A Database and Visualization of the Similarity of Contemporary Lexicons

Gábor Bella[1][0000−0002−3868−1740],
Khuyagbaatar Batsuren[2][0000−0002−6819−5444], and
Fausto Giunchiglia[1][0000−0002−5903−6150]

[1] University of Trento, via Sommarive, 5, 38123 Trento, Italy
{gabor.bella, fausto.giunchiglia}@unitn.it
[2] National University of Mongolia, Ulanbaatar, Mongolia
khuyagbaatar.b@gmail.com

Abstract. Lexical similarity data, quantifying the "proximity" of languages based on the similarity of their lexicons, has been increasingly used to estimate the cross-lingual reusability of language resources, for tasks such as bilingual lexicon induction or cross-lingual transfer. Existing similarity data, however, originates from the field of comparative linguistics, computed from very small expert-curated vocabularies that are not supposed to be representative of modern lexicons. We explore a different, fully automated approach to lexical similarity computation, based on an existing 8-million-entry cognate database created from online lexicons orders of magnitude larger than the word lists typically used in linguistics. We compare our results to earlier efforts, and automatically produce intuitive visualizations that have traditionally been hand-crafted. With a new, freely available database of over 27 thousand language pairs over 331 languages, we hope to provide more relevant data to cross-lingual NLP applications, as well as material for the synchronic study of contemporary lexicons.

Keywords: lexical similarity · cognate · language diversity · lexicostatistics · visualization

1 Introduction

The notion of *lexical similarity*, also known as *lexical distance*, refers to a quantified comparison of the proportion of words shared across languages. It is defined by *The Ethnologue* as *"the percentage of lexical similarity between two linguistic varieties is determined by comparing a set of standardized wordlists and counting those forms that show similarity in both form and meaning."*[3] Computaton methods are typically based on the amount of *cognates*—words of common origin with (more or less) similar pronunciation and meaning—found for a given language pair. The resulting similarity data is used in comparative linguistics to

[3] https://www.ethnologue.com/about/language-info

infer or back up hypotheses of phylogeny among languages. In computational linguistics, lexical similarity has also been used in bilingual lexicon induction and, more generally, in the context of the cross-lingual transfer of language processing tools and resources, in order to estimate the differing performance of specific language pairs or directly as input features [7, 11, 13]. Graphical visualizations of lexical similarity—beyond their popularity among the general public—are useful for a quick qualitative interpretation of the similarity data.

The typical approach in comparative linguistics has been to use a small number (typically less than 100) of carefully selected words with equivalent meanings in each language studied. The word meanings are deliberately chosen from the core vocabularies, and comparisons are made strictly on phonetic representations, also taking sound changes into account in historical studies.

Because these methods have been carefully tuned to the needs of language genealogy, they are less adapted to studies characterizing contemporary vocabularies. For the purposes of computational linguistics or the synchronic study of language diversity [8], similarity information computed on "everyday" written lexicons is more representative than data deliberately tuned for historical studies. English, for instance, borrowed a significant portion of its vocabulary from (the otherwise only distantly related) French. Due to the relative lexical homogeneity of the Romance family, these French borrowings bring the English lexicon closer to Spanish, Portuguese, or Romanian as well. While such evidence of lexical proximity can be useful for computational applications, similarity data from comparative linguistics does not provide this type of insight as they consider borrowings as "noise" over phylogenetic evidence and exclude them by design.

We investigate a different approach based on the the free online *CogNet* database[4] of 8.1-million cognate pairs covering 338 languages, itself computed from large-scale online lexicons. CogNet can be considered reliable (with a precision evaluated to 96%) and is based on a permissive interpretation of the notion of cognacy that includes loanwords, and as such it is well suited to practical cross-lingual applications. From CogNet we compute pairwise similarities among 331 languages, that we make freely downloadable for downstream uses in computational linguistics, e.g. cross-lingual NLP applications. We also provide visualizations of our results that provide an immediate qualitative interpretation of the similarity data and that, contrary to prior work, are computed fully automatically.

The rest of the paper is organized as follows. Section 2 presents the state of the art with respect to known lexical similarity databases and computation methods, as well as existing visualization techniques. Section 3 describes our lexical similarity computation method. Section 4 compares our results quantitatively against existing lexicostatistical similarity data. Section 5 presents our visualization method and results, as well as providing a qualitative visual interpretation of historic versus contemporary lexical similarity. Section 6, finally, provides conclusions.

