Gonzalo Navarro. 2001. A guided tour to approximate string matching. *ACM Comput. Surv.*, 33(1):31–88.

Ester Pantaleo, Vito Walter Anelli, Tommaso Di Noia, and Gilles Sérasset. 2017. Etytree: A graphical and interactive etymology dictionary based on wiktionary. In *Proc. 26th Int'l Conf. on World Wide Web Companion, 2017*, pages 1635–1640. ACM.

Evelyn Rothstein and Andrew S. Rothstein. 2008. *English grammar instruction that works!: developing language skills for all learners*. Corwin Press.

Gilles Sérasset. 2015. DBnary: Wiktionary as a Lemon-based multilingual lexical resource in RDF. *Semantic Web*, 6(4):355–361.