[4] `http://cognet.ukc.disi.unitn.it`

2 State of the Art

The comparison of lexicons has a methodology established in the framework of *lexicostatistics*, with the underlying idea of inferring the phylogeny of languages from their lexicons considered in diachrony [17, 14]. Studies typically span a large number (hundreds or even thousands) of languages, using a small but fully meaning-aligned vocabulary selected from each language. To be able to consider phonetic evolution spanning millennia, very basic words are used—such as *water*, *sun*, or *hand*—and only in phonetic representations, such as from the well-known *Swadesh list* [16]. While such data are of the highest possible quality, they are scarce and only reflect a tiny fraction of the lexicon. Thus, while well-suited for diachronic studies, by design they provide less information about the present state of lexicons and the more recent linguistic and cultural influences to which they were subjected.

There are many examples of popular graph-based visualizations of such data.[5] While informative to non-experts, they are typically human-drawn based on only a handful of language pairs, and therefore are prone to subjective and potentially biased emphasis on certain languages or relationships. For example, for Estonian, the second graph listed in the footnote highlights its two distant European phylogenetic relatives (Hungarian and Finnish), as well as Latvian from the neighboring country, while it does not say anything about its significant Germanic and Slavic loans.

The most similar project we know of is *EZ Glot*.[6] They used a total of roughly 1.5 million contemporary dictionary words taken from overall 93 languages, mined from resources such as *Wiktionary*, *OmegaWiki*, *FreeDict*, or *Apertium*. The precision of their input evidence was self-evaluated to be about 80%.

While our work is also based on comparing online lexicons, we took as our starting point a high-quality cognate database, CogNet [1, 2], evaluated through multiple methods to a precision of 96% and covering 338 languages. CogNet employs etymologic and phonetic evidence, as well as transliteration across 40 scripts, expanding the language pairs covered. In terms of visualization, in contrast to hand-produced graphs, our approach is entirely automatic and free from the bias of manual cherry-picking, favoring a global optimum as it is computed over the entire similarity graph.

3 Automated Similarity Computation

Our input data, v2 of the CogNet database, consists of over 8 million sense-tagged cognate pairs. CogNet was computed from the *Universal Knowledge Core*

[5] To cite a few: `https://en.wikipedia.org/wiki/Romance_languages`, `https://elms.wpcomstaging.com/2008/03/04/lexical-distance-among-languages-of-europe/`, `https://alternativetransport.wordpress.com/2015/05/05/34/`

[6] `http://www.ezglot.com`

resource [9], itself built from wordnets, Wiktionary, and other high-quality lexical resources [6, 4, 12, 3].

The identification of cognate pairs having already been done by CogNet, we compute the cognate-content-based similarity between the lexicons of languages A and B as follows:

$$S_{AB} = \frac{\sum_{\forall <c_i^A, c_i^B>} \alpha + (1-\alpha)\text{sim}(c_i^A, c_i^B)}{\frac{2|L_A||L_B|}{|L_A|+|L_B|}}$$

where $< c_i^A, c_i^B >$ is the i^{th} cognate word pair retrieved from CogNet for the languages A and B and $\text{sim}(c_i^A, c_i^B)$ is a string similarity value:

$$\text{sim}(w_1, w_2) = \frac{\max(l_{w_1}, l_{w_2}) - \text{LD}(w_1, w_2)}{\max(l_{w_1}, l_{w_2})}$$

where LD is the Levenshtein distance and l_w is the length of word w, our hypothesis being that the more similar the cognate words between two languages, the closer the languages themselves to each other. In case w_1 and w_2 use different writing systems, we compare their Latin transliterations, also provided by CogNet. The smoothing factor $0 < \alpha < 1$ lets us avoid penalizing dissimilar cognates excessively, while $\alpha = 1$ cancels word similarity and simplifies the numerator to cognate counting.

The denominator of S_{AB} normalizes the sum by the harmonic mean of the lexicon sizes $|L_A|$ and $|L_B|$: these can range from tens to more than a hundred thousand word senses. Normalization addresses lexicon incompleteness, in order to avoid bias towards larger lexicons that obviously provide more cognates. The harmonic mean we use is lower than the arithmetic and geometric means but higher than the minimum value (i.e. the size of the smaller lexicon). This choice is intuitively explained by the fact that the amount of cognates found between two lexicons depends on the sizes of both, but is more strongly determined by the smaller lexicon.

Another source of bias is the presence of specialized vocabulary inside lexicons. Even though CogNet was built solely from general lexicons, some of them still contain a significant amount of domain terms (such as binomial nomenclature or medical terms), as the boundary between the general and the specialized vocabulary is never clear-cut. Domain terms such as *myocardiopathy* or *interferometer* tend to be shared across a large number of languages. Due to the tendency of domain terminology to be universal and potentially to grow orders of magnitude larger than the general lexicon, their presence in our input lexicons would have resulted in the uniformization of the similarities computed.

In order to exclude domain terms, we filtered our input to include only a subset of about 2,500 concepts that correspond to *basic-level categories*, i.e. that are neither too abstract nor too specialised and that are the most frequently used in general language. Note that the core vocabulary words used in comparative linguistics are also taken from basic-level categories, representing everyday objects and phenomena. Thus, in our case, *dog* or *heart* would remain in our input while

Table 1. Evaluation results with respect to ASJP data (root mean square error, standard deviation, and correlation), for three robustness levels (full dataset including all language pairs, pairs with medium or high robustness, and pairs with high robustness). We also provide comparisons to EZ Glot over the 27 language pairs it supports.

Dataset	Size	Difference w.r.t. ASJP RMSE	σ	R
CogNet full data	6,420	9.61	8.26	**0.61**
CogNet high+medium robustness	3,975	8.83	7.61	**0.65**
CogNet high robustness	1,399	10.72	8.94	**0.69**
CogNet over EZ Glot language pairs	27	23.01	16.01	**0.69**
EZ Glot	27	30.07	17.27	**0.48**

the too specific—and from our perspective irrelevant—*Staffordshire bullterrier* or *myocardial infarction* would be filtered out. As an existing list of basic-level categories, we used the *BLC* resource developed as part of the development of the Basque wordnet [15]. From this resource we used the broadest, frequency-based category list, as it also takes corpus-based frequencies into account and is therefore more representative of general language.

Finally, we annotated our similarity scores in terms of the robustness of supporting evidence as *low*, *medium*, or *high*, depending on the lexicon sizes used to compute cognates: robustness is considered low below a harmonic mean of 1,000 senses, and high above 10,000.

Our final result is a database of 27,196 language pairs, containing language names, ISO 639-3 language tags, similarity values, and a robustness annotation for each similarity value. 11.4% of all similarities are highly robust while 34.9% have medium robustness).

4 Comparison to Results from Lexicostatistics

In lexicostatistics, the standard benchmark is the ability of similarity data to predict well-established phylogenetic classifications. As we have different goals and work with different input data (e.g. we do not restrict our study to the core historical vocabularies), we cannot consider phylogeny as a gold standard against which to evaluate our results. Instead, we have quantified the difference between our similarity data and recent results from lexicostatistics, as produced by the state-of-the-art *ASJP* tool (based on the latest v19 of the ASJP Database)[7]. We have also compared the (more scarce) symmetric similarity data that was available from EZ Glot to ASJP data.

The intersection of our output with ASJP contained 6,420 language pairs, and 27 European language pairs with EZ Glot. After linearly scaling similarities to fall between 0 and 100, We computed the Pearson correlation coefficient R, the standard deviation σ, as well as the root mean square error RMSE with respect to ASJP, for both CogNet and EZ Glot. Among these three measures,

[7] `https://asjp.clld.org`

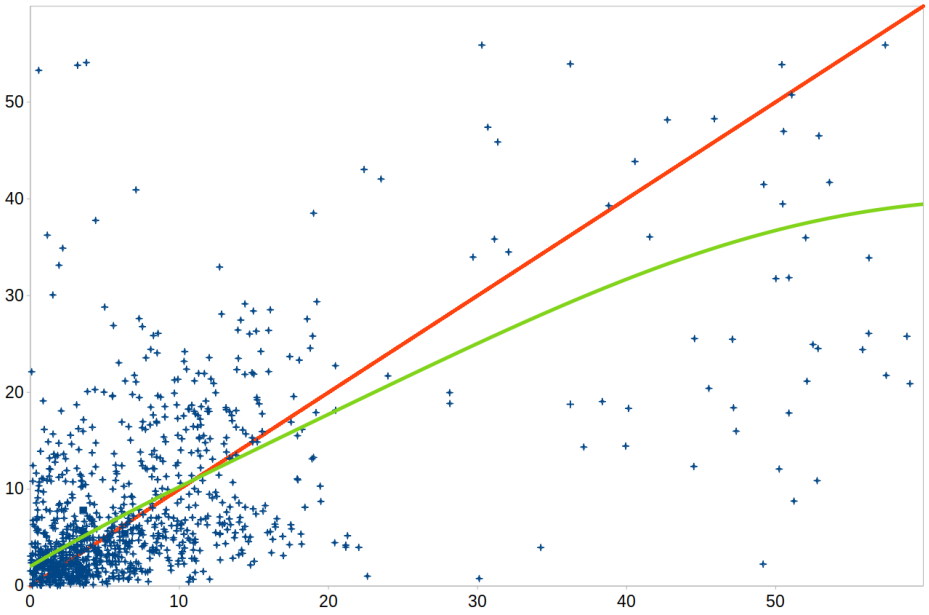

Fig. 1. Comparison of our similarity results from the high-robustness dataset (y axis) with the corresponding language pairs from ASJP (x axis). With respect to the core historical vocabularies covered by ASJP, the green trendline shows a generally higher similarity among genetically unrelated languages and a lower similarity among strongly related ones.

we consider correlation to be the most robust, being invariant to linear transformations such as how data is scaled. We generated three test sets: one restricted to high-robustness result (consisting of 1,399 pairs), one containing both high and medium results (3,975 pairs), and finally the full dataset (6,420 pairs).

The results are shown in Table 1 and in the scatterplot in Fig. 1. From both we see significant variance with respect to ASJP results. Yet, correlation with ASJP remains generally strong and is clearly increasing with robustness (from 0.61 up to 0.69). This result suggests that our robustness annotations are meaningful. EZ Glot results are more distant from ASJP and are more weakly correlated ($R = 0.48$, while for CogNet $R = 0.69$ over the same 27-language-pair subset). We experimentally set $\alpha = 0.5$, although its effect was minor, for instance over the full dataset $R(\alpha = 0) = 0.590$, $R(\alpha = 0.5) = 0.610$, while for simple cognate counting $R(\alpha = 1) = 0.597$.

The green trendline on the scatterplot shows that, on the whole, we compute higher similarities than ASJP for genetically unrelated languages (bottom left, $S < 10$) and lower similarities for genetically strongly related ones (top right). We attribute these non-negligible differences in part to borrowings across contemporary globalized lexicons and, in particular, to *universal words* (e.g. *"tennis"*, *"sumo"*, or *"internet"*) that increase the similarity of otherwise unrelated lan-

guages. On the other hand, *language change*—well known to affect the lexicon to a greater extent than it affects grammar—explains why the vocabularies of historically related languages generally show a higher dissimilarity today. This is our interpretation of the slope of the green trendline that always stays below unity.

5 Automated Visualization

Our aim was to reproduce the popular graph-based visualization of lexical similarities in a fully automated manner and based on the entire similarity graph, as opposed to human-produced illustrations based on cherry-picked data. We used the well-known *Sigma* graph visualization library,[8] combined with the JavaScript implementation of the *ForceAtlas2* algorithm [10]. The latter applies a physical model that considers graph edges to be springs, with tensions being proportional to the edge weights. We modeled languages as nodes and their lexical similarities as weighted edges, resulting in more similar languages displayed closer together. Because of the nature of the solution based on a physical tension-based model that dynamically evolves towards a global equilibrium, our visualizations favor a global optimum as opposed to locally precise distances. Thus, the visualizations produced give a realistic view of the "big picture", but distances of specific language pairs should be interpreted qualitatively rather than quantitatively.

Figure 2 shows a small portion of the graph computed.[9] In order to keep the graph compact, we restricted it to high-robustness similarities, covering about a hundred languages. In order to get an intuitive idea of the effect of both phylogeny and geography on the similarity of contemporary lexicons, we created two versions of the graph: in the first one (top), nodes are colored according to language families, as it is usually done in comparative linguistics that focuses on phylogenetic relationships. In the second version (bottom), we colored the nodes according to the approximate geographic position of language speakers, taking into account latitude–longitude coordinates as well as continents. Simply put, speakers of similar-colored languages live closer together. Both phylogenetic and geographic metadata were retrieved from the *World Atlas of Language Structures* [5].[10]

The visualization tool can also be used to display similarity data from different sources (provided that it is converted to the input format expected by the Sigma library). In particular, we used the tool to obtain a visual impression of the difference between our contemporary similarity data and those produced by the phylogeny-oriented ASJP tool. The result on ASJP data can be seen in Figure 3.

The visualizations in Figs. 2 and 3 provide remarkable insight into language change and the state of modern lexicons. In both the contemporary and the ASJP datasets, language families are clearly identifiable as their respective nodes tend

[8] `http://sigmajs.org`

[9] The full graphs are visible on the page `http://ukc.datascientia.eu/lexdist`.

[10] `http://wals.info`

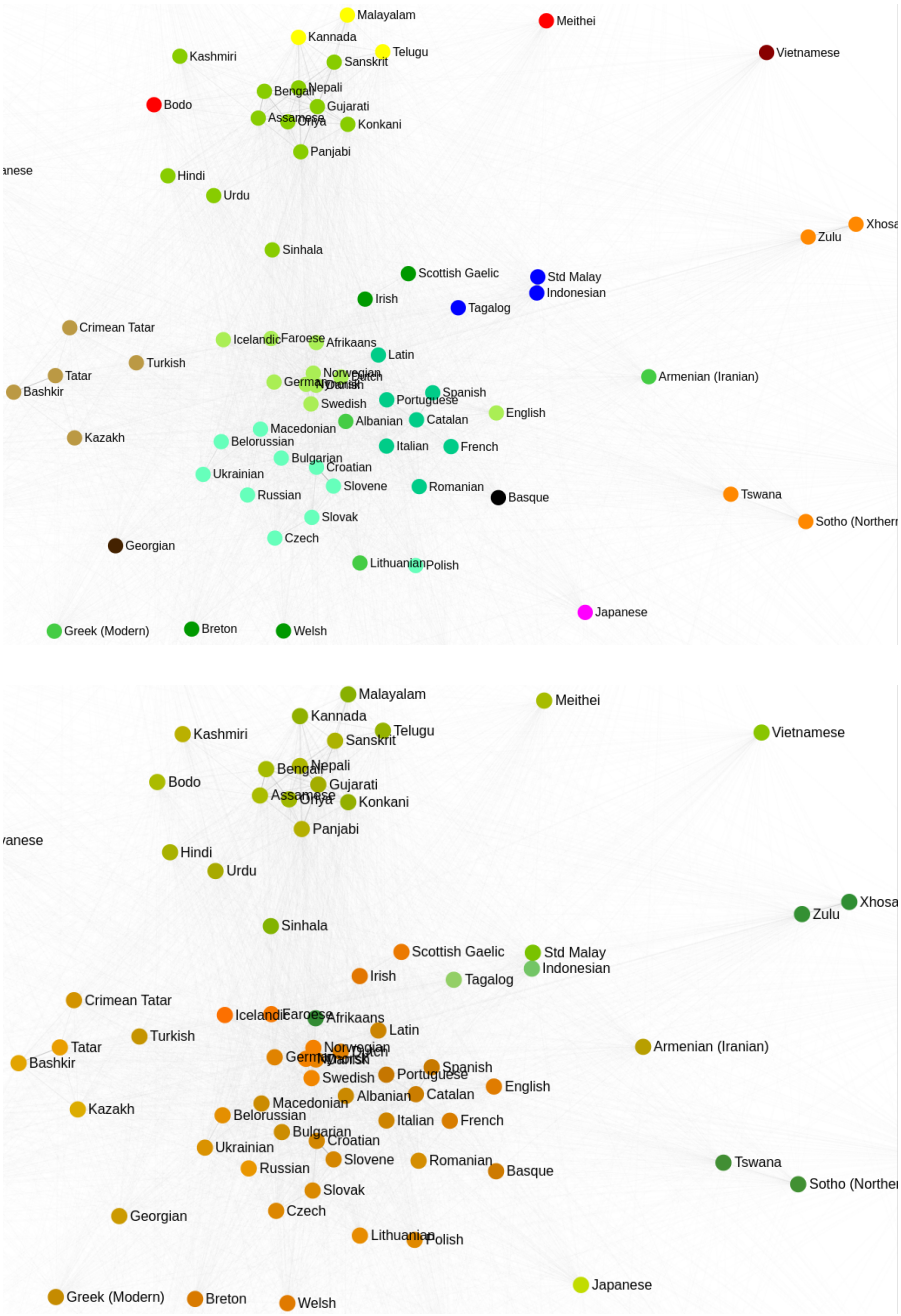

Fig. 2. Detail from the automatically-computed lexical similarity visualization, with colors corresponding to language families (top) and to the geographic location of speakers (bottom).

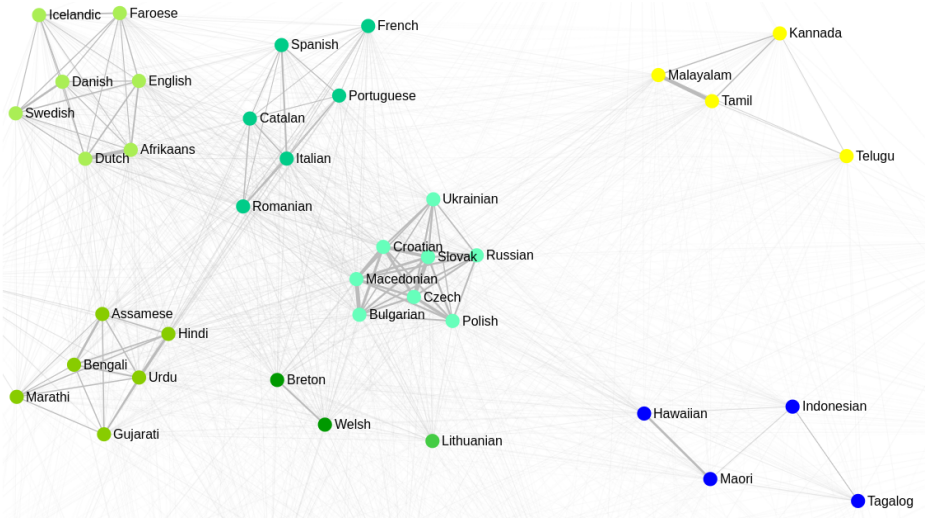

Fig. 3. Detail from the visualization of ASJP similarity data.

to aggregate together. The clusters are, however, much more salient in the ASJP data (Figure 3), where subfamilies within the Indo-European phylum form separate groups, all the while remaining within an Indo-European "macro-cluster". Dravidian languages from the Indian subcontinent (right-hand-side, in yellow) are far removed from the culturally and geographically close Indic group. Such a result was expected, as ASJP lexicons are optimized to highlight phylogenetic relationships. Due to borrowings, modern lexicons are less clearly distinguished from each other. This is evident, in Figure 2, from the proximity of the Germanic, Romance, and Slavic families, or from unrelated languages such as Japanese or Tagalog "approaching" the Indo-European families due to borrowings. Likewise, the fact that English is detached from the Germanic cluster to move closer to the Romance family reflects its massive French loanword content.

Further insights are gained on the effect of geography on contemporary lexicons. The bottom image in Figure 2 shows that even-colored (geographically close) nodes tend to group together (with self-evident exceptions such as Afrikaans). Remarkably, the languages of India aggregate into a single cluster far apart, despite the internal linguistic heterogeneity of the Indian subcontinent that is home to three fully distinct language families—Indic, Dravidian, and Sino-Tibetan—and despite the Indo-European relatedness of the Indic family. In this case, the effect of geography and culture seems stronger than phylogeny or English borrowings.

6 Conclusions and Future Work

We have found significant correlation between the similarity data obtained from large contemporary lexicons and from lexicostatistical databases geared towards language phylogeny research. At the same time, we have also found that, on the whole, large contemporary lexicons tend to resemble each other more. We believe that the uniformizing effect of globalized culture on languages plays a role in this observation.

Due to these differences, we consider our data to be more relevant to cross-lingual uses applied to contemporary language, such as machine translation, cross-lingual transfer, or bilingual lexicon induction, where—other things being equal—lexical similarities may predict efficiency over language pairs. On the other hand, our data is not suitable for use in historical linguistics that is based on a more strict definition of cognacy and on a more controlled concept set.

Our full lexical similarity data, as well as the dynamic visualizations, are made freely available online.[11]

Acknowledgments

This paper was partly supported by the *InteropEHRate* project, co-funded by the European Union (EU) Horizon 2020 programme under grant number 826106.

References

1. Batsuren, K., Bella, G., Giunchiglia, F.: Cognet: a large-scale cognate database. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 3136–3145 (2019)
2. Batsuren, K., Bella, G., Giunchiglia, F.: A large and evolving cognate database. Language Resources and Evaluation (2021). https://doi.org/10.1007/s10579-021-09544-6
3. Batsuren, K., Ganbold, A., Chagnaa, A., Giunchiglia, F.: Building the mongolian wordnet. In: Proceedings of the 10th Global Wordnet Conference. pp. 238–244 (2019)
4. Bella, G., McNeill, F., Gorman, R., Donnaile, C.O., MacDonald, K., Chandrashekar, Y., Freihat, A.A., Giunchiglia, F.: A major wordnet for a minority language: Scottish gaelic. In: Proceedings of The 12th Language Resources and Evaluation Conference. pp. 2812–2818 (2020)
5. Comrie, B.: The world atlas of language structures. Oxford University Press (2005)
6. Dellert, J., Daneyko, T., Münch, A., Ladygina, A., Buch, A., Clarius, N., Grigorjew, I., Balabel, M., Boga, H.I., Baysarova, Z., et al.: Northeuralex: a wide-coverage lexical database of northern eurasia. Language resources and evaluation **54**(1), 273–301 (2020)
7. Garcia, M., Gómez-Rodríguez, C., Alonso, M.A.: New treebank or repurposed? on the feasibility of cross-lingual parsing of romance languages with universal dependencies. Natural Language Engineering **24**(1), 91–122 (2018)

[11] `http://ukc.datascientia.eu/`

8. Giunchiglia, F., Batsuren, K., Bella, G.: Understanding and exploiting language diversity. In: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17). pp. 4009–4017 (2017)
9. Giunchiglia, F., Batsuren, K., Freihat, A.A.: One world—seven thousand languages. In: Proceedings 19th International Conference on Computational Linguistics and Intelligent Text Processing, CiCling2018, 18-24 March 2018 (2018)
10. Jacomy, M., Venturini, T., Heymann, S., Bastian, M.: Forceatlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software. PloS one **9**(6), e98679 (2014)
11. Lin, Y.H., Chen, C.Y., Lee, J., Li, Z., Zhang, Y., Xia, M., Rijhwani, S., He, J., Zhang, Z., Ma, X., et al.: Choosing transfer languages for cross-lingual learning. arXiv preprint arXiv:1905.12688 (2019)
12. Nair, N.C., Velayuthan, R.S., Batsuren, K.: Aligning the indowordnet with the princeton wordnet. In: Proceedings of the 3rd International Conference on Natural Language and Speech Processing. pp. 9–16 (2019)
13. Nasution, A.H., Murakami, Y., Ishida, T.: Constraint-based bilingual lexicon induction for closely related languages. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). pp. 3291–3298 (2016)
14. Petroni, F., Serva, M.: Measures of lexical distance between languages. Physica A: Statistical Mechanics and its Applications **389**(11), 2280–2283 (2010)
15. Pociello, E., Agirre, E., Aldezabal, I.: Methodology and construction of the basque wordnet. Language resources and evaluation **45**(2), 121–142 (2011)
16. Swadesh, M.: Towards greater accuracy in lexicostatistic dating. International journal of American linguistics **21**(2), 121–137 (1955)
17. Wichmann, S., Müller, A., Velupillai, V., Brown, C.H., Holman, E.W., Brown, P., Sauppe, S., Belyaev, O., Urban, M., Molochieva, Z., et al.: The asjp database (version 13). URL: http://email. eva. mpg. de/˜ wichmann/ASJPHomePage. htm **3** (2010)